

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 28, 2020 — DRAFT!**

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

Aalto University  
School of Science

Master's Programme in Computer, Communication and  
Information Sciences

ABSTRACT OF  
MASTER'S THESIS

<b>Author:</b>	Ananth Mahadevan		
<b>Title:</b>	Inferring Voting Networks in Online Elections		
<b>Date:</b>	March 14, 2020	<b>Pages:</b>	37
<b>Major:</b>	Computer Science	<b>Code:</b>	SCI3042
<b>Supervisor:</b>	Professor Aristides Gionis		
<b>Advisor:</b>	Blank M.Sc. (Tech.)		
abstract			
<b>Keywords:</b>	signed networks, balance, status, elections, Wikipedia, voting, graphs		
<b>Language:</b>	English		

Aalto-yliopisto

Perustieteiden korkeakoulu

Tieto-, tietoliikenne- ja informaatiotekniikan maisteriohjelma

DIPLOMITYÖN

TIIVISTELMÄ

<b>Tekijä:</b>	Ananth Mahadevan		
<b>Työn nimi:</b>	Äänestysverkkojen päätelmät online-vaaleissa		
<b>Päiväys:</b>	20. maaliskuuta 2020	<b>Sivumäärä:</b>	37
<b>Pääaine:</b>	Tietotekniikka	<b>Koodi:</b>	SCI3042
<b>Valvoja:</b>	Professori Aristides Gionis		
<b>Ohjaaja:</b>	Diplomi-insinööri Blank		
Finnish Abstract			
<b>Asiasanat:</b>	Finnish Keywords		
<b>Kieli:</b>	Englanti		

Aalto-universitetet

Högskolan för teknikvetenskaper

Magisterprogrammet i data-, kommunikations- och infor- SAMMANDRAG AV  
mationsteknik DIPLOMARBETET

<b>Utfört av:</b>	Ananth Mahadevan		
<b>Arbetets namn:</b>	Avsluta omröstningsnätverk i onlineval		
<b>Datum:</b>	Den 20 mars 2020	<b>Sidantal:</b>	37
<b>Huvudämne:</b>	Datateknik	<b>Kod:</b>	SCI3042
<b>Övervakare:</b>	Professor Aristides Gionis		
<b>Handledare:</b>	Diplomingenjör Blank		
Swedish abstract			
<b>Nyckelord:</b>	Swedish Keywords		
<b>Språk:</b>	Engelska		

# Acknowledgements

Espoo, March 14, 2020

Ananth Mahadevan

# Contents

<b>1</b>	<b>Introduction</b>	<b>8</b>
1.1	Thesis Outline . . . . .	9
<b>2</b>	<b>Graph Theory</b>	<b>10</b>
2.1	Preliminaries . . . . .	10
2.1.1	Undirected Graphs . . . . .	10
2.1.2	Directed Graphs . . . . .	11
2.2	Signed Graphs . . . . .	12
2.2.1	Balance Theory . . . . .	14
2.2.2	Status Theory . . . . .	15
2.3	Hierarchy in directed networks . . . . .	17
2.4	Link and Sign Prediction . . . . .	19
<b>3</b>	<b>Vote Prediction</b>	<b>21</b>
3.1	Result versus Vote Prediction . . . . .	21
3.2	Voting and Signed networks . . . . .	22
3.3	Linear Combination of Graphs . . . . .	23
3.4	Local Signed Network . . . . .	25
<b>4</b>	<b>Wikipedia</b>	<b>27</b>
4.1	Structure and hierarchy in Wikipedia . . . . .	27
4.2	Elections in Wikipedia . . . . .	27
<b>5</b>	<b>Experiments</b>	<b>28</b>
5.1	Datasets . . . . .	28
5.2	Graphs . . . . .	28
5.3	Models . . . . .	29
5.3.1	Linear Combination of Graphs . . . . .	29
5.3.2	Iterative Mode . . . . .	29
5.4	Evaluation . . . . .	29

<b>6</b>	<b>Results and Discussion</b>	<b>30</b>
6.1	Linear Combination of Graphs . . . . .	30
6.2	Local Signed Network . . . . .	30
6.3	Comparison . . . . .	31
<b>7</b>	<b>Conclusions and Future Work</b>	<b>32</b>

# 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, for example, voting for bills in the United States Congress [27] or electing administrators in Wikipedia [9, 26, 32], is of great importance in selecting successful policies or candidates. Voting is a classic problem and has been studied in the fields of game theory and political science [29, 41, 49]. More recently, there is a focus on using information from the interaction network of voters to model their behaviour. This provides an insight into these interactions and effect of influence of other individuals on voters within a community.

Votes can be represented as a *signed* network with positive or negative links. Finding groups using clustering techniques [6, 12, 35] and predicting signed links [13, 33, 34] in these networks is well researched. These approaches provide understanding of 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. Moreover, 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 [26, 27].

In this thesis, we propose a model that creates a local signed network consisting of the neighbours of the current voter and the votes already cast in the election in question. It will then predict the vote that 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. Firstly, we discuss the background relating to signed graphs and hierarchy in directed networks in Chapter 2. Next, in Chapter 3 we describe the vote prediction problem and approaches to solving it. Chapter 4 provides a comprehensive view of Wikipedia and the election process for administrators. In Chapter 5 we explain the datasets used, construction of the model and evaluation criteria. After that, we report our findings in Chapter 6 and discuss their implications. Finally, we conclude the thesis and present future work in Chapter 7.

