Journal of Universal Computer Science, vol. 18, no. 4 (2012), 487-506
submitted: 24/9/11, accepted: 14/12/11, appeared: 28/2/12 © J.UCS

Uncovering the Social Dynamics of Online Elections

John Boaz Lee, Gerard Cabunducan, Francis George C. Cabarle
Raphael Castillo, and Jasmine A. Malinao
(University of the Philippines - Diliman, Quezon City, Philippines
{jtlee4, gscabunducan, fccabarle, rscastillo, jamalinao}@up.edu.ph)

Abstract: Past work analysing elections in online domains has largely ignored the underlying social networks present in such environments. Here, the Wikipedia Request for Adminship (RfA) process is studied within the context of a social network and several factors influencing different stages of the voting process are pinpointed. Machine-learning problems were formulated to test the identified factors. The different facets explored are: election participation, decision making in elections, and election outcome. Our results show 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. Additionally, detailed analyses on different classes of voters and candidates were made. Finally, the structural properties corresponding to networks of election participants were analysed and these networks were found to exhibit higher degrees of community structure versus graphs of participants selected at random.

Key Words: social voter, social network analysis, logistic regression, social influence, election analysis
Category: H.3.0, J.1, J.4, M.7

1 Introduction

Much like institutions in the real world, social media sites are often guided by a group of dedicated users who are engaged in various administrative duties. Although these sites are shaped and driven by the aggregate contributions of their users, a smaller group of dedicated users usually wield most of the power and are responsible for making decisions on issues of critical importance to these sites. One such media site is Wikipedia, an online encyclopedia which has seen significant growth in terms of its content and community of users over the years of its development. Although it is collaboratively edited, its quality, based on evidences, is comparable to that of the Encyclopedia Britannica [Giles 00]. This quality is maintained by its administrators who perform various maintenance tasks on its content. The administrators act as custodians of the encyclopedia

A preliminary version of this work titled *"Voting Behavior Analysis in the Election of Wikipedia Admins"* appeared in proceedings of the 3rd IEEE International Conference on Advances in Social Networks Analysis and Mining (ASONAM '11), 2011.

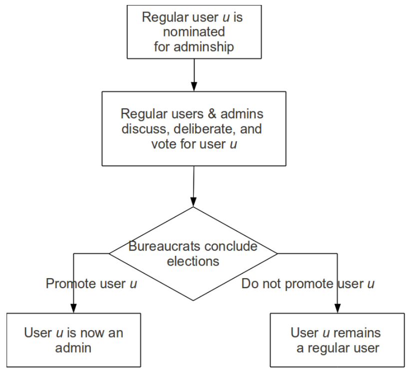

Figure 1: The Wikipedia RfA process. Note that the second block from the top, voting more specifically, can span a week on average. During the week long voting a user can change votes as well as view votes of other users.

and its community of contributors.

Since certain privileges are afforded this core 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.

In this paper, the RfA process is studied with special emphasis on the effects of the underlying social network on the voting process. An RfA begins when a regular user is nominated to become an *admin* - a user with special administrative privileges such as deletion of Wikipedia articles and entries. After the nomination, a period of discussion and deliberation ensues in which the community, composed of regular users and admins, votes on the eligibility of the candidate for adminship. A voter casts either a support (positive), oppose (negative), or neutral vote for a candidate. Once the voting period expires, a special class of admins called *bureaucrats* review the results of the voting and conclude with a final decision - whether to promote a user or not [Wikipedia 12]. The election process is shown graphically in Figure 1.

This process of deliberation can be viewed as a general form of election, similar to those conducted in offline settings as well as other online settings, wherein the goal is to achieve group consensus. However, a few things that distinguish the Wikipedia RfA from other known elections are: (1) an election spans a week on the average and voters are not required to vote simultaneously, (2) voters can observe and discuss the votes of others who voted before them, and (3) voters are allowed to change their votes.

Although the dynamics of election has been studied extensively in the literature, both in offline [Greenwald et al. 09][Rand et al. 09] and online settings

[Burke and Kraut 04][Leskovec et al. 10a][Brzozowski et al. 08], these studies are usually done in an environment where the underlying social network among participants is largely unobserved.

A social network based on *communication* between users is constructed and the network's properties are used to answer questions related to the voting process. The contributions of this paper are as follows:

- Results show that a user's tendency to participate in an election is influenced by his contacts' participation. Additionally, communication between a user and a candidate increases the likelihood of the former's participation in the latter's election.
- A voter's decision is also shown to be influenced by the actions of the voter's contacts.
- Network properties (e.g. degree, centrality, etc.) of participants in an election are analysed and it is shown that these properties can help explain the outcome of an election.
- Voters are further analysed by dividing them into classes based on the frequency of election participation and remarkable differences between the classes are discovered. Network properties of candidates are also studied and a correspondence between these properties and the success of the candidate is identified.
- The structural properties of networks comprised of election participants are studied and it is discovered that these networks exhibit a higher degree of community structure versus networks of participants chosen at random.

The rest of the paper is organized as follows. Related work are discussed in Section 2. In Section 3, the dataset is described and necessary concepts are introduced. The paper then proceeds, in Sections 4, 5, and 6, with an explanation of the proposed methods and discussion of experimental results. Finally, the conclusion and some directions for future work are given in Section 7.

2 Related Work

In his seminal work, Granovetter suggested that the behaviour of members in a social network is governed by the actions of their co-members [Granovetter 78]. Evidence was found to suggest that individuals are influenced into a certain behaviour once the threshold for that behaviour is exceeded. In relation to this, recent studies [Krebs et al. 05] on voter behaviour now place emphasis on *social voters* - citizens who do not make decisions in a social vacuum.

