Voting Behavior Analysis in the Election of Wikipedia Admins

Gerard Cabunducan[a], Ralph Christopher Castillo[a], and John Boaz Lee[b]

[a]Computer Vision and Machine Intelligence Laboratory, Department of Computer Science, University of the Philippines Diliman, Q.C. 1101

[b]University of the Philippines Information Technology Training Center, University of the Philippines Diliman, Q.C. 1101

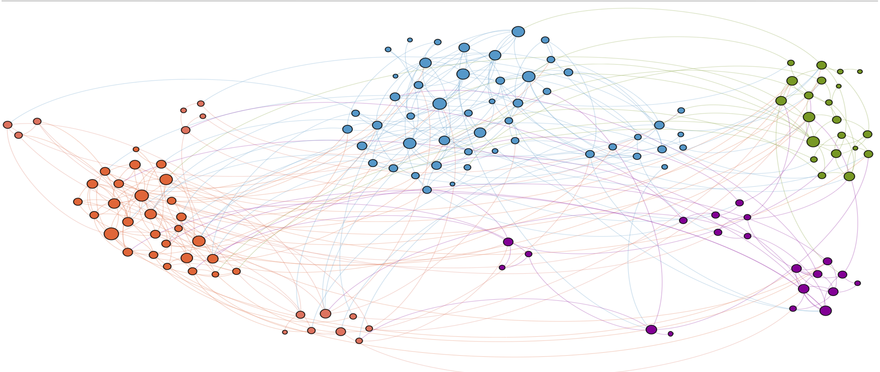

figure 1. A subgraph of the acquaintance network, defined by user communication, which we used to analyze voting behavior. The full network, comprised of 6,231 nodes and 265,155 edges, is connected, and has average node degree of 85.11 and diameter 5.

what we did

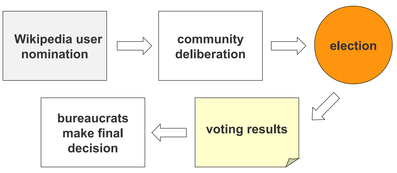

figure 2. Illustration of the Wikipedia Request for Adminship (RfA) process.

the Wikipedia Request for Adminship

When a Wikipedia user is nominated to become an admin, the community, composed of regular users and admins, deliberates and votes on the eligibility of the candidate for adminship. A voter casts either a support, oppose, or neutral vote for a candidate. The election period usually spans a week, during which the votes of prior voters can be reviewed.

we studied voting behavior of participants from a social network perspective

We constructed a social network based on communication between users and used its properties to help us analyze the voting process.

we discovered the following things

- Voters tend to participate in elections that their acquaintances have also participated in.
- Voters are influenced by the decisions of their acquaintances.
- Candidates that secure the support of "influential" nodes in the network usually succeed.

methodology

data is parsed from a complete dump of the English Wikipedia and further preprocessed to remove redundant and incomplete parts

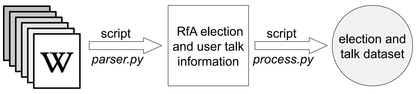

figure 3. The dataset construction process. The final dataset consisted of 2,587 elections, 1,242 of which were successful. 1,097,223 instances of communication were logged.

we answered questions related to the voting process by modeling them as machine-learning problems

We formulated machine-learning problems to study the factors influencing (1) participation in election, (2) decision-making during election, and (3) success of a candidate's adminship bid.

the logistic regression classifier is used for these reasons

- It is well studied and is used for classifying dichotomous elements.
- Each coefficient describes the contribution of its corresponding feature to the probability of the occurrence of an outcome.

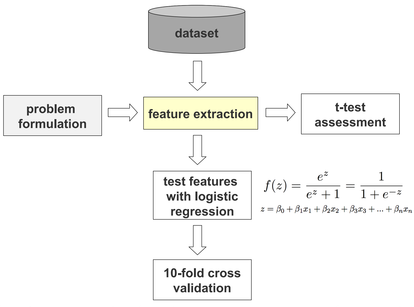

figure 4. Each experiment was formulated as a machine-learning problem, for which a set of features was extracted. Samples are then taken with the selected features and these are tested on a logistic regression classifier.

validation and testing

We used balanced datasets [1] in the experiments, these are datasets composed of classes with an equal number of samples. 10-fold cross validation is performed on all experiments. The features used in each experiment are also tested for their statistical significance.

our results

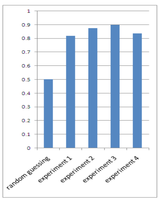

figure 5. The AUC scores of each experiment.