## Chapter 2

# Graph Theory

Change Chapter to Background and include signed link prediction theory here as well

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 and 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 social theories of balance and status in signed networks in Sections 2.2.1 and 2.2.2. Lastly, we explain techniques of finding hierarchies in directed networks and the concept of agony in Section 2.3.

Add  
line  
for  
link  
pre-  
dic-  
tion

### 2.1 Preliminaries

In this section, we define the various types of graphs and their basic properties. The notation and terminologies used closely follow those used in Diestel [16]. 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 possess a direction. 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 only with *simple graphs*, i.e. no 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 graph and is denoted by  $n = |G|$ . In turn, the *size* of a graph is the number of edges denoted by  $m = \|G\|$  or  $m = |E|$ . A vertex  $u$  is *adjacent* to  $v$  if they are the end points of an edge,  $(u, v) \in E$ . All the vertices adjacent to a 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  $w : E \rightarrow \mathbb{R}^+$  is a function that maps an edge  $e$  to a positive real weight  $w(e)$ . Now an *unweighted 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 vertex  $v$  in a weighted 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  $\text{src} : E \rightarrow V$  such that  $\text{src}(e) = u$  and a destination function  $\text{dst} : E \rightarrow V$  where  $\text{dst}(e) = v$ . 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  $v$  if 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 be defined as  $P(v) = \{u \mid (u, v) \in E\}$ . 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.  $N(v) = S(v) \cup P(v)$ . 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  $w : E \rightarrow \mathbb{R}^+$  maps each edge  $e$  to a weight  $w(e)$ . The indegree of a vertex  $v$  is defined as the sum of the edge weights from the predecessors of  $v$  and is denoted as  $d_{\text{in}}(v) = \sum_{u \in P(v)} w((u, v))$ . Similarly, the outdegree of a vertex  $v$  is defined as the sum of the edge weights to the successors of  $v$  and is denoted by  $d_{\text{out}}(v) = \sum_{u \in S(v)} w((v, u))$ . 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

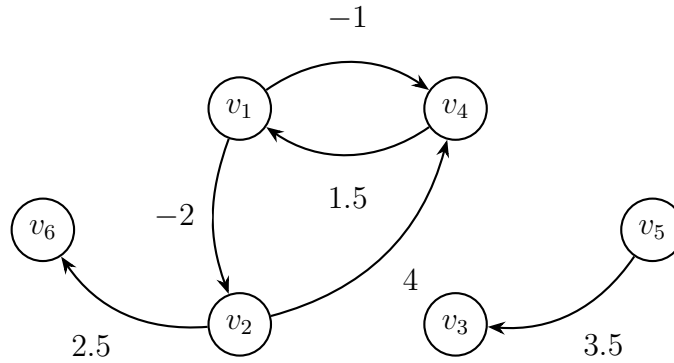
The simple weighted graphs we have defined so far only have non-negative edge weights that can represent similarity or closeness. In the 1950s, social psychologists found it desirable to express liking, disliking or indifference in

social interactions. This was formalized by Harary [22] using graphs with weights  $(-1, 0, 1)$ . These graphs are therefore called *signed graphs*, where a negative edge weight can denote dissimilarity between a pair of vertices. In this thesis we will use notations and terms similar to Gallier [17], Kunegis et al. [31], Hou [24] and Zaslavsky [48].

A signed graph is a triple  $G = (V, E, w)$ , where  $V$  is the set of vertices,  $E$  is the set of pairs of vertices and the weight function  $w : E \rightarrow \mathbb{R}$ . The weight function now takes an edge  $e$  and maps it to a signed weight  $w(e)$ . We can partition the edges into positive and negative edges,  $E = E^+ \cup E^-$ , where  $E^+ = \{e \mid w(e) > 0\}$  and  $E^- = \{e \mid w(e) < 0\}$ . Similar to Zaslavsky [48], we consider undirected signed graphs and directed signed graphs as distinct and separate entities. We can see some examples of signed graphs in Figure 2.3.



(a) A undirected signed graph



(b) A directed signed graph

Figure 2.3: Examples of Signed Graphs

We can now proceed to define a few more terms. The degree of a vertex  $v$  is now the sum of the absolute edge weight of its neighbours, called the

*signed degree* and is defined as

$$\bar{d}(v) = \sum_{u \in N(v)} |w((u, v))|$$

This can also be extended to *signed indegree* and *signed outdegree* denoted by  $\bar{d}_{\text{in}}(v)$  and  $\bar{d}_{\text{out}}(v)$  and defined as

$$\bar{d}_{\text{in}}(v) = \sum_{u \in P(v)} |w((u, v))|$$

$$\bar{d}_{\text{out}}(v) = \sum_{u \in S(v)} |w((v, u))|$$

We can create a  $n \times n$  square weight matrix  $W$ , where each entry  $w_{ij}$  is defined as

$$w_{ij} = \begin{cases} w((v_i, v_j)) & \text{if } (v_i, v_j) \in E \\ 0 & \text{if } (v_i, v_j) \notin E \end{cases}$$

The signed degree matrix  $\bar{D}$  is a diagonal matrix consisting of the signed vertex degrees,  $\bar{D} = \text{diag}(\bar{d}(v_1), \dots, \bar{d}(v_n))$ . We can now define the *signed Laplacian*,  $\bar{L}$  as

$$\bar{L} = \bar{D} - W$$

The signed Laplacian along with results from spectral analysis of signed graphs [24, 31], will be useful for balance theory of signed graphs.

