To Thread or Not to Thread: The Impact of Conversation Threading on Online Discussion

Pablo Aragón,[*†] **Vicenç Gómez**[*] **Andreas Kaltenbrunner**[*†]
[*]Universitat Pompeu Fabra, Barcelona, Spain
[†]Eurecat, Barcelona, Spain

Abstract

Online discussion is essential for the communication and collaboration of online communities. The reciprocal exchange of messages between users that characterizes online discussion can be represented in many different ways. While some platforms display messages chronologically using a simple linear interface, others use a hierarchical (threaded) interface to represent more explicitly the structure of the discussion. Although the type of representation has been shown to affect communication, to the best of our knowledge, the impact of using either one or the other has not yet been investigated in a large and mature online community.

In this work we analyze Menéame, a popular Spanish social news platform which recently transitioned from a linear to a hierarchical interface, becoming an ideal research opportunity for this purpose. Using interrupted time series analysis and regression discontinuity design, we observe an abrupt and significant increase in social reciprocity after the adoption of a threaded interface. We furthermore extend state-of-the-art generative models of discussion threads by including reciprocity, a fundamental feature to explain better the structure of the discussions, both before and after the change in the interface.

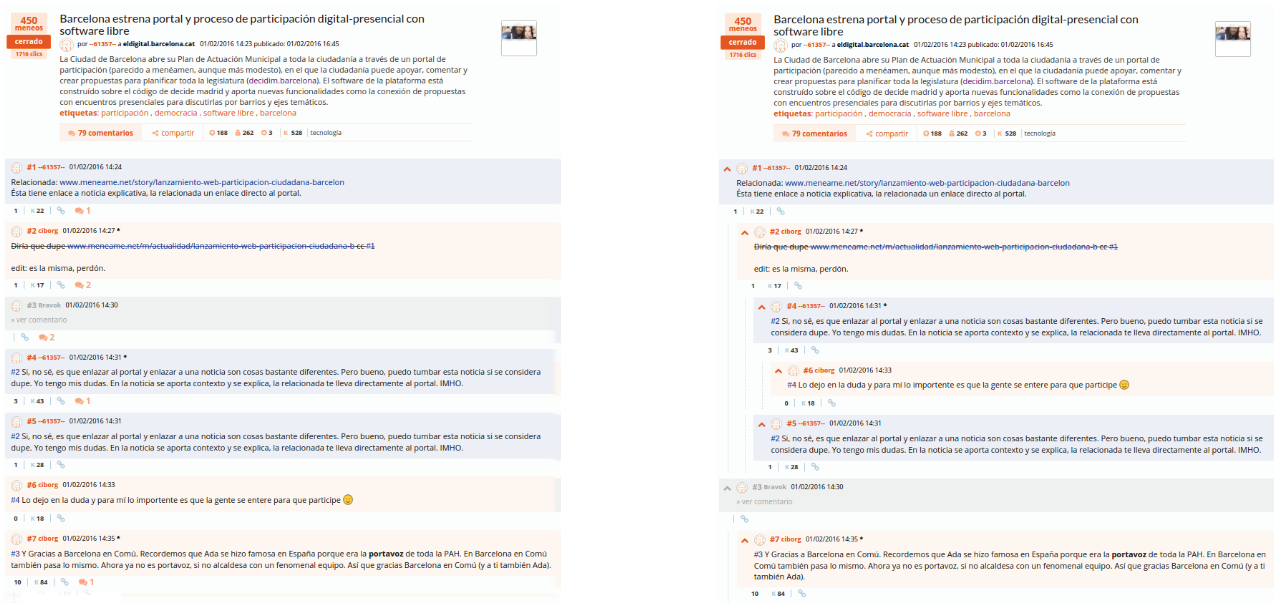
Introduction

The interaction between users in social media platforms has enabled the emergence of online communities. In these communities, online discussion is essential for the communication and collaboration. Since they are commonly built by strangers, trust between users is only possible when reciprocity occurs (Seabright 2010), for example in the form of a strong exchange of messages between users. Such reciprocity has been traditionally seen as a sign of an inward focus and vigorous debate (Fisher, Smith, and Welser 2006) and some theories have also suggested a relationship between reciprocity and captivating/engaging communication (Rafaeli and Sudweeks 1997). Furthermore, reciprocity is a necessary condition for deliberative purposes because it allows to gain knowledge of the perspectives of others (Habermas 1985). Thus, many approaches to measure deliberation include reciprocity (Schneider 1997; Jensen 2003; Graham and Witschge 2003) in order to quantify the degree to which a conversation is a real discussion (Janssen and Kies 2005).

Although online discussions are simply characterized by an exchange of messages, there are many ways in which a discussion can be presented to a user. Discussion threads are collections of messages posted as replies to previous messages. Therefore, many platforms like email clients and online forums have adopted a *hierarchical* view, also known as *conversation threading*, i.e. messages are arranged close to their replies in a tree-like structure. With this type of view, reciprocal interactions between users are explicitly shown. In contrast, some other platforms show messages regardless of reply relationships with a *linear* view. The sorting criteria of messages with this view is typically chronological to indicate how a discussion thread evolves over time.

Previous work has examined the performance of experimental tools with a specific form of view, either linear or hierarchical. Results usually indicate some benefits of using the hierarchical view, favoring knowledge construction (McVerry 2007) or providing better local context (Venolia and Neustaedter 2003). Conversation threading mitigates the so-called co-text loss problem (Fuks, Pimentel, and De Lucena 2006), i.e. the inability of readers to "identify which of the previous messages provides the elements that are necessary to understand the message that is being read" (Pimentel, Fuks, and de Lucena 2003). Co-text loss occurs when interactions are presented separately (e.g. with a linear view) and users are not able to distinguish the earlier message to which a particular message is replying to. A comparative study of both views in an experimental chat found that coherence was also improved thanks to the hierarchical view but, in contrast, participants reported better user experience when interacting with the linear view (Smith, Cadiz, and Burkhalter 2000).

These previous studies are based on small groups of recruited participants instead of an existing community, and they do not address how reciprocity is affected. Furthermore, they do not include a modeling approach, thus their theoretical insights about the observed behavioral differences are limited. Generative models of discussion threads have been proposed to explain the structure and growth of online discussion by means of behavioral features, such as popularity or novelty of messages (Kumar, Mahdian, and McGlo-



Barcelona estrena portal y proceso de participación digital-presencial con software libre

por --61357-- a eldigital.barcelona.cat 01/02/2016 14:23 publicado: 01/02/2016 16:45

La Ciudad de Barcelona abre su Plan de Actuación Municipal a toda la ciudadanía a través de un portal de participación (parecido a menéamen, aunque más modesto), en el que la ciudadanía puede apoyar, comentar y crear propuestas para planificar toda la legislatura (decidim.barcelona). El software de la plataforma está construído sobre el código de decide madrid y aporta nuevas funcionalidades como la conexión de propuestas con encuentros presenciales para discutirlas por barrios y ejes temáticos.

etiquetas: participación , democracia , software libre , barcelona

#1 --61357-- 01/02/2016 14:24

Relacionada: www.meneame.net/story/lanzamiento-web-participacion-ciudadana-barcelon
Ésta tiene enlace a noticia explicativa, la relacionada un enlace directo al portal.

