



**HAL**  
open science

# Social Interactions vs Revisions, What Is Important for Promotion in Wikipedia?

Romain Picot-Clémente, Cécile Bothorel, Nicolas Jullien

► **To cite this version:**

Romain Picot-Clémente, Cécile Bothorel, Nicolas Jullien. Social Interactions vs Revisions, What Is Important for Promotion in Wikipedia?. IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2015: ), Aug 2015, Paris, France. pp.888 - 893, 10.1145/2808797.2810063 . hal-01243183

**HAL Id: hal-01243183**

**<https://hal.science/hal-01243183>**

Submitted on 9 Oct 2023

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Social Interactions vs Revisions, What Is Important for Promotion in Wikipedia?

Romain Picot-Clémente  
UMR CNRS 3192 Lab-STICC  
Institut Mines Telecom Bretagne  
Brest, France

Cécile Bothorel  
UMR CNRS 3192 Lab-STICC  
Institut Mines Telecom Bretagne  
Brest, France

Nicolas Jullien  
ICI-M@rsouin  
Institut Mines Telecom Bretagne  
Brest, France

Romain.PicotClemente@telecom-bretagne.eu Cecile.Bothorel@telecom-bretagne.eu Nicolas.Jullien@telecom-bretagne.eu

**Abstract**—In epistemic communities, people are said to be selected on their contribution in knowledge to the project (articles, codes, etc.). However, the socialization process is an important factor for inclusion, sustainability as a contributor, and promotion. Finally, what matters for being promoted? Being a good contributor? Being a good animator? Knowing the boss? We explore this question by looking at the election process for administrators in the English Wikipedia. We used the candidates' revisions and/or social attributes to construct a predictive model of promotion success, based on the candidates' past behavior and a random forest algorithm. Our model explains 78% of the results, which is better than the former models. It also helps to refine the explanation of the election process.

- A study of the promotion process considering social and revision activities.
- A predictive model, based on the random forest algorithm, considering social and contribution activities to predict whether a wikipedian can be elected administrator or not. We give indicative thresholds for the most important attributes to maximize election probabilities.

The article is organized as follows: in Section 2, we will provide a review of the literature used to construct our framework of investigation; in Section 3, we present our data collection strategy and the model. Section 4 presents the results, which are discussed in Section 5, before concluding.

## I. INTRODUCTION

One of the key elements to develop a successful and sustainable (online) community, as explained a quarter of century ago [24], is to attract enough highly competent and “committed/committing” contributors, usually named “lead users”, “core”, or “big” contributors [16]. How and on which criteria these core contributors are recruited to become managers is still a matter of research, and what we want to address here.

If we agree with the theory of epistemic community [4], which stresses that those communities are project-oriented communities of experts, evaluated on their contribution in terms of knowledge, the main criterion for promotion in the different steps of their career will be their knowledge production. In Open Source projects, “developers who were making greater technical contributions and who were more engaged in organization building were more likely to become members of the leadership team”. [18, part II, p. 1096]. [6] found the same result regarding the IETF community.

However, [19], for Wikipedia, [26], for open source, have argued that becoming a big contributor may be an additional step from being a regular contributor, an additional commitment, which would occur for reasons developed during the attendance of the project. If we follow their argument, social interactions with peers may be an additional requirement for being promoted [21].

This article discusses whether contribution in knowledge or social connections matter more for being promoted in an online epistemic community, by looking at the electing process of the administrators (admin) in the English Wikipedia.

The contributions of this article to the literature are:

## II. LITERATURE REVIEW

These projects are made possible by the aggregation of various motivations and levels of involvement ([25, 10], and more recently [5]). This is something classic social theories, like the critical mass theory, about the construction of collective action [17], or the theory of (knowledge) commons [12], have stressed long time ago.

