The free encyclopedia that anyone can dispute: An analysis of the micro-structural dynamics of positive and negative relations in the production of contentious Wikipedia articles*

Jürgen Lerner
University of Konstanz
juergen.lerner@uni-konstanz.de

Alessandro Lomi
University of Italian Switzerland, Lugano
alessandro.lomi@usi.ch

Accepted manuscript, to appear in *Social networks*.
Online: https://doi.org/10.1016/j.socnet.2018.12.003.

Abstract

We consider two rival hypotheses on the emergence of organization in open production communities. According to the first ("reputation hypothesis"), patterns of agreement and disagreement among participants in open production communities are explained by differences in individual reputation for quality of contribution. The reputation hypothesis predicts that participants will tend to agree with more reputable others and disagree with less reputable others thus contributing to establish a stable open production community. According to the second hypothesis ("balance hypothesis"), patterns of agreement and disagreement are explained by membership in sub-communities of "friends" and "enemies." The balance hypothesis predicts that participants in open production communities will agree mainly with friends and disagree mainly with enemies, regardless of considerations about reputation for the quality of their contributions. In this paper, we examine which one of these hypotheses is more consistent with patterns of positive and negative interaction events observed during the production of the complete set of 1,206 English-language Wikipedia articles officially considered controversial. We specify and estimate new models for signed and weighted relational event networks predicting the probability that a user deletes the contributions of another user – thus expressing personal disagreement – and/or protects the contributions of another user against deletion from third parties – thus expressing personal agreement. In an analysis of positive and negative interaction among Wikipedia contributors consisting of more than 60 million observations, we find strong support for the balance hypothesis and for the predictions of the reputation hypothesis that are more consistent with alter-centric interpretations of social status as conferred by alters through observable acts of deference.

Keywords: Balance theory; Open production communities; Positive and negative relations; Relational event models; Self-organizing systems; Signed social processes; Social status; Weighted networks; Wikipedia

1 Introduction

Recent years have witnessed a considerable increase in the diffusion of, and interest in, new forms of peer production based on decentralized interaction within communities of independent participants (Benkler and Nissenbaum 2006; Benkler et al. 2015; Conaldi and Lomi 2013; Piskorski and Gorbatâi 2017; Lerner and Tirole 2002; Singh et al. 2011; von Hippel and von Krogh 2003). One of the main questions motivating the current interest in peer production concerns the emergence of order in the almost complete absence of hierarchical organizational structures and centralized coordination mechanisms. This issue is at the heart of

*This work has been supported by Deutsche Forschungsgemeinschaft (DFG Grant Nr. LE 2237/2-1) and Swiss National Science Foundation (FNS Project Nr. 100018_150126).

what Padgett and Powell (2012) identify as the problem of "emergence" – or how organizational and social structures arise out of "vortexes in the flow of social life" rather than being "buildings of stone."

Addressing questions about the emergence of organized order becomes particularly important – and difficult – when the absence of hierarchical conflict resolution mechanisms makes controversies among participants difficult to settle and potentially detrimental to the peer production process and its outcomes. How can order be achieved and maintained – and how can the production of anything collectively valuable be possible – under conditions of extreme decentralization and latent conflict that characterize peer production (Viégas et al. 2004; Kittur et al. 2007; Brandes and Lerner 2007; Suh et al. 2007; Sumi et al. 2011; Arazy et al. 2011; Yasseri et al. 2012; Tsvetkova et al. 2016)? Because conflict increases the private cost of producing public goods, these questions are at the heart of the current debate about the sustainability of peer production, and the survival of the open source production movement (Duguid 2006; Kreiss et al. 2011; Lerner and Tirole 2001; von Hippel and von Krogh 2006). Conflict in peer production organizations cannot just be settled by *fiat*, or by relying on hierarchical authority, but only by building collective consensus. In the near-absence of formal organizational structure, peer production projects are mostly regulated through informal networks arising from task-oriented interaction within communities of participants (O'Mahony and Lakhani 2011).

Our study is guided by two mutually non-exclusive hypotheses on the emergence of organizational order from decentralized text production and editing activities in Wikipedia – the "free encyclopedia that anyone can edit" which we have selected as the specific peer production organization of interest. Wikipedia may be considered as broadly representative of open peer-production projects where voluntary participants contribute and edit content that is made collectively and freely available (Lerner and Tirole 2001; von Hippel and von Krogh 2006). The hypotheses we forumulate allow us to express fundamental theoretical principles of social organization in terms of hypotheses on the evolutionary dynamics of signed event networks. According to the first hypothesis ("reputation hypothesis") positive and negative interaction (i. e., agreement and disagreement) are explained by the reputation of the target actor (Adler and de Alfaro 2007; Javanmardi et al. 2010). The reputation hypothesis predicts that more reputable actors are more likely to receive agreement, while disagreement flows toward less reputable actors. According to the second hypothesis, ("balance hypothesis"), expressions of agreement and disagreement are organized according to membership in latent communities of "friends" and "enemies."[1] The balance hypothesis predicts that agreement is expressed mainly towards friends and disagreement mainly towards enemies – regardless of their reputation.

Balance theory (Heider 1946; Cartwright and Harary 1956) explains the formation of signed networks, but empirical evidence for it has been mixed (Yap and Harrigan 2015). Leskovec et al. (2010) suggested status theory as an alternative which can explain the structure and evolution of signed networks. Status theory predicts that negative relations tend to point away from actors with high status and toward actors with low status, while positive relations tend to flow from low to high status actors. The predictions derived from the reputation hypothesis are a subset of the predictions that can be derived from status theory. More precisely, the reputation hypothesis makes only predictions related to *in-coming* relational events, such that (i) actors that received many positive events in the past are more likely to receive positive events and less likely to receive negative events in the future, and (ii) actors that received many negative events in the past are less likely to receive positive events and more likely to receive negative events in the future. The reputation hypothesis reflects an "alter-centric" conception of status as a social position conferred to ego ("receiver") by alters ("senders") through acts of deference (Podolny 2010; Torlò and Lomi 2017). This view of status is considered alter-centric because "deference cannot be seized by an actor but rather is something that is awarded by others" (Sauder et al. 2012, p.273). The predictions of the reputation hypothesis are consistent with results produced by studies of dominance hierarchies in animal societies (Chase et al. 1994).

These two alternative perspectives on the micro-mechanisms of network formation imply different network macro-structures. According to the reputation hypothesis, actors are assigned reputation values from a one-dimensional scale that influences probabilities to receive (dis-)agreement, regardless of the sending actor. The ratio of incoming positive ties over incoming negative ties would increase with higher reputation. In

[1]The terms *friend* and *enemy* are employed for readability. We say that two actors are friends if they are connected by positive events and that they are enemies if they are connected by negative events.

contrast, the balance hypothesis posits the emergence and progressive crystallization of a polarized network in which two groups (factions) mutually fight each other (Cartwright and Harary 1956). Membership in these groups explains the probabilities to receive (dis-)agreement – but only if we take into account the group membership of the sending actor. According to balance theory, actors would be more likely to agree with members of their own group but more likely to disagree with members of the other group. Thus, members from opposing factions would assess contributions of the same third user differently.

We test these hypotheses in an analysis of networks of signed relational events among the contributing users of the 1,206 Wikipedia articles that are labeled as controversial.[2] Controversial articles are those which "regularly become biased" and "are likely to suffer future disputes," as defined by Wikipedia. We extend currently available relational event models and analyze patterns of agreement (positive relations) and disagreement (negative relations) among contributing users of Wikipedia given the full history of their previous interaction. The models we specify and estimate include explanatory mechanisms encoding effects consistent with both balance and reputation hypotheses. We focus on the subset of controversial Wikipedia articles because we expect this context to be uniquely useful for identifying and illuminating the coordination mechanisms underlying the hypotheses that we have outlined and because controversial articles involve a high level of interaction among contributors almost by definition. Because our study covers the complete lifetime of Wikipedia, this feature of controversial articles gives rise to a very large sample of relational events.

The bipartite structure directly connecting contributing users to text in Wikipedia, dually connects contributors (Breiger 1974). For this reason, observable expressions of agreement or disagreement connecting users to text through acts of editing may be interpreted as agreement or disagreement between users. Thus, contributors interact through their joint involvement in the production of text – the raw input of Wikipedia articles. We are interested in identifying and interpreting patterns in the emergent social order in Wikipedia resulting from this signed event network.

Signed network data have been collected and analyzed throughout the history of social network analysis (Labianca 2014), and the analysis of signed network relations is currently experiencing a surge of renewed interest (Everett and Borgatti 2014; Labianca and Brass 2006). Linking the dynamics of production relations in Wikipedia to explicit models for signed networks allows us to illuminate fundamental general issues in the analysis of peer production.

To foreshadow the results of our analysis, we find strong support for balance theory: the micro-dynamics of positive and negative events seem to support a balanced, and hence polarized, macro-structural social order. An additional contribution of our study is to show that the "alter-centric" status implications of the reputation hypothesis, namely the predictions related to *in-coming* relations, receive strong empirical support; but, on the other hand, the predictions of status theory related to *out-going* relations are not supported. Specifically, we show that actors initiating many negative events do not necessarily have high status, and actors initiating many positive events do not necessarily have low status. Likewise, we find no empirical evidence for the anti-reciprocity of positive and negative relations predicted by status theory. Thus the empirical predictions of status theory, which received strong support in the context of signed networks of relations such as like/dislike or high/low esteem (Leskovec et al. 2010; Yap and Harrigan 2015), have to be restricted in our setting to the "alter-centric" components of status that are captured by our reputation hypothesis.

Above and beyond these substantive insights, our paper also makes a clear methodological contribution by extending current relational event models for signed, weighted, and directed social interaction data. Building on previous models (Lerner and Lomi 2017), we make several methodological improvements that are necessary to deal with large networks of relational events and propose an indicator of user reputation that proves to be one of the strongest and most reliable predictors for future positive or negative events.

After presenting the theoretical background in more detail in Section 2, we describe the empirical setting in Section 3. The construction of edit networks and relational event models for these is detailed in Section 4. Results are presented and discussed in Section 5. We close by discussing implications of our work and outline possibility for future work in Section 6.

[2] `https://en.wikipedia.org/wiki/Wikipedia:List_of_controversial_issues`

2 Background and hypotheses

Balance theory and the reputation hypothesis. Balance theory provides one possible theoretical connection between patterns of local interaction among individuals and the global network structure in which they are embedded (Doreian et al. 1996). Balance theory may be traced back to the work of Heider (1946) and has been generalized by Cartwright and Harary (1956) establishing that a signed network is balanced if every cycle has an even number of negative ties and proving that a network is balanced if and only if it decomposes into two groups such that all positive ties are within groups and all negative ties are between groups. Balance theory can be derived from the theory of cognitive dissonance (Festinger 1962) with the additional assumption that signs of ties can be multiplied along paths in a network. If different paths connecting the same two nodes have different signs, then there is dissonant information which persons try to avoid, according to the theory of cognitive dissonance.

Specifically in the context of signed event networks derived from collaborative editing in Wikipedia, if two contributing users A and B are connected by many positive editing events, then they are likely to be members of the same group – for instance, they might be on the same side of an ideological dispute. If a third user C makes contributions of B undone, then this provides a signal for A that C and B are on opposite sides and, hence C is also on the opposite side of A. To avoid cognitive dissonance, A would be tempted to consider C's edit as inappropriate. Thus, A would be more likely to undo C's edit than the edit of another user who is not an "enemy of a friend" of A. In contrast, A would be tempted to agree with contributions of B. Thus, if the balance hypothesis holds, editing decisions would not be purely based on the quality of contributions but, at least in part, depend on perceived group membership of other users.

We note that, as it is it typical for online peer production, Wikipedia is characterized by a high level of social transparency (Stuart et al. 2012): contributing users may learn the identity of their peers easily and at no cost while observing their (inter-)actions. Assuming that repeated interaction through text editing will eventually encourage participants to learn about each other's identity seems like a minimalist assumption – an assumption that may be easily substantiated by examining Wikipedia "talk pages" where active users discuss their and other's text editing activities. It is likely that participants will find it useful to know who they are agreeing or disagreeing with, as this piece of information will shape their expectation about the valence and character (concordant or discordant) of future interaction. Depersonalized – yet repeated – interaction through text editing activities seems to be a stronger assumption.