1 | K 22 | 1

#2 ciborg 01/02/2016 14:27 •

~~Diría que dupe www.meneame.net/m/actualidad/lanzamiento-web-participacion-ciudadana-b cc #1~~

edit: es la misma, perdón.

1 | K 17 | 2

#3 Bravok 01/02/2016 14:30

» ver comentario

| 2

#4 --61357-- 01/02/2016 14:31 •

#2 Si, no sé, es que enlazar al portal y enlazar a una noticia son cosas bastante diferentes. Pero bueno, puedo tumbar esta noticia si se considera dupe. Yo tengo mis dudas. En la noticia se aporta contexto y se explica, la relacionada te lleva directamente al portal. IMHO.

3 | K 43 | 1

#5 --61357-- 01/02/2016 14:31

#2 Si, no sé, es que enlazar al portal y enlazar a una noticia son cosas bastante diferentes. Pero bueno, puedo tumbar esta noticia si se considera dupe. Yo tengo mis dudas. En la noticia se aporta contexto y se explica, la relacionada te lleva directamente al portal. IMHO.

1 | K 28 |

#6 ciborg 01/02/2016 14:33

#4 Lo dejo en la duda y para mí lo importente es que la gente se entere para que participe 🙂

0 | K 18 |

#7 ciborg 01/02/2016 14:35 •

#3 Y Gracias a Barcelona en Comú. Recordemos que Ada se hizo famosa en España porque era la **portavoz** de toda la PAH. En Barcelona en Comú también pasa lo mismo. Ahora ya no es portavoz, si no alcaldesa con un fenomenal equipo. Así que gracias Barcelona en Comú (y a ti también Ada).

10 | K 84 | 1

(a) Linear

Barcelona estrena portal y proceso de participación digital-presencial con software libre

por --61357-- a eldigital.barcelona.cat 01/02/2016 14:23 publicado: 01/02/2016 16:45

La Ciudad de Barcelona abre su Plan de Actuación Municipal a toda la ciudadanía a través de un portal de participación (parecido a menéamen, aunque más modesto), en el que la ciudadanía puede apoyar, comentar y crear propuestas para planificar toda la legislatura (decidim.barcelona). El software de la plataforma está construído sobre el código de decide madrid y aporta nuevas funcionalidades como la conexión de propuestas con encuentros presenciales para discutirlas por barrios y ejes temáticos.

etiquetas: participación , democracia , software libre , barcelona

#1 --61357-- 01/02/2016 14:24

Relacionada: www.meneame.net/story/lanzamiento-web-participacion-ciudadana-barcelon
Ésta tiene enlace a noticia explicativa, la relacionada un enlace directo al portal.

1 | K 22 |

#2 ciborg 01/02/2016 14:27 •

~~Diría que dupe www.meneame.net/m/actualidad/lanzamiento-web-participacion-ciudadana-b cc #1~~

edit: es la misma, perdón.

1 K 17 |

#4 --61357-- 01/02/2016 14:31 •

#2 Si, no sé, es que enlazar al portal y enlazar a una noticia son cosas bastante diferentes. Pero bueno, puedo tumbar esta noticia si se considera dupe. Yo tengo mis dudas. En la noticia se aporta contexto y se explica, la relacionada te lleva directamente al portal. IMHO.

3 | K 43 |

#6 ciborg 01/02/2016 14:33

#4 Lo dejo en la duda y para mí lo importente es que la gente se entere para que participe 🙂

0 | K 18 |

#5 --61357-- 01/02/2016 14:31

#2 Si, no sé, es que enlazar al portal y enlazar a una noticia son cosas bastante diferentes. Pero bueno, puedo tumbar esta noticia si se considera dupe. Yo tengo mis dudas. En la noticia se aporta contexto y se explica, la relacionada te lleva directamente al portal. IMHO.

1 | K 28 |

#3 Bravok 01/02/2016 14:30

» ver comentario

|

#7 ciborg 01/02/2016 14:35 •

#3 Y Gracias a Barcelona en Comú. Recordemos que Ada se hizo famosa en España porque era la **portavoz** de toda la PAH. En Barcelona en Comú también pasa lo mismo. Ahora ya no es portavoz, si no alcaldesa con un fenomenal equipo. Así que gracias Barcelona en Comú (y a ti también Ada).

10 K 84 |

(b) Hierarchical (conversation threading)

Figure 1: The two types of conversation views in Menéame for an example thread: (a) linear view, before the platform change in January 2015, (b) hierarchical view, after the change. In both views, every reply to a comment contains the symbol # followed by the id of the comment it replies to. Blue comments are written by the post's author.Comments scored negatively are shown in white without text unless clicked upon.

hon 2010; Wang, Ye, and Huberman 2012; Gómez et al. 2013). They are language-independent approaches and can thus successfully reproduce many of the structural patterns observed in online discussions of very diverse nature. However, despite reciprocity patterns commonly emerge in online discussion networks, the state-of-the-art models do not incorporate this as a feature. Therefore, it is unclear whether reciprocity is either a behavioral feature or a resulting effect of discussion dynamics.

In this work we want to increase our understanding of the impact of conversation threading on online discussion. For that, we first measure how the specific type of conversation view affects reciprocity, and then how a modeling approach can capture the interplay between conversation view, structure of discussions, and reciprocity. The reciprocation of interactions plays a primary role since users are motivated to contribute to the community expecting useful help and information in return (Kollock and Smith 2002; Plickert, Cote, and Wellman 2007; Gray, Ward, and Norton 2014). Existing theories have established that reciprocity is a defining attribute of online communities (Wellman and Gulia 1999) and a behavioral indicator for their emergence (Herring et al. 2004). Its absence leads communities to fail (Harasim 1993). Given that reciprocity is essential in online communities and conversation threading makes explicit reciprocal interactions between users, as opposed to a linear view, our research questions are:

- **RQ1**: *How does conversation threading affect the reciprocity within the discussion of an online community?*
- **RQ2**: *Is reciprocity a key behavioral feature when modeling the structure and growth of discussion threads?*
- **RQ3**: *How does conversation threading affect the behavioral features when modeling the structure and growth of discussion threads?*

Answering these questions represents a methodological challenge, mainly because of the difficulties and limitations of performing a controlled experiment. We overcome this challenge using data from Menéame[1], the most popular Spanish social news networking service (the 2nd most visited site of this type in Spain after Reddit[2]). The website interface changed in January 2015. The original conversation view presented the comments of a thread linearly in a chronological order (see Figure 1a). Since that change, the comments are displayed by default hierarchically following a tree-structure (see Figure 1b). This platform intervention occurred in isolation, which allows us to analyze the impact of such a change with a reduced influence of possible confounders that may also affect the community and the originated discussions. For this reason, Menéame becomes an ideal opportunity to measure the impact of the two types of conversation view on a real and large online community.