In open source software, long-term participants enjoyed programming and interacting with the rest of the community, whereas short-term participants were typically driven by an immediate need for software [23]. Theoretical analyses of incentives, in software projects [7] or in wikis [8], estimate that the other main vector for participation is the quest for reputation. Applied works on Wikipedia [28], professional electronic networks [27], and open source software [22] confirm that peer recognition, whether being professional or community recognition, is a main motive for contribution amongst the main contributors, in addition to intrinsic factors (personal enjoyment and satisfaction from helping by sharing their knowledge).

Beyond a certain level of contribution, one can even wonder if these social interactions matter more than additional knowledge production for receiving peer recognition. More precisely, we wonder if it is possible to predict the promotion of a user according to her activity, separating this activity into knowledge production (i.e. “edits” of the articles, or “revisions”, for Wikipedia, “commits” for open source software), and social activity. This is what we tested here, using the English Wikipedia as a case study.

Wikipedia has become one of the most successful knowledge production projects ever, with more than 4 million articles

for the English version and more than one million visits per day, and is seen as a model for knowledge management theory [9].

This project has a quite competitive process of election for the managing position called “administrator”<sup>1</sup>, where social connections and knowledge production skills seem to matter. Any user can request to be promoted administrator. First, the candidate creates a page dedicated to the request. During seven days which can be assimilated to a campaign, the candidate is questioned and his characteristics are studied by anybody from the community. Every user can vote for or against the candidate, and can change her mind at any moment during this period. At the end of the campaign, following the votes and the discussions on the Request for Adminship (RfA) page, special users called “bureaucrats” give their verdict on accepting or rejecting the RfA. There is no objective threshold on the percentage of support votes needed to be elected by bureaucrats. Nevertheless, it appears that a candidate is more likely to pass if she achieves at least an 80% support. With less than a 70% support, the candidate is generally not promoted.

From 2001 to 2008, on the English Wikipedia, 2794 users have requested to become administrators. Among them, 1248 requests have been accepted, so that the success rate is about 44.7%.

Previous studies shed light on this entanglement of the knowledge production skills with the social skills for being promoted. Considering the knowledge skills, voters are more likely to give positive votes when candidates are more active than them in terms of “edits”, but also to those they have talked to before [15]. Voters are also more likely to participate to elections involving their contacts, and influent users participating in a vote can influence the final result [14]. The probability of a positive vote, with an accuracy of 84%, is function of the intensity of the voter’s relations (i.e. co-editions, discussions, co-reverts<sup>2</sup>) with the candidate [13].

Closer to what we want to study here, Burke and Kraut [3] proposed a model to predict RfA results, mainly based on counting attributes for modeling the candidate’s activity, according to the criteria put forward by the Guide to RfAs<sup>3</sup>: Strong edit history, Edit summaries (explaining what they did when editing), Varied experience, User interaction, Helping with chores (i.e. already working on admin tasks such as discussing articles for delation), Trustworthiness, Observing consensus, and having various experiences in terms of editing, user interaction, etc. Their model’s accuracy reached 75.6%. However, they did not measure the respective influence of the edits and of the social interaction on RfA results. Neither did they separate social networking with administrators from social networking with everyone, whereas an administrator (or a bureaucrat) may be more influential than an unknown user on an RfA result [14].

1. Administrators have more rights than normal users on Wikipedia; they can (un)block specific users from editing pages, they can do some special actions on pages like (un)protecting from editing, (un)deleting, renaming, reporting vandalism, etc.

2. The co-reverts acting negatively.

3. <http://en.wikipedia.org/wiki/Wikipedia:GRFA>

### III. DATA COLLECTION AND MODEL

To measure the respective importance of contribution in knowledge and socialization for being promoted, we separated the profile of a candidate in two parts. A revision part, which focuses on the revision activities of the candidate, and a social part which is based on her social activities.

#### A. Dataset

To be able to benchmark our results with the previous studies we presented, we used a dataset given by the Stanford Large Network Dataset Collection<sup>4</sup> on Wikipedia. We focused on the period of time from 2006-01-01 to 2007-10-01, because this period contained an important number of RfAs, and because there were sufficient activities before to construct the social networks based on user talks. In this considered period, we removed the RfAs done several times in a month by a same user (all of which failed) that are likely to be noisy data (bots, mistakes). Hence, the resulting number of RfAs in the dataset was 1,617, with a 49.2% rate of success.