### 2.2.1 Balance Theory

In the 1940s, Heider [23] proposed that when there are either *positive relations* (friendship, love, support) and *negative relations* (enmity, hate, oppose) in a group, the group tends towards *balance*. Balance was defined as there being all positive relations or one positive and two negative relations for a group of three people. Harary and Cartwright [10] generalized this notion of balance by using signed graphs. As these relations are typically symmetric, balance is usually defined for undirected signed graphs. This can be seen in Figure 2.4 where solid edges are positive and dashed edges are negative. Social psychology also showed that balanced triads  $B_1$  and  $B_2$  mirror aphorisms such as "the friend of my friend is also a friend" and "the enemy of my enemy is a friend" respectively. Davies [14] also offers an alternate definition of *weak balance* where we don't consider the triads of type  $B_2$  to predict relations as they are less common in real social networks.

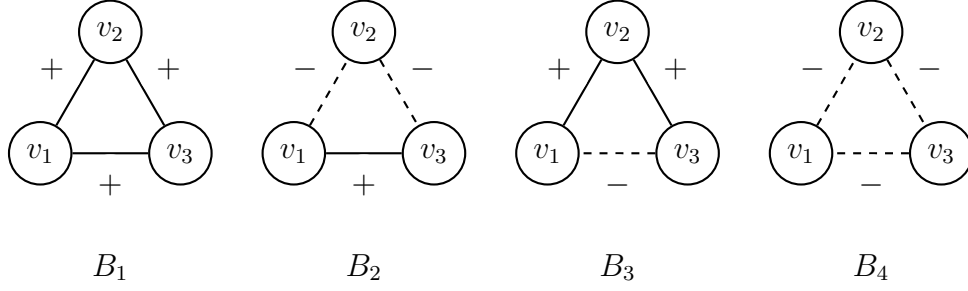


Figure 2.4: Triads in undirected signed graphs.  $B_1$  and  $B_2$  are *balanced* triads as they have even number of negative edges.  $B_3$  and  $B_4$  are *unbalanced* as they have odd number of negative edges.

The concept of balance can be generalized to state that an undirected signed graph  $G = (V, E, w)$  is balanced iff every cycle in  $G$  has an even number of negative edges. This leads to result from Harary [22] that states that if a graph  $G$  is balanced, then there is a partition of the vertices  $V = V_1 \cup V_2$  such that edges within the vertices of each set is positive and edges between the sets are negative. This means that we can have a bipartite graph when we delete the positive edges and negative edges span between the two sets of vertices. An example of a balanced signed graph is shown in Figure 2.5. Here the partition of the vertex set is  $V_1 = \{v_1, v_3, v_4, v_7, v_8\}$  and  $V_2 = \{v_2, v_5, v_6, v_9\}$ .

The signed Laplacian matrix  $\bar{L}$  of a signed graph  $G$  is always positive-semidefinite and  $\bar{L}$  is positive-definite iff  $G$  is *unbalanced* [24, 31, 48]. If the smallest eigenvalue of a graph  $G$  is denoted by  $\lambda_1(G)$ , then  $G$  is balanced iff  $\lambda_1(G) = 0$ . Hou [24] provides bounds on the value of  $\lambda_1(G)$  and Li et al. [36] show that  $\lambda_1(G)$  can be used as a measure of how far the signed graph  $G$  is from being balanced.

### 2.2.2 Status Theory

Guha et al. [20] mention implicitly that a signed edge from  $u$  to  $v$  can be interpreted in a asymmetric manner different from "friend" or "enemy". Leskovec et al. [33, 34] introduce the concept of *status* to contextualize directed signed edges. A positive edge  $u \xrightarrow{+} v$  indicates that  $v$  has a higher status than  $u$  and a negative edge  $u \xrightarrow{-} v$  means that  $v$  has a lower status than  $u$ . This concept of relative status can be propagated transitively along multiple steps which might lead to contradictions with balance theory [34].

Given three vertices  $v_1, v_2$  and  $v_3$ , the presence of an edge  $v_1 \xrightarrow{+} v_2$  in-



Figure 2.5: A balanced signed graph. Solid blue edges are positive and dashed red edges are negative. Every cycle in this graph contains an even number of negative edges.

indicates that  $v_1$  thinks  $v_2$  has higher status, the edge  $v_2 \xrightarrow{+} v_3$  indicates that  $v_2$  thinks  $v_3$  has higher status. Now we wish to close this triad with an edge from  $v_3$  to  $v_1$ . Status theory would say that through transitivity  $v_1$  has lower status than  $v_3$ , therefore the prediction is  $v_3 \xrightarrow{-} v_1$ . Whereas, in balance theory we would predict a positive edge  $v_3 \xrightarrow{+} v_1$  to make the cycle have even number of negative edges. This example can be seen in triads  $S_1$  and  $S_2$  shown in Figure 2.6.