The organization of the paper is as follows. In the following section we describe the dataset of discussion threads collected from Menéame. We then present our statistical analysis to measure the impact of the change of view on the reciprocity of replies. Next, we follow a modeling ap-

[1] https://www.meneame.net/

[2] http://www.alexa.com/siteinfo/meneame.net

proach and extend state-of-the-art generative models of discussion threads by incorporating an authorship model and reciprocity. This extension is critical to reproduce the structure and evolution of threads accurately and to measure how the user behavior is affected by the change of view. Finally, we discuss the implications of our findings for the design and research of online discussion platforms.

Dataset

Menéame is the most popular Spanish social news networking service. Social news websites, like Reddit, Slashdot or Digg, feature user-posted stories which are discussed in threads, and voted to be ranked based on their popularity within the community. The selection process of featured stories is made by an open source collaborative filtering algorithm similar to the one in Reddit. Besides the change of the conversation view (from linear to hierarchical), some other reasons make Menéame a platform of interest in our study:

- The community of Menéame consists of thousands of users who daily debate hundreds of stories (links to news and blog posts).
- The platform was released in 2005 and therefore Menéame is a large and mature community of users which have developed their own culture of practices.

We collected all the stories which were promoted to the front page between 2011 to 2015 (both years included) and every comment from the discussion thread of each story. The reasons for focusing on the promoted stories is because they are more appealing to the community of Menéame and to guarantee a sufficiently large volume of comments per story. In total, we obtained 72,005 posts and 5,385,324 comments.

For each comment, we kept the associated meta-data such as the id, the id of the post/comment it is being replied to, the url, the user name, and the time-stamp. We should remark that, as shown in Figure 1, both the linear and the hierarchical interface display at the beginning of every reply to a comment contains the symbol # followed by the id of the comment it replies to. Therefore, discussions threads in Menéame can *always* be mapped into a tree, which is implicit in the linear view and becomes explicit when the view is hierarchical.

Measuring conversation threading effects

In this section we present our statistical analysis and results on the dataset of Menéame. We first describe a preliminary analysis and then introduce our methodology based on regression discontinuity design (RDD). We then define how to characterize mathematically reciprocity and describe our results.

Preliminary analysis

To better understand the evolution of discussions in Menéame, we first examined the temporal profile of some global activity indicators of the platform. Results are shown in Figure 2 with a vertical line indicating the change from a linear to a hierarchical view in January 2015.

We observe that, although the number of stories in the front page (Figure 2a) decreases over time, the total number of comments (Figure 2b) first decreases from 2011 to 2014 but then increases from 2014 to 2016. The number of unique users (Figure 2c) also decreases from 2011 to 2014 but then remains stable. These trends are coupled with a seasonal pattern with activity drops during summer and winter holidays. These cyclic patterns are corrected when one normalizes the binned data by the number of threads. Interestingly, the average number of comments per thread (Figure 2d) and unique users per thread (Figure 2e) show a sustained increase with an apparent abrupt change in the beginning of 2015, i.e. when the conversation view was modified from linear to hierarchical.

Impact of conversation threading on reciprocity

To quantify the impact of the change of the conversation view, we apply regression discontinuity design (RDD). RDD is a statistical test used in econometrics to estimate treatment effects in a quasi-experimental setting, where treatment is determined by whether an observed *assignment* variable exceeds a known cutoff point (Thistlethwaite and Campbell 1960; Lee and Lemieux 2010). This technique has been applied recently in previous studies to measure how the design and technical features of a given platform constrain, distort, and shape user behavior on that platform (Hale et al. 2014; Malik and Pfeffer 2016).

In this work, we use RDD to assess statistically the impact of conversation threading, since we only have observational data in a non-experimental setting. We start from temporal measurements of our variable of interest, in our case, the reciprocity, which we define mathematically in the next paragraph. RDD fits two different functions to this temporal data, before and after the cutoff point (when conversation threading was adopted in Menéame) and allows to quantify the break between both fitted lines at the cutoff. The null hypothesis is that the reciprocity is not affected by the release of the new conversation view. For further mathematical details and assumptions of the method, we refer the reader to the appendix.

To formally characterize reciprocity, we consider the directed network of replies between users in each discussion thread. In this network, each node correspond to a user and a directed edge between user u and v exists if user u replied to user v in the discussion. The weight of that edge is the number of times u replied to v in that thread. Given a directed network of N nodes, reciprocity is traditionally defined as follows:

$$r = \frac{E^{\leftrightarrow}}{E}, \qquad (1)$$

where $E^{\leftrightarrow}$ corresponds to the number of bidirectional edges and E corresponds the total number of edges. This approach is limited in the sense that it does not consider the relative difference of reciprocity in comparison to a random network with the same number of nodes and edges. The definition by Garlaschelli and Loffredo (2004) overcomes this problem

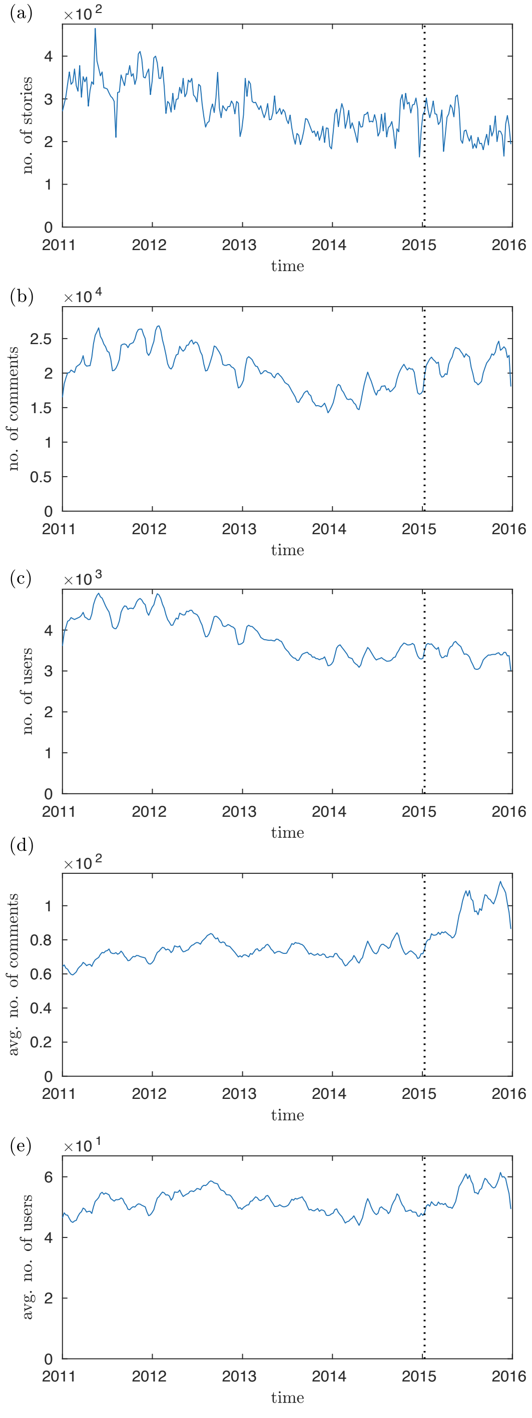


Figure 2: Number of stories (a), comments (b), unique users (c), average number of comments per story (d), and unique users per story (e). The vertical line indicates the change of the conversation view (from linear to hierarchical).

