Testing for the signature of policy in online communities

Alberto Cottica, Guy Melançon, Benjamin Renoust

Testing for the signature of policy in online communities

Alberto Cottica [1], Guy Melançon [2], and Benjamin Renoust [3]

Abstract Most successful online communities employ professionals, sometimes called "community managers", for a variety of tasks including onboarding new participants, mediating conflict, and policing unwanted behaviour. We interpret the activity of community managers as network design: they take action oriented at shaping the network of interactions in a way conducive to their community's goals. It follows that, if such action is successful, we should be able to detect its signature in the network itself. Growing networks where links are allocated by a preferential attachment mechanism are known to converge to networks displaying a power law degree distribution. Our main hypothesis is that managed online communities would deviate from the power law form; such deviation constitutes the signature of successful community management. Our secondary hypothesis is that said deviation happens in a predictable way, once community management practices are accounted for. We investigate the issue using empirical data on three small online communities and a computer model that simulates a widely used community management activity called *onboarding*. We find that the model produces in-degree distributions that systematically deviate from power law behavior for low-values of the in-degree; we then explore the implications and possible applications of the finding.

1 Introduction

Organizations running online communities typically employ community managers, tasked with encouraging participation and resolving conflict [18]. Only a small number of the participants (one or two members in the smaller communities) will recognize some central command, and carry out its directives. We shall henceforth call such directives *policies*. Putting in place policies for online communities is costly, in terms of recruitment, training, and software tools. This raises the question of what benefits organizations running online communities expect from policies; and why they choose certain policies, and not others.

Online communities can be modeled as social networks of interactions across participants, and organizations can be modeled as economic agents maximizing

[1]University of Alicante, Alicante, Spain & Edgeryders, Brussels, Belgium *alberto@cottica.net*, [2]University of Bordeaux, LaBRI CNRS UMR 5800, Bordeaux, France, *Guy.Melancon@u-bordeaux.fr*, [3] National Institute of Informatics & JFLI CNRS UMI 3527, Tokyo, Japan *renoust@nii.ac.jp*

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no 688670.

some objective function (*e.g.* profit, welfare). Hence the topology of the interaction network affects the ability for participants to contribute to the maximization of the target variable. For example, Facebook is constantly rewiring the interaction network across its users to ensure better targeted and more effective advertising, therefore enhancing their revenue [21].

Such organizations choose their policies such as community managers could take action to change the network towards maximizing their objective function.

All this implies that the decision to deploy a particular policy on an online community is a network design exercise. An organisation decides to employ a community manager to shape the interaction network of its community in a way that helps ist own ultimate goals. And yet, interaction networks in online communities cannot really be designed; they are the result of many independent decisions, made by individuals who do not respond to the organization's command structure. An online community management policy is then best understood as an attempt to "influence" emergent social dynamics; to use a more synthetic expression, it can be best understood as the attempt to design for emergence. Its paradoxical nature is at the heart of its appeal.

We are interested in detecting the mathematical signature of specific policies in the network topology. We consider a simple policy called *onboarding* [18, 19]. As a new participant becomes active (*e.g.* by posting her first post), community managers are instructed to leave her a comment that contains (a) positive feedback and (b) suggestions to engage with other participants that she might share interests with.

We model online conversations as social networks, and look for the effect of onboarding on the topology of those networks. We proceed as follows:

1. We initially examine data from three small online communities. Only two of them deploy a policy of *onboarding*. We observe that, indeed, the shape of the degree distribution of these two differs from that of the third.
2. We propose an experiment protocol to determine whether onboarding policies can explain the differences observed between the degree distributions of the first two online communities and that of the third one.
3. Based on the generalized model [10] we simulate the growth of online communities. Variants to the model cover the relevant cases: the absence of onboarding policies and their presence, with varying degrees of effectiveness.
4. We run the experiment protocol against the degree distributions generated by the computer model, and discuss its results.

Section 2 briefly examines the two strands of literature that we mostly draw upon. Section 3 presents some data from real-world online communities; it then proceeds to describe our main experiment, a computer simulation of interaction in online communities with and without onboarding. Section 4 presents the experiment's results. Section 5 discusses them.

2 Related work

Collective intelligence [15] scholars confirmed importance of online community management practices, indeed, they have tried to systematize it [9] and produce

technological innovation to support it [20, 8]. These tools are meant to facilitate and encourage participation to online communities, to make it easier for individuals to extract knowledge from them. Studying human communities is a traditional focus of network science [5, 6], for which easily available datasets of online communities make an ideal ground for structural analysis: friendship in Facebook [16, 17], following/retweet/mentions for Twitter [13, 12, 11], or vote and comments in discussions [11, 14, 23, 22].