The RfA process has already been studied from several different perspectives. Burke and Kraut [Burke and Kraut 04] focused on the identification and analysis of candidate characteristics that improve the likelihood of promotion. However, their analysis focused on factors at the level of an individual while this study is based on the network-level characteristics of a candidate's supporters.

In [Leskovec et al. 10a], the authors 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. They also studied the temporal dynamics of the elections and found no evidence of herding or information cascades. We also study the temporal dynamics of an election, albeit at a finer level, by observing how the cumulative decisions made by a user's contacts affect the user's decision.

In another paper, Leskovec et al. [Leskovec et al. 10b] observed that the presence of triads which are implicit within the social network can explain voting behaviour. While they make use of a social network in their analysis, the distinction of this work from their work is that communication is used in this work to define the network while the network used in the previous paper was based on votes.

Another work that uses a social network to analyse the RfA process is [Turek et al. 11]. In the second experiment of the previous paper, edit history was used to derive a social network to analyse user behaviour during RfA.

The network structure of discussion pages in Wikipedia were analysed in [Laniado et al. 11] and specific assortativity profiles were derived in an attempt to differentiate article discussions from personal conversations. A special class of discussion pages were also analysed in this work to model user interaction.

Similar to the methodology in [Leskovec et al. 10b], a social network of election participants is constructed and properties taken from the network are used as features in prediction problems. The machine-learning approach allows the authors to come up with concrete formulations of questions regarding voting behaviour and is a way to approach the goal of uncovering the social dynamics of voters in online elections.

Network metrics taken from the collaborative network of software engineers were also used in [Meneely et al. 08] in order to predict software failures. This shows the vitality of using structural measures in determining factors that contribute to a particular outcome.

In this paper, the logistic regression is used in the prediction problems, a similar paper that uses a linear model to weigh factors that influence a particular outcome is done by [Canini et al. 11]. They used coefficients from fitted linear models in assessing the influence of a set of factors on credibility judgements.

A preliminary version of this work appeared in [Cabunducan et al. 11], this work differs from the previous version by including a detailed analysis of different classes of voters. Furthermore, this work helps reinforce the findings in the previous work by the providing a structural analysis of graphs of election participants.

3 Basic Definitions and Notations

3.1 Dataset

Data used in this work was scraped from the January 2008 dump of the English version of Wikipedia which contained the complete edit history of all pages between September 17, 2004 and January 6, 2008. 2,587 elections were obtained after elections that were either incomplete or turned down by the nominee were removed. 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 were succesful (around 48%) while 1,345 were unsuccessful (around 52%). 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 corresponding votes, as well as the time each vote was cast.

In addition, information about the communication between users that participated in the elections were collected. A total of 1,097,223 instances of communication between 265,155 distinct pairs of users were observed.

In all of the analyses performed, the preprocessing in [Leskovec et al. 10a] was followed and neutral votes were removed. In the rare occassion that a user changes his vote, the final vote is considered as the user's vote to avoid ambiguity.

3.2 The Social Network Based on Talk Pages

An undirected graph that describes the social network of Wikipedia users in terms of their talk page communication is denoted by $G = (E, V)$. Each $u \in V$ corresponds to a user that has participated at least once in the RfA process, and each edge $(u, u') \in E$, for $u \neq u'$, 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 single user, general communication between users are usually done on their talk pages. Heretowith, a user u is considered to be a "contact" of user u' (and vice versa) if an edge exists between their corresponding nodes. In this work, the terms "communication" and "talk" are used interchangebly.

It is clear that a user's attempt at communication can be unreciprocated and edges can be directed; furthermore, the amount of words exchanged or the frequency of posts can be used to add weight to the edges. In this work, however, the authors only deal with the general case wherein edges are undirected and

Notations	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
$t_j(u)$	From a set of users V, $t_j(u) : V \to n \in \mathbb{N}^+ \cup \{\infty\}$ such that u is the nth voter in election j (e.g. if user u voted first in election j and was followed by user v while user w was the seventh voter, then $t_j(u) = 1$, $t_j(v) = 2$ and $t_j(w) = 7$; $t_j(u) = \infty$ if user u did not participate in election j)
$\mathcal{P}_u^j$	$\{u' \in N_u \| t_j(u') < t_j(u)$ and u' voted positively$\}$ The set of user u's contacts who voted positively before u in election j
$\mathcal{N}_u^j$	$\{u' \in N_u \| t_j(u') < t_j(u)$ and u' voted negatively$\}$ The set of user u's contacts who voted negatively before u in election j
$can(j)$	The candidate of election j

Table 1. Table of notations.

unweighted. Also, it is important to note that the element of time is not incorporated in the network i.e. an edge (u, u') is present in the network as long as users u and u' have communicated once. However, in the elections being studied, the votes are ordered in a sequence according to the time each vote was cast. The resulting graph G is connected, and has average node degree of 73.34 and diameter 5.

Based on the elections and the underlying social network, the following notations in Table 1 were defined. These notations are used in subsequent sections of the paper.

4 Experimental Setup

In this work, three different facets of the RfA process are explored: (1) election participation, (2) decision making in elections, and (3) election outcome. Binary classification problems are formulated to help the authors gain insight into these areas. To understand election participation, the authors define a problem to classify real participants from non-participants. The second is a problem to predict the sign of a user's vote (positive or negative) while the third is a problem to identify the successful candidates from the unsuccessful ones. Relevant network-based features are then selected for each of the machine-learning problems. The results of the experiments help the authors gain insight into the role of one's relationships and position in the network in influencing voting behaviour and outcome.