and defines the *corrected* reciprocity as

$$\rho = \frac{r - \bar{a}}{1 - \bar{a}}, \tag{2}$$

where $\bar{a}$ is the network density, i. e. the ratio between the number of existing edges and the total number of possible edges $\bar{a} = E/(N(N-1))$. The previous definitions of reciprocity do not take into account the weighted nature of edges in the reply network, i.e. the number of times that two users interchange messages within a thread. Squartini et al. (2013) proposes the following definition of reciprocity for weighted networks

$$r_w = \frac{W^{\leftrightarrow}}{W} = \frac{\sum_u \sum_{v \neq u} w_{uv}^{\leftrightarrow}}{\sum_u \sum_{v \neq u} w_{uv}}, \tag{3}$$

where u, v are nodes indexes, w_{uv} is the weight of the edge from u to v, and $w_{uv}^{\leftrightarrow}$ is the minimum weight between the edge from u to v and the edge from v to u.

We construct one network of replies between users for each conversation and compute the three previous reciprocity indicators in each of these networks. In the following, we omit results using r because they are indistinguishable from the results using ρ. This is explained because the constructed reply networks are very sparse and the density $\bar{a}$ is low. We then average these indicators at a time resolution of one month, which defines the bin-size in our analysis. The bin size is an arbitrary choice, we experimented with several sizes but observed no significant differences.

We show in Figure 3 how both corrected reciprocity ρ and weighted reciprocity r_w change over time, together with the results of the RDD test. We first note that both reciprocity measures show a sustained increasing trend, which suggests that captivating/engaging communication increases over time. Furthermore, if reciprocity is a defining attribute of an online community, as proposed in Wellman and Gulia (1999), the increasing trend can be interpreted as a positive indicator of the performance of Menéame. The weighted measure is slightly higher than the non-weighted metric, which suggests that the frequency of replies between the same users is important. However, both profiles are very similar, so this frequency is not qualitatively determinant.

We should remark that we use an F-test in every point in the time series to establish the most significant cutoff. Our analysis identifies January 2015 as the optimal cutoff, which corresponds exactly with the transition of the interface. This is indicated in Figure 3 by a black dashed line that separates the data before (in red) and after (in blue) the cutoff. The results show a notable impact for the both corrected reciprocity and weighted reciprocity (see the appendix for the numerical details of the results). This means that the null hypothesis can be rejected and, therefore, there is a significant effect in reciprocity when Menéame transitioned from a linear to a hierarchical conversation view. It is also important to mention that the slope increased after the change, indicating that reciprocity, not only changes abruptly after the adoption of conversation threading, but also increases at a higher speed during the period of available data considered. We will further discuss the impact of these findings in the discussion section.

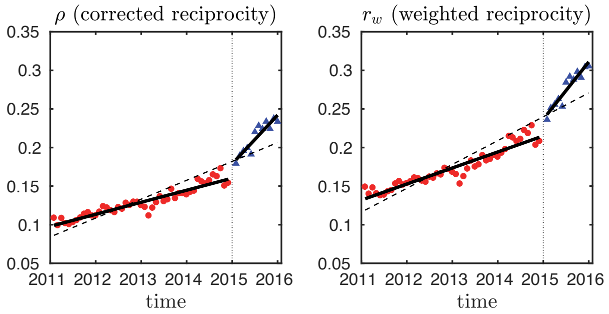


Figure 3: Regression discontinuity design for the metrics of reciprocity in the discussions (bin size = one month). Vertical line is the cutoff obtained through an F-test. Red circles are points before the cutoff, blue triangles data points after the cutoff. Solid line is the discontinuous linear regression, the dashed line is the continuous linear regression of the null model. The analysis shows that the two reciprocity metrics present a break and an increase of the slope after the cutoff.

Modeling reciprocal online discussions

We now take a modeling approach to gain understanding of the interplay between the structure and the evolution of the discussions, the reciprocity as an abstract feature, and the type of representation. In the next subsection, we characterize informally the discussion threads with special emphasis on their network structure and disregarding the content of the messages. We then describe an existing generative model of online discussions and present our extension which incorporates an authorship model and reciprocity. We show that our proposed extension better explains the observed data. Finally, we perform RDD within the model features.

Structure and growth of discussion threads

To illustrate the typical structure of discussion threads in Menéame, we use a thread visualization tool (Aragón, Gómez, and Kaltenbrunner 2016). Note that these networks differ from the reply networks analyzed in the previous section, since nodes here corresponds to comments, instead of users.

In Figure 4, we present two popular discussion threads that took place before and after the platform change. The first one is from 2013 (left) and the second one is from 2015 (right). Node color follows the criteria: black (root of the thread, i.e the story), gray (first level comments) and random color (replies to comments). We observe that every reply written by the same user gets an identical random color. These criteria allow us to observe that both threads share some similarities, such as long chains of two users that alternate reciprocal interactions (i.e. chains of nodes of two alternating colors). This finding is consistent with previous work on modeling the structure and evolution of discussion cascades using data from Menéame (Gómez, Kappen, and Kaltenbrunner 2011). Node size corresponds to the number of received comments (except for the root) and shows that replies (colored nodes) in the thread of 2015 often attract themselves many replies and originate new sub-discussions within the thread. This effect is not that pronounced in the 2013 thread, in which comments usually belong to chains of two users and rarely trigger a discussion cascade. In summary, we observe that the thread from 2013 is closer to a star-like structure (i.e. contains many more direct comments to the original post) while the thread from 2015 is more complex with higher branching probability at deep levels of the discussion.

A generative model of discussion threads

To measure the impact of using a hierarchical view in the evolution of the discussion threads, we build on the model introduced in Gómez et al. (2013), which has proven to be successful in capturing the structural properties and the temporal evolution of discussion threads present in very diverse platforms, e.g. Slashdot, Barrapunto, Wikipedia and the same Menéame (before conversation threading was adopted). It is a parametrized mathematical model that generates growing trees in discrete time. At each time-step, a new comment (node) arrives to the thread and each of the following structural features is considered for each node in the discussion:

- The *popularity* or number of replies. A node will *attract* replies proportionally with factor α to the number of replies received so far.
- The *novelty* or the elapsed time since it was written. Recent comments will tend to be more replied than old comments. Novelty decays exponentially with parameter τ.
- The *root-bias* or tendency to write more comments to the root node. This differentiates between the original post (root node), which attracts replies with factor β, and ordinary comments (non-root nodes).