Starting in the 2000s, online communities became the object of another line of enquiry, stemming from network science. Network representation of relationships across groups of humans has yielded considerable insights in social sciences since the work of the sociometrists in the 1930s, and continues to do so; phenomena like effective spread of information, innovation adoption, and brokerage have all been addressed in a network perspective [5, 6]. As new datasets encoding human interaction became available, many online communities came to be represented as social networks. This was the case for social networking sites, like Facebook [16, 17]; microblogging platform like Twitter [13, 12, 11]; news-sharing services like Digg [11]; collaborative editing projects like Wikipedia [14]; discussion forums like the Java forum [23]; and bug reporting services for software developers like Bugzilla [22]. Generally, such networks represent participants as nodes. Edges represent a relationship or interaction. The nature of interaction varies across online communities: one edge can stand for friendship for Facebook; follower-followed relationship, retweet or mention in Twitter; vote or comment in Digg and the Java forum; talk in Wikipedia; comment in Bugzilla.

In contrast to collective intelligence scholars, network scientists typically do not address the issue of community management, and treat social networks drawn from online interaction as fully emergent. In this paper, we employ a network approach to investigate the issue of whether the work of community managers leaves a footprint detectable by quantitative analysis. To our knowledge, no other work attempted this investigation. In particular, we exploit a result from the theory of evolving networks, from seminal work by Barabási and Albert [2] showing that the assumption of growth and preferential attachment, when taken together, result in a network whose degree distribution converges to a power law ([1, 3]). The model was later generalized in various ways and tested across a broad range of networks, including social networks [10].

We use this generalization as a baseline state. The degree distribution of the interaction network in an online community follows a power law by default. The action of online community managers, as they attempt to further the goals of the organisation that runs the online community, will result in its degree distribution deviating from the baseline power law in predictable ways. Such deviation can be interpreted as the signature that the policy is working well.

The most important difficulty with this method is the absence of a counterfactual: if a policy is enacted in the online community, the baseline degree distribution corresponding to the absence of the policy is not observable, and viceversa. This rules out a direct proof that the policy "works". Hence our choice to combine empirical data and computer simulations.

	Innovatori PA	**Edgeryders**	**Matera2019**
Policy	*"no special policy"*	*"onboard new users"*	*"onboard new users"*
In existence since	December 2008	October 2011	March 2013
Accounts created	10,815	2,419	512
Active participants (nodes)	619	596	198
Number of edges (weighted)	1,241	4,073	883
Average distance	3.77	2.34	2.51
Maximum degree	155	238	46
Average degree	2.033	6.798	4.454
Goodness-of-fit for $k \geq 1$			
exponent	1.611	1.477	1.606
p-value	0.21	0.00 (reject)	0.00 (reject)
Goodness-of-fit for $k \geq k_{min}$			
k_{min}	2	5	6
exponent	1.834	2.250	2.817
p-value	0.76	0.45	0.94

Table 1: Comparing interaction networks of the three online communities and testing for goodness-of-fit of power functions to degree distributions. "Exponent" refers to the power law's scaling parameter. "p-value" to the result of the test that the degree distribution of the community was generated by a power law with that exponent.

3 Materials and methods

In this section we introduce the empirical data, the experiment protocol and the simulation model we use in the experiment.

3.1 Empirical data

We examine data from three real-world online communities: InnovatoriPA is a community of (mostly) Italian civil servants discussing how to introduce and foster innovation in the public sector. It does not employ any special onboarding or moderation policy. Edgeryders is a community of (mostly) European citizens, discussing public policy issues from the perspective of grassroot activism and social innovation. It adopts the onboarding of new members policy. Matera 2019 is a community of (mostly) citizens of the Italian city of Matera and the surrounding region, discussing the city's policies. It also adopts the onboarding policy.

The communities are modeled as interaction networks (summarized in Table 1) in which nodes are users and edges represent directed comments from A to B, weighted by the number of comments written. A glance at their respective visualizations (Figure 1) suggests that the networks of the three communities have very different topologies. Innovatori PA displays more obviously visible hubs than the other two.