Each of the problems are tested with a logistic regression classifier. 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 $x = (x_1, x_2, ..., x_n)$ is the vector of independent variables or features for each observation.

There are two reasons that motivate the use of the logistic regression. First, the method is well-studied and is used for classifying dichotomous elements [Leskovec et al. 10b][Hosmer and Lemeshow 00]. 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 an absolute value close to zero has little influence on the probability of the outcome.

In the experiments in the succeeding sections, the assumptions in this work are tested on a logistic regression model and the AUC score for each experiment as well as the learned logistic regression coefficients of the features are provided. The values are derived from a 10-fold cross validation. Even though both area under the curve (AUC) and accuracy are used as measures for evaluating the predictive ability of learning algorithms, only AUC is provided in this paper because [Leskovec et al. 10b] have shown that the overall pattern of performance does not change. Moreover, as shown by [Huang and Ling 08], AUC is statistically more consistent and discriminating than accuracy in evaluating learning algorithms for binary classification tasks both in balanced and imbalanced datasets. Furthermore, their results show that AUC has a higher degree of consistency in balanced dataset than in imbalanced ones although a lower degree of discriminancy is observed in balanced datasets.

Balanced datasets are used in the experiments. Balanced datasets, as used in [Guha et al. 04], 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. Using balanced datasets also ensures a baseline score of 0.5 for a classification algorithm based on random guessing.

Finally, statistical analyses are conducted on the features used in each experiment and the derived regression model. A *t*-test assessment is used for testing the statistical significance of each feature.

The prediction problem tackled in each experiment and the features used therein are discussed in detail in the succeeding section.

5 Prediction Problems

5.1 Factors that Motivate Participation

The first problem tackled is a problem analogous to the edge prediction problem [Gomez-Rodriguez et al. 10][Liben-Nowell and Kleinberg 07][Backstrom et al. 10][Jung 10][Jung 12][Juszczyszyn et al. 11]. Given a balanced dataset where half of the voters participated in an election while the other half did not, an attempt is made to distinguish real voters from *pseudo-voters* - participants of other elections that are tested against an actual voter. Here, the factors that motivate participation in the RfA process are studied.

Samples of an actual voter and a pseudo-voter are taken and compared by using their respective social networks. The comparison is based on two features: (1) the number of contacts that participated in the election before the sampled voters, and (2) the presence of communication between them and the candidate. An attempt is then made to distinguish the actual voter from the pseudo-voter using this information.

5.1.1 Features

To construct the balanced dataset, each voter $u \in V$ is considered and the set E_u of elections that voter u participated in is examined. For every election $j \in E_u$, if $t_j(u) \geq 2$, another voter u' who has participated in the same number of elections as voter u was selected at random - u' did not participate in this particular election j. The first voter in an election is not considered because it is not possible for that voter to observe anybody else. The number of u's contacts who participated before him in the election is denoted by $f_j(u)$, similarly, $f_j(u')$ is the number of u''s contacts who participated before u in the election. Each corresponding u and u' pair are logged as a positive and a negative observation 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 the second feature, communication between the candidate and the voter is considered. 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', communication between u' and $can(j)$ is also considered.

Experiment 1		Experiment 2		Experiment 3	
Feature	Coefficient	Feature	Coefficient	Feature	Coefficient
$f_j(u) - f_j(u')$	0.1907	$\mathcal{P}_u^j$	0.0651	$\mathcal{P}_u^j$	0.0551
talk	0.3189	$\mathcal{N}_u^j$	-1.4013	$\mathcal{N}_u^j$	-1.3684
				talk	0.6277

Table 2. The regression coefficients corresponding to the selected features in the first three experiments.

5.1.2 Results and Discussion

The method scored an AUC of 0.8183. It is remarkable that a gain of 0.3 over random guessing is achieved by considering features in the immediate neighborhood of a user alone. Table 2 Experiment 1 lists the coefficients learned by the logistic regression method. Both the participation of a user's contacts ($f_j(u) - f_j(u')$) and communication (talk) between the user and candidate is seen to 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 observation may be due to the fact that in the dataset, 80% of the votes cast are support votes and voters are inclined to support candidates with whom they have established communication with. The first feature also has a positive coefficient which is indicative of the fact that users seem to be influenced to participate in an election if they observe their contacts' participation.

5.1.3 Analysis on Different Types of Voters

Similar to [Leskovec et al. 10a][Jung 11], the voters are divided into two groups: (1) "frequent voters" - voters that have participated in more than 90 elections, and (2) "infrequent voters" - those who have participated in less than 91 elections. The voters are split in this way since users from the two groups participated in roughly the same number of elections. The number of voters in each group who joined the election after observing i contacts are counted, for $i \in \{0, 1, 2, ..., 9\}$.