1) *The variables*: In this part, we describe the different features we consider for modeling the candidates.

a) *The knowledge contribution activities*: We extract the number of revisions/editions made (variable: *Revision*), the number of distinct pages edited (*Pages*) and the number of distinct categories edited (*Categories*), and finally, the repartition of the revisions (*Revision<sub>repartition</sub>*) in order to take into account both the volume and the variety of the revisions<sup>5</sup>.

We assume that the talks on the articles’ discussion pages were related to the revision activities, and we add three attributes for this activity: the number of distinct pages the candidates have talked on (*TalkPages*), the total number of talks on the articles’s discussion pages (*PageTalks*). We consider *PageTalks<sub>repartition</sub>* as the repartition of the talks on these pages.

b) *The social part*: We focus on the conversations on the users’ pages, to assess the impact of what happens beside the discussion on the edits, and more generally, beside the interactions regarding the production of knowledge. To evaluate the hypothesis that socialization with future “peers” has more influence on the promotion than indistinctive socialization, we create three weighted and oriented graphs, based on the social interaction: a general one, named *userSN*, in which the nodes are the considered candidates and all the wikipedians, and two specific ones:

- a graph, named *adminSN*, in which the nodes are the considered candidates and all the (already) admins;
- a graph, named *burSN*, in which the nodes are the considered candidates and all the (already) bureaucrats.

For each graph, we compute the attributes that described the characteristics of each node “candidate”. These attributes are described below. As they are the same for each graph, we only give one name for each type of attribute, and add a suffix which is the name of the related graph.

4. <https://snap.stanford.edu/data/##wikipedia>

5. For this, for each candidate, we calculate the Gini coefficient on the number of revisions by pages. This attribute allows to quantify the inequalities in a distribution. A value close to 1 indicates a focus on few pages among the whole set of revised pages. A value close to 0 indicates an equal revision behavior on every page.

The first attribute is the degree of the node (*Degree*), without taking into account the orientation of edges. Then, for more details, we considered 1) the outdegree of the node (*outDegree*), which represents the number of distinct users/ admins/bureaucrats to whom the candidate posted a message, on their user page, 2) the indegree of the node (*inDegree*), which is the number of distinct users (resp. admins, bureaucrats) that posted a message on the candidate’s page. Then, we considered the total number of messages posted and received by the candidate (*TalksNumber*). It is different from the *Degree* since weights are used here. The graph being oriented, we take into account the total number of messages posted to users/admins/bureaus pages (*outTalksNumber*) and the total number of messages received by the candidate (*inTalksNumber*). Then, we compute multiple centrality measures on the graphs:

- the closeness centrality attribute (*Closeness*);
- the PageRank centrality (*PageRank*);
- the betweenness centrality (*Betweenness*).

Finally, we compute the Gini coefficient for both the number of messages posted by the candidates (*outTalks<sub>repartition</sub>*) and the number of messages received by them (*inTalks<sub>repartition</sub>*). These attributes allow to quantify the repartition of the messages from or for the candidates. As previously mentioned, a low value (0) means a dispersed behavior whereas a high value (1) means a focused one.

### B. The models

We create multiple predictive models of RfA success based on the random forest algorithm. This algorithm is a learning method for classification (and regression) that operates by constructing a multitude of decision trees [20] during training time, and outputting the class that is the dominant value (mode) of the classes output by individual trees. For more details, see [2].

