Aalto University
School of Science
Master's Programme in Computer, Communication and Information Sciences

Ananth Mahadevan

Inferring Voting Networks in Online Elections

Master's Thesis
Espoo, March 14, 2020

DRAFT! — April 14, 2020 — DRAFT!

Supervisor: Professor Aristides Gionis
Advisor: Blank M.Sc. (Tech.)

Aalto University
School of Science

Aalto University
School of Science
Master's Programme in Computer, Communication and Information Sciences

ABSTRACT OF MASTER'S THESIS

Author:	Ananth Mahadevan		
Title:	Inferring Voting Networks in Online Elections		
Date:	March 14, 2020	**Pages:**	23
Major:	Computer Science	**Code:**	SCI3042
Supervisor:	Professor Aristides Gionis		
Advisor:	Blank M.Sc. (Tech.)		
abstract			
Keywords:	signed networks, balance, status, elections, Wikipedia, voting, graphs		
Language:	English		

Aalto-yliopisto
Perustieteiden korkeakoulu
Tieto-, tietoliikenne- ja informaatiotekniikan maisteriohjelma

DIPLOMITYÖN TIIVISTELMÄ

Tekijä:	Ananth Mahadevan		
Työn nimi: Äänestysverkkojen päätelmät online-vaaleissa			
Päiväys:	20. maaliskuuta 2020	**Sivumäärä:**	23
Pääaine:	Tietotekniikka	**Koodi:**	SCI3042
Valvoja:	Professori Aristides Gionis		
Ohjaaja:	Diplomi-insinööri Blank		
Finnish Abstract			
Asiasanat:	Finnish Keywords		
Kieli:	Englanti		

Aalto-universitetet
Högskolan för teknikvetenskaper

Aalto-universitetet
Högskolan för teknikvetenskaper
Magisterprogrammet i data-, kommunikations- och informationsteknik

SAMMANDRAG AV DIPLOMARBETET

Utfört av:	Ananth Mahadevan		
Arbetets namn: Avsluta omröstningsnätverk i onlineval			
Datum:	Den 20 mars 2020	**Sidantal:**	23
Huvudämne:	Datateknik	**Kod:**	SCI3042
Övervakare:	Professor Aristides Gionis		
Handledare:	Diplomingenjör Blank		
Swedish abstract			
Nyckelord:	Swedish Keywords		
Språk:	Engelska		

Acknowledgements

Espoo, March 14, 2020

Ananth Mahadevan

Contents

Chapter 1

Introduction

In recent years, researchers have become increasingly interested in understanding the behaviour of voters in social networks. Knowledge of the factors that motivate voters is of great importance in selecting successful policies or candidates. This is a classic problem and has been studied in the fields of game theory and political science [8, 13, 14]. More recently, there is a focus on using information from the network of voters to model their behaviour. This provides an insight into the interactions and effect of influence on voters in a community. For example, voting for bills in the United States Congress [7] or electing administrators in Wikipedia [2, 6, 9].

Votes can be represented as a *signed* network with positive or negative links. Finding groups using clustering techniques [1, 3, 12] and predicting signed links [4, 10, 11] in these networks is well researched. These approaches provide an ability to understand the group dynamics at play and predict votes and in such a network. However, they do not consider the iterative and chronological nature of the voting that takes place in these networks. In cases where research does focus on voter models, they rely on external features to build machine learning models that are task-specific and static [6, 7].

In this thesis, we propose a model that creates a local signed network consisting of the neighbours of the current voter and the preceding votes. It will then predict the vote, which when added to the network will comply the most with concepts of balance and status in signed networks. After all the votes are cast in a session, the model can be easily updated to improve quality and is, therefore, iterative and dynamic. The model is also flexible and can incorporate external features to build the local signed network of a voter. The results for Wikipedia administrator elections shows that our model outperforms machine learning based models and traditional signed link prediction solutions.

1.1 Thesis Outline

The rest of the thesis is organized in the following manner. We discuss the background relating to signed graphs, hierarchy in directed networks in Section 2. In Section 3 we describe the vote prediction problem and approaches to solving it. Section 4 provides a comprehensive view of Wikipedia and the election process for administrators. In Section 5 we explain the datasets used, construction of the model and evaluation criteria. We report our findings in Section 6 and discuss their implications. Finally, we conclude the thesis and present future work in Section 7.

Chapter 2

Graph Theory

In this chapter we will provide the fundamentals of the graph theory concepts required to understand the rest of the thesis. In Section 2.1 we cover the basic definitions, terminologies used to describe different types of graphs. Then we define a signed graph and discuss its unique properties in Section 2.2. We outline the theory of balance in signed networks and methods to measure it in Section 2.3.Next, we discuss the theory of status and illustrate the differences to balance in Section 2.4. Lastly, we explain techniques of finding hierarchies in directed networks and the concept of agony in Section 2.5.