Since the average degree of each node in the communication network is only 73 and there are 7,231 nodes in the network, assuming one is the second to vote, the probability of observing a contact before you in the same election is $\frac{73}{7{,}230} \approx 0.0101$. If the users are equally likely to join an election, from the probability, it can be inferred that a higher fraction of voters should have observed only i contacts before joining an election versus $i+1$ contacts, for all i. However, it is remarkable that the graph (b) in Figure 2 shows that infrequent voters are more likely to participate in an election after a few contacts have joined. This indicates that infrequent voters are more likely to be influenced to join an election by their contacts' participation. Frequent voters, on the other hand, are less

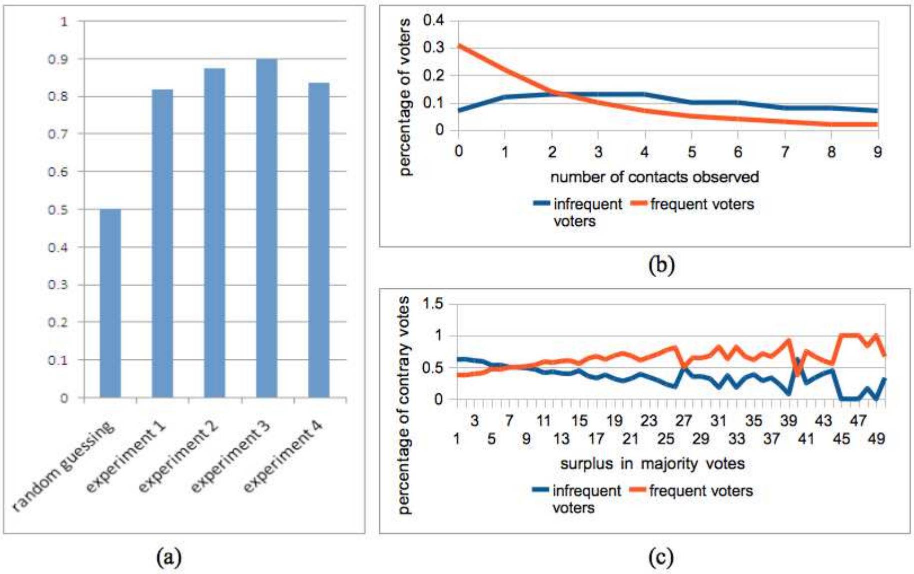

Figure 2: (a) AUC scores of each experiment. (b) Percentage of voters in a group who voted versus the number of contacts who voted before them. (c) Percentage of voters who voted contrary to their contacts' consensus versus the number of surplus votes in the consensus.

affected by the actions of their contacts. However, from the gradual decrease of the curve corresponding to frequent voters, it can still be concluded that contacts do affect the decision of both classes of voters in participating, albeit at different degrees.

In the next analyses, a different set of classes is considered: "more-frequent voters" and "less-frequent voters". The more-frequent voters consist of users who have participated in at least 50 elections, while the less-frequent voters consist of users who have participated in only 5 elections or less.

There are 494 more-frequent voters and 3,373 less-frequent voters. Among the more-frequent voters, there are 55,563 candidate-voter instances; 46.48% of these show evidence of communication between the voter and the candidate through the Wikipedia talk page. On the other hand, out of the 5,394 voting instances among the less-frequent voters, 54.26% have evidence of candidate-voter communication. It is remarkable that the proportion of instances with candidate-voter communication or "talk" evidence is higher in less-frequent voters than in more-frequent voters. It is possible that in the case of less-frequent voters, they are more inclined to participate in the election if they have "talked" with the

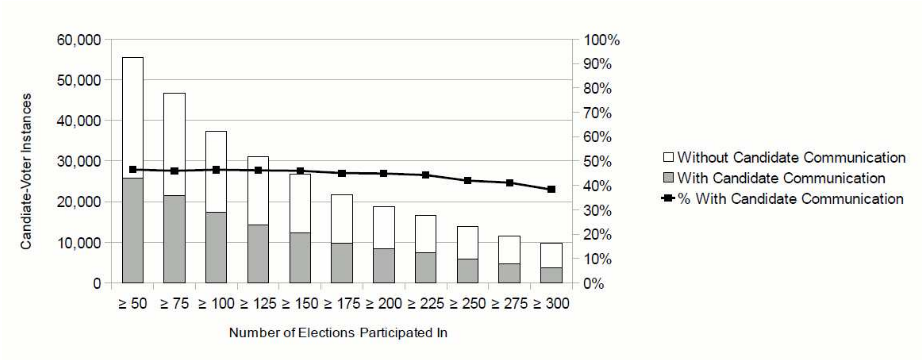

Figure 3: Number of elections participated in and the proportion of communication between voter and candidate.

candidate; while in the case of more-frequent voters the consideration is lower, in fact lower than simple majority, which may be indicative that there are other factors that drive the more-frequent voters in their participation.

Findings in this work are further reinforced by obtaining the local clustering coefficients [Latapy et al. 08] associated with the more-frequent and less-frequent voters. The results show that the average local clustering coefficient of more-frequent voters is 0.1683, while in less-frequent voters, it is 0.2416. A possible explanation is that less-frequent voters tend to be part of a more clustered community, and they tend to participate in an election if the candidate is part of their community. This conclusion is reinforced by the proportion of the "talk" evidence mentioned above. Also, by observing the maximum clustering coefficient, the authors find that a local clustering coefficient of 1.0000 exists in the less-frequent voters, while in more-frequent voters the highest is only 0.3975. However, the mean degree of nodes corresponding to less-frequent voters is 5 while the mean degree of more-frequent voters is 256. This seems to suggest that more-frequent voters are part of larger but sparser communities while less-frequent voters are usually part of small but tight-knit communities. Figure 3 shows the downward trend of the proportion of voting instances with candidate-voter communication. Figure 4, on the other hand, shows the trend of the local clustering coefficient values of voters.

5.1.4 Statistical Significance of Features

Acquiring the t-test statistic with a p-value of $p < 0.000$ for each feature, it can be said with 95% confidence that the features in this experiment are statistically significant.

