# CS-E4870
Research Project in Machine Learning and Data Science

Ananth Mahadevan

Department of Computer Science, Aalto University
ananth.mahadevan@aalto.fi

## Abstract

## 1 Introduction

Wikipedia is the largest online encyclopedia containing over 5 million pages of content. It is one the most popular websites on the Internet. Wikipedia has a diverse collection of articles from many different topics and is constantly being updated. Although Wikipedia started out as an open platform where anyone could create and articles, this lead to many factual errors and biased articles. Wikipedia started to incorporate elements hierarchy gradually over time. In the English version of Wikipedia all editors need to have a registered account and pages that are controversial and of a sensitive nature are protected by administrators.

Administrators are editors who are given access to tools such as blocking and unblocking other users, deleting and undeleting pages, protecting and renaming pages etc. Any user can **Request for Adminship**(RfA) in which the Wikipedia community participates. The RfA spans over seven days, during which any editor can comment and discuss the candidate. Editors scrutinize the candidate's contributions and credentials as well their conduct in the online discussion and overall experience. They can then state either their support or opposition to the candidate along with comments. At the end of seven days a Bureaucrat (an editor higher up in the hierarchy) decides on the consensus of the election and declare the outcome. Consensus is not a direct majority voting scheme and the final call rests with the Bureaucrat.

The RfA is a very intense and selective process, there are only 1400 total administrators of which only 500 are currently active[1]. This is out of 38 million registered editors with only around 130 thousand are regular contributors. This small group of active administrators and editors are responsible for creating and maintaining all articles on Wikipedia.

Therefore the RfA process can give us valuable insight into the dynamics of social interactions and elections in an online platform. In this paper we will first discuss the existing work on studying the RfA elections and other such similar online processes. Next we provide an overview of the data collected and used from Wikipedia in this paper. We then present our main contribution, the use of a *Viscous Democracy* to model the RfA election process. We discuss the results and possible extensions of this framework to other online elections systems.

## 2 Literature review

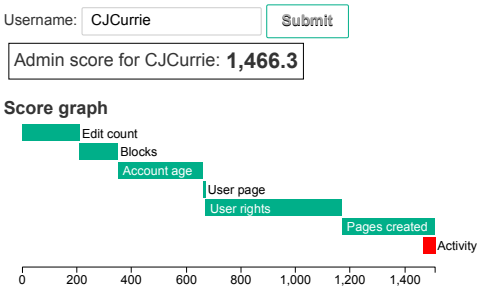
The Wikipedia RfA process has been widely studied in various domains from many different perspectives such as those of the candidate, the voters, the community etc. In this section we discuss the existing work in this field.

Administrator is a highly coveted status on Wikipedia and there are many features than can be used to determine the worthiness of a candidate. Wikipedia themselves provide tools and guides[2] to help potential candidate assess their own electability. Wikipedia's *admin score tool* as seen in Figure 1 uses features such as edit counts, pages created, age of account etc. Similarly, Burke et al. [1] utilized past RfAs to find features that correlate highly with success such as presence of edit summaries, politeness in user interactions and varied experience. Such tools and models are useful for finding potential nominees and understanding what the community values and respects. This however doesn't offer

[1] all data as of March 2020 for English version Wikipedia

[2] http://en.wikipedia.org/wiki/Wikipedia:GRFA

any insights into the dynamics that might play out in any particular election




**Score components**

| Component | Data | Score |
|---|---|---|
| Edit count | 62,616 | +210.95 |
| Blocks | never blocked | +140.00 |
| Account age | 5,808.3 days (15.91 years) | +310.04 |
| User page | exists | +10.00 |
| User rights | sysop | +500.00 |
| Pages created | 5,697 article-space pages created | +340.36 |
| Activity | 14.4 edits per month, on average (over the last year) | -45.00 |
| CSD log | missing | +0.00 |

Figure 1: Admin score tool for user CJCurrie and it's breakdown

Leskovec et al. provide a thorough analysis of the election from the perspective of the voter. They show that the voters make decisions based on *relative assessment* of merit and degree of correspondence with the candidate. Elections do not follow a *herd mentality* and standard information cascades. We see an interesting result that voters have diverse personal response functions as well as admin and non-admin patterns of voting differ. [3] We get a detailed picture of the temporal dynamics in a RfA.

As the votes in an RfA election can be positive or negative they can form a *signed network* which has been studied and analyzed in great detail. We see that the Wikipedia RfA network has more compliance with status theory compared to balance theory in Leskovec et al. [5]. When Leskovec et al. try and use these signed structural properties to predict edges in [4], they see that the predictive accuracy is poor for Wikipedia RfA network compared to the other networks used. However as signed edge prediction methods are designed to work with any generic signed network, they tend to discard information that RfAs are elections and play out in a timely manner. Also predicting a single edge i.e a vote does not increase the accuracy in predicting the result of an election.