Formally, the discussion thread at time-step t is represented as a vector of parent nodes $\pi_{1:t} = (\pi_1, \pi_2 \dots, \pi_t)$, where π_t indicates the parent of the node written at time t. When a new comment arrives to the discussion, it is attached to an existing node $j \in 1, \dots, t$ with probability proportional to its *attractiveness* function $\phi_j(\cdot)$, defined as a combination of the features $\theta = (\alpha, \tau, \beta)$

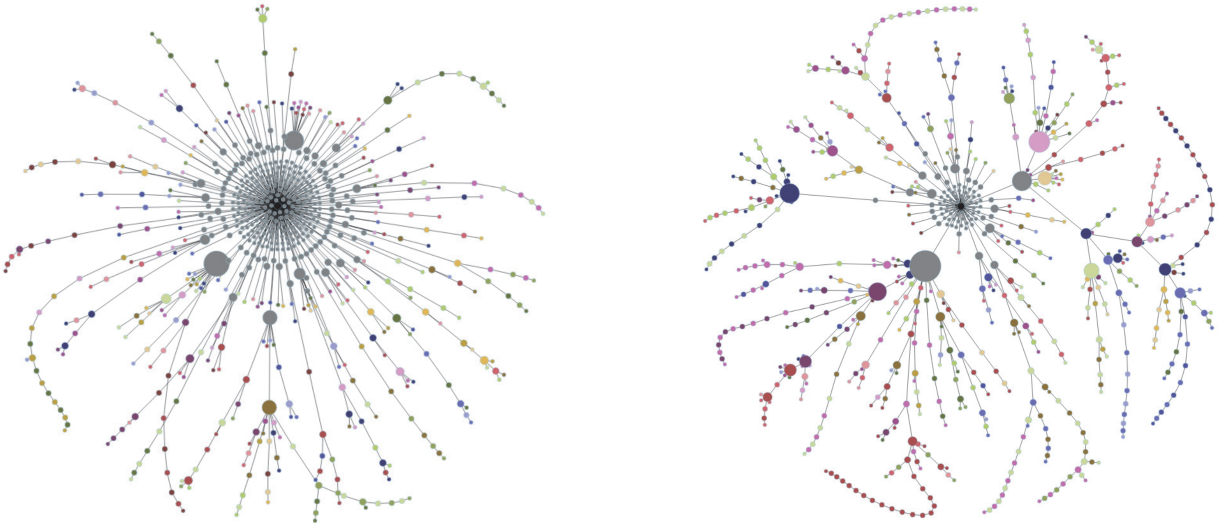
$$\phi_j(\pi_{1:t}; \theta) := \alpha \deg_j(\pi_{1:t}) + \tau^{t+1-j} + \beta \delta_{j,1}$$

$$p(\pi_{t+1} = j | \pi_{1:t}; \theta) \propto \phi_j(\pi_{1:t}; \theta), \qquad (4)$$

where $\deg_j(\pi_{1:t})$ is the degree of node j in the tree $\pi_{1:t}$ and δ is the Kronecker delta function, i.e. β is only relevant for the root node.

The model parameters are estimated through maximum likelihood given a dataset composed of M threads $\mathcal{D} = \{\pi^{(1)}, \dots, \pi^{(M)}\}$ corresponding to a particular period of time.

The previous generative model may fail in describing some structural properties, such as the average depth of a comment, which tends to be underestimated, as noted in Gómez et al. (2013). This is actually the case in Menéame, which is characterized by very deep threads with long chains of messages between two alternating users, as shown in Figure 4. We postulate that the original model fails to capture precisely that commenting behavior tends to be reciprocal,



(a) Thread in 2013.
https://www.meneame.net/story/1860558

(b) Thread in 2015.
https://www.meneame.net/story/2484585

Figure 4: Visualization of two example threads before (a) and after (b) the conversation view was modified. Black node is the root of the thread (the post). Gray nodes are first level comments. The other nodes are replies to comments where comments written by the same user get the same color. Node size corresponds to the number of received comments, except for the root.

i.e. users tend to reply comments that are replies to their previous comments. In the next section, we extend the original model with an authorship model and introduce a new feature: the reciprocity.

Extending the model

We now represent a conversation thread with the parent vector $\pi_{1:t}$ together with a vector of respective authors $a_{1:t} = (a_1, a_2, \dots, a_t)$. The authorship vector will grow depending on the structure of the discussion, which in turn will depend on the authorship of the messages.

Our author model does not allow two consecutive comments to be written by the same user. Furthermore, a user cannot self-reply a comment made by herself. Let U denote the number of different users that participated in the conversation so far. At time $t+1$, a new comment is originated from a new user with id $U+1$ with probability p_{new}, or otherwise from an existing user v chosen according to how many times user v has been replied in the thread, r_v. Our author model is described as

$$p(a_{t+1} = v | a_{1:t}, \pi_{1:t}) = \begin{cases} p_{new}, & \text{for } v = U+1 \\ \frac{(1-p_{new})2^{r_v}}{\sum_{i=1}^{U} 2^{r_i}}, & \text{for } v \in 1, \dots, U \end{cases} \quad (5)$$

We set p_{new} empirically to $p_{new} = t^{-1/k}$ and estimate k from the data ($k \approx 7$). Notice that the preferential attachment process that selects authors is multiplicative. This is required to capture well the probability distribution of the number of comments per unique author in a thread. Once the author a_{t+1} is decided, the new comment is attached to an existing comment j proportionally to the extended attractiveness function $\phi'_j(\cdot)$, which now depends on the vector of authors $a_{1:t}$ and the parameters $\theta' = (\alpha, \tau, \beta, \kappa)$

$$\phi'_j(\pi_{1:t}, a_{1:t}; \theta') := \phi_j(\pi_{1:t}; \theta) + \kappa\delta_{a_{\pi_j}, a_{t+1}}$$
$$p'(\pi_{t+1} = j | \pi_{1:t}, a_{1:t}; \theta') \propto \phi'_j(\pi_{1:t}; \theta'), \quad (6)$$

where the additional term $\kappa\delta_{a_{\pi_j}, a_{t+1}}$ is non-zero for reciprocal comments only and $\phi_j(\cdot)$ is the original (author-independent) attractiveness function given in Equation (4).

The new parameter κ determines how strong reciprocal comments are weighted. Only those replies to comments authored by the selected author, i.e. $a_{\pi_j} = a_{t+1}$, will contribute to the κ-term. Thus, for $\kappa = 0$ the new feature will play no role in the evolution of the thread whereas very large values of κ will make all comments of corresponding users reciprocal. The additional parameter κ can be optimized using maximum likelihood together with α, β and τ[3].

We first compare the original model and the proposed extension and then we analyze how the change in the interface affects the model parameters. To show that the extended model not only reproduces better the depths, we also compare the two models using the same indicators as in Gómez et al. (2013). Figure 5 shows that the distribution of the number of replies, subthreads sizes and the relation between the thread sizes and depths are reproduced significantly better thanks to the authorship model and the reciprocity feature.