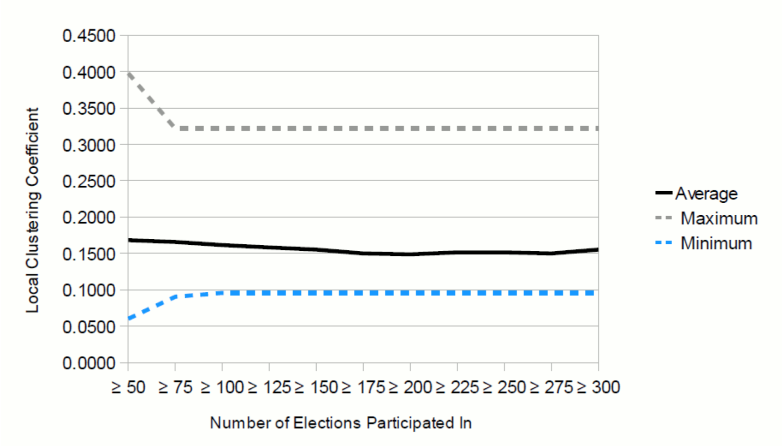

Figure 4: Number of elections participated in and the local clustering coefficient.

5.2 Factors that Influence Voting

Next, the problem of predicting the sign of a vote in the dataset is considered. The problem stated here is a variant of the one described by Leskovec et al. [Leskovec et al. 10b]: Given a full network where access to the voting behaviour of each individual's contacts for any particular election is available, the aim is in predicting the sign of the individual's vote in that election. This is done by assessing the votes of the voter's contacts who participated before him. Here, communication between the candidate and the voter is also taken into consideration. In essence, the goal is to discover if the votes of an individual's contacts have any influence on the voter's decision.

5.2.1 Features

The logistic regression model is tested on two different sets of features. The first set is based solely on the decisions of a voter's contacts. Specifically, for each voter $u \in V$ and for each election $j \in E_u$, $\mathcal{P}_u^j$ and $\mathcal{N}_u^j$ are considered as the features for this set. In other words, the vote of the user is inferred by simply observing the number of contacts who voted positively or negatively before u.

In the second set of features, in addition to the first two features defined previously, a third binary variable is included. The variable holds the value 1 if the edge $(u, u') \in E$ for a voter u participating in election j where $u' = can(j)$, and 0 otherwise. The third feature indicates whether the voter and the candidate have communicated.

5.2.2 Results and Discussion

Figure 2 graph (a) experiments 2 and 3 shows the AUC scores obtained using the two different feature sets and in Table 2 Experiments 2 and 3 the coefficients corresponding to each feature in the two experiments are provided. The test without the communication feature scored a total AUC of 0.8740 while the second test scored 0.8996. Again, from these results, it is remarkable that it is already possible to explain voting behaviour by just examining the direct neighborhood of a voter.

While the first two features were assigned coefficients that aligned with our intuition, it is interesting to note that the presence of contacts who have voted negatively weighs more heavily compared to contacts who voted positively. In fact, in the dataset, not a single voter can be found who voted positively after majority of his contacts voted negatively. It is also worth noticing, in this context, that communication between the candidate and the voter seems to contribute more strongly to the probability of a positive vote than a single contact's support vote.

5.2.3 Analysis on Different Types of Voters

Again, the frequent voters are compared against infrequent voters and a contrast is drawn between their contacts' influence on their votes. In the tests performed, the two groups of voters voted contradictorily to the consensus of their contacts roughly the same number of times across elections. The authors counted the number of times they voted contradictorily when i more contacts voted differently, for $i \in \{1, 2, 3, ..., 49\}$. Since 99% of all voters in the dataset observed less than 50 contacts, the upper-bound for i is set to 49.

Interestingly, it can be seen, from Figure 2 graph (c) that once an overwhelming number (greater than 6) of their contacts vote a certain way, infrequent voters are more likely to follow suit while frequent voters are more likely to "stand their ground". This observation seems to tell us that "new" voters seem to rely more on their peers during decision making while "mature" voters rely on a different set of measures.

Similar to the findings in [Leskovec et al. 10a], it is found that both more-frequent and less-frequent voters tend to vote positively for candidates they have communicated with; 60% of the time, less-frequent voters voted positively for a candidate that they have communicated with. For more-frequent voters, this ratio is slightly lower at 51%.

5.2.4 Statistical Significance of Features

In both tests, all features received p-values of $p < 0.000$. It can be said with 95% confidence that the features used in these experiments are statistically significant.

5.3 Influential Voters in the Social Network

Finally, the authors study the network metrics of a candidate's supporters as well as those in the opposition. The authors attempt to identify the more "influential" of the two groups of voters - the supporters and the opposers - and analyse whether this information is telling of the outcome of the election. Given an election and the set of network characteristics of the voters, the factors that contribute to the success or failure of an election are assessed.

The analysis is done by obtaining all the voters of a specific election. The voters are then divided into two groups, wherein the mean value of their respective social network characteristics are obtained. This information is then used to infer the success or failure of the election.

5.3.1 Features

The voters in an election can be divided into two general camps, the support and the opposition camp. For each election, the following social network characteristics of the participants are gathered: degree, closeness centrality, betweenness centrality, authority, hub, PageRank, clustering coefficient, and eigenvector centrality. Please refer to [Jackson 08] if unfamiliar with the terms. It was shown, through a sampling method, that influential individuals can be approximately identified through multiple centrality measures such as betweenness, closeness, PageRank, and eigenvector centrality [Maiya and Berger-Wolf 10]. Also, structural significance of a node can be derived by computing purely structure-based properties such as degree, hub, and authority [Desikan and Srivastava 06]. A similar work that employed centrality measures in feature analysis to investigate social network profiles is done by [Musial et al. 09].