2.1 Preliminaries

In this section we define the various types of graphs and their basic properties. The notation and terminologies used closely follows those used in Diestel [5]. Graphs are structures that describe relationships between entities. These entities are called *vertices* and entities related to one another are joined by edges. The terms graph, vertices and edges can be used interchangeably with *network*, *nodes* and *links* respectively.

Graphs can be classified broadly into two types based on whether the edges posses a direction or not. We now go on to define them in detail.

2.1.1 Undirected Graphs

An undirected graph is pair $G = (V, E)$, where V is the set of vertices and E is the set $E \subseteq \{(u, v) \mid u, v \in V\}$ of unordered pairs of vertices called edges. In this thesis we will deal with only *simple graphs*, i.e. no are so self loops $(u, v) \in V \times V,\ u \neq v$ and there is at most one edge between vertices u and v.

The number of the vertices in a graph is called the *order* of the graphs is denoted by $n = |G|$ and the *size* of a graph is the number of edges denoted by $m = \|G\|$ or $m = |E|$. A vertex u is *adjacent* to v is they are the end points of an edge, $(u, v) \in E$. All the vertices adjacent to vertex v is called the *neighbourhood* of v and is denoted by $N(v)$. The *degree* of a vertex v is the number of nodes adjacent to that vertex and is denoted by $d(v) = |N(v)|$.

The edges of an undirected graph can also have an associated value. This value can indicate the distance or similarity between a pair of vertices. These values are called *weights* and the corresponding graph is called a *weighted undirected graph*. Therefore, a weighted graph is defined as a triple $G = (V, E, w)$, where $w : E \to \mathbb{R}^+$ is a function that maps an edge e to a positive real weight $w(e)$. Now an *unweighed graph* is simply a weighted graph where the function w is defined as: if $e \in E$ then $w(e) = 1$ else $w(e) = 0$. The degree of a a vertex v in a weighed graph is the sum of the weights to all the neighbours of v and is defined as $d(v) = \sum_{u \in N(v)} w((u, v))$ An example of a weighted undirected graph is shown in Figure 2.1.

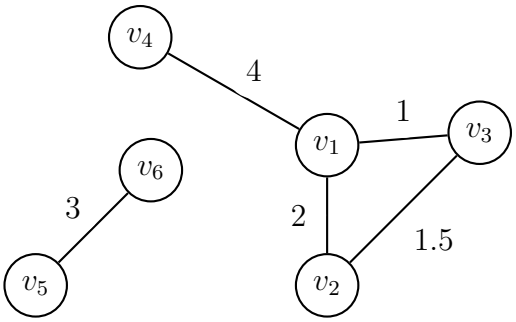

Figure 2.1: An example of a weighted undirected graph

2.1.2 Directed Graphs

The main distinction regarding a *directed graph* (or *digraph*) is that the edges are ordered pairs, i.e.$(u, v) \neq (v, u)$. Therefore, a directed graph has a similar definition: a pair $G = (V, E)$, where V is the set of vertices and E is the set of *ordered* pairs of vertices. Now given an edge $e = (u, v)$ we can define a source function $\text{src} : E \to V$ such that $\text{src}(e) = u$ and a destination function $d : E \to V$ where $\text{dest}(e) = v$. These functions classify the vertices in an edge e as either the source or the destination. In this thesis, we deal only with *simple directed graphs*, i.e. no self-loops and there can be at most one edge from u to v.

As the edges now have an inherent direction we can define the *successors* and *predecessors* of a node v. A vertex u is called the *successor* of a node v is there exists a directed edge from v to u, therefore the set of successors for a vertex v can be defined as $S(v) = \{u \mid (v, u) \in E\}$. A *predecessor* of a node v is a vertex u such that there exists a directed edge from u to v, the set of predecessors for a vertex v can de defined as $P(v) = \{u \mid (u, v) \in E\}$. We now a vertex u that is either a successor or a predecessor of a vertex v can be called a neighbour of the vertex v. Therefore, we define the *neighbourhood* of a vertex v as the set of vertices in the union of successors and predecessor, i.e $N(v) = S(u) \cup P(v)$. This definition is also compatible with undirected graphs because if $(u, v) \in E$ then $(v, u) \in E$.

Directed graphs can also have values associated with each directed edge called a *weight*. A *weighted directed graph* can be defined as a triple $G = (V, E, w)$, where the weight function $w : E \to \mathbb{R}^+$ that maps each edge e to a weight $w(e)$. The indegree of a vertex v is defined as the sum of the weights from the predecessors of v and is denoted as $\mathrm{d_{in}}(v) = \sum_{u \in P(v)} w((u, v))$. Similarly, the outdegree of a vertex v is defined as the sum of the weights to the successors of Figure 2.2 shows an example of a weighted directed graph.