Figure 6 shows the empirical probability distributions (pdf) of the depth of a comment calculated from the real threads and from synthetic ones generated from both models after optimizing their respective parameters. Whereas

[3] The source code for estimating the model parameters given a collection of threads can be found here:
git@bitbucket.org:vicengomez/threads.git

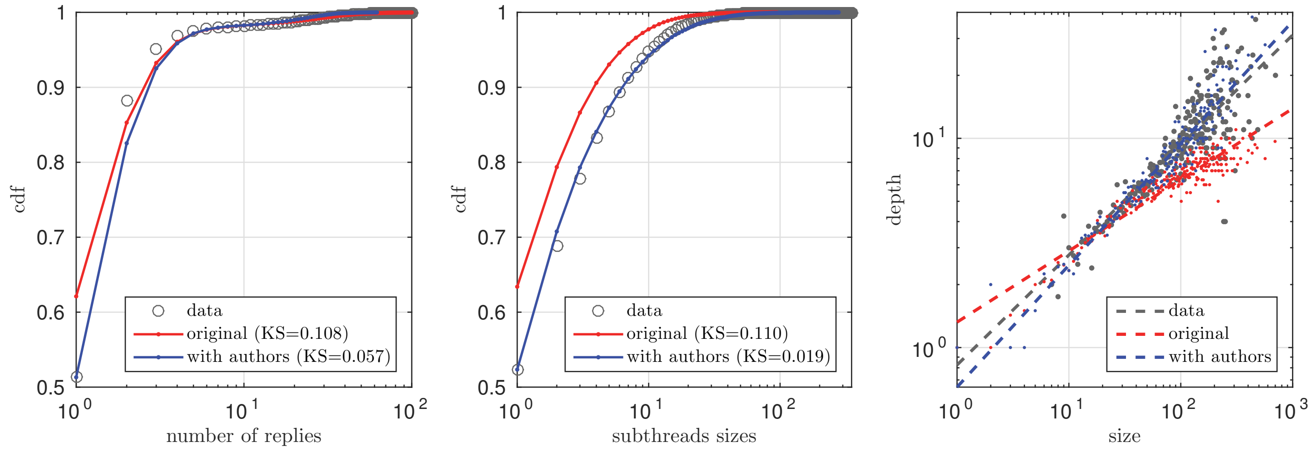


Figure 5: Comparison between the original model (in red) and the extended model with authorship and reciprocity (in blue) in terms of how well they reproduce the real discussion threads (gray circles). The plots show the cumulative distribution function (cdf) of the degrees (left), subtree sizes (center) and the correlation between depth and number of comments (right). The curves where obtained from $2 \cdot 10^3$ threads generated from both models after optimization of their respective parameters. Dashed lines in the right subplot correspond to linear fits in the logarithmic domain. KS indicates Kolmogorov-Smirnov test value (the lower the better).

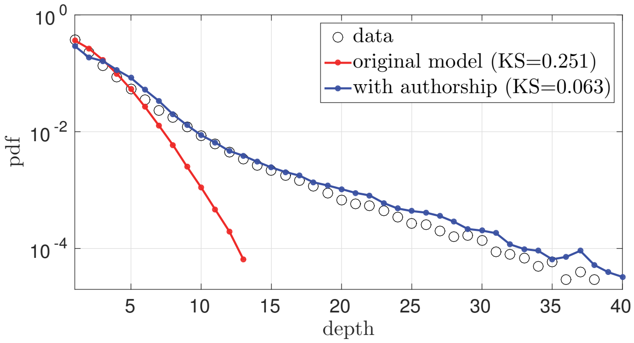


Figure 6: Probability distribution of the comment's depths. The original model fails to capture the long tail created by reciprocal message chains whereas the proposed model is able to reproduce the data accurately. The curves where obtained from $2 \cdot 10^3$ threads generated from both models after optimization of their respective parameters. KS indicates Kolmogorov-Smirnov test value (the lower the better).

the resulting depths using the original model are underestimated (red curve), the extended model is able to generate deeper threads and to reproduce better the depth distribution. In particular, it captures the tail behavior accurately and the observed discrepancies are only minor. The KS test accepts the model hypothesis at the 5% confidence level (p-value = 0.0041). The synthetic threads also contain chains of messages with alternating users, as in the original data.

We thus conclude that by increasing minimally the complexity of the model with the authorship model and the reciprocity, the overall descriptive power of the model is greatly improved.

Impact of conversation threading on behavioral features

We now analyze how the platform change affected the evolution of the threads by fitting the extended model to data from different periods of time. In Figure 7 we present the results of the RDD on the four estimated parameters, each of them corresponding to one of the features (see the appendix for the numerical details of the results).

Globally, we observe notable increases in all the parameters after the platform change. The most noticeable change corresponds to the reciprocity feature, parameterized by κ (see the change of order of magnitude in Figure 7). Once the hierarchical view is active, users behave significantly more reciprocally and tend to engage more in dialogues. These findings are consistent with the above one for the corrected and the weighted reciprocity metrics.

The other features also show an abrupt increase after the platform change, but to a lesser extent. We emphasize that the interplay between the features may be nontrivial, even mediated by a hidden, not modeled feature, since the relative weights differ between the two conditions. Nevertheless, since reciprocity is only relevant at the later stages of the discussions, where comments are written from existing authors that have already been replied, their relevance is also increased after the platform change. Finally, it is interesting to mention that the same analysis performed in the original model was unable to detect a significant change in parameters β and α at the time of the platform intervention.

Discussion

We have presented a study about the impact of conversation threading in online discussions. While previous studies in this field (McVerry 2007; Fuks, Pimentel, and De Lucena 2006; Venolia and Neustaedter 2003; Smith, Cadiz, and

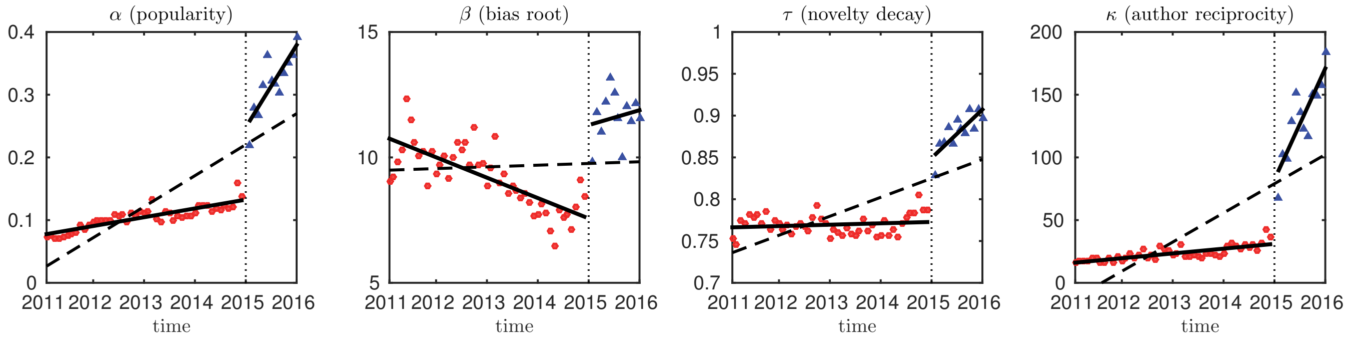


Figure 7: Regression discontinuity (RD) analysis for the metrics of the structural properties of discussions (bin size = one month). Vertical line indicates the cutoff obtained through the F-test. Red circles are data points before the cutoff and blue triangles data points after the cutoff. The solid line is the discontinuous linear regression, the dashed line is the continuous linear regression corresponding to the null model. The analysis shows that the four model parameters present a break and an increase of the slope after the cutoff.