Empirical work on signed networks, in general, has found mixed evidence for balance theory and emphasized the need to consider, or control for, other effects explaining signed network formation when testing balance theory (Hummon and Doreian 2003; Doreian and Mrvar 2009, 2014; Leskovec et al. 2010; Yap and Harrigan 2015; Lerner 2016). Of particular interest for this paper is the work of Leskovec et al. (2010) who proposed status theory claiming that a negative relation from A to B indicates that A has higher status than B and a positive relation from A to B indicates that A has lower status than B. In our paper we claim that in the context of peer production involving controversial tasks, only a subset of the predictions of status theory will hold.

We test the predictions of balance theory with models incorporating an alternative hypothesis (dubbed the *reputation hypothesis*) on the evolutionary dynamics of signed networks. According to the reputation hypothesis, actors are characterized by varying degrees of reputation (Adler and de Alfaro 2007; Javanmardi et al. 2010), that are consistently assessed by others in the course of interaction. In this sense the reputation hypothesis, is closely related to the concept of popularity in signed networks (Yap and Harrigan 2015). Actors with higher reputation are less likely to see their work disputed and undone (i. e., less likely to receive negative events) and more likely to have their work redone or supported, if previously contested or deleted (i. e., more likely to receive positive events).

Specifically, in the case of Wikipedia, the reputation hypothesis predicts the presence of a latent variable indicating the user's reputation. This variable, in turn, will have a consistent effect on the probability that a user's contributions are made undone. In contrast to patterns derived from balance theory, the direction of the effect of the reputation variable does not depend on the source actor (that is, the user who potentially performs the undo). While membership in one group decreases (according to balance theory) the probability to receive negative events from members of the same group, it increases the probability to receive negative

events from members of the opposite group. In contrast, having high reputation will decrease (according to the reputation hypothesis) the probability to receive a negative event from any actor in the network. The reputation hypothesis is consistent with the view that edit decisions are based on the quality of content. On the other hand, agreement with balance theory might point to polarization where contributions are assessed based on their ideological orientation and not, or not only, based on their quality.

There is a crucial difference between status theory and the reputation hypothesis. A positive event from actor A to B implies, according to status theory, that B has higher status than A (Blau 1955; Leskovec et al. 2010). According to the reputation hypothesis the same event merely implies that B has relatively high reputation but not necessarily higher than A. Thus, according to the reputation hypothesis, a highly reputable actor might send a positive event to an actor with a lower, but sufficiently high reputation, while according to status theory an actor with high status should not send a positive event to an actor with lower status. In particular, and in contrast to status theory, the reputation hypothesis does not predict anti-reciprocity of positive and negative events. It also does not assign reputation based on out-going relational events, so that sending many positive events is not an indicator for low reputation (while it is an indicator for low status according to status theory) and sending many negative events is not an indicator for high reputation (while it is an indicator for high status in status theory).

These modifications of status theory, leading to the reputation hypothesis, are motivated by theoretical considerations as well as by findings from previous work. Theoretically it seems to be implausible that an actor preserving or restoring the work of another one considers herself less reputable than the recipient of the positive event. This would exclude the possibility of a benevolent senior team member supporting less senior actors with lower but sufficiently high reputation. Empirically, previous work on modeling edit events in Wikipedia has shown that the assumption that positive redo events point from lower to higher status is not supported (Lerner and Lomi 2017).

Considering balance theory and the reputation hypothesis jointly in the same model is particularly useful since the two theories offer alternative predictions for the network structure of peer production that is more likely to be observed, given the coordination micro-mechanisms connecting individual participants. The reputation hypothesis assumes that actors consistently assess the reputation of others, and act rationally so that they tend to preserve the contributions of more reputable actors and tend to undo the contributions of less reputable actors. All that is needed for the reputation hypothesis to work is the assumption that text editing event sequences provide signals of quality that are interpreted by participants and influence their decisions. We note that this assumption is uncontroversial in organizational sociology research on status (Benjamin and Podolny 1999; Sauder et al. 2012).

Balance theory, on the other hand, predicts socially motivated behavior in which actors classify others as friends or enemies and delete or preserve edits based on the membership of contributors in opposing social groups, rather than their reputation. Similar to the reputation hypothesis, the balance hypothesis stands on the assumption that participants observe and interpret the behavior of others as a signal of membership in social groups, and that this information affects their future evaluation and editing decisions. As before, given the high level of social transparency of Wikipedia, this assumption is routinely made in sociological studies of how social identities conferred by category membership affect social evaluation (Negro et al. 2014; Zuckerman 2012).

Even though the relation between network effects and the quality of the final product is not empirically tested in this paper, we suggest that teams acting according to the reputation hypothesis will be more likely to produce high-quality output, while behavior consistent with balance theory will be less likely to support high-quality contributions. We discuss this hypothetical relation further in Sect. 6. We further note that the predictions of balance theory and the reputation hypothesis are not mutually exclusive.

Hypotheses Both balance and reputation hypotheses share the assumption that interaction among users produce signals that these users themselves, but also bystanders, observe, interpret, and act upon. As we have discussed in the prior section, the assumption that observable patterns of association between actors (or between actors and social groups, or actors and artifacts) contain information that observers interpret and evaluate is foundational both for theories of status (Podolny 2001) and reputation (Kilduff and Krackhardt

1994), and for contemporary theories of identity as conferred by membership in social categories (Hannan 2010; Negro et al. 2014). Therefore, it seems that the main difference between theories of reputation and balance involves the interpretation of signals produced by events connecting users. These differences are important because (i) they sustain different (although partially overlapping) predictions about the micro-mechanisms that generate patterns of agreement (positive relations) and disagreement (negative relations) among users, and because (ii) such patterns imply different macro-structures.

According to the reputation hypothesis, contributions that are not disputed by others signal, and at the same time increase, the reputation of the associated contributor. According to the balance hypothesis, the valence of interaction (positive or negative) signals relative group affiliation of the sending and receiving actor. As we have discussed, the unique value of Wikipedia as an empirical setting is that these signals are generally observable.

The effects predicted by balance theory can be derived from the core assumption that participants are members in separate communities such that, if A and B share membership in the same community, a positive event in the dyad (A, B) will be more likely to be observed. Conversely, if A and B belong to different communities, then a negative event in (A, B) will be more likely. Balance theory predicts the following configurations of event sequences. At the dyadic level it predicts that a positive event from A to B increases the probability of a future positive event in (A, B) and in the reverse dyad (B, A) and it decreases the probability of future negative events in both, (A, B) and (B, A). Conversely, a negative event from A to B increases the probability of a future negative event in (A, B) and in the reverse dyad (B, A) and it decreases the probability of future positive events in those dyads. Thus, positive and negative relational events are predicted to exhibit *repetition* and *reciprocation* within the same sign and *anti-repetition* and *anti-reciprocation* across signs. We note that the predictions about reciprocation and anti-reciprocation are contrary to those from status theory. On the triadic level, balance theory predicts that two actors A and B are more likely to interact positively, and less likely to interact negatively, if they have a common friend or if they have a common enemy. Conversely, they are less likely to interact positively, and more likely to interact negatively, if there is a third actor that is a friend of A and an enemy of B, or vice versa.

The reputation hypothesis predicts that the more reputable an actor A, the more likely it is that A receives positive events, and the less likely it is that A receives negative events. For the reputation based on in-coming events, this implies the following indegree effects. The higher the positive indegree of an actor A (that is the aggregate weight of past positive events received by A), the more likely it is that A will receive future positive events, and the less likely it is that A will receive future negative events. Conversely, the higher the negative indegree of an actor A (that is the aggregate weight of past negative events received by A), the less likely it is that A will receive positive events, and the more likely it is that A will receive negative events in the future. We also hypothesize that the share of a user's contributions that survive in the past serves as an indicator of reputation (explained in more detail in Sect. 4) and will increase the probability to receive positive events in the future and decrease the probability to receive negative events in the future. Furthermore, the reputation hypothesis predicts, as does balance theory, repetition of positive and negative events, since actors should evaluate others consistently. However, the reputation hypothesis makes no predictions about reciprocation or anti-reciprocation. We note that the models we fit in this paper include more effects than just those predicted merely by balance theory or the reputation hypothesis, among them all variations of pure and mixed, signed outdegree and indegree effects.

3 Empirical setting: controversial articles in Wikipedia

One of the most successful peer production projects is our empirical setting, Wikipedia. It is an open encyclopedia where contributing users can make incremental edits to an article that other users can then evaluate and preserve, or – as the case may be – dispute and even undo. Third parties might disagree with the undoing at any time, and take action to re-instate deleted contributions they consider valuable. In this way, contributors express agreement toward some of their peers and disagreement toward others triggering bursts of sudden conflict (Yasseri et al. 2012; Kittur et al. 2007) that may evolve into veritable "edit wars" (Sumi et al. 2011). Eventually, such edit conflicts may induce the emergence of "pecking orders" (Chase

1980)– stable hierarchical arrangements based on acts of interpersonal aggression (Lerner and Lomi 2017).

Analysis of controversy in Wikipedia. Since its early years, Wikipedia researchers have been interested in how contributing users act in the presence of conflict and controversy and this trend continues in current studies (Viégas et al. 2004; Kittur et al. 2007; Brandes and Lerner 2007; Suh et al. 2007; Sumi et al. 2011; Arazy et al. 2011; Yasseri et al. 2012; Tsvetkova et al. 2016). While the general quality of Wikipedia articles is already surprisingly high (Giles 2005), given the near-absence of a formal organizational structure, it is even more astonishing that the Wikipedia community succeeds in writing high-quality articles about controversial topics. Indeed, the 1,206 Wikipedia articles labeled as controversial include 61 featured articles (the highest assessment grade in Wikipedia's quality scale)—a rate that is more than 50 times as high as the overall proportion of featured articles in the whole of Wikipedia (5,029 featured articles out of 5,412,810 articles). Thus, Wikipedia contributors seem to manage controversies in a quite productive way.

Controversy in Wikipedia has another interesting aspect that distinguishes it from other online communities. Online discussion, in general, seems to have a tendency to form "echo chambers," that is, separate clusters consisting of users with similar opinions that thereby avoid interacting with dissonant information (Gilbert et al. 2009; Barberá et al. 2015; Jasny et al. 2015). By design, Wikipedia is different in this respect; users contributing to an article about a controversial topic such as global warming must work on one single common product, namely the article with that title, and factions with diverging opinions cannot simply create their own version of an article about global warming. Relatedly, a recent study found that Wikipedia users are more likely to contribute to articles with the opposite "slant" (Republican-leaning vs. Democrat-leaning) than to articles with their own slant (Greenstein et al. 2016). Earlier, Neff et al. (2013) reported a similar tendency against the formation of echo chambers in Wikipedia. Thus, conflicting parties in Wikipedia usually do not avoid each other but they have to interact and to find a consensus—or fail to do so. In this aspect, contributing users of Wikipedia are faced with a similar situation as actors in certain offline task-oriented groups such as members of parliament who must deal with actors from opposing parties. Wikipedia provides a unique opportunity to study completely observed and fine-grained data about social interaction in production teams working on controversial tasks.

Wikipedia is not completely void of formal organizational structure as it has groups with additional user rights, such as "administrators" or "bureaucrats". These users with special rights work on important administrative tasks and play a crucial role, for instance, in dealing with deviant behavior, such as vandalism or other violations of Wikipedia policies. On the other hand, administrators have no special power in settling content disputes. For example, Wikipedia policies on administrator conduct state: "*Administrators should not normally use their tools in matters in which they are personally involved (for example, in a content dispute in which they are a party).*[3]" We therefore claim that *content* disputes are normally not settled via top-down decision making by users with special rights, but rather through consensus arising through, if at all, informal interaction networks of coordination and control. There are two major types of such networks in Wikipedia: the discussion network resulting from talk pages and the edit network arising from co-editing articles, analyzed in this paper and defined more precisely below. While we do not question the importance of discussion in Wikipedia for settling content disputes, the analysis of the discussion network, separately or jointly with the edit network, is out of scope for this paper. Note that the discussion network can only have an indirect effect on the article's content, while the edit network directly determines and results from the evolution of content.