The vector that represents the mean of each characteristic for support voters is denoted by s where $s = (s_1, s_2, s_3, s_4, s_5, s_6, s_7, s_8)$, corresponding to the order of characteristics previously stated. Similarly, $o = (o_1, o_2, o_3, o_4, o_5, o_6, o_7, o_8)$ denotes the vector of means for the different characteristics of oppose voters. The feature vector $f = (f_1, f_2, f_3, f_4, f_5, f_6, f_7, f_8)$ is then defined as $f_i = s_i - o_i$ for $1 \leq i \leq 8$. This is done 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, the data is normalized using z score normalization. For testing, since the number of successful and unsuccessful elections are almost equal, a separate balanced dataset is no longer created.

Experiment 4			
Top 4	Coefficient	Bottom 4	Coefficient
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 3. The regression coefficients for the features used in the election outcome prediction problem grouped by their absolute weights.

5.3.2 Results and Discussion

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 3.

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 voting behaviour may have an impact on the result of the election.

5.3.3 Candidate Analysis

The candidates are also analysed based on their degrees and their local clustering coefficients. Graphs (a) and (b) in Figure 5 show the distribution of the candidates with respect to the above mentioned metrics. Based on graphs (a) and (b) in Figure 6, it can be noted that candidates with more neighbors or contacts are relatively more successful than those with less contacts. There are two curves; one curve represents the proportion of success with respect to the total number of RfAs, and the other is with respect to the total number of candidates. However, candidates that have higher clustering coefficients tend to be less successful. Related to the discussion in section 4.1.3, this may be due to the fact that candidates that have high clustering coefficients are part of very small tight-knit communities while candidates with lower clustering coefficients are part of large but sparse communities thus the latter group of candidates are more influential in the community at large.

5.3.4 Statistical Significance of Features

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

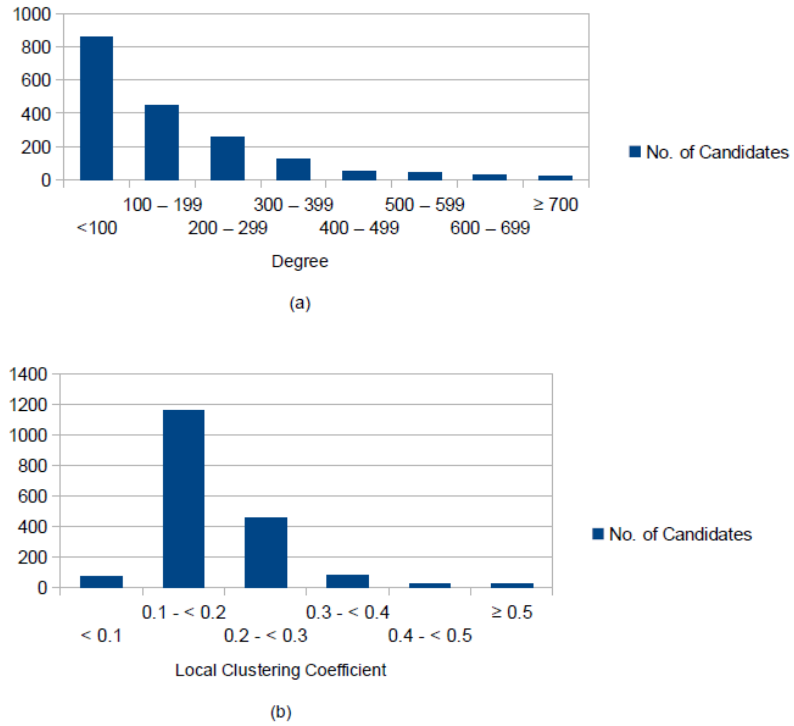

Figure 5: (a) Distribution of candidates by degree. (b) Distribution of candidates by local clustering coefficient.

6 Structural Analysis of Voter Graphs

6.1 Communities of voters

For each election j, $1 \leq j \leq 2,587$, a graph $G_j = (E_j, V_j) \subseteq G = (E, V)$ is created. The vertex set V_j contains only nodes corresponding to participants of election j, and $E_j = \{(e, j) \in E | e \in V_j \text{ and } j \in V_j\}$. For each G_j, another graph G'_j whose vertex set also contains $|V_j|$ number of nodes is also created. The random participants are selected with probability $\frac{|E_u|}{\# \text{ of votes in total}}$ so that the selection is biased towards voters who participated in more elections.

The Clustering Coefficient [Latapy et al. 08], which is a measure of the degree that nodes in a graph tend to cluster together, is then calculated for the random graphs and the graphs of real election participants. The average Clustering Coefficient of the real graphs is 0.4136 while the random graphs only scored 0.2591. It is interesting that the graphs of real voters are more clus-