We fitted power laws in-degree distributions of these three online communities, as of early December 2014. Next, we tested the hypothesis that degree distributions follow a power law, as predicted by [10]. To do so, we first fitted power functions

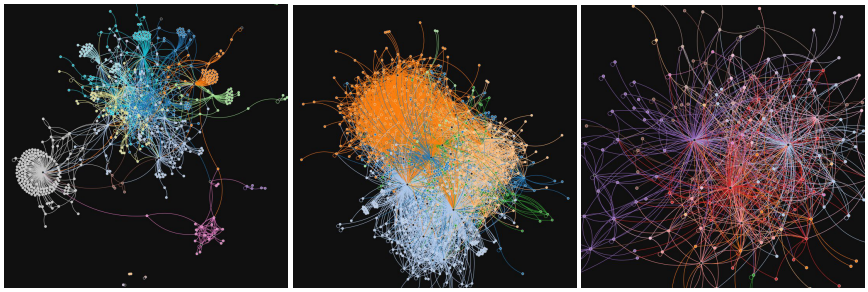


Fig. 1: Interaction networks of three small online communities. Innovatori PA (left) does not have an onboarding policy in place, whereas the two others do (Edgeryders: center, Matera: right).

to the entire support of each in-degree distribution[1]. We next fitted power functions to the right tail of each in-degree distribution, *i.e.* for any degree $k(n) \geq k_{min}$, where k_{min} is the in-degree that minimizes the Kolmogorov-Smirnov distance (hereafter denoted as D) between the fitted function and the data with in-degree $k \geq k_{min}$.

Finally, we ran goodness-of-fit (hereafter GoF) tests for each in-degree distribution and for fitted power functions. The method we followed throughout the paper is borrowed from Clauset *et al* [7]. The null hypothesis tested is that the observed distribution is generated by a power function with exponent α. We compare the D statistic of the observed distribution with those of a large number of synthetic datasets drawn by the fitted power function. Such comparison is summarized in a p-value, that indicates the probability of the D statistic to exceed the observed value conditional to the null hypothesis being true. p-values close to 1 indicate that the power function is a good fit for the data: the null hypothesis is not rejected. p-values close to zero indicate that the power function is a bad fit for the data, and reject the null hypothesis. The rejection value is set, conservatively, at 0.1. Results are summarized in Table 1.

As we consider the interval $k \geq 1$, we find that the in-degree distribution of the Innovatori PA network – the unmoderated one – is consistent with the expected behavior of an evolving network with preferential attachment. We cannot reject the null hypothesis that it was generated by a power law. For other two communities, both with onboarding policies, the null hypothesis is strongly rejected. On the other hand, when we consider only the tail of the degree distributions, i.e. $k \geq k_{min}$, all three communities display a behavior that is consistent of a setting with preferential attachment.

These results are consistent with the objectives of the onboarding policy, consisting in helping newcomers find their way around a community that they don't know yet. A successfully onboarded new user will generally have some extra interaction with existing active members. All things being equal, we can expect extra edges to appear in the network, and interfere with the in-degree distribution that would appear in the

[1] We emphasize in-degree, as opposed to out-degree, because directedness is implicit in the idea of preferential attachment, and because the in-degree distribution is the one to follow a power law in online conversation networks ([10]).