Each predictive model considers a different modeling of candidate profiles, taking into account subsets of features from the modeling proposed in the previous section. Since we want to understand the contribution of the social attributes in the RfA result, we first create two predictive models, one based on the revision attributes (Model 1) and one based on the social attributes (Model 2). Then, we consider a model using all the attributes (Model 3). The different types of profiles for each model are described below:

- 1) Model 1 (revisions):  
*Revisions, Pages, Categories, TalkPages, PageTalks, Revision<sub>repartition</sub>, PageTalks<sub>repartition</sub>*
- 2) Model 2 (social) with 32 social attributes<sup>6</sup> computed in *adminSN, burSN* and *userSN* networks  
*Degree, outDegree, inDegree, TalksNumber, outTalksNumber, inTalksNumber, Closeness, PageRank, Betweenness, outTalksRepartition, inTalksRepartition.*
- 3) Model 3 (revisions + social) with the whole set of variables.

For each predictive model, we separate the dataset in a training and a test sets. The training set consists in a random 70% of all the candidates and the test set contains the

remaining 30%. Then, the predictive model is trained on the first one and applied on the second one to predict RfA success. We compare those predictions to the test set’s real RfA success value, to deduce the accuracy value of each predictive model (ratio of the number of good predictions over the number of predictions). This process is done 100 times to smooth out the extreme cases. We present the boxplots of the results in Figure 1.

There are many attributes in Model 3 and some of them may be useless, being very correlated to others. Hence, to process Model 4, we calculated the Pearson correlation on all the pairs of attributes and we removed one element of every pair whose the absolute value of the correlation was over 0.8, which is as good in prediction and better in variance (Figure 1). In this model, we remove *Degree, TalksNumber, Closeness* and *Betweenness* in the 3 networks. We also remove *Pages, PageTalks*, and both *outTalksNumber* and *inTalksNumber* in the networks *userSN* and *adminSN*. All these attributes are very correlated to the ones we keep.

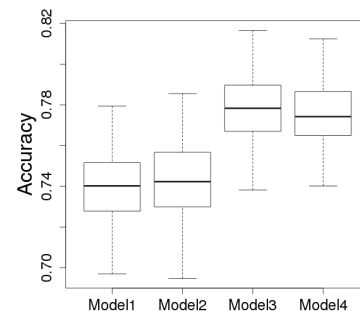


Figure 1: Prediction accuracy for the four models

Table I details the accuracy of Models 1, 2, and 4 in predicting either an unsuccessful (0) or a successful (1) promotion, giving the confusion matrix.

	Confusion Matrix			Accuracy
	0	1		
Model 1 (Revisions)	0	169	77.5	68.6%
	1	48	190	79.8%
Model 2 (Social)	0	168	80	67.7%
	1	46.5	192	80.5%
Model 4 (Revisions + light social)	0	177	70	70.6%
	1	37	200	84.4%

Table I: Confusion matrix and accuracy.

## IV. RESULTS

Figure 1 and Table I show that the predictive models based on either revision attributes (Model 1) or social attributes (Model 2) are almost equivalent in terms of quality of prediction.

Aggregating social and revisions (Model 4) improves the median prediction accuracy up to 77.8%, while the accuracy is 84.4% for predicting successful promotions and 70.6% for predicting unsuccessful ones. According to these results,

6. 32 and not 33 attributes: the betweenness centrality has not been computed on the general graph *userSN* because of limited computer capacity.

Successful promotions	Unsuccessful promotions
outDegree_userSN	Revisions
Revisions	TalkPages
TalkPages	inDegree_userSN
outDegree_adminSN	Categories
Categories	outDegree_adminSN

Table II: The most important attributes to predict successful and unsuccessful promotions

social and revision attributes seem to be complementary for predicting the RfA results.

The random forest method gives values for quantifying the importance of each attribute for the quality of the prediction. For this purpose, it computes the average decrease of accuracy for each tree in the forest when a given attribute is not used. The higher the value, the more important the attribute for the prediction is. Figure 2 shows that the most important attributes in general are *Revisions*, *TalkPages*, *outDegree<sub>userSN</sub>*.