Burkhalter 2000) had relied on experiments recruiting small groups of participants, our findings are observed in an existing, large and mature community with over five years of online discussion data.

We first analyze how the implementation of conversation threading affects the reciprocity in the discussion of an online community (*RQ1*). One would expect reciprocity to increase since a hierarchical conversation view emphasizes the exchange of messages between users. Indeed, although we already observe a natural increase of reciprocity over time, as suggested in Fisher, Smith, and Welser (2006), the adoption of this type of interface triggers an additional boost leading to even higher levels of reciprocity. This is a positive behavioral indicator for online communities (Wellman and Gulia 1999; Herring et al. 2004; Plickert, Cote, and Wellman 2007), and it is aligned with previous work on the benefits of hierarchical views for constructing knowledge (McVerry 2007), providing better context of the discussion (Fuks, Pimentel, and De Lucena 2006; Venolia and Neustaedter 2003) and improving coherence (Smith, Cadiz, and Burkhalter 2000). Our results have implications for the characterization of user roles in online discussion. Reciprocity has been used to distinguish different types of users in online forums; e.g *taciturns* with low tendency to reciprocate interactions and *grunts* with relatively higher levels of reciprocity (Chan, Hayes, and Daly 2010). Given that users in social media change their role over time (García-Gavilanes et al. 2014), this opens interesting research directions like assessing whether the distribution of user roles is affected by changing the conversation view.

The relevance of reciprocity in online discussion leads us to reflect on the role of this behavioral pattern in the formation of discussion threads. The existing generative models of discussion threads (Kumar, Mahdian, and McGlohon 2010; Wang, Ye, and Huberman 2012; Gómez et al. 2013) include features from messages like the popularity (number of incoming replies) or the novelty (timestamp). However, the tendency of users to reply to the replies to their messages has only been considered indirectly, *a posteriori*. For example, although the authorship model of Kumar, Mahdian, and McGlohon (2010) establishes authors of messages to promote the reciprocity of replies, it does it after the discussion thread is generated and, therefore, reciprocity is ignored during the growth of the discussion. Furthermore, all of these models fail in modeling accurately the depth of discussion threads which can be explained by the occurrence of long chains of reciprocal interactions between two users, as postulated in Pelaprat and Brown (2012) and empirically shown in Figure 4. This leads to our second research question (*RQ2*) about whether reciprocity can improve the descriptive power of models of discussion threads. To answer this question we extend the model in Gómez et al. (2013) by incorporating authorship and establishing reciprocity as a behavioral feature. This is an important difference from previous models such as Kumar, Mahdian, and McGlohon (2010), in which the structure of a thread does not depend explicitly on the authorship. To the best of our knowledge, the presented model is the first in which the structure and authorship co-evolve jointly. The results on discussion threads from Menéame show that our approach not only captures better the distribution of the number of replies and sizes of subthreads, it also reproduces more accurately the temporal evolution of the discussion threads in view of the distribution of the depth of discussion threads.

The model extension includes reciprocity together with the existing behavioral features of popularity, novelty, and root-bias. This allows us to answer our third research question (*RQ3*) which analyzes whether modeling discussion threads can quantify the impact of the conversation view on behavioral features. On the one hand, our results show that the hierarchical view induced more reciprocal behavior, which is consistent with the findings from the regression discontinuity design. On the other hand, we also observe that the transition to threaded discussion makes popular comments to attract more replies and slows down the decay of novelty, i.e. comments take longer to be ignored. This second effect can be explained by the fact that the hierarchical view on Menéame does not apply comment folding and, therefore, branches of comments are always fully expanded. With this type of interface, conversation threading gives preference to the first comments and their replies, i.e. branches (and sub-branches) are ordered chronologically. Although it is true that reciprocity increases and online deliberation requires reciprocity (Schneider 1997; Jensen 2003), new contributions with no connection with previous arguments will be less visible to the community. Given that deliberation also requires users gaining knowledge of the perspectives of others (Habermas 1985), additional mechanisms (e.g. comment folding, branch sorting) must receive special attention in the design of online discussion platforms.

Our methodology is based on the structural properties of the discussions and is language-independent. Therefore, it can be easily applied to other platforms. For this reason, modeling approaches like the ones applied here can also be used to assess the impact of other features in online discussion platforms and to compare the model parameters in different environments and communities. Moreover, it might be of interest to extend these models to further explore content-based features from the messages of the discussions. Recent studies have suggested that linguistic indications of reciprocity can measure the chance of success of individual requests in online communities (Althoff, Danescu-Niculescu-Mizil, and Jurafsky 2014). Also, hierarchical comment threads have been noted to represent a topical hierarchy in online discussions (Weninger, Zhu, and Han 2013). Therefore, future work might explore whether the transition from a linear to a hierarchical conversation view can also affect the narrative structure and the distribution of topics in online discussion.

Acknowledgments

This work is supported by the Spanish Ministry of Economy and Competitiveness under the Maria de Maeztu Units of Excellence Programme (MDM-2015-0502) and the EU project D-CENT (FP7/CAPS 610349). We would like to thank Helena Gallego for her suggestions that helped to improve this manuscript.

References

Althoff, T.; Danescu-Niculescu-Mizil, C.; and Jurafsky, D. 2014. How to ask for a favor: A case study on the success of altruistic requests. In *Eighth International AAAI Conference on Weblogs and Social Media*.

Aragón, P.; Gómez, V.; and Kaltenbrunner, A. 2016. Visualization Tool for Collective Awareness in a Platform of Citizen Proposals. In *ICWSM-16 - 10th International AAAI Conference on Web and Social Media*. The AAAI Press.

Chan, J.; Hayes, C.; and Daly, E. M. 2010. Decomposing discussion forums and boards using user roles. *ICWSM* 10:215–218.

Fisher, D.; Smith, M.; and Welser, H. T. 2006. You are who you talk to: Detecting roles in usenet newsgroups. In *Proceedings of the 39th Annual Hawaii International Conference on System Sciences (HICSS'06)*, volume 3, 59b–59b. IEEE.

Fuks, H.; Pimentel, M.; and De Lucena, C. J. P. 2006. RU-Typing-2-Me? Evolving a chat tool to increase understanding in learning activities. *International Journal of Computer-Supported Collaborative Learning* 1(1):117–142.