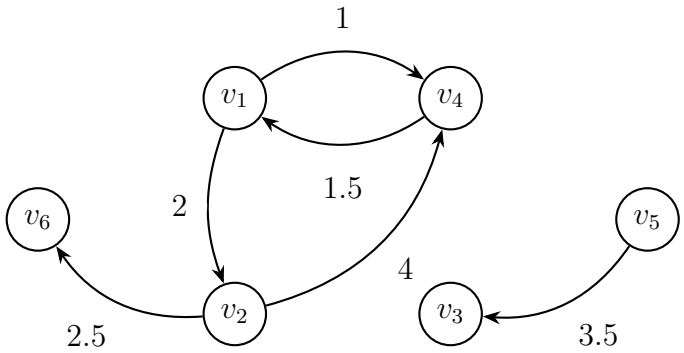

Figure 2.2: An example of a weighted directed graph

2.2 Signed Graphs

- Signed graphs and restrictions
- Explain relevance in real world settings

2.3 Balance Theory

- Explain balance theory origin and significance.
- Illustrate with triads and examples
- Define mathematical background to measure balance through the Eigen decomposition of the graph Laplacian

2.4 Status Theory

- Describe the nature of the directed setting
- Illustrate the differences to Balance theory
- Mention existing ways to measure violation to status in a network

2.5 Hierarchy in directed networks

- Discuss the hierarchy in DAGs
- Explain concept of Agony
- Provide existing algorithms to find the ranking of nodes.

Chapter 3

Vote Prediction

In this section we first provide the motivation of choosing independent vote prediction as our target and the differences from predicting the result of an election.Next we describe the available techniques and methods to predict individual votes or signed edges in a network and how it relates to the problem at hand. We then provide two novel methods of using user information long with past election results to predict votes.

3.1 Election versus Vote Prediction

- Discuss existing election result predictions schemes
- Discuss the limitations in understanding election dynamics through just predicting election results
- Describe the process as an information cascade, discuss the potential Game Theory settings
- Show the two parts of the problem from an information cascading perspective
  - Who is going to vote next
  - How they are going to vote
- Discuss the assumptions in usual Independent Cascade (IC) models
- Explain the difficulty of both aspects in the domain of an election
- Motivate the selection of the problem as an **Independent Vote Prediction**

3.2 Signed Edge Prediction

- Discuss the existing edge predictions work
- Directly using signed triads as features
- Using triads along with network features
- Using user information and interaction data for predicting votes and/or elections
- The main drawbacks in these methods when considering an election setting

3.3 Linear Combination of Graphs

- Describe the linear combination of graphs derived from user and election data
- Explain topic similarity, follows network, interaction networks and other features
- How it can also incorporate signed features as additional features in prediction

3.4 Local Signed Network

- Explain the concept of the local signed network for a particular user
- Motivate the definition with respect to elections and influence
- Describe how to use balance and status theory to predict the vote
- Clarify the differences to signed edge prediction efforts
- Mention Agony as a way to measure status compliance here?

Chapter 4

Wikipedia

In this section we provide an overview of how Wikipedia is structured, the hierarchy that exists withing editors. We then explain the election process of getting administrator rights in Wikipedia.

4.1 Structure and hierarchy in Wikipedia

4.2 Elections in Wikipedia

- Explain Editors and Administrators in Wikipedia
- Describe the Request for Administrator(RfA) process
- Discuss general trends and patters
- Mention research interest and possible current works?

Chapter 5

Experiments

In this section we first describe the datasets that will be used in building our vote prediction models. Then we discuss the various linear and graphical models that we consider and their implementations details. Lastly we define the metrics and other means of evaluating the models and the results.

5.1 Datasets

- Maybe a short description of existing SNAP datasets and their limitations
- The details of the *Wiki-RfA* data and the *User-Contribution* datasets

5.2 Graphs

- Discuss the process of extraction of the various graphs discussed in the previous sections
- **Agree Graphs and Follows Graph**, where we measure the degree to which one user agrees and follows another user in previous elections
- **Topic similarity** from the top 100 articles edited for each user and the pairwise Jaccard similarity
- **Talk and Interaction graphs**, measures communication between users on their respective user talk pages
- **Signed Graphs**, triad encoding and extracting the triad counts for each voter

5.3 Models

5.3.1 Linear Combination of Graphs

- Discuss the various linear models considered for Graph Combinations
  - Linear Regression
  - Support Vector Classifier
  - Extreme Gradient Boosting (XGBOOOST)
- Discuss how each graph contributes features and the problem is a linear classification problem