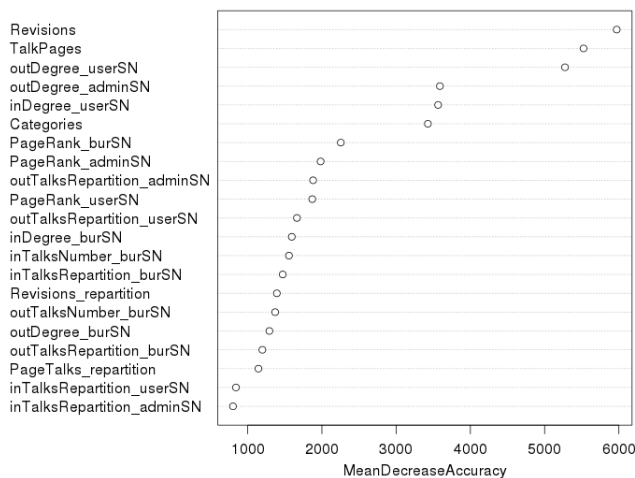


Figure 2: Importance of attributes in Model 4

To predict the unsuccessful promotions, two of these three attributes stand out: *Revisions* and *TalkPages*. But these attributes which are also standing out to predict a successful promotion, are not in the same order: *outDegree<sub>userSN</sub>*, *Revisions*, *TalkPages*, giving more importance to *outDegree<sub>userSN</sub>* (see Table II).

Unfortunately, these results do not give any information on the preferred values for each attribute: high or small *outDegree<sub>userSN</sub>* for example. A high *outDegree<sub>userSN</sub>* means discussions with a lot of different wikipedians (neither administrators nor bureaucrats). If it is the high value of this attribute which is relevant to predict election, it would mean that the commitment in the whole community would be a sign to become a good administrator. To explore this, we first compare the density of probabilities of each attribute between accepted candidates and rejected ones. For each of the main 9 attributes, but the PageRank, there are significant

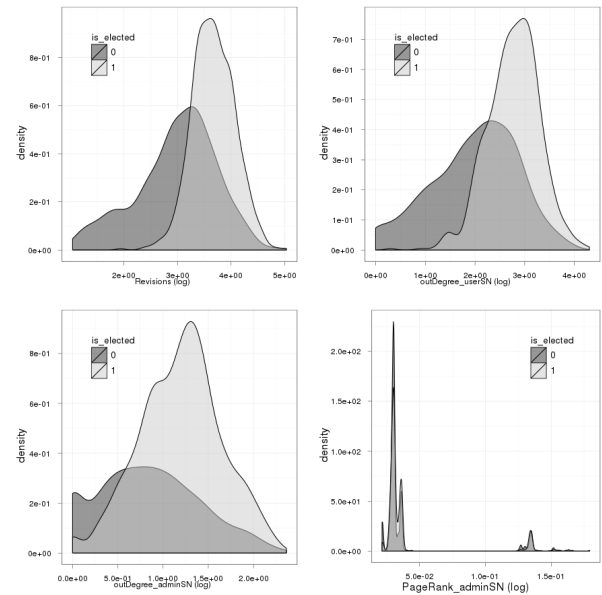


Figure 3: Density of probabilities of each attribute for promoted and non-promoted candidates

behavioral differences between the promoted and the non-promoted candidates: the interdecile range of the probability density for the promoted candidates is smaller than the one for non-promoted candidates (the curve is flatter). Figure 3 shows the densities of probabilities for the three first attributes in terms of accuracy importance (the talk page curves have the same shape as the revision curves), and the PageRank curve.

This suggests that the promoted candidates behave more similarly than the non-ones. This could explain why the models predict better promotion than non-promotion. Moreover, we can see in the first three plots that the peak of density for the promoted candidates has a higher value than those for the non-promoted: the successful candidates are more active than the unsuccessful ones (for example a candidate with a greater *outDegree<sub>userSN</sub>* has more chance to be elected).

To estimate these differences, we compute the estimated probability of being promoted knowing only one attribute, one by one, for the top nine attributes.