García-Gavilanes, R.; Kaltenbrunner, A.; Sáez-Trumper, D.; Baeza-Yates, R.; Aragón, P.; and Laniado, D. 2014. Who are my audiences? A study of the evolution of target audiences in microblogs. In *International Conference on Social Informatics*, 561–572. Springer.

Garlaschelli, D., and Loffredo, M. I. 2004. Patterns of link reciprocity in directed networks. *Physical review letters* 93(26):268701.

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(5-6):645–675.

Gómez, V.; Kappen, H. J.; and Kaltenbrunner, A. 2011. Modeling the structure and evolution of discussion cascades. In *Proceedings of the 22nd ACM conference on Hypertext and hypermedia*, 181–190. ACM.

Graham, T., and Witschge, T. 2003. In search of online deliberation: Towards a new method for examining the quality of online discussions. *Communication, Sankt Augustin Then Berlin* 28(2):173–204.

Gray, K.; Ward, A. F.; and Norton, M. I. 2014. Paying it forward: Generalized reciprocity and the limits of generosity. *Journal of experimental psychology: General* 143(1):247.

Habermas, J. 1985. *The theory of communicative action*, volume 2. Beacon press.

Hale, S. A.; John, P.; Margetts, H. Z.; and Yasseri, T. 2014. Investigating political participation and social information using big data and a natural experiment.

Harasim, L. M. 1993. *Global networks: Computers and international communication*. MIT Press.

Herring, S. C.; Barab, S.; Kling, R.; and Gray, J. 2004. An approach to researching online behavior. *Designing for virtual communities in the service of learning* 338.

Janssen, D., and Kies, R. 2005. Online forums and deliberative democracy. *Acta política* 40(3):317–335.

Jensen, J. L. 2003. Public spheres on the internet: Anarchic or government-sponsored–a comparison. *Scandinavian political studies* 26(4):349–374.

Kollock, P., and Smith, M. 2002. *Communities in cyberspace*. Routledge.

Kumar, R.; Mahdian, M.; and McGlohon, M. 2010. Dynamics of conversations. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, 553–562. ACM.

Lee, D. S., and Lemieux, T. 2010. Regression discontinuity designs in economics. *Journal of economic literature* 48(2):281–355.

Malik, M. M., and Pfeffer, J. 2016. Identifying platform effects in social media data. In *Tenth International AAAI Conference on Web and Social Media*.

McVerry, J. G. 2007. Forums and functions of threaded discussions. *New England Reading Association Journal* 43(1):79.

Pelaprat, E., and Brown, B. 2012. Reciprocity: Understanding online social relations. *First Monday* 17(10).

Pimentel, M. G.; Fuks, H.; and de Lucena, C. J. P. 2003. Co-text loss in textual chat tools. In *Modeling and Using Context*. Springer. 483–490.

Plickert, G.; Cote, R. R.; and Wellman, B. 2007. It's not who you know, it's how you know them: Who exchanges what with whom? *Social networks* 29(3):405–429.

Rafaeli, S., and Sudweeks, F. 1997. Networked interactivity. *Journal of Computer-Mediated Communication* 2(4):0–0.

Schneider, S. M. 1997. *Expanding the Public Sphere through Computer-Mediated Communication: Political Discussion about Abortion in.* Ph.D. Dissertation, Massachusetts Institute of Technology.

Seabright, P. 2010. *The company of strangers: A natural history of economic life*. Princeton University Press.

Smith, M.; Cadiz, J. J.; and Burkhalter, B. 2000. Conversation trees and threaded chats. In *Proceedings of the 2000 ACM conference on Computer supported cooperative work*, 97–105. ACM.

Squartini, T.; Picciolo, F.; Ruzzenenti, F.; and Garlaschelli, D. 2013. Reciprocity of weighted networks. *Scientific reports* 3(2729).

Thistlethwaite, D. L., and Campbell, D. T. 1960. Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational psychology* 51(6):309.

Venolia, G. D., and Neustaedter, C. 2003. Understanding sequence and reply relationships within email conversations: a mixed-model visualization. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, 361–368. ACM.

Wang, C.; Ye, M.; and Huberman, B. A. 2012. From user comments to on-line conversations. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, 244–252. ACM.

Wellman, B., and Gulia, M. 1999. Net surfers don't ride alone: Virtual communities as communities. *Networks in the global village* 331–366.

Weninger, T.; Zhu, X. A.; and Han, J. 2013. An exploration of discussion threads in social news sites: A case study of the reddit community. In *Advances in Social Networks Analysis and Mining (ASONAM), 2013 IEEE/ACM International Conference on*, 579–583. IEEE.

Appendix: Regression Discontinuity Design

Regression Discontinuity Design (RDD) is a statistical test that measures causal effects in cases where an arbitrarily strict cutoff along one covariate exists. In the linear case the regression is:

$$Y_i = \omega_0 + \omega_1 \cdot x_i + \omega_2 \cdot \mathbb{1}(x_i > c) + \omega_3 \cdot x_i \cdot \mathbb{1}(x_i > c) + \epsilon_i,$$

where x_i is the time-stamp of a bin, Y_i is the average value of bin i (bin size = one month), ω_i are the coefficients of the regression, ϵ_i is a random error term, and c is the cutoff.

Linear RDD fits two different linear functions, before and after the cutoff, and allows to quantify the break between both fitted lines at the cutoff. The null hypothesis is that there is no discontinuity (the metric is not affected by the release of the new conversation view), i.e. $\omega_2 \approx 0$ and $\omega_3 \approx 0$.

The cutoff in classical RDD is the intervention given in the experiment. In the context of platform effects for our study, the cutoff is expected to be the time when the conversation view was modified in Menéame. However, we should note that, by definition, a discontinuous regression with a cutoff at midpoint of the time series is likely to better fit data than a continuous regression. Therefore, to enhance the robustness of our analysis and to prove the statistical significance of the change of the conversation view, we use an F-test, as suggested in Lee and Lemieux (2010), to set the cutoff as the most significant point in the time series.

In all RDD reported results, we prevented biased estimates of the treatment effect by checking that the linear model represented a good model using a statistical analysis of the residuals.

In our first experiment visualized in Figure 3 the obtained values in the RDD for the corrected reciprocity ρ are $break = 0.019, \omega_2 = -0.171$ and $\omega_3 = 0.004$. The corresponding values for the weighted reciprocity r_w are $break = 0.021$, $\omega_2 = -0.192$ and $\omega_3 = 0.004$.

In our second experiment, results shown in Figure 7, we obtained the following RDD values for the reciprocity feature κ (break = 51.28; $\omega_2 = -287.78$; $\omega_3 = 7.06$), the popularity α (break = 0.12; $\omega_2 = -0.35$; $\omega_3 = 0.01$), the novelty τ (break = 0.08 ; $\omega_2 = -0.15$; $\omega_3 = 0.004$), and the root-bias β (break = 3.72 ; $\omega_2 = -1.90$; $\omega_3 = 0.12$).