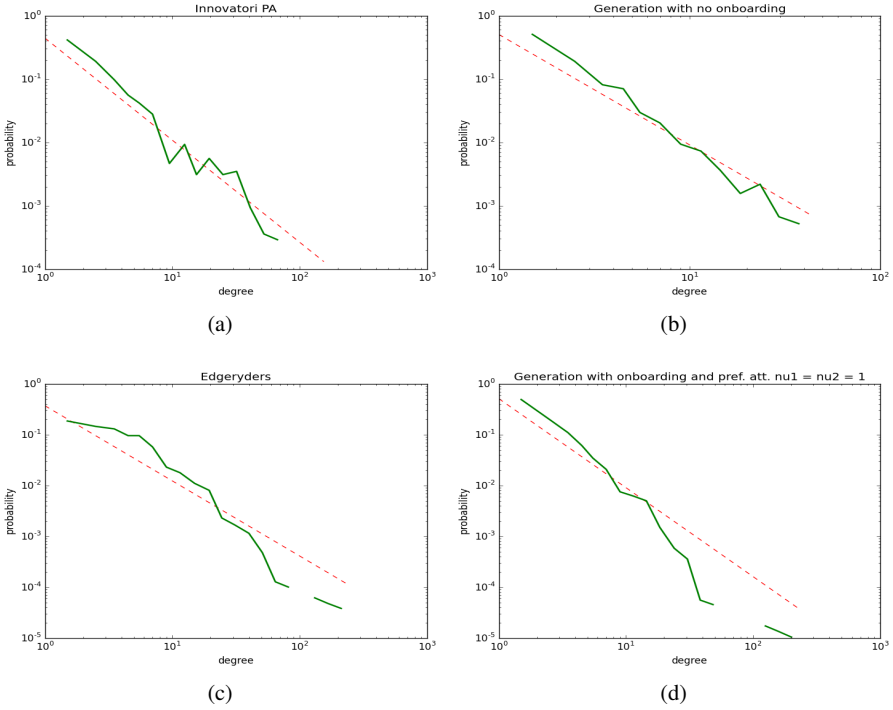


Fig. 2: (log - log) Probability density function from the degree distributions of: (a) the Innovatori PA network without onboarding policy in place versus (b) a simulated network with preferential attachment and no onboarding. (c) The Edgeryders network with onboarding and preferential attachment versus (d) a simulated network with preferential attachment and fully effective onboarding ($\nu_1 = \nu_2 = 1$).

absence of onboarding – explaining the non-power law distribution of Edgeryders and Matera2019. Extra edges target mostly low connectivity nodes: onboarding targets newcomers, and focuses on helping them through the first few successful interactions. Highly active (therefore highly connected) members do not need to be onboarded. This may explain why all three communities display power law behavior in the upper tail of their in-degree distributions, regardless of onboarding.

3.2 Experiment protocol

The difference observed between the two communities with onboarding policies and the one without might be caused not by the policy itself, but by some other unobserved variable. To explore the policy's effects, we generate and compare computer simulations of interaction networks in online communities that are identical except for the presence and effectiveness of onboarding policies.

Communities are assumed to grow over time, with new participants joining them in sequence. At each point in time, new edges appear; their probability of targeting an

existing node grows linearly with that node's in-degree. Additionally, communities might have or not have onboarding policies. See section 3.3 below for a specification of onboarding in the model.

We generated 100 communities with no onboarding policy (control group), 100 communities for each couple of ν_1 and ν_2 in $\{0.0, 0.2, 0.4, 0.6, 0.8, 1.0\}$ (treatment group), and computed their in-degree distribution. Next, we tested two hypotheses for the 3700 networks generated.

- *Hypothesis 1*. The in-degree distribution of C is generated by P for any $k \geq 1$.
- *Hypothesis 2*. The in-degree distribution of C is generated by P for any $k \geq k_{min}$.

Where C is the synthetic network; $k(s)$ is the in-degree of a node s; k_{min} is the in-degree that minimizes the Kolmogorov-Smirnov distance D between the fitted function and the data over $k \geq k_{min}$; and P is the best-fit power-law model for the in-degree distribution of C. We expect non-rejection of both hypotheses for the control group; and rejection of Hypothesis 1, but not of Hypothesis 2, in case of effective onboarding (high ν_1) in the treatment group.

3.3 Simulation

We simulated the growth of network in an online community with and without onboarding following preferential attachment [2] in the generalized model [10].

Without onboarding: A network is initialized with two reciprocally connected nodes. At each step a new node (new user) is introduced, and m new edges (comments) are also created, with a uniformly random picked source. The probability that the new edge points to a node s is proportional to $k(s) + A_s$ where $k(s)$ is the in-degree of node s and A_s is a parameter representing additional attractiveness of the node.

With onboarding: Network initialization and growth are as in the case of no onboarding. Additionally, an edge targeting the newly created node is added at each step. This edge represents the action of the community manager, addressing a welcome message to the newcomer. At this point of each step, with probability $\nu_1 \in [0,1]$, a new edge is added with source as the new node (the newcomer becomes active). The edge's target is chosen by preferential attachment, as described previously[2]. Next (still in the same step), with probability $\nu_2 \in [0,1]$, another edge is added with a uniformly picked source and the newcomer node as target. This represent the online community acknowledging the newcomer by addressing her a comment.

We call ν_1 *onboarding effectiveness*. It is the probability of the newcomer to react to the community manager's onboarding activity. We call ν_2 *community responsiveness*. It is the probability for the new participant to have attracted the attention of other participants and engage in a conversation. We set network size to 2000 nodes; $m = 1$; and $A_s = 1$ for all nodes, in the tradition of [2] and [10].

[2] The source of the new edge is irrelevant to the model's results, since we only study in-degree. We specify it in the text to help exposition, since the expected result of onboarding is the activation of newcomers.