The estimated probability of being elected according to the number of revisions (*Revisions*) is less than 50% beyond about 2,500 revisions and is about 70% beyond 6000 revisions<sup>7</sup>.

There is a similar behavior pattern (probability of 50% beyond a threshold (T) and extreme values with not enough case studies) for the following attributes: *Categories* ( $T \approx 1700$ ), *TalkPages* ( $T \approx 130$ ), *outDegree<sub>userSN</sub>* ( $T \approx 450$ ), *outDegree<sub>adminSN</sub>* ( $T \approx 17$ ), *inDegree<sub>userSN</sub>* ( $T \approx 140$ ).

The attribute *outTalksRepartition<sub>adminSN</sub>* shows that candidates that are too dispersed in terms of number of talks with admins (thus in the *adminSN* graph) have a reduced chance for being promoted.

7. It reaches extreme values (0%, 100%) beyond 40000 revisions because there are not enough candidates with such a big number of revisions. And these few cases are not easily identified: are they active contributors or vandals?

No explicit trend has emerged from the probability according to the centrality attributes. In other words, the direct, first order interactions matter more than the position in the network in terms of information diffusion. This is rather consistent with what Wikipedia is (a network of cooperation more than a network to diffuse information). This leads us to the discussion of our results.

## V. DISCUSSION

Our results are consistent with the Guide to RfA, previous results and the theories on epistemic communities. Regarding the guide, we provide much more precise figures of how many contributions and interactions are needed to have a high probability for being elected. We show that there are quite narrow windows in terms of number of contributions and discussions, in which the chances of being elected are maximized. Simplifying the measures proposed by Burke and Kraut [3] regarding the edit activities, and adding the activity on social networks leads to a better evaluation of the chances to be elected (75.6% of good prediction for their model, 78% for our model), while keeping the number of explanatory variables reasonably low.

Zhu et al.'s study [29] on the differences between administrative persons ("admin" or "sysop") and project leaders in the English Wikipedia, showed that local project leaders leave more task oriented messages when administrators are more invested in social exchanges, sending more personal messages to users' personal pages. We confirm that the social activity in the whole project (*inTalks*, *outTalks*, *inDegree* and *outDegree* variables on the *userSN* network) is very important to become an administrator. We also refine Antin et al. [1]'s findings that people involved from the beginning in more diverse revision activities are more likely to take administrative responsibilities. To be elected, a candidate has to be involved in various talkpages, in various articles, but not too many! Being too dispersed (not focused) in terms of revisions or being excessive in the number of contacts lead to a failure.

We also contribute to the more general discussion about how communities are created and how core developers are recruited in open source software communities Rullani and Haefliger [21]. As supposed for an epistemic community, the contribution in knowledge (*Revision*) is the first criterion to be considered as a good candidate, and a shortage of contribution is often synonymous to failure. But, once the candidates have proven their competence (production of knowledge) and their willingness to do the job (interacting with people), knowing and being known by these core members, who are the future peers, makes the difference. In our case, the variable *outDegree<sub>adminsSN</sub>* is the fifth most important explanatory variable for a positive election, and *inDegree<sub>burSN</sub>* the seventh.

There are obvious limitations to our work. First, our study addresses only one project (the English Wikipedia) and should be extended to other languages and other epistemic communities. However, as we already pointed out, our results are consistent with those found in open source software communities. Second, if our model is good at forecasting the elections (more than 80% of accuracy), it is not as good for the non-elections (around 70%). Dropping the extreme cases (people who are not elected because they talked too

much, maybe because they fought with the administrators) may improve the prediction<sup>8</sup>.

## VI. CONCLUSION

In this article, we discussed the functioning of the epistemic communities, and what is important for being recognized in such communities. We proposed to consider the social behavior of the candidates in addition to the usual knowledge contribution behavior to predict the success in being promoted. We did so looking at the election of administrators in the English Wikipedia.

We compared three predictive models using the random forest algorithm, a revision-based model, a social-based model and a mixed social and revision based model. We showed that combining social and revision behaviors, increases the accuracy rate, up to 77.8%, which is better than the 75.6% accuracy given in Burke and Kraut [3].