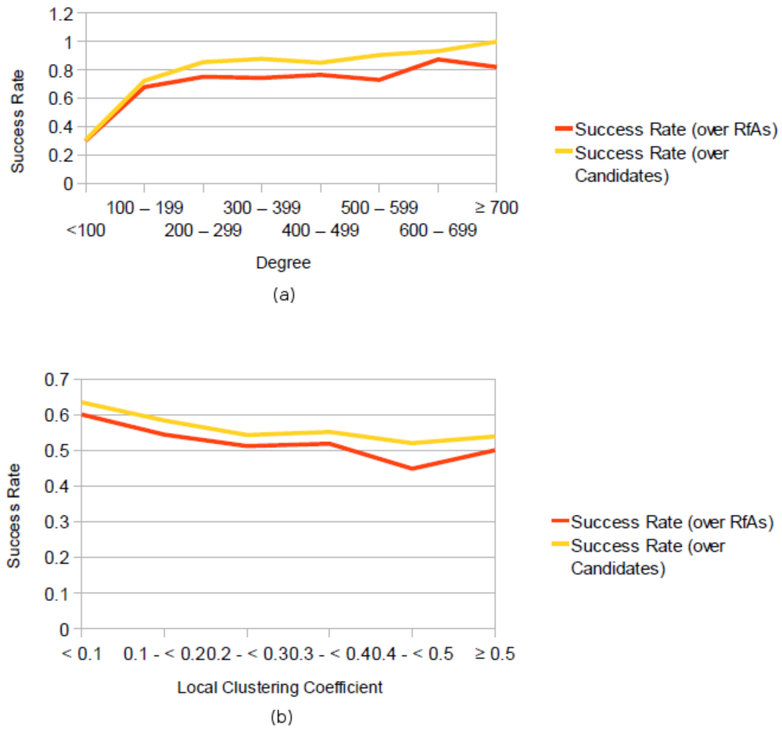

Figure 6: (a) Probability of success given the degree. (b) Probability of success given the local clustering coefficient.

tered than expected. Two possible theories may explain the clustering: (1) voters influence those around them to participate in an election, or (2) a candidate may appeal to a certain subgroup of voters. This conforms to the well known fact that real world social networks usually exhibit community structure [Newman and Park 03][Girvan and Newman 02].

Figure 7 shows an example of an election where majority of the participants form a single giant community while only three single participants did not know anybody else in the election.

7 Conclusions and Future Work

The voting process of Wikipedia has been studied from a social network perspective and factors that influence voting behaviour at different stages of the election have been discovered in this work. Viewing the election at the perspective of the

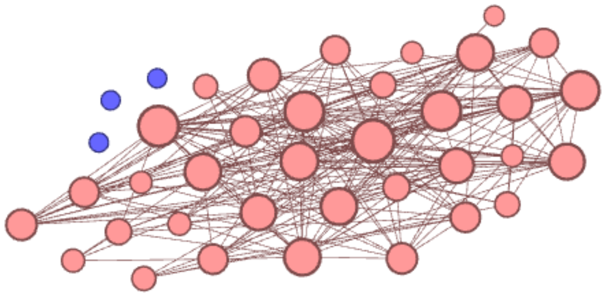

Figure 7: A subgraph of participants in a particular RfA. The Clustering Coefficient for this particular network is 0.6144.

voter's network, the authors were able to identify factors that influence voting behaviour in different stages of the election. Evidence from the results of this work show that voters tend to participate in elections where their contacts have already participated in. Results indicate that the actions of contacts have an impact on a voter's vote. The network properties of an election's participants were also studied and it has been found that the participation of voters that are relatively more influential than the other group of participants can impact the outcome of an election.

Given the identified features, high AUC scores, based on 10-fold cross validation, were obtained. All the experiments posted a gain of at least 0.3 over random guessing. All of the features used in the experiments have also been shown to be statistically significant. Although the model already performs well with the identified features, in future work the authors intend to identify additional features that will yield further insights on the dynamics of online elections.

Within the context of social network analysis, there are still areas that can be explored. An interesting consideration would be to construct a directed weighted social graph to aid in finer level examination of the data. Treating the social graph as a dynamic network that evolves over time is another area that could yield further insights. Studying the reasons that cause users to change their vote is an area of interest as well.

For the problem of predicting the outcome of an election, it is also an interesting research direction to study more subtle properties for prediction when support and opposition groups share near equal dominance.

Acknowledgement

G. Cabunducan, F.G.C. Cabarle, and J.A. Malinao would like to thank the Engineering Research and Development for Technology (ERDT) program of the

Department of Science and Technology (DOST) of the Philippines for research funding. J.B. Lee is supported in part by the University of the Philippines Information Technology Training Center. We thank Erlo Robert Oquendo and Henry Adorna for their invaluable suggestions and insights.

References

[Backstrom et al. 10] Backstrom, L., Leskovec, J.: "Supervised Random Walks: Predicting and Recommending Links in Social Networks"; Proc. of 4th WSDM, (2011), 635-644.

[Brzozowski et al. 08] Brzozowski, M.J., Hogg, T., Szabo, G.: "Friends and Foes: Ideological social networking"; Proc. of CHI, ACM, (2008), 817-820.

[Bullmore and Sporns 09] Bullmore, E., Sporns, O.: "Complex brain networks: graph theoretical analysis of structural and functional systems"; Nature Reviews Neuroscience, 10, (2009), 186-198.

[Burke and Kraut 04] Burke, M., Kraut, R.: "Mopping up: Modeling wikipedia promotion decisions"; Proc. of CSCW, (2008), 27-36.

[Cabunducan et al. 11] Cabunducan, G., Castillo, R., Lee, J.B.: "Voting Behavior Analysis in the Election of Wikipedia Admins"; Proc. of 3rd ASONAM, IEEE, (2011), 545-547.

[Canini et al. 11] Canini, K.R., Suh, B., Pirolli, P.L.: "Finding credible information sources in social networks based on content and social structure"; Proc. of 3rd International Conference on Social Computing, IEEE, (2011), 1-8.

[Desikan and Srivastava 06] Desikan, P., Srivastava, J.: "Mining Temporally Changing Web Usage Graphs"; Advances in Web Mining and Web Usage Analysis, 3932, (2006), 1-17.