Regression Coefficients

Experiment 1		Experiment 2		Experiment 3		Experiment 4			
Feature	Coefficient	Feature	Coefficient	Feature	Coefficient	Top 4	Coefficient	Bottom 4	Coefficient
number of acquaintances	0.1907	positive votes by acquaintances	0.0651	positive votes by acquaintances	0.0551	closeness	1.0619	degree	0.2020
voter-candidate talk	0.3189	negative votes by acquaintances	-1.4013	negative votes by acquaintances	-1.3684	PageRank	0.3536	authority	0.2014
				voter-candidate talk	0.6277	Eigenvector centrality	0.2264	betweenness	-0.1245
						hub	0.2041	clustering	-0.0411

table 1. The regression coefficients corresponding to the different features used in the different experiments.

factors that motivate participation

First, we tackle a problem analogous to the edge prediction problem [2]. Given a balanced dataset where half of the voters participated in an election while the other half did not, we attempt to distinguish the real voters from pseudo-voters - participants of other elections that are tested against an actual voter.
Both voters participated in same number of elections overall.

We tag each observation with the following features: (1) number of acquaintances who participated before the voter, and (2) the presence of an edge between the voter and the candidate.

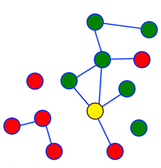

figure 6. The green, red, and yellow nodes denote respectively participants, non-participants, and the candidate of an election.

- *Scored an AUC of 0.818.*
- *Communication with candidate weigh more heavily.*
- *Both user-candidate talk and participation of acquaintances motivate user to join an election.*

factors that influence voting

Next, we consider the problem of predicting the sign of a vote in our dataset. The problem we state here is a variant of the one described by Leskovec et al. [3]: Given a full network where we have access to the voting behavior of each individual's friends for any particular election, we are interested in predicting the sign of the individual's vote in that election. We gather two sets of features.

The first set consists of two features. For each election, we tally separately the number of a voter's acquaintances who voted positively and negatively prior to the voter. For the second set, we include voter-candidate communication.

- *AUC scores are 0.874 and 0.900.*
- *Negative votes carry more weight.*
- *Candidate-voter communication has more weight than a positive vote.*
- *A voter is more likely to vote the same way as his acquaintances.*

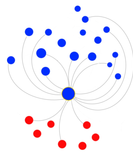

figure 7. We try to infer the vote of the center node by studying the votes of his acquaintances. Blue nodes voted positively.

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 attempt to identify the more "influential" of the two groups of voters and analyze whether this information is telling of the outcome of the election.

The features for this experiment are the difference of the mean of several network metrics of supporters and opposers. The metrics are: degree, closeness centrality, betweenness centrality, authority, hub, PageRank, clustering coefficient, and Eigenvector centrality.

- *The method scored an AUC of 0.837.*
- *Different measures of influence or importance like closeness, Pagerank, and Eigenvector centrality have prominent weights.*
- *A prominent coalition can influence election outcome.*

other matters

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.

acknowledgement

G. Cabunducan would like to thank the Engineering Research and Development for Technology (ERDT) for his graduate scholarship. J.B. Lee would like to thank ITTC for partially funding his graduate studies. We also thank the anonymous reviewers for their insightful comments.

cited literature

[1] R. V. Guha, R. Kumar, P. Raghavan, and A. Tomkins. Propagation of trust and distrust. In Proc. of WWW 2004.
[2] D. Liben-Nowell and J. Kleinberg. The link-prediction problem for social networks. Journal of the American Society for Information Science and Technology, 58(7):1019–1031, 2007.
[3] J. Leskovec, D. Huttenlocher, and J. Kleinberg. Predicting positive and negative links in online social networks. In Proc. of WWW 2010.

write us

gscabunducan@up.edu.ph | rscastillo@up.edu.ph | jblee@ittc.up.edu.ph