Norm enforcement in Wikipedia. Of particular relevance for the current study is the work by Piskorski and Gorbatâi (2017) who analyze factors explaining the likelihood of users committing or experiencing "norm violations," the probability that these are punished, and the probability that punishing norm violations will be rewarded within the community of active Wikipedia participants. They showed that punishments and rewards for these are more frequent in denser communities – making norm violations more infrequent. When comparing our work to Piskorski and Gorbatâi (2017) we have to be aware of several differences. While the

[3] `https://en.wikipedia.org/wiki/Wikipedia:Administrators#Misuse_of_administrative_tools`, accessed on 20th January, 2018

starting point of Piskorski and Gorbatâi (2017) are "undos" (operationalized as reverts, that is, edits that restore exactly to a previous version) that constitute vandalism, we are concerned with all editing events – not only those that are norm violations. In our paper, an "undo event" occurs if one user makes part of, but not necessarily all of, the contributions of another user undone. These undo events are not restricted to norm violations; or their punishment. For instance, an edit that replaces acceptable content by improved content is neither a norm violation nor does it revert a norm violation. On a more technical level, we analyze editing at a finer level of granularity. If a user adds new text comprising, say, 100 words and another user deletes 30 of these in the next revision then we consider this as an undo event of weight 30 and our models seek to explain the undo ratio of 30/100. On the other hand, such an edit would not be a revert and, thus, would not be analyzed in Piskorski and Gorbatâi (2017). With respect to differences in the analysis we do not aggregate over time intervals and we model *dyadic* undo and redo probabilities, that is, probabilities that potentially depend on the combination of a source with a target actor. Piskorski and Gorbatâi (2017) analyze the probability to commit or experience norm violations (or punishment for committing those, or reward for those who punish) but these probabilities do not depend on the combination of source and target actor – as is necessary to analyze agreement with the predictions of balance theory. In summary: norm violation studied by Piskorski and Gorbatâi (2017) happens at the *node level*. As social mechanism, balance can only be defined at the *dyadic and triadic level*. The predictions of balance theory cannot be tested at the node level.

These differences notwithstanding, we discuss if and how the mechanisms analyzed by Piskorski and Gorbatâi (2017) might impact our analysis. Vandals, that is, users intentionally making contributions that violate Wikipedia policies, would in our model be assigned very low reputation scores (since their contributions will be soon and completely made undone). In consequence, their future contributions, if any, will also have a very high probability to be made undone. This would correspond to "punishment" for norm violations considered in Piskorski and Gorbatâi (2017). On the other hand, our models do not assess whether those who punish norm violators are "rewarded" by the community. Furthermore, we do not assess whether network density has an impact on the incidence of norm violations, or on the incidence of receiving punishment for such violations.

Controversial articles in Wikipedia We analyze all articles in the English-language edition of Wikipedia that are labeled as controversial. For identifying these articles we started with all articles linked from the "List of controversial issues.[4]" According to the definition given on that page, controversial articles are those that "regularly become biased" or that "are likely to suffer future disputes." Thus, in order to qualify as a controversial article it is not enough to experience a dispute among editors a single time. Rather, these controversial articles are about subjects of a "divisive nature [...] reflecting the debates of society as a whole." Some of the articles linked from the page of controversial issues redirect to other articles. We followed the redirects, if necessary over several steps, and included the redirect target in the list of articles (the articles that are redirects are discarded since redirects are no longer edited). Through this procedure we obtained the list of 1,206 *controversial articles* that we analyze. We extracted the complete edit histories of all controversial articles from the Wikipedia database dump[5] from January 1st, 2017. From these histories we compute, separately for each article, the article's edit network as described in Sect. 4.1.

Controversial articles tend to have much longer histories and many more contributors than articles in general. The mean number of edits of controversial articles is 3,416, the median number is 2,525, the mean number of contributors is 1,852 and the median number of contributors is 1,477. By comparison, averaging over all 5.4 million Wikipedia articles gives a mean number of 86 edits, a median number of 28 edits, a mean number of 42 contributors, and a median number of 16 contributors. Clearly, controversial articles attract more attention by Wikipedia users. This also means that for most controversial articles (aside from a few exceptions reported later) we have enough data to fit relational event models separately to their edit event networks.

[4] `https://en.wikipedia.org/wiki/Wikipedia:List_of_controversial_issues`
[5] `https://dumps.wikimedia.org/`

4 Data, variables, and models

4.1 The edit event network

For each controversial article separately, we compute a network of signed and weighted relational events, denoted as the *edit network* (Brandes et al. 2009a), by successively comparing the text of every revision with the text of the previous one. Obtaining fine-grained information about user interaction from comparing the text of successive revisions is quite established in Wikipedia research (Adler and de Alfaro 2007; Brandes et al. 2009a; Javanmardi et al. 2010; Maniu et al. 2011; Flöck and Acosta 2014; Lerner and Lomi 2017), and is claimed to give a more complete picture of the collaborative editing than considering only complete reverts. To compute edit events we use the method described in Lerner and Lomi (2017), with slight modifications; a short general description of this method and the differences from prior research is given below.

The *history* of a Wikipedia article is the sequence of its revisions $r_1, \ldots, r_N$, in increasing order by time, where each revision codes the user uploading it and the text of the article after the edit has been done. The nodes of the edit network associated with a given Wikipedia article after revision r_i are the users uploading at least one of the revisions $r_1, \ldots, r_i$. Thus the number of nodes increases over time. These actors are connected by relational events – resulting from text modifications – of the form

$$(t, A, B, x, w) \ ,$$

where t is the time of the edit in which the text modification takes place (in our study the time is taken as the revision number), A is the source actor of the event (i. e., the user who uploads the revision at time t), B is the target actor of the event (i. e., the user whose previous edits get modified at time t) which is by definition different from A,[6] x is the event type, and w is the event weight (these two latter components are explained below). We note that uploading one revision of the article by user A can, in general, generate several dyadic events linking A to different target actors $B_1, \ldots, B_k$ by events of different types and weights.

Considering two of the three types of events proposed by Lerner and Lomi (2017), we define that the type x of the event can be either *undo* or *redo*.[7] An event of type undo from A to B encodes that A makes edits of B undone. Such an event can result from three kinds of text modifications: (1) A deletes text that has been authored by B, (2) A restores text that has been previously deleted by B, or (3) A deletes text that has been previously restored by B. The weight w of the undo event indicates how much (operationalized by counting the number of modified words) of B's contribution are made undone in the edit. Events of type undo are considered as negative events revealing that user A disagrees with (some of) the contributions of user B. An event of type redo from A to B encodes that A re-does edits of B that have been made undone before. Such an event can result from three kinds of text modifications: (1) A restores text that has been authored by B (and has been deleted in the meanwhile), (2) A deletes text that has been previously deleted by B (and has been restored in the meanwhile), or (3) A restores text that has been previously restored by B (and has been re-deleted in the meanwhile). Similar as for undo events, the weight w of the redo event indicates how much (operationalized by counting the number of words) of B's contribution are redone in the edit. Events of type redo are considered as positive events that reveal that user A agrees with (some of) the contributions of user B.

We also write $undo_i(A, B)$ for the weight of the undo event from A to B at revision r_i and $redo_i(A, B)$ for the weight of the redo event from A to B at revision r_i. The variables $undo_i(A, B)$ [or $redo_i(A, B)$] are defined to be zero if there is no undo [redo] event from A to B at revision r_i.

Weights of events of the same type are added dyadwise over the history of the article to encode past interaction among actors. Following Lerner and Lomi (2017) we define the *cumulative undo* and *cumulative*

[6] We ignore events where users are interacting with their own previous edits.

[7] In this paper, we do not consider events of type *third-party edit* that have been proposed by Lerner and Lomi (2017).

redo after revision r_i by[8]

$$\begin{aligned} cum.undo_i(A,B) &= \sum_{j=1}^{i} undo_j(A,B) \\ cum.redo_i(A,B) &= \sum_{j=1}^{i} redo_j(A,B) \ . \end{aligned}$$

We note that there is an implicit additional type of positive evaluation of others' contributions: if, say, B contributed some text and A, who performs a subsequent edit, does not delete this text, then this can also be interpreted as a positive evaluation of B's contributions by A. In the models that we introduce below it is indeed considered as an observation if A could delete some text of B but decides not to do so. Thus, an event that could have happened, but did not, constitutes an observation (compare Butts (2008); Butts and Marcum (2017)). A similar remark applies to observations of non-events where an actor A could redo contributions of B but decides not to do so.

The discussion in the preceding paragraph makes it clear that we have to keep track not only of the events that actually happen but also of the possibility for such events. Such variables have already been proposed in Lerner and Lomi (2017) under the names of *potential undo*, denoted by $pot.undo_i(B)$, which is the maximal weight of an undo event that *could be* received by B at revision r_i and *potential redo*, denoted by $pot.redo_i(B)$, which is the maximal weight of a redo event that *could be* received by B at revision r_i.

Here we introduce the first small but important addition to the method proposed in Lerner and Lomi (2017): additionally to the potential undo and potential redo mentioned above, we also maintain event potentials with a decay. To get an idea of the necessity of this potential with a decay consider the following example. Assume that a Wikipedia user B makes a contribution to a controversial article in the early days of Wikipedia by adding some new text. Some other user considers B's contribution as being of low quality and deletes it in the next revision. This means that B's text could be restored in any subsequent revision—until it is actually restored (the potential redo with target user B is equal to the number of words of the deleted text that could still be restored). Assume, however, that B's text never gets restored in the next few years. Without a decay function, our models would still seek to explain after thousands of intermediate revisions why B's text does not get restored, although it could. There are two negative implications of this: first, such a model would be simply implausible and, second, it would be time- and space-consuming since this approach generates an enormous number of observed non-events. (The latter argument is indeed crucial when analyzing controversial articles that have up to tens of thousands of revisions and thousands of contributors.)

We define for a decay factor α (a real number between zero and one), in addition to the un-scaled event potentials, the α-scaled potentials $pot.undo_i^{(\alpha)}(B)$ and $pot.redo_i^{(\alpha)}(B)$. When going from revision r_{i-1} to r_i, we first multiply the previous scaled potential by α and then update if some specific text editing events happen in revision r_{i-1}. (Note that the updates defining these potentials at revision r_i indeed depend on edits in the previous revision r_{i-1}.) Concretely, the variables $pot.undo_1^{(\alpha)}(B)$ and $pot.redo_1^{(\alpha)}(B)$ are initialized with zero (indicating that no undo or redo event is possible at the first revision). Then the values for $i > 1$ are iteratively defined by the equations

$$\begin{aligned} pot.undo_i^{(\alpha)}(B) &= \alpha \cdot \left[pot.undo_{i-1}^{(\alpha)}(B) - decrease.pot.undo_{i-1}^{(\alpha)}(B) \right] + increase.pot.undo_{i-1}(B) \\ pot.redo_i^{(\alpha)}(B) &= \alpha \cdot \left[pot.redo_{i-1}^{(\alpha)}(B) - decrease.pot.redo_{i-1}^{(\alpha)}(B) \right] + increase.pot.redo_{i-1}(B) \ . \end{aligned}$$

In the notation above $increase.pot.undo_{i-1}(B)$ is the number of words added by B in revision r_{i-1}, plus the number of words of which B is the author that are restored in revision r_{i-1}, plus the number of words deleted

[8] Past interaction relative to time point i, such as $cum.undo_i$ and $cum.redo_i$, is defined as a function of the event history derived from revisions $r_1, \ldots, r_i$. For readability we do not write this history in the argument of the function. This comment applies accordingly to all subsequent functions encoding aspects of past interaction.

by B in revision r_{i-1}, plus the number of words restored by B in revision r_{i-1} and $decrease.pot.undo_{i-1}^{(\alpha)}(B)$ is given by

$$decrease.pot.undo_{i-1}^{(\alpha)}(B) = undo_{i-1}(A,B) \cdot \frac{pot.undo_{i-1}^{(\alpha)}(B)}{pot.undo_{i-1}(B)} \ ,$$

where A is the user uploading revision r_{i-1}. Similarly, the value $increase.pot.redo_{i-1}(B)$ is the number of words of which B is the author that are deleted in revision r_{i-1}, plus the number of words of which B is the last deleter that are restored in revision r_{i-1}, plus the number of words of which B is the last restorer that are deleted in revision r_{i-1} and $decrease.pot.redo_{i-1}^{(\alpha)}(B)$ is given by

$$decrease.pot.redo_{i-1}^{(\alpha)}(B) = redo_{i-1}(A,B) \cdot \frac{pot.redo_{i-1}^{(\alpha)}(B)}{pot.redo_{i-1}(B)} \ ,$$

where A is the user uploading revision r_{i-1}.