5.3.1.1 Iterative Mode

- Discuss the motivation behind an iterative model versus a static prediction model
- Describe how balance is derived from the Agree Graph in a local signed network
- Discuss how the Agree graph is updated in terms of Balance
- Describe how status is derived from the Follows graph in a local signed network
- Discuss how the Follows graph is updated after every election
- Describe how to make the predictions
  - Deterministic : just decide based on eigen value or agony as support or oppose
  - Probabilistic : provide a probability for predicting a support vote

5.4 Evaluation

- Discuss the issues with the imbalance in the datasets
- Illustrate the issues with pure measures of accuracy
- Define Precision, Recall and Macro F1 score
- Discuss ROC AUC and Precision Recall curves for probability based predictions

Chapter 6

Results and Discussion

In this section we will present the results of the models and discuss their implications.

6.1 Linear Combination of Graphs

- Present results for each linear classifier
- Discuss the different splits of the dataset to check for robustness and chronological consistency
- Show the feature importances and discuss their relevance
- Compare the raw accuracy versus the macro f1 scores
- Highlight the difficulty of predicting negative votes

6.2 Local Signed Network

- Present the Iterative Balance model results
- Discuss quality of predictions using evaluation metrics
- Mention the difference between deterministic and probabilistic prediction accuracies
- Explain the Iterative Status model results
- Discuss the issues with local model of status and the potential reasons for lower score and quality

6.3 Comparison

- Compare results from signed edge prediction and Iterative signed models
- Discuss Static Linear combination predictions versus Iterative signed predictions
- Discuss the assumptions used in the models and limitations

Chapter 7

Conclusions and Future Work

- Explain the quality of results with the election perspective
- Future work is to extend this to other election settings and investigate generality of this approach
- Possible future work in congressional voting data
- Can also tackle the other problem in information cascade theory of how to predict who is most likely to vote next
- This can lead to a complete model of election dynamics and could incorporate elements of game theory and network inference

Bibliography

[1] Brito, A. C. M., Silva, F. N., and Amancio, D. R. A complex network approach to political analysis: Application to the brazilian chamber of deputies. *PLOS ONE 15*, 3 (2020).

[2] Cabunducan, G., Castillo, R., and Lee, J. B. Voting behavior analysis in the election of wikipedia admins. In *2011 International Conference on Advances in Social Networks Analysis and Mining* (2011), pp. 545–547.

[3] Chiang, K.-Y., Hsieh, C.-J., Natarajan, N., Dhillon, I. S., and Tewari, A. Prediction and clustering in signed networks: a local to global perspective. *Journal of Machine Learning Research 15*, 1 (2014), 1177–1213.

[4] Chiang, K.-Y., Natarajan, N., Tewari, A., and Dhillon, I. S. Exploiting longer cycles for link prediction in signed networks. In *Proceedings of the 20th ACM international conference on Information and knowledge management* (2011), pp. 1157–1162.

[5] Diestel, R. *Graph Theory*. 1997.

[6] Jankowski-Lorek, M., Ostrowski, L., Turek, P., and Wierzbicki, A. Modeling wikipedia admin elections using multidimensional behavioral social networks. *Social Network Analysis and Mining 3*, 4 (2013), 787–801.

[7] Karimi, H., Derr, T., Brookhouse, A., and Tang, J. Multifactor congressional vote prediction. In *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (2019), pp. 266–273.

[8] Kearns, M. J., Judd, J. S., Tan, J., and Wortman, J. Behavioral experiments on biased voting in networks. *Proceedings of the National*

Academy of Sciences of the United States of America 106, 5 (2009), 1347–1352.

[9] Lee, J. B., Cabunducan, G., Cabarle, F. G., Castillo, R., and Malinao, J. A. Uncovering the social dynamics of online elections. *Journal of Universal Computer Science 18* (2012), 487–505.

[10] Leskovec, J., Huttenlocher, D., and Kleinberg, J. Predicting positive and negative links in online social networks. In *Proceedings of the 19th international conference on World wide web* (2010), pp. 641–650.

[11] Leskovec, J., Huttenlocher, D., and Kleinberg, J. Signed networks in social media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2010), pp. 1361–1370.

[12] Levorato, M., and Frota, Y. Brazilian congress structural balance analysis. *Journal of Interdisciplinary Methodologies and Issues in Sciences* (2016).

[13] Tal, M., Meir, R., and Gal, Y. K. A study of human behavior in online voting. *adaptive agents and multi-agents systems* (2015), 665–673.

[14] Zou, J., Meir, R., and Parkes, D. Strategic voting behavior in doodle polls. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing* (2015), pp. 464–472.