There are also cases when status theory is ambivalent to the edge that closes a triad. Consider the example when we have the edges  $v_1 \xrightarrow{+} v_2$  and  $v_2 \xrightarrow{-} v_3$ . If we were to indicate the status of a vertex  $v$  using the function  $s(v)$ , then the edges describe the following:  $s(v_2) > s(v_1)$  and  $s(v_2) > s(v_3)$ . Therefore, we have no knowledge of the relative difference in status between the vertices  $v_1$  and  $v_3$ . Hence, both edges  $v_3 \xrightarrow{+} v_1$  ( $s(v_3) > s(v_1)$ ) and  $v_3 \xrightarrow{-} v_1$  ( $s(v_1) > s(v_3)$ ) are equally valid for status theory. Balance theory on the other hand can only predict  $v_3 \xrightarrow{-} v_1$  to keep the triad balanced. This case is shown in Figure 2.6 as triads  $S_3$  and  $S_4$ .

Each positive link inwards ( $d_{\text{in}}^+(v)$ ) and negative link outwards ( $d_{\text{out}}^-(v)$ ) increases status. Each positive link outwards ( $d_{\text{out}}^+(v)$ ) and negative link inwards ( $d_{\text{in}}^-(v)$ ) decreases status. Therefore,  $\sigma(v) = d_{\text{in}}^+(v) + d_{\text{out}}^-(v) -$



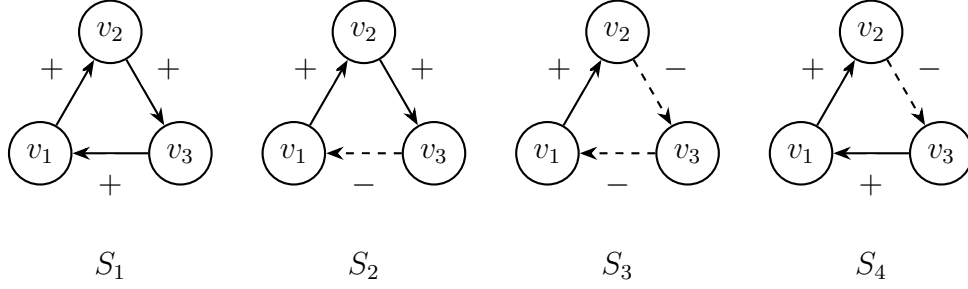


Figure 2.6: Triads in directed signed graphs. Triads  $S_2, S_3$  and  $S_4$  are compliant with status theory. Only triads  $S_1$  and  $S_3$  are compliant with balance theory.

$d_{\text{out}}^+(v) - d_{\text{in}}^-(v)$  is a heuristic for the status of a node [33]. An interesting fact is that the edge  $u \xrightarrow{-} v$  can be converted into positive edge in the opposite direction  $u \xleftarrow{+} v$ . This fact reduces the number of unique triads that can be formed using status theory and will be used in edge prediction tasks that will be discussed in coming chapters.

## 2.3 Hierarchy in directed networks

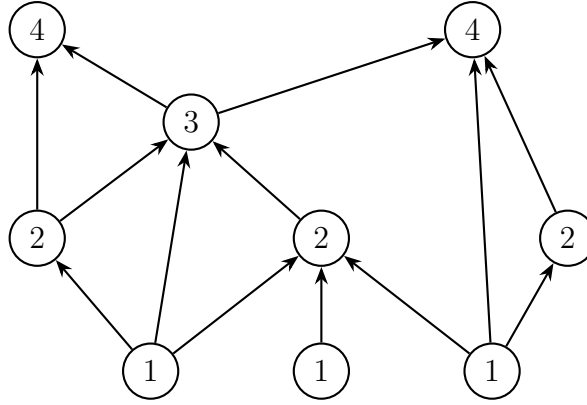
Hierarchies exist in all social structures, from the explicit levels found in nature such as the food chain or organizational structures in businesses to more implicit stratification that occurs on social media or online networks. A common method to represent such hierarchies is through a tree, for example, the chain of command in the military or within governments. Trees have well defined levels and a single person at the top. If we generalize this structural concept then we get a *Directed Acyclic Graph* (DAG) which represents a partial ordering set. DAGs have perfect hierarchy while structures such as cycles tend to have no hierarchy. Other directed graphs occur somewhere between these two extremes.

Gupte et al. [21] provide a method to discern the levels of stratification present in a given directed network when no prior information of the hierarchy exists. They define a measure of the hierarchy of a given directed network  $G$  as  $h(G)$  along with a polynomial algorithm to find the largest hierarchy in that network. They define a concept of *social agony* which posits that agony is present when a person having a higher rank in network interacts with a person who has a lower rank. Therefore, if we define the rank of a node in graph  $G$  as the function  $r : V \rightarrow \mathbb{N}$ , then a directed edge  $u \rightarrow v$  causes agony when  $r(u) \geq r(v)$ . The agony for the edge can be quantified as

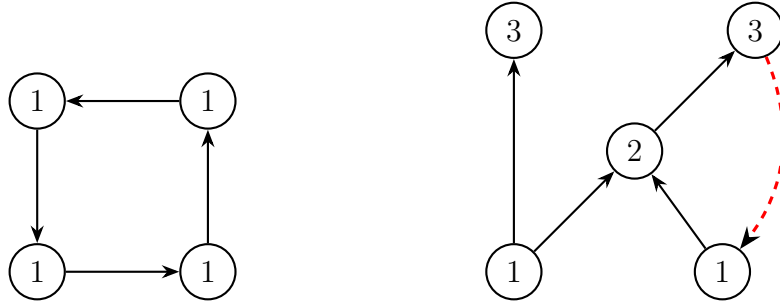
$\max(r(u) - r(v) + 1, 0)$ . The agony of the graph  $G$  wrt to rank function  $r$  is defined as

$$A(G, r) = \sum_{(u,v) \in E} \max(r(u) - r(v) + 1, 0)$$

As nodes in a graph tend to minimize the overall agony, the agony of a network  $G$  is the smallest possible agony over all possible ranking for  $r$ ,  $A(G) = \min_{r \in \text{Rankings}} A(G, r)$ . The hierarchy of a given network  $G$ , denoted by  $h(G)$  can now be expressed in terms of the agony of the network,  $h(G) = 1 - \frac{1}{m} A(G)$ , where  $m$  is the number of edges. We can see examples of hierarchy in unweighted directed graph in Figure 2.7.



