Voting Behavior Analysis in the Election of Wikipedia Admins

Gerard Cabunducan, Ralph Castillo
Department of Computer Science
University of the Philippines Diliman
Quezon City, Philippines
Email: {gscabunducan, rscastillo}@up.edu.ph

John Boaz Lee
Information Technology Training Center
Department of Computer Science [ACLab]
University of the Philippines Diliman
Quezon City, Philippines
Email: jblee@ittc.up.edu.ph

***Abstract*—Past work analyzing elections in online domains has largely ignored the underlying social networks present in such environments. Here, we study the Wikipedia Request for Adminship (RfA) process within the context of a social network and pinpoint several factors influencing different stages of the voting process. The facets explored are: election participation, decision making in elections, and election outcome. We find that voters tend to participate in elections that their *contacts* have participated in. Furthermore, there is evidence showing that an individual's decision-making is influenced by his contacts' actions. The properties of voters within the social graph were also studied; results reveal that candidates who gain the support of an influential coalition tend to succeed in elections.**

I. Introduction

Wikipedia's quality is maintained by its admins who perform various maintenance tasks. The admins act as custodians of the encyclopedia and its community of contributors. Since certain privileges are granted to this group of users, membership to this group is usually deliberated upon by the community to ensure that a person seeking membership is qualified. In Wikipedia, the RfA process is instituted to give regular users administrative privileges.

An RfA begins when a user is nominated to become an admin. After the nomination, the community deliberates on the qualification of the nominee and then finally votes on the eligibility of the candidate for adminship. A voter casts either a support (positive), oppose (negative), or neutral vote. Once the voting period expires, members of a special class of admins called bureaucrats review the results of the voting and conclude with a final decision - whether to promote the nominee or not. A few things that distinguish the Wikipedia RfA from other known elections are: (1) voters can change their votes, (2) an election spans a week on the average, and (3) voters can observe the votes of others.

Although the dynamics of election has been studied extensively in the literature, both in the offline [2] and the online [1], [7] setting, these studies are usually done in an environment where the underlying social network among participants is largely unobserved.

In this paper, we construct a social network based on communication between users and use its properties to answer questions related to the voting process.

II. Related Work

The RfA process has already been studied from several different perspectives. Burke and Kraut [1] focused on the analysis of candidate characteristics that improve the likelihood of promotion. However, their analysis focused on factors at the level of an individual while our study concentrated on the network-level characteristics of a candidate's supporters.

In [7], Leskovec et al. studied the assessment strategies employed by voters and found that certain forms of *relative* assessments which are based on the relation of the voter to the candidate helped shape a voter's decision. In another paper [6], Leskovec et al. observed that the presence of triads which are implicit within the social network can explain voting behavior. Although they make use of a social network in their analysis, the distinction of our work from theirs is that we use communication, instead of votes, to build our network. In contrast to [6], we use an undirected graph since we are only concerned with the presence of communication between users.

We adopt the methodology in [6] which constructed a set of features from network information. These features are used in prediction problems. Even though both area under the curve (AUC) and accuracy are used as measures of evaluating the predictive ability of learning algorithms, we only considered AUC in this paper because it was shown in [6] that the overall pattern of performance does not change; moreover, Huang et al. [4] have shown theoretically and empirically that AUC is a better measure than accuracy.

III. Dataset and Background

A. Dataset

We scraped our data from the January 2008 dump of the English version of Wikipedia which contains the complete edit history of all pages between September 17, 2004 and January 6, 2008. We obtained a total of 2,587 elections after removing those that were either incomplete or turned down by the nominee. The elections contained a total of 22,143 negative votes, 83,141 positive votes, and 6,640 neutral votes. Out of the 2,587 elections, 1,242 resulted in successful promotions (around 48%). A total of 7,231 users participated at least once in the RfA process, either as candidates or voters. For each election, we take note of the candidate, the voters and their

978-0-7695-4375-8/11 $26.00 © 2011 IEEE
DOI 10.1109/ASONAM.2011.42

Symbol	Meaning
$n(u)$	$\{u' \in V \| (u', u) \in E\}$ The set of user u's contacts
$e(u)$	The set of elections that user u participated in
$t_j(u)$	The timewise order of user u's vote in election j, infinity is returned if user u did not participate in election j
$pos_j(u)$	$\{u' \in n(u) \| t_j(u') < t_j(u)$ and u' voted positively$\}$ The set of user u's contacts who voted positively before him in election j
$neg_j(u)$	$\{u' \in n(u) \| t_j(u') < t_j(u)$ and u' voted negatively$\}$ The set of user u's contacts who voted negatively before him in election j
$can(j)$	The candidate of election j
f_j^u	The number of u's contacts who participated before him in election j

Table 1. Table of symbols.

corresponding votes, as well as the time each vote was cast.