When the re-scaled potentials are smaller than one after these update steps, we round them down to zero. In our analysis we use a decay factor of $\alpha = 0.8$.

An example might help to understand these α-scaled potentials. Assume that B adds 100 words in revision r_{i-1}, that B did not contribute anything else, and that $\alpha = 0.8$. Then, assuming that this text gets never deleted, we have

$$\begin{aligned}
pot.undo_{i}^{(\alpha)}(B) &= 100 \\
pot.undo_{i+1}^{(\alpha)}(B) &= 0.8 \cdot 100 = 80 \\
pot.undo_{i+2}^{(\alpha)}(B) &= (0.8)^2 \cdot 100 = 64 \\
&\cdots \\
pot.undo_{i+k}^{(\alpha)}(B) &= (0.8)^k \cdot 100
\end{aligned}$$

where the last equation holds as long as $(0.8)^k \cdot 100 \geq 1$, which holds for $k < 21$. If these 100 words survive 21 subsequent revisions, the α-scaled potential drops below one and, thus, gets rounded down to zero. This has the effect that we discard any future deletion of this text that might happen later. Thus, our models explain only modifications of text that has been edited not too long ago in the past.

Continuing with our example, assume that B adds 100 words in revision r_{i-1}, that B did not contribute anything else before or after, and that $\alpha = 0.8$. Further, assume that 50 words of this text get deleted in r_{i+1} and 20 of these deleted words get restored in r_{i+2}. Then we have

$$\begin{aligned}
pot.undo_{i}^{(\alpha)}(B) &= 100 \\
pot.undo_{i+1}^{(\alpha)}(B) &= 0.8 \cdot 100 = 80 \\
pot.undo_{i+2}^{(\alpha)}(B) &= 0.8 \cdot \left[80 - 50 \cdot \frac{80}{100}\right] = 32 \\
pot.undo_{i+3}^{(\alpha)}(B) &= 0.8 \cdot 32 + 20 = 45.6
\end{aligned}$$

Note that the deletion of 50 words in the example above meant that half of the actual text has been deleted. However, at the time of deletion, this text had a weight of only 80 word units. Thus, the deletion was treated as the deletion of 40 words out of 80.

4.2 Representing reputation

We propose a time-varying variable for user *reputation*. This variable is one of the two ways to operationalize tests for the reputation hypothesis (the other way is via signed indegrees, as we explain below). Intuitively, reputation encodes the share of a user's contributions that remains stable, that is, those contributions that

are not made undone. It therefore measures to what extent the user's contributions have an impact on the content of the article. Considering that the production of encyclopedic articles is the core purpose of Wikipedia, it is of utmost importance to know who shapes their content. The reputation variable that we define in this paper is similar to, but not identical, with reputation scores defined in Adler and de Alfaro (2007); Javanmardi et al. (2010). The precise definition makes use of the decay mechanism introduced above. Thus, if B's contributions are quickly undone, this will decrease the reputation of B more than if they survive some time and are only undone after several intermediate revisions.

To define the reputation variable we add through the history of a Wikipedia article for each user B the values of the α-scaled potential undo variable, resulting in the α-scaled cummulative potential undo defined by

$$cum.pot.undo_i^{(\alpha)}(B) = \sum_{j=1}^{i} pot.undo_j^{(\alpha)}(B) \ .$$

The value of $cum.pot.undo_i^{(\alpha)}(B)$ is a measure how much of B's contributions could have been made undone in revisions $r_1, \ldots, r_i$, taking into account the decay mechanism introduced above. Some of these contributions that can potentially be undone might be actually undone in one of these revisions. Again we have to scale these actual undo events appropriately. More precisely we define the α-scaled weight of an undo event from A to B at revision r_j by

$$undo_j^{(\alpha)}(A,B) = undo_j(A,B) \cdot \frac{pot.undo_j^{(\alpha)}(B)}{pot.undo_j(B)} \ .$$

Then we define the α-scaled cummulative undo received by B in revisions $r_1, \ldots, r_i$ as

$$cum.undo_i^{(\alpha)}(B) = \sum_{j=1}^{i} undo_j^{(\alpha)}(A_j, B) \ ,$$

where A_j is the actor uploading revision r_j.

Finally, the reputation of user B after revision r_i is defined to be the fraction of B's contribution that has not been made undone. In formulas, we define

$$reputation_i(B) = \frac{cum.pot.undo_i^{(\alpha)}(B) - cum.undo_i^{(\alpha)}(B)}{cum.pot.undo_i^{(\alpha)}(B)} \ ,$$

where we resolve 0/0 to be equal to zero. (Note that users with $cum.pot.undo_i^{(\alpha)}(B) = 0$ have never made any contributions so far.) In the equation above, the numerator is the total amount of B's contributions that could have been made undone minus the total amount of B's contributions that has been made undone, giving the surviving amount of B's contributions. This quantity is divided by the total amount of B's contributions that could have been made undone, giving the share of B's contributions that survive.

4.3 Relational event models for edit networks

Model overview. A user A, who uploads the next revision r_i of a given Wikipedia article, can undo $[pot.undo_i(B)]$ many words of every other user B and A can redo $[pot.redo_i(B)]$ many words. While editing, A will decide to undo or redo a certain share of these potential undos or redos. Our models, which are modifications of the model proposed by Lerner and Lomi (2017), specify these dyadic probabilities with logit models. That is, the observed dyadwise ratios of words that are undone or redone in revision r_i

$$\begin{aligned} ratio.undo_i(A,B) &= undo_i(A,B)/pot.undo_i(B) \\ ratio.redo_i(A,B) &= redo_i(A,B)/pot.redo_i(B) \ , \end{aligned}$$

are explained by latent probabilities specified by logistic regression models:

$$prob.undo_i(A,B) = \text{logit}^{-1}\left(\sum_\ell \theta_\ell^{(undo)} \cdot s_\ell(i-1;A,B)\right) \quad (1)$$

$$prob.redo_i(A,B) = \text{logit}^{-1}\left(\sum_\ell \theta_\ell^{(redo)} \cdot s_\ell(i-1;A,B)\right) \ . \quad (2)$$

In the equations above, the *statistics* $s_\ell(i-1;A,B)$ are real values describing how the dyad (A,B) is embedded in the network of past interaction after the $i-1$'th revision and the real values $\theta_\ell^{(undo)}$ (respectively $\theta_\ell^{(redo)}$) are the parameters of the undo model (redo model). A high value for $prob.undo_i(A,B)$ indicates that A tends to have a rather negative opinion towards B's contributions. Thus, a positive (negative) estimated parameter $\theta_\ell^{(undo)}$ associated with a statistic s_ℓ reveals that the higher the value of $s_\ell(i-1;A,B)$ the more (less) likely is A to undo contributions of B, so that a high value of $s_\ell(i-1;A,B)$ implies a more (less) negative attitude of A towards B's contributions. Likewise, a high value for $prob.redo_i(A,B)$ indicates that A tends to have a rather positive opinion towards B's contributions. Thus, a positive (negative) estimated parameter $\theta_\ell^{(redo)}$ associated with a statistic s_ℓ reveals that the higher the value of $s_\ell(i-1;A,B)$ the more (less) likely is A to redo contributions of B, so that a high value of $s_\ell(i-1;A,B)$ implies a more (less) positive attitude of A towards B's contributions.

Conditioning on the active user. An important aspect of our model is that we condition on the users uploading revisions. More precisely, we do not model the activity rates of users ("who is the next user who edits the article?") but, given that some user A is the one who uploads the next revision r_i, we model the probabilities $prob.undo_i(A,B)$ and $prob.redo_i(A,B)$ for every potential target user B different from A. On the other hand, we do not condition on the event targets. Our choice for conditioning, thus, lies between the model proposed by Butts (2008) (where both, the source and target of events, are explained by the model) and the "conditional event type" part of the models from Brandes et al. (2009b); Lerner et al. (2013) (where the event type is specified conditional on the observed source *and* the observed target). Models for chosing the target of links are also part of the actor-oriented models proposed by Snijders (2005); Stadtfeld (2010); Stadtfeld and Geyer-Schulz (2011); Stadtfeld et al. (2017); Stadtfeld and Block (2017). See Lerner (2016) for further discussion on the implications of conditioning on the presence of dyadic interaction, when modeling signed networks.

Note that a user A can send an edit event towards several different users $B_1,\ldots,B_k$ in one revision. In our models we assume that these different simultaneous dyadic observations are conditionally independent, given the network of previous interaction. (Note that dyadic events are modeled as a function of events that happen earlier – not on events that originate from the same revision and therefore happen simultaneously.) In a more sophisticated approach we might model mutual dependencies between such simultaneous dyadic observations in a similar way as dyadic dependencies are modeled in exponential random graph models (Lusher et al. 2013). We do not consider this in our paper, and assume instead conditional independence of simultaneous observations, in order to deal with the size and number of the event networks at hand.

The weight of observations. When estimating the probabilities $prob.undo_i(A,B)$ and $prob.redo_i(A,B)$ we down-weight the impact of editing events that modify older text using the decay mechanism described above. This is best explained by referring to the concrete software and function that we apply for estimating logistic regression models which is the `glm` function of the R environment for statistical computing (R Core Team 2017). The response variable of a dyadic observation is the observed ratio of modified words (either $ratio.undo_i(A,B)$ or $ratio.redo_i(A,B)$) and the information that these ratios come from varying numbers of word-unit observations can be specified with the `weights` argument of `glm`. If, for instance, user B adds 100 words to an article and in the next revision user A decides to delete 30 of them, then the response variable is 0.3 and the weight of this observation is 100. (Such an observation is equivalent to observing 100 binomial experiments, 30 of which yield '1' and 70 of which yield '0'.) If, however, A deletes the 30 words not in the

revision immediately following the addition of text by B but one revision later, then we give this observation a weight of $\alpha \cdot 100$ (the response variable is still 0.3). With every additional intermediate revision between adding and deleting text, we decay the weight of this observation by a factor α and if less than one word-unit is left, then we round the scaled undo-potential down to zero so that we will obtain no further observation from this text. In other words, we take the α-scaled potentials $pot.undo_i^{(\alpha)}(B)$, respectively $pot.redo_i^{(\alpha)}(B)$, as the observation weight for the response variable $ratio.undo_i(A, B)$, respectively $ratio.redo_i(A, B)$. Note that it is an observation if user A could delete, say, 100 words, added by user B in the preceding revision, but decides to delete none of them. From such a *non-event* we obtain an observation with response variable 0.0 and weight 100.

Dealing with non-event dyads: case-control sampling. If revision r_i is done by user A, then, in the undo model, we get observations for all users $B \neq A$ for which $pot.undo_i^{(\alpha)}(B)$ is positive. (A similar point can be made for the redo model). Most of these observations are non-events, that is, user A does not undo any contribution of most potential target users. Since our networks are rather large and since we are analyzing more than a thousand networks, computing and storing explanatory variables for all these observations would be too costly.