(a) DAG has perfect hierarchy,  $h(G) = 1$  and agony of each edge is 0



(b) Cycle has no hierarchy,  $h(G) = 0$  and each edge has agony of 1 (c) Graph with some hierarchy  $h(G) = \frac{2}{5}$ . Red dashed edge has agony of 3 and solid black edges have 0 agony.

Figure 2.7: Examples of hierarchy in unweighted directed graphs. Numbers inside nodes indicate the rank of vertex.

Therefore, finding a ranking of the nodes that minimizes the agony of the network gives us the optimal hierarchy present in that network. Gupte

et al. [21] and Tatti [43] present a polynomial algorithm that can solve the dual of the agony minimization problem to obtain the optimal ranking  $r$  for unweighted graphs. Tatti [44] provides an alternate approach using a capacitated circulation solver that can handle weighted digraphs as well as additional cardinality constraints. These algorithms allows us to find the levels of hierarchy present in any given directed social network and analyse the interaction between members belonging to different strata in that community.

We will explain in future chapters how hierarchies in social network is intrinsically linked to the theory of status in directed signed networks. We show one can use the concept of agony of a directed signed graph  $G$  to quantify the violation to status theory and use it as a metric to predict an unknown signed edge.

## 2.4 Link and Sign Prediction

The *link prediction problem* is defined by Liben-Nowell and Kleinberg [37] as inferring possible future edges between vertices in a social network. The datasets used was split into training edges and testing edges which had a common core set of vertices. They showed that topological features such as number of common neighbours and Katz's centrality index can be used in a unsupervised setting to accurately predict edges.

Leskovec et al. [33] extended the link prediction problem to signed networks. They also introduced the problem of *edge sign prediction*: predict the sign of a given edge  $(u, v)$  using the existing signed network  $G$ . A supervised model for the task is proposed that uses network features such as indegree and outdegree of a node along with *triad* features. The edge nodes  $u$  and  $v$  and a common neighbour  $w$  form a triad. The directed edge between  $(u, w)$  and  $(w, u)$  can be either forward or backwards and each edge is either positive or negative. Therefore there are 16 possible triad types for a given neighbour  $w$ . They analyse these triadic features from the trained model and show how they relate to balance and status theory. For the link prediction problem, the information from the negative edges in the signed network offers improvement in the overall accuracy of the model. This seminal paper inspired many more approaches to solving these problems for signed graphs.

Matrix factorization and latent space approaches facilitate link prediction and sign prediction tasks for multiple edges simultaneously in a signed network [1, 19, 25]. Supervised algorithms for both tasks can be improved using additional node features such as inverse square metric [2] or node rankings [40]. Models that use graph motifs [30, 38] generalize the concept of triadic

feature for link and sign prediction. Chiang et al. generalize balance theory for longer cycles and use it as features to improve link prediction [13]. Tang et al. [42] discuss the importance of predicting negative links and highlight methods to overcome the inherent imbalance present in signed network datasets. Cesa et al. [11] and Chiang et al. [12] utilize clustering techniques to solve sign prediction and link prediction respectively. Shuang et al. [45] and Karimi et al. [27] create bespoke models incorporating user behaviour and political party affiliation respectively to predict the sign of edges present in the signed networks.

## Chapter 3

# Vote Prediction

In this chapter, we cover the main motivation behind predicting the vote of an individual voter and present the methods that can be used to solve this task. We discuss the contrast in perspectives that is present when predicting the result compared to predicting a vote in Section 3.1. Next, we explain how the problem of vote prediction is intrinsically linked to the tasks of edge and sign prediction in signed network in Section 3.2. In Section 3.3, we describe a supervised machine learning framework that can use graphs features from voting and non-voting data. Lastly, we present our novel approach of constructing a signed graph from neighbours of the current voter and previous votes and using balance and status theory to predict the vote in Section 3.4.

### 3.1 Result versus Vote Prediction

In this thesis, we are interested in the voting behaviour for a collective action. In such cases, members of a community come together as *voters* to decide on a particular *candidate* item during a *session*. In a government parliament the voters are the elected members of the parliament and the candidate is usually a bill or policy matter. When it is promotion within a political party or an online community such as Wikipedia, the members vote on a candidate who has been nominated for the position. In all these cases we have two levels of decisions being made. The first the individual decisions that a voter makes with regard to the candidate. The second is the final decision that the group arrives to after everyone has voted. We refer to task of predicting the former as *vote prediction* and the latter as *result prediction*.

Result prediction provides a macro level perspective of the incentives of a community. We can create models based on the characteristics of a candidate to predict the result of a collective action. This will lead to understanding on

a communal level of what features are preferred and if there are voting blocks formed within the members based on the type of candidate. This translates to practical examples such as party level dynamics in a parliament, the topic of a bill or the credentials of a nominee [8, 46, 47].