Table 2: Average p-values (number of rejections) for *GoF* tests of power-law models to in-degree distributions of interaction networks in online communities. Control group communities have no onboarding (control group). Power-law models are estimated over all nodes with degree $k \geq 1$

Control group: 0.262688 (23)						
	$\nu_2 = 0.0$	$\nu_2 = 0.2$	$\nu_2 = 0.4$	$\nu_2 = 0.6$	$\nu_2 = 0.8$	$\nu_2 = 1$
$\nu_1 = 0.0$	0.0593 (83)	0.0601 (81)	0.0520 (83)	0.0479 (88)	0.0551 (82)	0.0514 (85)
$\nu_1 = 0.2$	0.0629 (78)	0.0797 (73)	0.0852 (70)	0.0834 (73)	0.0834 (73)	0.0796 (70)
$\nu_1 = 0{,}4$	0.1047 (66)	0.0970 (65)	0.0986 (61)	0.0831 (69)	0.0829 (76)	0.1157 (56)
$\nu_1 = 0.6$	0.0964 (59)	0.0855 (67)	0.1021 (63)	0.1269 (51)	0.0906 (70)	0.0797 (71)
$\nu_1 = 0.8$	0.1326 (55)	0.1152 (60)	0.1036 (66)	0.1091 (61)	0.1188 (60)	0.1228 (61)
$\nu_1 = 1$	0.1009 (65)	0.1207 (62)	0.1326 (54)	0.1164 (60)	0.1230 (54)	0.1205 (57)

4 Results

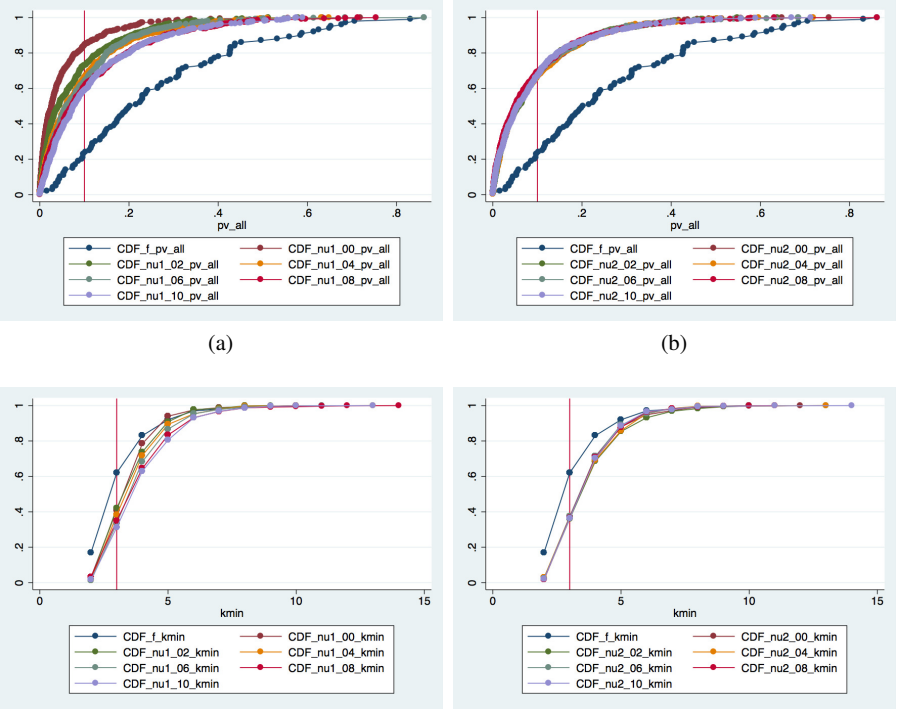
4.1 Goodness-of-fit of the power-law model

For each network evolved we computed two best-fit power-law models, one for $k \geq 1$ and the other for $k \geq k_{min}$ where k_{min} is the in-degree the minimizes D between the fitted function and the data over $k \geq k_{min}$. On each of these models, we ran a *GoF* test as in section 3.1, results are reported in Table 2.

We first examine the case in which $k \geq 1$. We conclude that onboarding seems to have some effect on the goodness-of-fit of the generated data to their respective best-fit power-law models. When onboarding is introduced, fewer degree distributions, out of our 100 runs, are power law-shaped; also, the average p-values returned by *GoF* tests are lower than those of the control group. Running t-tests of the null hypothesis that the average p-value in the control group is equal to the average p-values in the treatment group results in a strong rejection for any combination of ν_1 and ν_2.

We now turn to the question of the role played by ν_1 and ν_2 within the treatment group. Figure 3 (a, b) shows the cumulated density functions of the p-values in the control and treatment groups as ν_1 and ν_2 vary. Increasing onboarding effectiveness ν_1 pushes average p-values of the *GoF* tests down, making it less likely that Hypothesis 1 would be rejected. Increasing community responsiveness ν_2 seems not to play any role at all. This is somewhat surprising. Recall that we modeled onboarding as the command-and-control creation of an extra edge at each step, targeting newcomers to the online community. This has a strong negative effect on the p-value returned by the *GoF* test (compare any p-value in Table 2 with the p-value of the control group with no onboarding). When a responsive community adds a second edge, however, there is no additional effect on the p-value. This result is confirmed by regression analysis (not shown here).