A solution is proposed by Borgan et al. (1995) who observe in the context of sampling subjects for an epidemiologic study: "*Loosely, if the disease of interest is rare, the contribution of the nonfailures, in terms of the statistical power of the study, will be negligible compared to that of the failures. Thus cohort sampling methods which include all the failures and a portion of the nonfailures are highly desirable.*" (Borgan et al. 1995, p. 1750) In our case, 'failures' translates to dyads on which we observe some interaction (undo or redo) and 'nonfailures' translates to dyads on which we observe no interaction but where there could have been interaction. Thus, we will apply case-control sampling in our analysis where for every dyad on which we observe an event, we sample a certain number (see below) of dyads on which no event happened—but could have happened—and estimate the model on these observations. Thus, dyads on which an event happens will always be included in the analysis but we will include only a given number out of all non-event dyads. Case-control sampling in the context of relational event models has been proposed by Vu et al. (2015). Earlier, Butts (2008) proposed to approximate terms over the "support set" by sampling uniformly from this set.

The question of how many non-events should be sampled for every observed event (the number $m - 1$ from Borgan et al. 1995) is difficult to answer theoretically. We first make the choice that we sample a certain multiple (denoted by c and called the *CCS factor*[9]) of the number of events *plus one*. We computed explanatory variables and estimated models for a single network (stemming from the article `Global warming`, which is one of the articles with the largest number of revisions in our sample) with a CCS factor c ranging from one to 20 in increments of one. With few exceptions, our findings on this single article do not vary qualitatively for c equal to five or higher. (These results are not reported in this paper.) Therefore, in this paper we use a CCS factor c equal to five. To provide a concrete example, if a revision generates events on 10 dyads, then we include in our analysis these 10 dyads, plus $5 \cdot (10 + 1)$ randomly selected dyads on which no event happened but could have happened.

Explanatory variables: dyadic event statistics. The statistics $s_\ell(i; A, B)$ used in the specification of the undo probability (1) and in the specification of the redo probability (2) include variables that are related to our research questions and variables that control for other likely network effects which are not of primary interest in our study. Note that variables with first argument i (encoding past interaction after revision r_i) are used to estimate undo and redo probabilities at revision r_{i+1}.

We use the number of nodes in the network as a control variable. The statistic $number.of.nodes(i; A, B)$, thus gives the number of nodes in the network at the time of the i't revision; it is independent of the source A and the target B of the dyad. It could be expected that if the number of actors increases then the probability that a given actor interacts with a single given target actor decreases, so that we would obtain a negative parameter associated with $number.of.nodes$. However, since case-control sampling tends to stabilize density

[9] CCS for case-control sampling

of the analyzed observations independent of the number of null dyads, this effect might be less strong or might even go in the other direction.

The reputation variable defined in Sect. 4.2 is used to define two statistics:

$$\begin{aligned} reputation.of.source(i;A,B) &= reputation_i(A) \\ reputation.of.target(i;A,B) &= reputation_i(B) \ . \end{aligned}$$

The reputation hypothesis predicts that the parameter associated with *reputation.of.target* is negative in the undo model (contributions of more reputable users should not be undone) and positive in the redo model (contributions of more reputable users should be redone if they have been undone previously). It makes no predictions for *reputation.of.source*.

Our models further include a list of 16 statistics, dependent on the network of past interaction, which introduce hypothetical dyadic effects, degree effects, and triadic effects. These statistics are common in the literature on relational event models; specifically for signed relational events see, for instance, Brandes et al. (2009b).

Dyadic network effects are introduced by the four statistics

$$\begin{aligned} undo.repetition(i;A,B) &= cum.undo_i(A,B) \\ undo.reciprocation(i;A,B) &= cum.undo_i(B,A) \\ redo.repetition(i;A,B) &= cum.redo_i(A,B) \\ redo.reciprocation(i;A,B) &= cum.redo_i(B,A) \ . \end{aligned}$$

In the undo model balance theory predicts a positive parameter for *undo.repetition* and for *undo.reciprocation* and a negative parameter for *redo.repetition* and for *redo.reciprocation*. In the redo model balance theory reverses these predictions: a negative parameter for *undo.repetition* and for *undo.reciprocation* and a positive parameter for *redo.repetition* and for *redo.reciprocation*. The reputation hypothesis makes the same predictions for the repetition statistics (since actors should evaluate others consistently) but does not make any predictions for the reciprocation statistics. (Status theory predicts that the reciprocation statistics will have the opposite sign compared to balance theory.)

We define eight different statistics for degree effects obtained by varying the node for which we compute the degree (source or target of the next event), the direction of past events (in or out), and the sign of past events (positive or negative, that is, based on past redo or undo). In the formulas below, the summation index (A' or B') runs over all actors in the network at revision r_i.

$$\begin{aligned} undo.outdegree.source(i;A,B) &= \sum_{B' \neq A} cum.undo_i(A,B') \\ undo.indegree.source(i;A,B) &= \sum_{B' \neq A} cum.undo_i(B',A) \\ undo.outdegree.target(i;A,B) &= \sum_{A' \neq B} cum.undo_i(B,A') \\ undo.indegree.target(i;A,B) &= \sum_{A' \neq B} cum.undo_i(A',B) \\ redo.outdegree.source(i;A,B) &= \sum_{B' \neq A} cum.redo_i(A,B') \\ redo.indegree.source(i;A,B) &= \sum_{B' \neq A} cum.redo_i(B',A) \\ redo.outdegree.target(i;A,B) &= \sum_{A' \neq B} cum.redo_i(B,A') \\ redo.indegree.target(i;A,B) &= \sum_{A' \neq B} cum.redo_i(A',B) \ . \end{aligned}$$

Balance theory makes no specific prediction about degree effects. The reputation hypothesis predicts parameter signs for the indegree of target statistics in the sense that other users should consistently evaluate the target user. More specifically, in the undo model it predicts a positive parameter for *undo.indegree.target* and a negative parameter for *redo.indegree.target*. In the redo model it predicts a negative parameter for *undo.indegree.target* and a positive parameter for *redo.indegree.target*. The other degree effects serve mostly as control variables. Among others, they account for varying activity and popularity with respect to positive and negative events.

We define four statistics for triadic effects, obtained by varying the signs of past events connecting the source and target to common third actors. Since the predictions of balance theory do not depend on the direction of relations we use symmetrized weights: $sym.undo_i(A,B) = cum.undo_i(A,B) + cum.undo_i(B,A)$ and $sym.redo_i(A,B) = cum.redo_i(A,B) + cum.redo_i(B,A)$. With this notation we define (letting the summation index C run over all actors in the network at revision r_i)

$$\begin{aligned}
friend.of.friend(i;A,B) &= \sqrt{\sum_{C\neq A,B} sym.redo_i(A,C)\cdot sym.redo_i(C,B)} \\
friend.of.enemy(i;A,B) &= \sqrt{\sum_{C\neq A,B} sym.undo_i(A,C)\cdot sym.redo_i(C,B)} \\
enemy.of.friend(i;A,B) &= \sqrt{\sum_{C\neq A,B} sym.redo_i(A,C)\cdot sym.undo_i(C,B)} \\
enemy.of.enemy(i;A,B) &= \sqrt{\sum_{C\neq A,B} sym.undo_i(A,C)\cdot sym.undo_i(C,B)}\ .
\end{aligned}$$

Defining the structural balance statistics on symmetrized tie weights and scaling sums over signed two-paths by the square root has been proposed by Brandes et al. (2009b). Balance theory predicts for the undo model a negative parameter for $friend.of.friend$ and for *enemy.of.enemy* and a positive parameter for $friend.of.enemy$ and for *enemy.of.friend*. It further predicts for the redo model a positive parameter for $friend.of.friend$ and for *enemy.of.enemy* and a negative parameter for $friend.of.enemy$ and for *enemy.of.friend*. The reputation hypothesis makes no predictions for these triadic statistics.

Due to a skewed distribution of edge weights we apply the transformation $s \mapsto \log(1+s)$ to every statistic described above except to the intercept and to the two reputation statistics. Afterwards we standardize every statistic (except the intercept) by subtracting its mean and dividing by its standard deviation. This normalization allows some comparison of the effect size of the different variables: hypothetically increasing a statistic $s_\ell(i;A,B)$ by one standard deviation multiplies the estimated odds ratio $prob.undo_i(A,B)/[1 - prob.undo_i(A,B)]$ of the undo probability by $\exp(\theta_\ell^{(undo)})$; respectively it multiplies the estimated odds ratio $prob.redo_i(A,B)/[1 - prob.redo_i(A,B)]$ of the redo probability by $\exp(\theta_\ell^{(redo)})$.

5 Results

In this section we report results from estimating the undo model and the redo model separately for each of the 1,206 controversial articles. The estimation of the undo-model did not converge on the networks of 47 out of 1,206 controversial articles and for the redo-model we had non-convergence on 43 of 1,206 networks. The networks where the estimation did not converge are those that result from controversial articles with relatively short histories, where we do not have enough observations to identify the various effects. They are excluded in the results presented in this section.

5.1 The undo model

For every effect, the undo model has a potentially different parameter estimate for each of the 1,159 controversial articles for which the estimation converges. Table 1 summarizes these parameter estimates by

reporting, for each effect, the median z-value, the median parameter estimate, the number of networks yielding a parameter that is significantly negative at the 5% level, the number of networks yielding a significantly positive parameter, and the number of networks where the estimated parameter is not significantly different from zero at the 5% level. Checking the distributions (histograms) of the estimated parameters and estimated z-values (not provided in this paper) we see a clear concentration of estimates around the median. We do not observe a multi-modal distribution with distinct clusters of values. Estimating parameters separatelly for each network allows us to determine whether specific effects are near-universal, typical, or rather change from network to network.

	median z-value	median estimate	#negative	#positive	#not sig.
(Intercept)	-377.42	-2.18	1136	6	17
number_of_users	11.29	0.05	445	686	28
reputation_of_source	-49.30	-0.31	964	163	32
reputation_of_target	-406.35	-0.79	1153	6	0
undo_repetition	31.28	0.10	261	865	33
undo_reciprocation	21.54	0.04	351	768	40
redo_repetition	-31.66	-0.19	965	107	87
redo_reciprocation	-43.01	-0.14	975	134	50
undo_outdegree_source	51.09	0.93	118	1028	13
undo_indegree_source	-32.46	-0.47	887	240	32
undo_outdegree_target	50.17	0.35	302	843	14
undo_indegree_target	40.12	0.22	272	863	24
redo_outdegree_source	18.80	0.25	347	768	44
redo_indegree_source	-8.03	-0.06	636	477	46
redo_outdegree_target	-31.01	-0.21	775	352	32
redo_indegree_target	-37.88	-0.21	854	270	35
friend_of_friend	-60.93	-0.31	982	136	41
friend_of_enemy	-11.07	-0.09	686	435	38
enemy_of_friend	72.20	0.35	164	966	29
enemy_of_enemy	-11.16	-0.08	684	428	47

Table 1: Summary of parameter estimates for the undo model: median z-value and median parameter estimate over all networks, number of networks yielding a parameter that is significantly negative (z-value < -1.96), respectively positive (z-value > 1.96) at the 5% level, and number of networks yielding a non-significant parameter. The total number of dyadic observations (including non-event dyads that have been selected by case-control sampling) used to estimate the undo models is 33,726,502.

Owing to the sparsity of the event networks, the estimated intercept is negative for most networks (1,136). Note that the intercept value depends on case-control sampling and is only of minor interest. The number of users has a slight tendency to yield a positive parameter (on 686 networks). Although this is contrary to intuition, it might be an outcome of the application of case-control sampling that keeps the density of observations used in the estimation relatively constant, independent of the number of users in the network.

The reputation of the source actor A has a decreasing effect on the undo probability $prob.undo(A, B)$ in 964 networks, while it is positive in only 163, and insignificant in 32. We have no hypothesis related to this statistic. More interesting—and in strong support of the reputation hypothesis—is the finding that the reputation of the target actor B has a decreasing effect on the undo probability $prob.undo(A, B)$ in almost all networks (1,153 yield a negative parameter, while it is positive in only 6 networks, and insignificant on none). The median of the estimates for this statistic is -0.79 which is one of the largest in absolute value of all effects (apart from the intercept and the median parameter associated with the undo outdegree of the source actor). Thus, we find that the reputation variable is one of the strongest, and most universal, factors in determining whether the contributions of an actor are preserved or undone. Obtaining this finding on collaboration networks stemming from *controversial* articles seems to be a good sign for Wikipedia: even

teams that have disputes do not completely forget to respect the reputation of users. It might be one explanation for why Wikipedia is so successful in writing high-quality articles on controversial topics.