On the other hand, when we focus on the vote prediction problem, we get a deeper understanding of the dynamics between voters and the candidate. In fields such as game theory, this is well studied using frameworks such as *strategic voting models* and *momentum* [3, 5, 39, 41, 49]. These models are more theoretical and are studied under synthetic conditions. Nevertheless, they still provide a foundation on which we can construct practical models that can utilize additional external features. One such popular approach is using textual information from bills, speeches and legislature to predict votes of politicians in parliament [7, 18]. The next important step is to represent the voting data as networks and leveraging network features to understand and predict voter behaviour.

## 3.2 Voting and Signed networks

Votes by nature express a positive or negative relationship between a *voter* and a *candidate*. Therefore, *signed graphs* provide an intuitive way to structurally represent the voting pattern of members in a community. These signed voting networks can be used to develop models to solve the task of vote prediction and analyse voter behaviour.

Correlation clustering and community detection of signed voting graphs can discover trends and vote blocks in communities [4, 6]. Analysing the networks using social theories of balance and status provides knowledge of voter behaviour and features for prediction methods [15, 35]. The vote prediction task can be broken down into two subtasks which are analogous to link and sign prediction in signed networks.

The first subtask is to predict who will vote next given a candidate  $c$  and a set of previous votes. This subtask is similar to link prediction in a signed network where we aim to predict possible future edges of the type  $(v, c)$ . The complexity subtask depends on the format of voting that takes place. If there is a known voting order, such as in roll call in parliamentary proceedings or explicit timestamps then it is essentially solved. If the voting occurs simultaneously, the subtask can be reduced to predicting the possible subset of members who will vote in a given session. When the voting is iterative and there is no known underlying process of who votes next, then a separate model might be required to infer the voting sequence. This case can be combinatorially hard as we would need to find the correct ordering of

votes in a session.

The second subtask is to predict how a voter  $v$  will vote for a candidate  $c$  given the previous votes in the session. This task translates into predicting the sign of an edge  $(v, c)$  in the current session. We call this subtask the *independent vote prediction* problem. It is independent in nature as we are only interested in the decision of the voter  $v$  assuming that we have complete prior knowledge of how the previous votes have been cast. This problem can be framed as a supervised learning task, using features of the interaction between the current voter  $v$ , previous voters  $U$  and the candidate  $c$  to predict the sign of the edge  $(v, c)$ . We can utilize theories of balance and status in signed networks to create models similar to those by Karimi et al. [27] and Jankowski-Lorek et al. [26] to predict voter behaviour.

### 3.3 Linear Combination of Graphs

In this section we explain how the approaches outlined in Section 2.4 can be applied to solve the *independent vote prediction* problem. As discussed previously, the *edge sign prediction* task in signed network is analogous to vote prediction. The models proposed by Leskovec et al. [33] can be used to predict the sign of the edge  $(v, c)$ . Voting in a community takes place across many sessions in a chronological manner. Therefore, we must partition the training and testing datasets to avoid data leakage. We propose the following framework to gather features using a linear combination of the voting history and several auxiliary graphs.

We denote the directed signed graph for the current voting session as  $S = (V_S, E_S)$ . The current voter in consideration is denoted by  $v$  and the candidate of the session is  $c$ . The votes that have been cast prior to the current voter exists as edges  $(u, c)$  in the session  $S$  and the set of prior voters is denoted as  $U = \{u \mid (u, c) \in E_S\}$ . The history  $H = (V_H, E_H)$  is defined as the directed signed graph containing the votes from sessions chronologically prior to  $S$ . We also define a set of auxiliary graphs  $A = \{G_1, G_2, \dots, G_l\}$  based on external non-voting data. These auxiliary graphs can be either directed or undirected, weighted or unweighted, signed or unsigned. This is similar to the relational layer in *Multidimensional Social Networks* (MSN) described by Kazienko et al. [28] and Jankowski-Lorek et al. [26]. However, the auxiliary graphs capture different relations between a subset of the voting members which will be used to generate additional features for the vote prediction task.

Algorithm 1 describes how to generate a feature vector  $\mathbf{a}$  from the auxiliary graphs  $A$ . The algorithm finds the intersection of the prior voters  $U$  and

the neighbourhood of the current voter  $v$  in the graph  $G_i$  which we call the *voting neighbourhood*. Then the feature  $\mathbf{a}_i$  is computed as the weighted sum of the voting neighbourhood plus the edge weight to the candidate  $(v, c)$ . Figure 3.1 provides an example with three auxiliary graphs and two prior voters  $u_1$  and  $u_2$ . The dashed red edges are the votes cast in the current session  $S$  by the prior voters. We see that in the example  $G_1$  is a directed graph,  $G_2$  is an undirected graph and  $G_3$  is a signed directed graph. The current voter  $v$  has different relations to his voting neighbourhood in each auxiliary graph and therefore each feature is a different combination of edge weights.

In addition to the auxiliary feature vector  $\mathbf{a}$ , we can also create triad features based on the historical voting graph  $H$ . Similar to Leskovec et. al. [33] and Karimi et al. [27] we can form a set of unique triads  $T$ . Then, for each node  $u$  in the common neighbourhood of  $N_{vc}$  we can count the triad formed from the three vertices. Algorithm 2 describes this procedure.