In addition, we collected information about the communication between users that participated in the elections. We were able to gather 1,097,223 instances of communication between 265,155 distinct pairs of users.

In all of our analyses, we followed the preprocessing described in [6] and removed the neutral votes. To remove ambiguity, we only considered the final vote of a voter.

B. Basic Definitions and Notation

We denote by $G = (V, E)$ an undirected graph that describes the social network. Each $u \in V$ corresponds to a user that has participated at least once in the RfA process, and each edge $(u, u') \in E$ represents the presence of communication between users u and u'. Two users are considered to have communicated when either one edits the other's *talk page*. A talk page is a special page in Wikipedia that belongs to a user; general communication between users are usually done on their talk pages. Heretowith, we consider user u' to be a "contact" of user u, for $u \neq u'$, if an edge exists between their corresponding nodes. We summarize in Table 1 the different symbols used in the subsequent sections of this paper.

C. Experimental Setup

We test several problems within the context of the machine-learning paradigm. For each problem, we extract a different set of features and test these with a logistic regression classifier. The features for each problem will be discussed in detail in the succeeding section.

The logistic regression learns a model of the form

$$f(z) = \frac{e^z}{e^z + 1} = \frac{1}{1 + e^{-z}}$$

where $z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + ... + \beta_n x_n$. $\beta_0, ..., \beta_n$ are the coefficients or weights estimated by the logistic regression based on the training set while x is the vector of independent variables or features for each observation.

There are two reasons that motivate our use of the logistic regression. First, the method is well-studied and is used for classifying dichotomous elements. Second, and perhaps more interestingly, each coefficient describes the contribution of its corresponding feature to the probability of the occurrence of an outcome, giving us an idea of how a feature explains an outcome. A positive coefficient indicates that its corresponding feature increases the probability of the outcome while a negative coefficient means that the feature decreases the probability of the outcome. A coefficient with a large absolute value means that the feature strongly influences the probability of its corresponding outcome while a coefficient with a value close to zero has little influence on the probability of the outcome.

IV. Methodology and Results

In the succeeding experiments, we test our assumptions on a logistic regression model. We provide the AUC score for each experiment and the learned logistic regression coefficients. The values are derived from a 10-fold cross validation.

We use *balanced datasets* in the experiments. Balanced datasets, as used in [3], are datasets composed of classes with equal number of samples. This ensures that the a priori probability of sampling from the different classes is equal.

Finally, we use the t-test assessment to determine the statistical significance of each feature.

A. Factors that Motivate Participation

We attempt to distinguish actual voters from *pseudo-voters* by using these features: (1) number of contacts that participated in the election before the sampled voters, and (2) the presence of communication between them and the candidate.

1) Features: To construct our balanced dataset we examine for each voter $u \in V$ the set $e(u)$ of elections that voter u participated in. For every election $j \in e(u)$, if $t_j(u) \geq 2$, we select at random another voter u' who has participated in the same number of elections as voter u - but did not participate in this particular election j. The first voter in an election was not considered because it is not possible for that voter to observe anybody else. Each voter u and pseudo-voter u' are logged as positive and negative observations respectively. The first feature for the positive observation is $f_j^u - f_j^{u'}$, similarly $f_j^{u'} - f_j^u$ is used as the negative observation's feature.

For our second feature, we consider communication between the candidate and the voter. Communication is represented as a binary variable which holds the value 1 if the edge $(u, can(j))$ exists in E for a voter u participating in election j. The variable has a value of 0 if the edge does not exist. Similarly, for the pseudo-voter u', we also consider communication between u' and $can(j)$.

2) Results: The method scored an AUC of 0.8183. It is remarkable that a gain of 0.3 over random guessing is achieved by only considering features based on the immediate (i.e. one-hop) neighborhood of users $u \in V$. Table 2 lists the coefficients learned by the logistic regression method. We can see that both the participation of contacts and communication between the user and candidate contribute positively to the probability of a user's participation in an election, with the user's communication with the candidate weighing more heavily. This may be due to the fact that voters are inclined

Features	Coefficient
Number of contacts	0.1907
Voter-candidate talk	0.3189

Table 2. The regression coefficients corresponding to the selected features in the election participation prediction problem.

test	$pos_j(u)$	$neg_j(u)$	voter-candidate talk
without talk	0.0651	-1.4013	n/a
with talk	0.0551	-1.3684	0.6277

Table 3. The regression coefficients corresponding to the two feature sets used in the vote sign prediction problem.

to support candidates with whom they are acquainted with. Another explanation would be the proportion of support votes in our data, which comprise about 80% of the votes.

3) Analysis: Acquiring the t-test statistic with p-values of $p < 0.000$ for each feature, we can say with 95% confidence that the features in this experiment are statistically significant.