Findings for the dyadic network effects (repetition and reciprocation), however, are in strong support of balance theory and suggest that two actors clearly categorize each other as either friends or enemies. More specifically, if more past undos happened in a dyad (A, B), then typically the future undo probability in the same dyad (A, B) is higher (*undo.repetition* yielding a positive parameter on 865 networks) and typically the future undo probability in the reverse dyad (B, A) is higher as well (*undo.reciprocation* yielding a positive parameter on 768 networks). The effects of past redos on the future undo probability reinforces this result. The more past redos happened in a dyad (A, B) the lower the future undo probability in the same dyad (*redo.repetition* being negative on 965 networks) and the lower the future undo probability in the reverse dyad (*redo.reciprocation* being negative on 975 networks). These results suggest that users who disagree while contributing to controversial Wikpedia articles are unlikely to ever reach consensus *dyadwise*—rather they continue to undo each other's work. This emphasizes the importance of third users evaluating the relative quality of the contributions of two quarreling users— a result also suggested in a recent study by Lerner and Lomi (2017). We note that the findings for both reciprocation variables contradict the predictions of status theory (Leskovec et al. 2010; Yap and Harrigan 2015) which claims that past negative interaction in (A, B) should decrease future negative interactions in the reverse dyad (B, A), while past positive interaction in (A, B) should increase future negative interactions in the reverse dyad (B, A).

Turning to the degree effects, the parameter of *outdegree.source* is positive on the vast majority of networks (1,028). This implies that users who undid many contributions in the past typically continue to undo others' contributions. This is a form of generalized repetition where past undo begets future undo by the same source actor that is not necessarily directed to the same target actor. Can we say that actors having a high undo-outdegree are therefore high-status actors? The result that a higher outdegree of an actor B also leads to a higher future undo probability on dyads with *target* B (the parameter associated with *undo.outdegree.target* is positive on 843 networks) suggests that this is not the case. A more plausible explanation is that users with a high undo-outdegree perform the specific role of watching others' contributions and undo those they think are of low quality. Sometimes other users disagree with that assessment and undo the undo operation, making these watchers the target of many undo events.

The parameter associated with *undo.indegree.source* is negative on 887 networks. This suggests that users receiving many undo events (that is, users who are corrected by others) seem not to have the tendency to retaliate against the whole team. This result has to be considered jointly with the positive effect of undo reciprocation: users do have the tendency to fight back if their contributions are undone but, importantly, they just fight the particular user who undid their work—not other team members. Thus, we find a tendency against generalized reciprocation of undo events (although we do find dyadic reciprocation).

The parameter associated with the undo-indegree of the target is positive on 863 networks. This means that, typically, actors who have seen a lot of their work undone in the past are likely to see it undone again in the future. Thus, teams seem to agree, in general, about the users responsible for low-quality contributions. This finding provides further support to the reputation hypothesis.

The redo-outdegree of a source actor A has an increasing effect on the undo probability on dyads (A, B) for any target actor B on 768 networks. This seems to document an activity effect across network signs: those who initiated a lot of positive events in the past tend to undo many contributions in the future. We find such activity effects also restricted negative events (parameter of *undo.outdegree.source* in the undo model, described above), restricted to positive events (parameter of *redo.outdegree.source* in the redo model, described below), and for past negative events inducing future positive events (parameter of *undo.outdegree.source* in the redo model, described below). Thus, we find that outdegrees are indicators of activity levels or self-assigned roles, rather than of high or low status: those users that have high positive or negative outdegree are more active in modifying edits of others, positively and negatively.

Users who see their work redone tend to perform fewer undo events on a (small) majority of networks (*redo.indegree.source* yielding a negative parameter on 636 networks). Users who initiate many redo events are less likely to receive undo events on 775 networks (parameter of *redo.outdegree.target* being negative).

Concluding our discussion of degree effects, the parameter associated with *redo.indegree.target* is neg-

ative on 854 networks suggesting that actors who are positively evaluated in the past are less likely to be negatively evaluated, by potentially other users, in the future. Thus, the community tends to agree, in general, on the users responsible for high-quality contributions. This result provides further support for the reputation hypothesis.

Turning to the triadic effects, we observe that three out of four effects support the predictions of balance theory. Users working on controversial articles have a tendency not to undo contributions of friends of friends on 982 networks. More specifically, this means that if two users A and B exchanged frequent past redo events with one or several common third actors, then the future undo probability on the dyad (A, B) is lower. Thus, friends of friends tend not to be enemies, as predicted by balance theory. Users have a tendency to undo contributions of enemies of friends on 966 networks. Thus if actor B exchanged frequent undo events with a third actor C and the same C, in turn, exchanged frequent redo events with actor A, then A is more likely to undo contributions from B. Thus, enemies of friends tend to be enemies, as predicted by balance theory. However, the same pre-condition leads to a decreased undo probability on the reverse dyad (B, A) on a (smaller) majority of 686 networks. Thus, users have a tendency not to consider the friends of enemies as enemies—contrary to the predictions of balance theory. Finally, users working on controversial articles have a tendency not to undo contributions of enemies of enemies on a small majority of 684 networks. Thus, enemies of enemies tend not to be enemies, supporting the predictions of balance theory.

5.2 The redo model

For every effect in the redo model we have a potentially different parameter estimate for each one of the 1,163 controversial articles for which the estimation converges. We summarize these parameter estimates by reporting in Table. 2, for each effect, the median z-value , the median parameter estimate, the number of networks yielding a parameter that is significantly negative at the 5% level, the number of networks yielding a significantly positive parameter, and the number of networks where the estimated parameter is not significantly different from zero at the 5% level. We checked the distributions (histograms) of the estimated parameters and z-values for each effect (not reported in this paper). Similarly to the undo model, the estimates of the redo model show a clear concentration around the median.

The intercept of the redo model is estimated to be negative in 908 networks, expressing that redo probabilities are typically below 0.5. The number of users has a decreasing effect on the redo probability on a small majority of 686 networks.

The reputation of the source actor A decreases the redo probability in any dyad (A, B) in 881 networks. Since we found that the reputation of A also decreases the undo probability in dyads (A, B), as discussed above, is seems that actors with higher reputation are less active in modifying the contributions of others, positively or negatively. They might instead be actors that focus on providing new (high-quality) content. The effect of the reputation variable on the target user is more important for this study: in a vast majority of 1,135 networks we find that the higher the reputation of the target user B the higher the probability that contributions of B are redone, if they have been made undone previously. This effect seems to be very strong with a median parameter estimate of 1.46 (the largest in absolute value). Thus, we obtain again strong support for the reputation hypothesis: if the contributions of more reputable users are made undone, then they are more likely to be restored subsequently.

The dyadic effects in the redo model (repetition and reciprocation) go in the same direction as those in the undo model with the understanding that an increased undo probability implies a negative relationship, while an increased redo probability implies a positive relationship. More specifically, the more past undo happened in a dyad (A, B) the lower the redo probability in the same dyad (A, B) (we obtained this finding in a vast majority of 1127 networks) and the lower the redo probability in the reverse dyad (B, A) (we obtained this result in 784 networks). Thus, dyads exchanging negative events will not engage in positive interaction. The parameter associated with *redo.repetition* is positive in 936 networks and the parameter for *redo.reciprocation* is positive in 689 networks. Thus, dyads that exchanged positive events in the past will do so in the future. All four effects provide further support for the predictions of balance theory that dyadic interaction is either consistently positive in both directions or consistently negative in both directions. Thus,

	median z-value	median estimate	#negative	#positive	#not sig.
(Intercept)	-88.53	-0.79	908	227	28
number_of_users	-18.00	-0.08	686	443	34
reputation_of_source	-30.30	-0.26	881	257	25
reputation_of_target	357.49	1.46	20	1135	8
undo_repetition	-93.01	-0.46	1127	18	18
undo_reciprocation	-15.29	-0.07	784	308	71
redo_repetition	34.05	0.19	171	936	56
redo_reciprocation	8.62	0.04	380	689	94
undo_outdegree_source	12.47	0.34	415	697	51
undo_indegree_source	-54.45	-1.14	1023	108	32
undo_outdegree_target	-88.87	-0.88	970	171	22
undo_indegree_target	70.89	0.59	187	959	17
redo_outdegree_source	57.93	1.12	157	976	30
redo_indegree_source	23.01	0.37	280	823	60
redo_outdegree_target	84.95	0.73	197	940	26
redo_indegree_target	56.16	0.25	241	900	22
friend_of_friend	60.06	0.50	151	965	47
friend_of_enemy	-33.15	-0.25	930	175	58
enemy_of_friend	-58.68	-0.50	967	147	49
enemy_of_enemy	0.89	0.01	544	574	45

Table 2: Summary of parameter estimates for the redo model: median z-value and median parameter estimate over all networks, number of networks yielding a parameter that is significantly negative (z-value < -1.96), respectively positive (z-value > 1.96) at the 5% level, and number of networks yielding a non-significant parameter. The total number of dyadic observations (including non-event dyads that have been selected by case-control sampling) used to estimate the redo models is 35,777,697.

two users categorize each other clearly as either friends or enemies. As it was the case in the undo model, the findings on the reciprocation effects in the redo model contradict the predictions of status theory.

The parameter of *undo.outdegree.source* is positive in 697 networks, pointing to activity effects across network signs (actors who initiate more negative events also initiate more positive events). Actors receiving many negative events are less likely to redo work of others (parameter associated with *undo.indegree.source* being negative in 1,023 networks). Actors undoing a lot of work of others are less likely to receive redo events (*undo.outdegree.target* begin negative in 970 networks). More remarkable is the finding that users receiving many undo events are also more likely to receive redo events (the parameter associated with *undo.indegree.target* is positive in 959 networks). This finding, which is the only contradiction to the predictions of the reputation hypothesis, might be due to users whose contributions are highly disputed (that is, there is no agreement on the quality of their contributions). Some users tend to undo their contributions but others tend to redo them.

Considering the degrees based on past redo events, it is striking that all four redo-degree effects tend to increase the probability of future redo in most networks (between 823 and 976)—independent of whether we consider out-degrees or in-degrees and independent of whether we consider the degree of the source or of the target. This seems to point to generalized repetition and generalized reciprocation in the sub-network composed of positive relations. Of particular interest for this paper is the finding that *redo.indegree.target* is positive in 900 networks, indicating that users who receive positive events in the past are likely to do so in the future by potentially different source actors. Thus, other users seem to agree, in general, that these users make high-quality contributions. This is again consistent with the predictions of the reputation hypothesis.

Turning to the balance effects, we observe that no result contradicts the predictions of balance theory (even though the number of networks on which the effect of *enemy.of.enemy* is positive for the redo probability is only slightly higher than the number of those on which the effect is negative). More specifically,

friends of friends are more likely to exchange redo events in 965 networks, that is, they tend to be friends themselves. Users are less likely to redo contributions of the friends of their enemies in 930 networks and they are less likely to redo contributions of the enemies of their friends in 967 networks, that is, they tend to be enemies. These three effects support the predictions of balance theory. The effect of having a common enemy, however, might lead to an increase of the redo probability (positive parameter found in 574 networks, as predicted by balance theory) but might also lead to a decreased redo probability (negative parameter found in 544 networks, contrary to the predictions of balance theory).

5.3 Summary of findings

Taking results from the undo model and the redo model together, it is striking that out of 16 predictions made by balance theory (8 for dyadic effects and 8 for triadic effects), we observed only one single contradicting result (the negative effect of being a friend of an enemy in the undo model) and one finding that yields no clear direction (the effect of being an enemy of an enemy in the redo model). Thus, balance theory explains well the behavior of actors working on contentious tasks. In this context, actors seem to classify others into "friends" and "enemies" and this classification has consequences for the way users evaluate the contributions of others in the collective production of Wikipedia articles. Our findings are consistent with the view that observed undo and redo events can serve as signals for membership in the in-group or out-group. This perceived group membership, in turn, influences the probability of future undo or redo events.