When $k \geq k_{min}$, the effect of introducing onboarding on the *GoF* disappears. Over 99% of the networks in the treatment group give rise to distributions that turn out to be a good fit for a power-law model when k_{min} is chosen so as to minimize D between the degree distributions themselves and their best-fit power-law models. We conclude that Hypothesis 2 cannot be rejected, regardless of whether onboarding is present or not.



(a) (b)

(c) (d)

Fig. 3: (a,b): CDF of p-values returned by *GoF* tests to the (best-fit) power-law models for in-degree distributions of the interaction networks in the control and treatment groups. 20% of the networks evolved without onboarding (dark blue) have degree distributions that test negatively for H1. When onboarding is introduced, it rises to between 50 and 90%. (a,c) the treatment group interaction networks have been grouped according to the value taken by ν_1. (b,d) they have been grouped according to the value taken by ν_2. (c,d) CDF of the average value of k_{min} that minimizes D between the in-degree distribution of each interaction network and its best-fit power-law model.

4.2 Lower bounds

We find a limited, albeit statistically significant, effect of onboarding on the value of k_{min}, the value of k that minimizes D between the data generated by the computer simulation and the best-fit power-law model. Figure 3(c,d) shows that over 60% of the in-degree distributions from interaction networks in the control group, vis-a-vis only 30 to 40% of those in the treatment group, fit a power-law model best for $k_{min} \leq 3$. Within the treatment group, some variability is associated to the increase of ν_1, whereas ν_2 does not seem to play a significant role. Regression analysis (not shown here) shows that, once we control for the presence of onboarding, neither parameter is significant.

Table 3: Average values of the power-law model's exponent α in the control group and in the treatment group by values of ν_1 and ν_2, computed over the whole support $k \geq 1$ (top) and $k \geq k_{min}$ (bottom). The number in parenthesis is the p-value associated to a t-test that $\alpha(treatment) = \alpha(control)$; they were omitted for $k \geq k_{min}$ as they are all smaller than 0.001.

$k \geq 1$ Control group: 1.752						
	$\nu_1 = 0.0$	$\nu_2 = 0.2$	$\nu_2 = 0.4$	$\nu_2 = 0.6$	$\nu_2 = 0.8$	$\nu_2 = 1$
$\nu_1 = 0.0$	1.89 (0.00)	1.89 (0.00)	1.89 (0.00)	1.89 (0.00)	1.89 (0.00)	1.89 (0.00)
$\nu_1 = 0.2$	1.85 (0.00)	1.85 (0.00)	1.85 (0.00)	1.85 (0.00)	1.85 (0.00)	1.85 (0.00)
$\nu_1 = 0.4$	1.82 (0.00)	1.82 (0.00)	1.82 (0.00)	1.82 (0.00)	1.82 (0.00)	1.82 (0.00)
$\nu_1 = 0.6$	1.79 (0.00)	1.79 (0.00)	1.79 (0.00)	1.79 (0.00)	1.79 (0.00)	1.79 (0.00)
$\nu_1 = 0.8$	1.77 (0.00)	1.77 (0.00)	1.77 (0.00)	1.77 (0.00)	1.77 (0.00)	1.77 (0.00)
$\nu_1 = 1$	1.75 (0.21)	1.75 (0.20)	1.75 (0.26)	1.75 (0.43)	1.75 (0.24)	1.75 (0.19)
$k \geq k_{min}$ Control group: 2.419						
	$\nu_2 = 0.0$	$\nu_2 = 0.2$	$\nu_2 = 0.4$	$\nu_2 = 0.6$	$\nu_2 = 0.8$	$\nu_2 = 1$
$\nu_1 = 0.0$	2.985	2.989	2.868	3.000	3.004	3.015
$\nu_1 = 0.2$	2.855	2.852	2.868	2.834	2.821	2.854
$\nu_1 = 0.4$	2.746	2.727	2.735	2.725	2.739	2.749
$\nu_1 = 0.6$	2.661	2.655	2.632	2.650	2.656	2.623
$\nu_1 = 0.8$	2.562	2.602	2.571	2.553	2.554	2.553
$\nu_1 = 1$	2.496	2.527	2.518	2.514	2.514	2.499

4.3 Exponents

Introducing onboarding to an online community has a positive and significant effect on the value of the exponent of the best-fit power-law model for the in-degree distribution of its interaction network. This is consistent with previous studies ([10]). This result holds when the best-fit power-law models is computed over $k \geq k_{min}$, where k_{min} is the value of k that minimizes D between the simulated in-degree distribution and its best-fit power-law model. When it is computed over the whole support of the in-degree distribution ($k \geq 1$), it also holds, except for $\nu_1 = 1$. Table 3 illustrate the average value of the scaling parameter α, and the p-value of a t-test on the null hypothesis that such value is the same as the corresponding statistics in the control group, against the alternative hypothesis that the former is greater than the latter.