B. Factors that Influence Voting

Next, we consider the problem of predicting the sign of a user's vote in our dataset based on his contacts' votes. This is a variant of the problem described by Leskovec et al. [6].

1) Features: We test the logistic regression model on two sets of features. The first set is based solely on the decisions of a voter's social contacts. Specifically, for each voter $u \in V$ and for each election $j \in e(u)$ we consider $pos_j(u)$ and $neg_j(u)$ as the features for this set. In other words, by simply observing the number of contacts who voted positively or negatively before a user, we try to infer the vote of a user.

In the second set of features, we include a binary variable which represents communication between the candidate and the voter in addition to the features defined in the first set.

2) Results: The first test received an AUC of 0.8740 while the second test scored 0.8996. We find that we can already explain voting behavior by just examining the immediate neighborhood of a voter. In table 3 we provide the coefficients corresponding to each feature in the two tests. While all the features were assigned coefficients that aligned with our initial intuition, it is interesting to note that the presence of contacts who have voted negatively weighs more heavily compared to those who voted positively.

3) Analysis: In both tests, all features received p-values of $p < 0.000$. We can say with 95% confidence that the features used in these experiments are statistically significant.

C. Influential Voters in the Social Network

Finally, we study the network metrics of a candidate's supporters as well as those in the opposition. We consider all the voters of each election. The voters are then divided into two groups, wherein we obtain the mean value of their respective social network characteristics, and then use the information to infer the success or failure of the election.

1) Features: The voters in an election can be divided into two general camps, the support and the opposition camp. For each election, we evaluate the social network of participants using the following metrics: degree, closeness centrality, betweenness centrality, authority, hub, PageRank, clustering coefficient, and eigenvector centrality. Please refer to [5] if unfamiliar with the terms. The vector that represents the mean of each metric for support voters is denoted by s where $s = (s_1, ..., s_8)$, corresponding to the order of characteristics previously stated. Similarly, we denote by $o = (o_1, ..., o_8)$ the vector of means of the different metrics for oppose voters. The feature vector $f = (f_1, ..., f_8)$ is then defined as $f_i = s_i - o_i$ for $1 \leq i \leq 8$. We do this to measure the dominance of either side, negative f_is denote dominance of the opposition while positive values denote the opposite. Since the different characteristics are measured using different scales, we normalize the data using z score normalization. For testing, since the number of successful and unsuccessful elections are almost equal, we no longer create a separate balanced dataset.

Top 4	Score	Bottom 4	Score
closeness	1.0619	degree	0.2020
Pagerank	0.3536	authority	0.2014
Eigenvector cent.	0.2264	betweenness	-0.1245
hub	0.2041	clustering	-0.04106

Table 4. The regression coefficients for the features used in the election outcome prediction problem grouped by their weights.

2) Results: A total AUC score of 0.8368 was achieved in the test. This result shows that a group of influential supporters (or opposers) can skew an election in favor (or against) a candidate. The learned coefficients are displayed in table 4. It is interesting to observe that different measures of influence or importance like closeness, PageRank, and eigenvector centrality have prominent weights. This observation seems to suggest that decisions of influential nodes can affect the outcome of the RfA process. Although it was not studied in this paper, a possible explanation for this result is that influential users may sway other users to vote the same way and this aggregate behavior may have an impact on the result of the election.

3) Analysis: All features obtained p-values of $p < 0.000$, which means that, with 95% confidence, the features used are statistically significant.

V. CONCLUSION

We have studied the voting process of Wikipedia from a social network perspective and have discovered factors that influence voting behavior at different stages of the election.

REFERENCES

[1] M. Burke and R. Kraut. Mopping up: Modeling wikipedia promotion decisions. In *Proceedings of the CSCW*, 2008.

[2] A. G. Greenwald, C. T. Smith, N. Sriram, Y. Bar-Anan, and B. A. Nosek. Implicit race attitudes predicted vote in the 2008 u.s. presidential election. *Analyses of Social Issues and Public Policy*, 9:241–253, 2009.

[3] R. V. Guha, R. Kumar, P. Raghavan, and A. Tomkins. Propagation of trust and distrust. In *Proceedings of 13th WWW*, 2004.

[4] J. Huang and C. X. Ling. Using auc and accuracy in evaluating learning algorithms. In *proceedings of IEEE TKDE*.

[5] M. O. Jackson. *Social and Economic Networks*. Princeton University Press, 2008.

[6] J. Leskovec, D. Huttenlocher, and J. Kleinberg. Predicting positive and negative links in online social networks. In *Proceedings of WWW '10*.

[7] J. Leskovec, D. Huttenlocher, and J. Kleinberg. Governance in social media: A case study of the wikipedia promotion process. In *Proceedings of 4th ICWSM*, 2010.