However, the alternative reputation hypothesis – predicting that actors consistently evaluate the reputation of other users received equally strong support (with five out of six predictions receiving empirical support). Thus, no exclusive social mechanism accounts for the observed patterns of agreements and disagreement among Wikipedia contributors. But the *concatenation of mechanisms* that our models specify explain the data well (Gambetta 1998). It is useful to remark that these mechanisms involve both positive as well as negative relations. The findings related with the repuation hypothesis are consistent with the view that observed text editing events can serve as signals for the reputation of users (or for the typical quality of their contributions) and that this perceived reputation influences future editing behavior. Several of the predictions of status theory that go beyond the reputation hypothesis – among them the prediction that actors initiating many negative events have high status as well as the predicted anti-reciprocity of positive and negative events – are contradicted by our findings.

Some of the explanatory variables in our models revealed distinct user roles. For instance, we have found that users with high reputation (that is, users providing content recognized by others as being of high quality) are less likely to edit contributions of others – positively or negatively. Thus, highly reputable users are content providers and, in general, less involved in editing tasks. On the other hand, users with high positive or negative outdegree are more likely to modify contributions of others, positively or negatively, in the future. As a consequence, we did not find a distinction between "hostile" users (having a tendency to undo rather than redo) and "supportive" users (having a tendency to redo), but rather a distinction between editing-activity levels. Users that initiate many positive events, in general, also initiate many negative events and vice versa.

We also estimated joint undo and redo models on the aggregate observations from all 1,206 networks (these results are not reported in this paper). With few exceptions, findings of these joint models are consistent with those from estimating parameters separately for each of these networks. Indeed, for all but a small number of effects, the parameters of the joint model have the same sign as the majority of the respective parameters estimated separately. Moreover, to rule out that temporal heterogeneity distorts our results, we perform an additional analysis by spliting the aggregate observations into two sub-sequences, where we cut at the median event time. Then we fit the joint models separatelly to these subsequences (the results of this analysis are not reported in this paper). Our results do not change qualitatively: conclusions related with our hypotheses can be obtained from analyzing any of these sub-sequences, or their concatenation.

6 Conclusions

Active participants to the Wikipedia community contributing to an article about a controversial topic have to resolve their differences in the near absence of hierarchical conflict resolution mechanisms based on authority. While participants to the Wikipedia community might appeal to conflict resolution routines in clear and evident cases of misconduct, Wikipedia's own policies impose that content disputes be settled not by impersonal hierarchical rules,[10] but by participants themselves. How can the production of public goods be possible under conditions of latent conflict among peers that cannot be resolved by central authority?

Building on previous research, we argued that organized order emerges endogenously from the evolving network of positive and negative user interactions. This happens because individual acts of editing produce signals that participants interpret and take into account when evaluating contributions of others. We argued, further, that meaningful interaction happens through article co-editing activities since this process ultimately determines the articles' content and hence the quantity and quality of text collectively produced by the Wikipedia community. We analyzed which patterns in the signed event networks stemming from co-editing controversial articles determine the dyadic probabilities that particular users preserve, undo, or redo, the contributions of particular other users. We have established a systematic mapping of these various kinds of events onto signed (positive and negative) relations among contributing Wikipedia users. Our work demonstrates the usefulness – and in fact importance – of modeling both positive as well as negative relations. The concatenation of mechanisms that explain the observations could not be specified only in terms of positive relations. As Labianca (2014, p. 239) observes: "Most network research in organizations assumes away the dissociative forces instantiated in negative ties, instead pursuing ties that reflect only associative forces, to the detriment of understanding organizational networks." We have shown that keeping into account both positively as well as negatively signed interaction greatly expands our understanding of large-scale organizational phenomena.

Understanding how peer production might effectively work under conditions of decentralized collaboration and diffuse conflict is an issue as important as it is insufficiently understood. Indeed, how organizations deal with polarization and conflict, online and offline, is one of the most crucial questions posed by new media and new interactive technologies (Del Vicario et al. 2016; Saperstein 2004). In this paper we outlined alternative explanations based on fundamental relational principles.

One of the main empirical results of our study is that balance has a significant effect on the collective production of controversial articles: contributing users to contentious Wikipedia articles seem to clearly partition others into friends (those who have the same opinion on a given topic) and enemies (those who hold the opposite opinion) and decisions to undo or preserve work of others are influenced by these assignments. This is an important contribution to the literature on balance theory, and signed networks in general, where concerns have been raised that "*evidence for this theory is mixed, at very best*" (Yap and Harrigan 2015, p. 103). The behavior of Wikipedia contributors in the presence of controversy seems to be well-explained by balance theory.

The other main empirical result of our paper is that status theory, often proposed as an alternative to balance theory in explaining the evolution of signed networks (Leskovec et al. 2010), has to be significantly modified and restricted to receiving activities as suggested by alter-centric theories of status (Torlò and Lomi 2017). The reputation hypothesis, which leads to the subset of predictions of status theory that are only related to in-coming relations, posits that actors have different reputation (Adler and de Alfaro 2007; Javanmardi et al. 2010), and assumes that other actors consistently assess this reputation, and act accordingly, so that actors with higher reputation are less likely to have their work undone (i. e., less likely to receive negative events) and more likely to have their work redone, if deleted previously (i. e., more likely to receive positive events). Thus, the reputation hypothesis assumes that undo or redo events serve as signals for the reputation of (or typical quality of the text provided by) users. Those who receive many positive events and few negative events are considered as highly reputable by the community. This latent reputation

[10] For example, Wikipedia's policies on administrator conduct state that "*Administrators should not normally use their tools in matters in which they are personally involved (for example, in a content dispute in which they are a party).*" See `https://en.wikipedia.org/wiki/Wikipedia:Administrators#Misuse_of_administrative_tools`, accessed on 20th January, 2018.

score, in turn, influences future editing behavior so that more reputable users are less likely to have their contributions undone. In general, we found strong support for the reputation hypothesis. Thus, the two theories are not mutually exclusive, but might be jointly invoked to explain actor behavior when working on contentious tasks.

Our results also provide further evidence that pairs of Wikipedia users – once they start undoing each others work – are unlikely to settle their dispute at the dyadic level. Consistent with findings from Lerner and Lomi (2017), we conclude that extra-dyadic effects, that is by-standing third parties intervening in disputes among two quarreling users, are of utmost importance – at least when writing about controversial topics.

Besides these substantive insights, our paper also makes several methodological contributions to modeling networks of signed and weighted relational events. Three, in particular, deserve special mention. First we defined a decay mechanism whereby dyads gradually drop out of the risk set if the associated edits are not modified over a certain period. Second, we applied, and assessed the robustness of case-control sampling when estimating model parameters. Both of these contributions were necessary to deal with the large data size stemming from co-editing controversial articles. A third methodological contribution is the definition of the variable for user reputation, which turned out to be one of the strongest and most reliable factors explaining whether individual contributions are preserved or undone.

We claim that the general model for the Wikipedia edit network can also be specified to test predictions other than those implied by balance and reputation mechanisms. Since the ultimate goal of Wikipedia is the production of encyclopedic articles, modeling the probability that individual contributions to articles are preserved or undone is one of the most relevant aspects in understanding collaboration in this extraordinarily large and successful decentralized online community. Similar consideration would extend directly to different peer-productions like, for example, open source software.

6.1 Limitations and future work

One limitation of the analysis presented in this paper is that we generate weighted positive and negative events from the amount of modified text without considering the semantics of text. For instance, it is possible that a user deletes text contributed by another user and replaces it with different text expressing the same or similar ideas. Future work might use techniques from computational linguistics, such as *word embeddings* (Turian et al. 2010), to detect the similarity of old and new text which could then be used to re-weight undo or redo events accordingly.

We've already indicated before that the analysis in this paper is dependent on article writing but ignores discussions taking place on related talk pages in Wikipedia. Models for discussion on Wikipedia (e. g. Kaltenbrunner and Laniado 2012; Gómez et al. 2013) could in principle be combined with our model for the evolution of article content. In such a joint model, one could analyze network effects where editing might depend on discussion and vice versa. This could give a more complete picture of the collaborative article writing in Wikipedia.

We conditioned our models on the observed active users; that is, our models cannot predict who is going to edit which article and when. Given that a specific user is observed to contribute to a specific article at a given point in time, we modeled how this user is going to modify other's edits. It seems that models explaining user engagement in particular articles could be specified within the general relational event model framework (Butts 2008) – although the size of the complete user-article network in Wikipedia certainly poses computational challenges. In our work we analyzed each article separately; future work could go beyond this by taking into account that some users interact in co-editing more than one article. Such an approach would again involve a higher computational cost.

A further limitation of our work indicates a clear avenue for future research. The network effects in edit networks that our models have revealed could be related to the quality of the output, e. g., to the quality of the article produced by a team of Wikipedia contributors. Even though this relation has not been examined in our paper, we offer the conjecture that relational patterns consistent with the reputation hypothesis might affect positively the quality of the output of peer production. This may happen because the contributions of users with higher reputation have a higher probability to persist while contributions made by less reputable participants have a higher probability to be undone. Along a similar line of reasoning we conjecture that

network patterns consistent with balance theory might affect negatively the quality of the team output. This may happen because decisions about whose contributions are deleted or preserved might be based on membership of users to opinion groups, potentially ignoring the quality of contributions. Testing these hypothetical relations is clearly possible within the analytical framework that we have established, since community-based quality assessment of Wikipedia articles is readily available. However, such a test would require a different sample of articles, for instance, a set of high quality articles and a control set of comparable articles with lower quality.

Besides estimating parameters independently for each network and estimating a single parameter vector for the aggregate observations (where results of the latter estimation are not reported in this paper), an alternative approach would involve fitting a hierarchical relational event model to the observations from all networks (DuBois et al. 2013), also denoted as multilevel analysis of networks. In such a hierarchical model we would not require that parameters are identical across networks but could still propagate some information across networks. Importantly, we could also let network-level characteristics, such as the topic area of the article or its quality, influence the distribution of these parameters. Thus, we could analyze hypothetical causes and consequences of network-effect variation across articles. Such a hierarchical approach, however, implies a higher computational cost in model estimation, which is a crucial limitation in our context where we have to deal with relatively large data sets.

References

Adler, B. T. and de Alfaro, L. (2007), "A Content-Driven Reputation System for the Wikipedia," in *Proc. 16th Intl. Conf. WWW*, ACM, pp. 261–270.

Arazy, O., Nov, O., Patterson, R., and Yeo, L. (2011), "Information quality in Wikipedia: The effects of group composition and task conflict," *Journal of Management Information Systems*, 27, 71–98.

Barberá, P., Jost, J. T., Nagler, J., Tucker, J. A., and Bonneau, R. (2015), "Tweeting from left to right: Is online political communication more than an echo chamber?" *Psychological science*, 26, 1531–1542.

Benjamin, B. A. and Podolny, J. M. (1999), "Status, quality, and social order in the California wine industry," *Administrative science quarterly*, 44, 563–589.

Benkler, Y. and Nissenbaum, H. (2006), "Commons-based peer production and virtue," *Journal of Political Philosophy*, 14, 394–419.

Benkler, Y., Shaw, A., and Hill, B. M. (2015), *Peer production: A form of collective intelligence*, Cambridge, MA: MIT Press.

Blau, P. M. (1955), *The dynamics of bureaucracy*, vol. 26, Chicago: University of Chicago Press.

Borgan, Ø., Goldstein, L., and Langholz, B. (1995), "Methods for the analysis of sampled cohort data in the Cox proportional hazards model," *The Annals of Statistics*, 1749–1778.

Brandes, U., Kenis, P., Lerner, J., and van Raaij, D. (2009a), "Network analysis of collaboration structure in Wikipedia," in *Proc. 18th intl. conf. WWW*, ACM, pp. 731–740.