We can now create a feature matrix  $\mathbf{X}$  for all the sessions in a given dataset. Each row is the concatenation of the auxiliary feature vector and the triadic feature vector  $\mathbf{x} = [\mathbf{a}, \mathbf{t}]$ . The target vector  $\mathbf{y}$  is the vector of true votes. Now we train a linear machine learning model and use it to predict the votes in a test dataset.

---

**Algorithm 1:** Auxiliary feature vector for voter  $v$

---

**Input:** Voter  $v$ , Candidate  $c$ , Set of auxiliary graphs  $A$ , Current voting session  $S$  and Prior voters  $U$

**Result:** Auxiliary Feature vector  $\mathbf{a}$

Initialize  $\mathbf{a}$  of length  $|A|$

**foreach**  $G_i$  *in*  $A$  **do**

$Z = N_i(v) \cap U$                       // neighbours in  $G_i$  who have voted

$\mathbf{a}_i \leftarrow 0$

**foreach**  $z$  *in*  $Z$  **do**

        /\* vote in current session multiplied by the edge weight in  $G_i$  \*/

$\mathbf{a}_i += w_S((z, c)) \cdot w_i((v, z))$

**end**

$\mathbf{a}_i += w_i((v, c))$                       // Add edge weight to candidate

**end**

**return**  $\mathbf{a}$

---



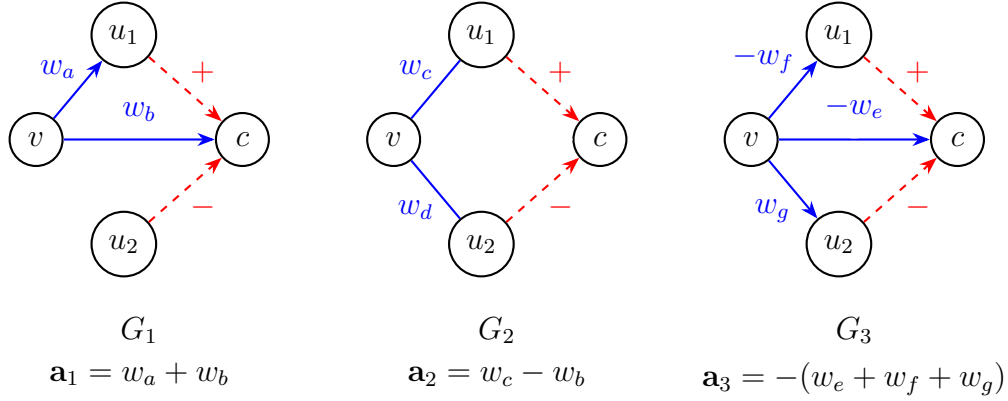


Figure 3.1: Example auxiliary features for  $v$  from combination of three graphs and two prior voters  $u_1$  and  $u_2$ . Dashed red lines are prior votes in the session. Solid blue lines are edge weights in auxiliary graph.  $\mathbf{a}_i$  is feature for voter  $v$  from auxiliary graph  $G_i$

---

**Algorithm 2:** Triad feature vector for voter  $v$

---

**Input:** Voter  $v$ , Candidate  $c$ , Set of unique triads  $T$ , Voting history graph  $H$

**Result:** Triad Feature vector  $\mathbf{t}$

$k \leftarrow |T|$

Initialize counters  $cnt_1, \dots, cnt_k$  to 0

$N_{vc} = N_H(v) \cap N_H(c)$                       // common neighbours in  $H$

**foreach**  $u$  in  $N_{vc}$  **do**

    Let  $\Delta$  be the triad formed by vertices  $\{v, u, c\}$

    Classify  $\Delta$  as the  $j$ th triad in  $T$

    Increment counter  $cnt_j$

**end**

$\mathbf{t} \leftarrow [cnt_1, \dots, cnt_k]$

**return**  $\mathbf{t}$

---

### 3.4 Local Signed Network

- Explain the concept of the local signed network for a particular user
- Motivate the definition with respect to voters and influence in a small network
- Describe how to use balance and status theory to predict the vote

- Balance is through creating a signed adjacency matrix and then computing the smallest eigenvalue and choosing the edge that makes the graph most balanced
- The status is measured using the concept of agony and in a similar way we choose the edge that has least agony when added.
- Use them as deterministic rules or confidence values, akin to Logistic Regression

## 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 (XGBOOOOST)
- Discuss how each graph contributes features and the problem is a linear classification problem

### 5.3.2 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] AGRAWAL, P., GARG, V. K., AND NARAYANAM, R. Link label prediction in signed social networks. In *Twenty-Third International Joint Conference on Artificial Intelligence* (2013).
- [2] AHMADALINEZHAD, M., AND MAKREHCHI, M. Sign prediction in signed social networks using inverse squared metric. In *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation* (2018), pp. 220–227.
- [3] ALI, S. N., AND KARTIK, N. A theory of momentum in sequential voting.
- [4] ARINIK, N., FIGUEIREDO, R., AND LABATUT, V. Signed graph analysis for the interpretation of voting behavior, 2017.
- [5] BANERJEE, A. V. A simple model of herd behavior. *The quarterly journal of economics* 107, 3 (1992), 797–817.
- [6] 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).
- [7] BUDHWAR, A., KUBOI, T., DEKHTYAR, A., AND KHOSMOOD, F. predicting the vote using legislative speech. In *Proceedings of the 19th Annual International Conference on Digital Government Research* (2018), p. 35.
- [8] BURKE, M., AND KRAUT, R. Mopping up: modeling wikipedia promotion decisions. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work* (2008), pp. 27–36.