[Giles 00] Giles, J.: "Internet encyclopeadias go head to head"; Nature, 438, 7070 (2000), 900-901.

[Girvan and Newman 02] Girvan, M., Newman, M.E.J.: "Community structure in social and biological networks"; Proceedings of the National Academy of Science 99, 12 (2002), 7821-7826.

[Gomez-Rodriguez et al. 10] Gomez-Rodriguez, M., Leskovec, J., Krause, A.:"Inferring Networks of Diffusion and Influence"; Proc. of 16th KDD, ACM, (2010), 1019-1028.

[Granovetter 78] Granovetter, M.: "Threshold Models of Collective Behavior"; The American Journal of Sociology 83, 6 (1978), 1420-1443.

[Greenwald et al. 09] Greenwald, A.G., Smith, C.T., Sriram, N., Bar-Anan, Y., Nosek, B.A.: "Implicit Race Attitudes Predicted Vote in the 2008 U.S. Presidential Election"; Analyses of Social Issues and Public Policy 9, (2009), 241-253.

[Guha et al. 04] Guha, R.V., Kumar, R., Raghavan, P., Tomkins, A.: "Propagation of trust and distrust"; Proc. of 13th WWW, ACM, (2004), 403-412.

[Hosmer and Lemeshow 00] Hosmer, D.W., Lemeshow, S.: "Applied Logistic Regression, 2nd edn."; Wiley, (2000).

[Huang and Ling 08] Huang, J., Ling, C.X.: "Using AUC and Accuracy in Evaluating Learning Algorithms"; TKDE 17, 3 (2005), 299-310.

[Jackson 08] Jackson, M.O.: "Social and Economic Networks"; Princeton University Press, (2008).

[Jung 10] Jung, J.J.: "Reusing Ontology Mappings for Query Segmentation and Routing in Semantic Peer-to-Peer Environment," Information Sciences, 180, 17 (2010), 3248-3257.

[Jung 11] Jung, J.J.: "Ubiquitous Conference Management System for Mobile Recommendation Services Based on Mobilizing Social Networks: a Case Study of u-Conference," Expert Systems with Applications, 38, 10 (2011), 12786-12790.

[Jung 12] Jung, J.J.: "Evolutionary Approach for Semantic-based Query Sampling in Large-scale Information Sources," Information Sciences, 182, 1 (2012), 30-39.
[Juszczyszyn et al. 11] Juszczyszyn, K., Musial, K., Budka, M.: "Link Prediction Based on Subgraph Evolution in Dynamic Social Networks"; Proc. of 3rd International Conference on Social Computing, IEEE, (2011), 27-34.
[Krebs et al. 05] Krebs, V. and Ratcliffe, M. and Lebkowsky, J.: "Extreme Democracy"; Lulu.com, (2005).
[Laniado et al. 11] Laniado, D., Tasso, R., Volkovich, Y., Kaltenbrunner, A.: "When the Wikipedians Talk: Network and Tree Structure of Wikipedia Discussion Pages"; Proc. of 5th ICWSM, AAAI, (2011).
[Latapy et al. 08] Latapy, M., Magnien, C., Del Vecchio, N.: "Basic notions for the analysis of large two-mode networks"; Social Networks 30, 1 (2008), 31-48.
[Leskovec et al. 10a] Leskovec, J., Huttenlocher, D., Kleinberg, J.: "Governance in Social Media: A case study of the Wikipedia promotion process"; Proc. of 4th ICWSM, AAAI, (2010).
[Leskovec et al. 10b] Leskovec, J., Huttenlocher, D., Kleinberg, J.: "Predicting Positive and Negative Links in Online Social Networks"; Proc. of 19th WWW, ACM, (2010), 641-650.
[Liben-Nowell and Kleinberg 07] Liben-Nowell, D., Kleinberg, J.: "The link-prediction problem for social networks"; Journal of the American Society for Information Science and Technology 58, 7 (2007), 1019-1031.
[Maiya and Berger-Wolf 10] Maiya, A.S., Berger-Wolf, T.Y.: "Online Sampling of High Centrality Individuals in Social Networks"; Proc. of 14th PAKDD, (2010), 91-98.
[Meneely et al. 08] Meneely, A., Laurie, W., Snipes, W., Osborne, J.: "Predicting failures with developer networks and social network analysis"; Proc. of 16th SIGSOFT, ACM, (2008), 13-23.
[Musial et al. 09] Musial, K., Kazienko, P., Brodka, P.: "User position measures in social networks"; Proc. of 3rd SNA-KDD, (2009), 6.
[Newman and Park 03] Newman, M.E.J., Park, J.: "Why social networks are different from other types of networks"; Phys. Rev. E 68, 036122 (2003).
[Rand et al. 09] Rand, D.G., Pfeiffer, T., Dreber, A., Sheketoff, R.W., Wernerfelt, N.C., Benkler, Y.: "Dynamic remodeling of in-group bias during the 2008 presidential election"; Proc. of National Academy of Sciences of the United States of America, (2009), 6187-6191.
[Turek et al. 11] Turek, P., Spychala, J., Wierzbicki, A., Gackowski, P.: "Social Mechanism of Granting Trust Basing on Polish Wikipedia Requests for Adminship"; Proc. of 3rd SocInfo, (2011), 212-225.
[Wikipedia 12] Wikipedia contributors, "Wikipedia:Requests for adminship"; Wikipedia, The Free Encyclopedia, `http://en.wikipedia.org/w/index.php?title=Wikipedia:Requests_for_adminship&oldid=468962866` (accessed Jan 3, 2012).