Brandes, U. and Lerner, J. (2007), "Visual Analysis of Controversy in User-generated Encyclopedias," in *Proc. IEEE Symp. Visual Analytics Science and Technology (VAST'07)*, IEEE, pp. 179–186.

Brandes, U., Lerner, J., and Snijders, T. A. (2009b), "Networks Evolving Step by Step: Statistical Analysis of Dyadic Event Data," in *Proc. 2009 Intl. Conf. Advances in Social Network Analysis and Mining (ASONAM)*, IEEE, pp. 200–205.

Breiger, R. L. (1974), "The duality of persons and groups," *Social forces*, 53, 181–190.

Butts, C. T. (2008), "A relational event framework for social action," *Sociological Methodology*, 38, 155–200.

Butts, C. T. and Marcum, C. S. (2017), "A relational event approach to modeling behavioral dynamics," in *Group Processes*, eds. Pilny, A. and Poole, M., Springer, pp. 51–92.

Cartwright, D. and Harary, F. (1956), "Structural balance: A generalization of Heider's theory," *The Psychological Review*, 63, 277–293.

Chase, I. D. (1980), "Social process and hierarchy formation in small groups: a comparative perspective," *American Sociological Review*, 905–924.

Chase, I. D., Bartolomeo, C., and Dugatkin, L. A. (1994), "Aggressive interactions and inter-contest interval: how long do winners keep winning?." *Animal Behaviour*, 48, 393–400.

Conaldi, G. and Lomi, A. (2013), "The dual network structure of organizational problem solving: A case study on open source software development," *Social Networks*, 35, 237–250.

Del Vicario, M., Bessi, A., Zollo, F., Petroni, F., Scala, A., Caldarelli, G., Stanley, H. E., and Quattrociocchi, W. (2016), "The spreading of misinformation online," *Proceedings of the National Academy of Sciences*, 113, 554–559.

Doreian, P., Kapuscinski, R., Krackhardt, D., and Szczypula, J. (1996), "A brief history of balance through time," *Journal of Mathematical Sociology*, 21, 113–131.

Doreian, P. and Mrvar, A. (2009), "Patitioning signed social networks," *Social Networks*, 31, 1–11.

— (2014), "Testing Two Theories for Generating Signed Networks Using Real Data," *Metodološki zvezki*, 11, 731–63.

DuBois, C., Butts, C. T., McFarland, D., and Smyth, P. (2013), "Hierarchical models for relational event sequences," *Journal of Mathematical Psychology*, 57, 297–309.

Duguid, P. (2006), "Limits of self-organization: Peer production and "laws of quality"," *First Monday*, 11.

Everett, M. G. and Borgatti, S. P. (2014), "Networks containing negative ties," *Social Networks*, 38, 111–120.

Festinger, L. (1962), *A theory of cognitive dissonance*, vol. 2, Stanford university press.

Flöck, F. and Acosta, M. (2014), "WikiWho: Precise and efficient attribution of authorship of revisioned content," in *Proc. 23rd Intl. Conf. WWW*, ACM, pp. 843–854.

Gambetta, D. (1998), "Concatenations of mechanisms," *Social mechanisms: An analytical approach to social theory*, 102–124.

Gilbert, E., Bergstrom, T., and Karahalios, K. (2009), "Blogs are echo chambers: Blogs are echo chambers," in *HICSS'09. 42nd Hawaii International Conference on System Sciences*, IEEE, pp. 1–10.

Giles, J. (2005), "Internet Encyclopaedias Go Head to Head," *Nature*, 438, 900–901.

Gómez, V., Kappen, H. J., Litvak, N., and Kaltenbrunner, A. (2013), "A likelihood-based framework for the analysis of discussion threads," *World Wide Web*, 16, 645–675.

Greenstein, S., Gu, Y., and Zhu, F. (2016), "Ideological segregation among online collaborators: evidence from Wikipedians," Tech. Rep. w22744, National Bureau of Economic Research, http://www.nber.org/papers/w22744.

Hannan, M. T. (2010), "Partiality of memberships in categories and audiences," *Annual Review of Sociology*, 36.

Heider, F. (1946), "Attitudes and cognitive organization," *The Journal of Psychology*, 21, 107–112.

Hummon, N. P. and Doreian, P. (2003), "Some dynamics of social balance processes: bringing Heider back into balance theory," *Social Networks*, 25, 17–49.

Jasny, L., Waggle, J., and Fisher, D. R. (2015), "An empirical examination of echo chambers in US climate policy networks," *Nature Climate Change*, 5, 782–786.

Javanmardi, S., Lopes, C., and Baldi, P. (2010), "Modeling user reputation in wikis," *Statistical Analysis and Data Mining*, 3, 126–139.

Kaltenbrunner, A. and Laniado, D. (2012), "There is no deadline: Time evolution of Wikipedia discussions," in *Proc. 8th Annual Intl. Symp. Wikis and Open Collaboration*, ACM.

Kilduff, M. and Krackhardt, D. (1994), "Bringing the individual back in: A structural analysis of the internal market for reputation in organizations," *Academy of management journal*, 37, 87–108.

Kittur, A., Suh, B., Pendleton, B. A., and Chi, E. H. (2007), "He says, she says: conflict and coordination in Wikipedia," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, ACM, pp. 453–462.

Kreiss, D., Finn, M., and Turner, F. (2011), "The limits of peer production: Some reminders from Max Weber for the network society," *New Media & Society*, 13, 243–259.

Labianca, G. (2014), "Negative ties in organizational networks," in *Contemporary perspectives on organizational social networks*, Emerald Group Publishing Limited, pp. 239–259.

Labianca, G. and Brass, D. J. (2006), "Exploring the social ledger: Negative relationships and negative asymmetry in social networks in organizations," *Academy of Management Review*, 31, 596–614.

Lerner, J. (2016), "Structural balance in signed networks: Separating the probability to interact from the tendency to fight," *Social Networks*, 45, 66–77.

Lerner, J., Bussmann, M., Snijders, T. A., and Brandes, U. (2013), "Modeling Frequency and Type of Interaction in Event Networks," *Corvinus Journal of Sociology and Social Policy*, 4, 3–32.

Lerner, J. and Lomi, A. (2017), "The Third Man: Hierarchy Formation in Wikipedia," *Applied Network Science*, 2, 24.

Lerner, J. and Tirole, J. (2001), "The open source movement: Key research questions," *European economic review*, 45, 819–826.

— (2002), "Some simple economics of open source," *The journal of industrial economics*, 50, 197–234.

Leskovec, J., Huttenlocher, D., and Kleinberg, J. (2010), "Signed Networks in Social Media," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, pp. 1361–1370.

Lusher, D., Koskinen, J., and Robins, G. (eds.) (2013), *Exponential Random Graph Models for Social Networks*, Cambridge University Press.

Maniu, S., Cautis, B., and Abdessalem, T. (2011), "Building a signed network from interactions in Wikipedia," in *Proc. Databases and Social Networks*, ACM, pp. 19–24.

Neff, J. J., Laniado, D., Kappler, K. E., Volkovich, Y., Aragón, P., and Kaltenbrunner, A. (2013), "Jointly they edit: Examining the impact of community identification on political interaction in Wikipedia," *PloS one*, 8, e60584.

Negro, G., Hannan, M. T., and Fassiotto, M. (2014), "Category signaling and reputation," *Organization Science*, 26, 584–600.

O'Mahony, S. and Lakhani, K. R. (2011), "Organizations in the shadow of communities," in *Communities and organizations*, Emerald Group Publishing Limited, pp. 3–36.

Padgett, J. F. and Powell, W. W. (2012), *The emergence of organizations and markets*, Princeton University Press.

Piskorski, M. J. and Gorbatâi, A. (2017), "Testing Coleman's Social-Norm Enforcement Mechanism: Evidence from Wikipedia," *American Journal of Sociology*, 122, 1183–1222.

Podolny, J. M. (2001), "Networks as the pipes and prisms of the market," *American journal of sociology*, 107, 33–60.

— (2010), *Status signals: A sociological study of market competition*, Princeton University Press.

R Core Team (2017), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

Saperstein, A. M. (2004), "'The Enemy of My Enemy Is My Friend' Is the Enemy: Dealing with the War-Provoking Rules of Intent," *Conflict Management and Peace Science*, 21, 287–296.

Sauder, M., Lynn, F., and Podolny, J. M. (2012), "Status: Insights from organizational sociology," *Annual Review of Sociology*, 38, 267–283.

Singh, P. V., Tan, Y., and Mookerjee, V. (2011), "Network effects: The influence of structural capital on open source project success," *MIS Quarterly*, 35, 813–829.

Snijders, T. A. (2005), "Models for longitudinal network data," in *Models and Methods in Social Network Analysis*, eds. Carrington, P. J., Scott, J., and Wasserman, S., Cambridge University Press.

Stadtfeld, C. (2010), "Who Communicates With Whom? Measuring Communication Choices on Social Media Sites," in *Proc. IEEE 2nd Intl. Conf. Social Computation (SocialCom)*, pp. 564–569.

Stadtfeld, C. and Block, P. (2017), "Interactions, Actors, and Time: Dynamic Network Actor Models for Relational Events," *Sociological Science*, 4, 318–352.

Stadtfeld, C. and Geyer-Schulz, A. (2011), "Analyzing event stream dynamics in two-mode networks: An exploratory analysis of private communication in a question and answer community," *Social Networks*, 33, 258–272.

Stadtfeld, C., Hollway, J., and Block, P. (2017), "Dynamic Network Actor Models: Investigating Coordination Ties through Time," *Sociological Methodology*, 47.

Stuart, H. C., Dabbish, L., Kiesler, S., Kinnaird, P., and Kang, R. (2012), "Social transparency in networked information exchange: a theoretical framework," in *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, ACM, pp. 451–460.

Suh, B., Chi, E. H., Pendleton, B. A., and Kittur, A. (2007), "Us vs. Them: Understanding Social Dynamics in Wikipedia with Revert Graph Visualizations," in *Proc. IEEE Symp. Visual Analytics Science and Technology (VAST'07)*, pp. 163–170.

Sumi, R., Yasseri, T., Rung, A., Kornai, A., and Kertész, J. (2011), "Edit wars in Wikipedia," in *3rd intl. conf. Privacy, Security, Risk and Trust (PASSAT) and 3rd intl. conf. Social Computing (SocialCom)*, IEEE, pp. 724–727.

Torlò, V. J. and Lomi, A. (2017), "The network dynamics of status: assimilation and selection," *Social Forces*, 1–34.

Tsvetkova, M., García-Gavilanes, R., and Yasseri, T. (2016), "Dynamics of Disagreement: Large-Scale Temporal Network Analysis Reveals Negative Interactions in Online Collaboration," *Scientific reports*, 6.

Turian, J., Ratinov, L., and Bengio, Y. (2010), "Word representations: a simple and general method for semi-supervised learning," in *Proceedings of the 48th annual meeting of the association for computational linguistics*, Association for Computational Linguistics, pp. 384–394.

Viégas, F. B., Wattenberg, M., and Dave, K. (2004), "Studying Cooperation and Conflict Between Authors with History Flow Visualizations," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 575–582.

von Hippel, E. and von Krogh, G. (2003), "Open source software and the "private-collective" innovation model: Issues for organization science," *Organization science*, 14, 209–223.

— (2006), "Free revealing and the private-collective model for innovation incentives," *R&D Management*, 36, 295–306.

Vu, D., Pattison, P., and Robins, G. (2015), "Relational event models for social learning in MOOCs," *Social Networks*, 43, 121–135.

Yap, J. and Harrigan, N. (2015), "Why does everybody hate me? Balance, status, and homophily: The triumvirate of signed tie formation," *Social Networks*, 40, 103–122.

Yasseri, T., Sumi, R., Rung, A., Kornai, A., and Kertész, J. (2012), "Dynamics of conflicts in Wikipedia," *PloS one*, 7, e38869.

Zuckerman, E. W. (2012), "Construction, concentration, and (dis) continuities in social valuations," *Annual Review of Sociology*, 38, 223–245.