5 Discussion and conclusion

We started this work in the hope of discovering a simple statistical test that could be used to assess the presence and effectiveness of online community management policies, onboarding among them. Enacting onboarding on an online community leads to a strong rejection of a power-law behaviour hypothesis on its degree distribution. So, indeed, we can test for *the presence* of onboarding by looking at the degree distribution itself, which is much simpler than analysing the network's whole topology. However, we did not find a monotonic relationship between onboarding's effectiveness and the distance of the resulting degree distribution from a pure power-law form. So, our simple test cannot tell the analyst *how effective* these policies are.

Our models incorporates two forces: preferential attachment and onboarding. The former is meant to represent the (emergent) rich-get-richer effect observed in many real-world social networks; the latter is meant to represent the (command-and-control) onboarding action of moderators and community managers. The former's effect is known to lead to the emergence of an in-degree distribution that approximates a power-law model. The latter's effect is more subtle, because it is in turn composed of two other effects. One consists in the direct action of the moderator, which always targets the newcomer; the other results of the consequences of a well-executed onboarding policy.

The direct action of the moderators creates edges pointing to nodes not selected by preferential attachment: this is definitional of onboarding, and of other online community management activities. What (non-moderator) participants in the online community do as a result of moderator activity is not as clear cut. In our simulation model, fully successful onboarding results in extra edges, some of which point to nodes selected by preferential attachment, others to nodes selected otherwise.

Also, onboarding only targets newcomers. As many online community management policies, it concerns weakly connected participants in the community: moderators have no need to engage with very active, strongly connected participants, who clearly need no help in getting a conversation going. By engaging weakly connected participants, moderators hope to help some shy newcomers turn into active community members. Once this process is under way, moderators have no reason to continue to engage with the same individuals. In terms of our model, this means that newcomers, after having being onboarded, are going to receive new edges by preferential attachment only. It is therefore reasonable to expect that the degree distributions generated by our model display a heavy tail, with the frequency of highly connected nodes following a reasonable approximation of a power law. The overall result of onboarding, then, is an in-degree distribution with power-law behavior for high values of in-degree k and non-power law behavior for low (close to 1) values of k. This is indeed what we observe.

Non-preferential attachment selection of edge targets leads to a poorer fit of power-law models to the in-degree distributions where onboarding is present. This effect takes three forms. The first one is that, fitting a power-law model to the network's in-degree distribution and then running goodness-of-fit tests return a lower p-value than the p-value returned by the same test when onboarding is absent. The second effect is that the value of k that minimizes D between the best-fit power-law model and the observed data tends to be higher than without onboarding. The third one is that the scaling parameter of the best-fit power law tends to be higher with onboarding: onboarding makes the allocation of incoming edges more equal.

Our specification of the model accounts for an apparent paradox: the deviation of the observed networks' degree distributions from power-law behavior is greater when onboarding is ineffective than when it is effective. Ineffective onboarding only adds edges directly created by moderators, *none* of which are allocated across existing nodes by preferential attachment. As onboarding gets more effective, even more edges are added; some are allocated by preferential attachment, and drive the degree distribution back towards a pure power-law behavior. This paradoxical response may

explain why our community responsiveness parameter ν_2 does not appear to impact the shape of the in-degree distribution.

5.1 Future work

Modeling online community management means accounting for the interplay of bottom-up forces (like preferential attachment) with top-down ones (like onboarding policies). This weaving of emergence and design is precisely what we wish to investigate. There are three obvious directions in which we plan to expand the present model. The most obvious one is a systematic exploration of the parameter space, with the goal of assessing our results' robustness with respect to model specification.