Beyond the predictive model, this article provided estimated probabilities of being elected according to each attribute. They highlight thresholds under which the candidates reduce their chances for being promoted, but also show that too active candidates in terms of contributions or social interactions may also find it difficult to be elected. Finally, the candidates who are too dispersed (not focused) in terms of revisions (many pages with few revisions) and also in terms of social talks (many talks with few users/admins) reduce their chances for being promoted.

Even if our results must be confirmed in other online epistemic communities, our article has very practical consequences for the managers and the people involved in those communities. Promoting and encouraging people to take responsibilities in voluntary organization is a major issue for the convenors of communities, being online or offline, and our article shows that the rules could be made more explicit. The candidates for responsibilities in those communities must be aware that beyond the "professional" skills requested to be considered for such promotion, taking responsibilities in those big communities means working in a team, and that the social skills, the knowledge of the incumbents, matter too, as social interaction and coordination are key for the team to be effective, and thus efficient. Our results suggest also that more precise criteria, such as thresholds of interactions and contributions, could be posted on the pages for explaining what is expected from the candidates to the administrative functions.

In the future, since we are interested in weighting social and contribution attributes, it would be interesting to qualify better the content of the discussions (socialization or work-related, fight or friendship exchanges). This should help to explain why some wikipedians that should be elected according to the predictive model are not. This would improve the predictive model. Finally, we could strengthen our results by studying more recent data (more recent elections) and other language projects than the English one. Finally, other statistical explorations may improve the results.

8. We actually run model 4 dropping those extreme cases. This increased the global accuracy of the prediction up to 78.8%, because it strongly increased the non-election prediction accuracy up to 76.7%. However it also decreased the election prediction accuracy to 79.9%.