The work of Desai et al. [2] is related closely with the contributions presented in this paper. They use linear models for regression and classification to identify a core of *influential voters* through feature selection. Using a set of 40 most influential voters they are able to predict the result of an election with a high accuracy. They also collect additional network features of the voters independent from the elections. Their results do not improve significantly in using the additional features in predicting election results. These results show that there are a group of influential voters that determine elections. This will be more evident when we analyze the dataset in coming sections.

# 3 Dataset

- explain RfA data collection
  - existing SNAP data and limitations
  - XML parsing
  - regex and string matching
  - date parsing
- Social interactions
  - User contributions
  - wealth and diversity of info
  - creating underlying network

We would like to have two different types of data to help build the election model in this paper. The first would be information of the votes cast in a RfA and the eventual result of the RfA. This gives us the users interactions in an online election process where they have to judge their peers. The second, is information on the interactions of users in other non-elections settings. In Wikipedia discussions occur in *Talk Pages*. Every type of Wikipedia page (articles, user pages, help pages etc) has a *Talk Page* where users can discuss the contents of that article or interact with the user or provide information to others. These are valuable data sources to gather more details on user activities.

In this section we will discuss the existing Wikipeida datasets from Stanford large network datasets (SNAP) [6] that satisfy our requirements and their inherent limitations. Next we will illustrate the process by which we collected newer data from Wikipedia. Lastly, we analyze the data to understand general trends and patterns.

### 3.1 Existing Datasets

For the first type of data we requore thre are two existing Wikipedia RfA datasets in SNAP namely *wiki-Elect* a nd *wiki-Rfa* . They both contain attributes of each vote in a RfA such as the source, target, vote , result of RfA, timestamp. The *wiki-Rfa* is a more recent version of the *wiki-Elect* dataset. It has RfAs till May 2013 and also has the comment text of voters. There are 11,000 users and around 190 thousand votes in total. Both of these datasets have been used in many previous works mostly as signed networks. There are a few limitations of these datasets when we would like to analyze them as vote cast in an election. There is more than 5% of *wiki-Rfa* votes that have no timestamp and almost 1% of votes that have no source. As most RfAs have fewer than 300 votes this is an issue when considering the sequence of votes as well as who has cast a vote.

The interactions between users outside of RfA elections is useful to understand behaviour and perceptions of others. Wikipedia users can directly interact with another user by writing on their *User Talk Page*. This can be a measure of how much correspondence exists between two users. We saw how this is a good indication of probability of supporting a candidate for an election [3]. The *wiki-Talk* dataset on SNAP contains a directed network where an edge from node *u* to *v* signifies that *u* has written in *v*'s talk page. This dataset is a large network containing more than 2 million nodes and 5 million edges. The limitation of this network data is that nodes do not have user id mapping and also the edges are not weighted. Without a node to user id mapping the network canot be used with the election data. Having weighted edges tells us how many times a user has interacted with someone else which is more informative.

Due to the limitations of the existing datasets we set out to collect our own data to build our election model.

### 3.2 Data Collections

## 4 Viscous Democracy

Brief explanation of viscous democracy

## 5 Proposed Model

Use viscous democracy models using heuristic delegation functions on social network to predict elections separately

## 6 Implementation

directed graph concepts and delegation function considerations. Agony and hierarchy. local and global top k delegates.

## 7 Results

The quality of predictions using local or global important editors.

## 8 Conclusions

How we can instead try and model individual voter behaviour. Find a more robust ML framework to learn an optimal delegation function.

## References


[1] Moira Burke and Robert Kraut. Mopping up: Modeling wikipedia promotion decisions. In *Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work*, CSCW '08, page 27–36, New York, NY, USA, 2008. Association for Computing Machinery.

[2] Nikhil Desai, Raymond Liu, and Catherine Mullings. Result prediction of wikipedia administrator elections based on network features. 2014.

[3] Jure Leskovec, Daniel Huttenlocher, and Jon Kleinberg. Governance in social media: A case study of the wikipedia promotion process. In *Fourth International AAAI Conference on Weblogs and Social Media*, 2010.

[4] Jure Leskovec, Daniel Huttenlocher, and Jon Kleinberg. Predicting positive and negative links in online social networks. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, page 641–650, New York, NY, USA, 2010. Association for Computing Machinery.

[5] Jure Leskovec, Daniel Huttenlocher, and Jon Kleinberg. Signed networks in social media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, page 1361–1370, New York, NY, USA, 2010. Association for Computing Machinery.

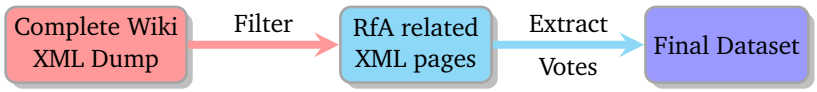



Figure 2: RfA Data Collection Process

[6] Jure Leskovec and Andrej Krevl. SNAP Datasets: Stanford large network dataset collection. `http://snap.stanford.edu/data`, June 2014.