A second direction for further research would be to attempt to make the model into a more realistic description of a real-world online community. Such an attempt would draw attention onto how some real-world phenomena, when incorporated in the model, influence its results. It would also carry the advantage of allowing online community management professional to more easily interact with the model and critique it. Several issues that could be investigated in this vein come to mind. For example, we could relax the assumption that the additional attractiveness parameter A_s is identical for all nodes, allowing for different nodes in the network to attract incoming edges at different rates (a phenomenon known as multiscaling [4]). Secondly, we could introduce a relationship between out-degree and in-degree: this would reflect the fact that, in an online community, reaching out to others (which translates in increasing one's own out-degree in the interaction network) is a good way to get noticed and attract incoming comments (which translates in an increase in one's in-degree). Finally, we could work with other community management policies.

A third direction for further research would attempt to gauge the influence of onboarding and other community management policies on network topology by indicators other than the shape of its degree distribution, such as the presence of subcommunities.

Additionally, we wish to obtain and analyse more empirical data from real-world online communities with and without onboarding policies.

References

1. Barabasi, A.L.: The origin of bursts and heavy tails in human dynamics. Nature **435**(7039), 207–211 (2005)
2. Barabási, A.L., Albert, R.: Emergence of scaling in random networks. science **286**(5439), 509–512 (1999)
3. Barabási, A.L., Albert, R., Jeong, H.: Mean-field theory for scale-free random networks. Physica A: Statistical Mechanics and its Applications **272**(1), 173–187 (1999)
4. Bianconi, G., Barabási, A.L.: Competition and multiscaling in evolving networks. EPL (Europhysics Letters) **54**(4), 436 (2001)
5. Borgatti, S.P., Mehra, A., Brass, D.J., Labianca, G.: Network analysis in the social sciences. science **323**(5916), 892–895 (2009)
6. Burt, R.S.: Structural holes: The social structure of competition. Harvard university press (2009)

7. Clauset, A., Shalizi, C.R., Newman, M.E.: Power-law distributions in empirical data. SIAM review **51**(4), 661–703 (2009)
8. De Liddo, A., Sándor, Á., Shum, S.B.: Contested collective intelligence: Rationale, technologies, and a human-machine annotation study. Computer Supported Cooperative Work (CSCW) **21**(4-5), 417–448 (2012)
9. Diplaris, S., Sonnenbichler, A., Kaczanowski, T., Mylonas, P., Scherp, A., Janik, M., Papadopoulos, S., Ovelgoenne, M., Kompatsiaris, Y.: Emerging, collective intelligence for personal, organisational and social use. In: Next generation data technologies for collective computational intelligence, pp. 527–573. Springer (2011)
10. Dorogovtsev, S.N., Mendes, J.F.: Evolution of networks. Advances in physics **51**(4), 1079–1187 (2002)
11. Hodas, N.O., Lerman, K.: The simple rules of social contagion. Scientific reports **4** (2014)
12. Java, A., Song, X., Finin, T., Tseng, B.: Why we twitter: understanding microblogging usage and communities. In: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis, pp. 56–65. ACM (2007)
13. Kunegis, J., Blattner, M., Moser, C.: Preferential attachment in online networks: measurement and explanations. In: Proceedings of the 5th Annual ACM Web Science Conference, pp. 205–214. ACM (2013)
14. Laniado, D., Tasso, R., Volkovich, Y., Kaltenbrunner, A.: When the wikipedians talk: Network and tree structure of wikipedia discussion pages. In: ICWSM (2011)
15. Levy, P.: Collective intelligence: Mankinds emerging world in cyberspace. Cambrigde, Mass.: Perseus Books (1997)
16. Lewis, K., Kaufman, J., Gonzalez, M., Wimmer, A., Christakis, N.: Tastes, ties, and time: A new social network dataset using facebook. com. Social networks **30**(4), 330–342 (2008)
17. Nick, B.: Toward a better understanding of evolving social networks. Ph.D. thesis (2013)
18. Rheingold, H.: The virtual community: Homesteading on the electronic frontier. MIT press (1993)
19. Shirky, C.: Here comes everybody: The power of organizing without organizations. Penguin (2008)
20. Shum, S.B.: The roots of computer supported argument visualization. In: Visualizing argumentation, pp. 3–24. Springer (2003)
21. Slegg, J.: Facebook news feed algorithm change reduces visibility of page updates (2014). URL http://searchenginewatch.com/sew/news/2324814/facebook-news-feed-algorithm-tweak-reduces-visibility-of-page-updates
22. Zanetti, M.S., Sarigol, E., Scholtes, I., Tessone, C.J., Schweitzer, F.: A quantitative study of social organisation in open source software communities. arXiv preprint arXiv:1208.4289 (2012)
23. Zhang, J., Ackerman, M.S., Adamic, L.: Expertise networks in online communities: structure and algorithms. In: Proceedings of the 16th international conference on World Wide Web, pp. 221–230. ACM (2007)