## REFERENCES

- [1] J. Antin, C. Cheshire, and O. Nov. Technology-mediated contributions: Editing behaviors among new wikipedians. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work, CSCW '12*, pages 373–382, New York, NY, USA, 2012. ACM.
- [2] L. Breiman. Random forests. *Machine learning*, 45(1): 5–32, 2001.
- [3] M. Burke and R. Kraut. Mopping up: Modeling wikipedia promotion decisions. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work*, pages 27–36. ACM, 2008.
- [4] P. Cohendet, F. Créplet, and O. Dupouet. Interactions between epistemic communities and communities of practice as a mechanism of creation and diffusion of knowledge. In J.-B. Zimmermann and A. Kirman, editors, *Interaction and Market Structure*. Springer, Londres, 2001.
- [5] K. Crowston. Lessons from volunteering and free/libre open source software development for the future of work. In *Proceedings of Researching The Future in Information Systems: IFIP Working Group 8.2 Working Conference, Future IS*, page 215, Turku, Finland, June 2011. Springer. URL [http://crowston.syr.edu/system/files/ifipwg82paper\\_final.pdf](http://crowston.syr.edu/system/files/ifipwg82paper_final.pdf).
- [6] L. Fleming and D. M. Waguespack. Brokerage, Boundary Spanning, and Leadership in Open Innovation Communities. *Organization Science*, 18(2):165–180, 2007.
- [7] D. Foray and J.-B. Zimmermann. L'économie du logiciel libre: organisation coopérative et incitation à l'innovation. 52:77–93, Oct. 2001.
- [8] A. Forte and A. Bruckman. Why do people write for Wikipedia? Incentives to contribute to open-content publishing. In *Proceedings of 41st Annual Hawaii International Conference on System Sciences (HICSS)*, pages 1–11, 2008.
- [9] H. Hasan and C. Pfaff. The wiki: an environment to revolutionise employees' interaction with corporate knowledge. In *Proceedings of the Australasian Computer-Human Interaction Conference*, Sydney, 2006. OZCHI 2006.
- [10] G. Hertel, S. Niedner, and S. Herrmann. Motivation of software developers in open source projects: an internet-based survey of contributors to the linux kernel. *Research Policy*, 32:1159–1177, 2003.
- [11] C. Hess and E. Ostrom, editors. *Understanding Knowledge as a Commons. From Theory to Practice*. MIT Press, december 2006.
- [12] C. Hess and E. Ostrom. Introduction: An Overview of the Knowledge Commons. In [11], editor, *Understanding Knowledge as a Commons. From Theory to Practice*, pages 3–26. 2006.
- [13] M. Jankowski-Lorek, L. Ostrowski, P. Turek, and A. Wierzbicki. Modeling wikipedia admin elections using multidimensional behavioral social networks. *Social Network Analysis and Mining*, 3(4):787–801, 2013.
- [14] J. B. Lee, G. Cabunducan, F. G. Cabarle, R. Castillo, and J. A. Malinao. Uncovering the social dynamics of online elections. *Journal of Universal Computer Science*, 18(4): 487–505, 2012.
- [15] J. Leskovec, D. Huttenlocher, and J. Kleinberg. Predicting positive and negative links in online social networks. In *Proceedings of the 19th international conference on World wide web*, pages 641–650. ACM, 2010.
- [16] D. Mahr and A. Lievens. Virtual lead user communities: Drivers of knowledge creation for innovation. *Research Policy*, 41(1):167 – 177, 2012.
- [17] P. Oliver, G. Marwell, and R. Teixeira. A theory of critical mass interdependence, group heterogeneity, and the production of collective action. *American Journal of Sociology*, 91(3):522–556, November 1985.
- [18] S. O'Mahony and F. Ferraro. The emergence of governance in an open source community. *Academy of Management Journal*, 50:1059–1106, 2007.
- [19] C. Pentzold. Imagining the Wikipedia community: What do Wikipedia authors mean when they write about their "community"? *New Media & Society*, 13(5):704–721, August 2011.
- [20] J. R. Quinlan. *C4. 5: programs for machine learning*, volume 1. Morgan kaufmann, 1993.
- [21] F. Rullani and S. Haefliger. The periphery on stage: The intra-organizational dynamics in online communities of creation. *Research Policy*, 42(4):941–953, 2013.
- [22] W. Scacchi. Free/open source software development: recent research results and emerging opportunities. In *ESEC-FSE companion 07: The 6th Joint Meeting on European software engineering conference and the ACM SIGSOFT symposium on the foundations of software engineering*, pages 459–468, New York, NY, USA, 2007. ACM. doi: <http://doi.acm.org/10.1145/1295014.1295019>.
- [23] S. K. Shah. Motivation, Governance, and the Viability of Hybrid Forms in Open Source Software Development. *Management Science*, 52(2):1000–1014, July 2006.
- [24] E. von Hippel. Lead users: a source of novel product concepts. *Management Science*, 32(7):791–805, July 1986.
- [25] G. Von Krogh, S. Spaeth, and K. R. Lakhani. Community, joining, and specialization in open source software innovation: A case study. *Research Policy*, 32(7):1217–1241, 2003.
- [26] G. von Krogh, S. Haefliger, S. Spaeth, and M. W. Wallin. Carrots and Rainbows: Motivation and Social Practice in Open Source Software Development. *MIS Quarterly*, 36(2):649–676, 2012.
- [27] M. M. Wasko and S. Faraj. Why Should I Share? Examining Social Capital and Knowledge Contribution in Electronic Networks of Practice. *MIS Quarterly*, 29(1):35–57, 2005.
- [28] H.-L. Yang and C.-Y. Lai. Motivations of Wikipedia content contributors. *Computers in Human Behavior*, 26(6):1377 – 1383, 2010.
- [29] H. Zhu, R. E. Kraut, Y.-C. Wang, and A. Kittur. Identifying shared leadership in Wikipedia. In *Proceedings of the 2011 annual conference on Human factors in computing systems, CHI '11*, pages 3431–3434, New York, NY, USA, 2011. ACM.