- [9] 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.
- [10] CARTWRIGHT, D., AND HARARY, F. Structural balance: a generalization of heider’s theory. *Psychological Review* 63, 5 (1956), 277–293.
- [11] CESA-BIANCHI, N., GENTILE, C., VITALE, F., AND ZAPPELLA, G. A correlation clustering approach to link classification in signed networks. In *Annual Conference on Learning Theory* (2012), vol. 23, Microtome, pp. 34–1.
- [12] 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.
- [13] 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.
- [14] DAVIS, J. A. Structural balance, mechanical solidarity, and interpersonal relations. *American Journal of Sociology* 68, 4 (1963), 444–462.
- [15] DERR, T., AND TANG, J. Congressional vote analysis using signed networks. In *2018 IEEE International Conference on Data Mining Workshops (ICDMW)* (2018), IEEE, pp. 1501–1502.
- [16] DIESTEL, R. *Graph Theory*. 1997.
- [17] GALLIER, J. Spectral theory of unsigned and signed graphs. applications to graph clustering: a survey. *arXiv preprint arXiv:1601.04692* (2016).
- [18] GERRISH, S., AND BLEI, D. M. Predicting legislative roll calls from text. In *Proceedings of the 28th International Conference on Machine Learning* (2011), pp. 489–496.
- [19] GU, S., CHEN, L., LI, B., LIU, W., AND CHEN, B. Link prediction on signed social networks based on latent space mapping. *Applied Intelligence* 49, 2 (2019), 703–722.

- [20] GUHA, R., KUMAR, R., RAGHAVAN, P., AND TOMKINS, A. Propagation of trust and distrust. In *Proceedings of the 13th international conference on World Wide Web* (2004), pp. 403–412.
- [21] GUPTE, M., SHANKAR, P., LI, J., MUTHUKRISHNAN, S., AND IFTODE, L. Finding hierarchy in directed online social networks. In *Proceedings of the 20th international conference on World wide web* (2011), pp. 557–566.
- [22] HARARY, F. On the notion of balance of a signed graph. *Michigan Mathematical Journal* 2, 2 (1953), 143–146.
- [23] HEIDER, F. Attitudes and cognitive organization. *The Journal of Psychology* 21, 1 (1946), 107–112.
- [24] HOU, Y. P. Bounds for the least laplacian eigenvalue of a signed graph. *Acta Mathematica Sinica* 21, 4 (2005), 955–960.
- [25] HSIEH, C.-J., CHIANG, K.-Y., AND DHILLON, I. S. Low rank modeling of signed networks. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining* (2012), pp. 507–515.
- [26] 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.
- [27] KARIMI, H., DERR, T., BROOKHOUSE, A., AND TANG, J. Multi-factor congressional vote prediction. In *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (2019), pp. 266–273.
- [28] KAZIENKO, P., MUSIAL, K., KUKLA, E., KAJDANOWICZ, T., AND BRÓDKA, P. Multidimensional social network: model and analysis. In *International Conference on Computational Collective Intelligence* (2011), Springer, pp. 378–387.
- [29] 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.

- [30] KHODADADI, A., AND JALILI, M. Sign prediction in social networks based on tendency rate of equivalent micro-structures. *Neurocomputing* 257 (2017), 175–184.
- [31] KUNEGIS, J., SCHMIDT, S., LOMMATZSCH, A., LERNER, J., LUCA, E. W. D., AND ALBAYRAK, S. Spectral analysis of signed graphs for clustering, prediction and visualization. In *SDM* (2010), pp. 559–570.
- [32] 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.
- [33] 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.
- [34] 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.
- [35] LEVORATO, M., AND FROTA, Y. Brazilian congress structural balance analysis. *Journal of Interdisciplinary Methodologies and Issues in Sciences* (2016).
- [36] LI, H. S., AND LI, H. H. A note on the least (normalized) laplacian eigenvectors of signed graphs. *Tamkang Journal of Mathematics* 47, 3 (2016), 271–278.
- [37] LIBEN-NOWELL, D., AND KLEINBERG, J. The link-prediction problem for social networks. *Journal of the Association for Information Science and Technology* 58, 7 (2007), 1019–1031.
- [38] LIU, S., XIAO, J., AND XU, X. Link prediction in signed social networks: from status theory to motif families. *IEEE Transactions on Network Science and Engineering* (2019), 1–1.
- [39] MEIR, R., GAL, K., AND TAL, M. Strategic voting in the lab: compromise and leader bias behavior. *Autonomous Agents and Multi-Agent Systems* 34, 1 (2020), 31.
- [40] SHAHRIARI, M., AND JALILI, M. Ranking nodes in signed social networks. *Social Network Analysis and Mining* 4 (12 2014).

- [41] 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.
- [42] TANG, J., CHANG, S., AGGARWAL, C., AND LIU, H. Negative link prediction in social media. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining* (New York, NY, USA, 2015), WSDM '15, Association for Computing Machinery, pp. 87–96.
- [43] TATTI, N. Faster way to agony discovering hierarchies in directed graphs. In *ECMLPKDD'14 Proceedings of the 2014th European Conference on Machine Learning and Knowledge Discovery in Databases - Volume Part III* (2014), pp. 163–178.
- [44] TATTI, N. Tiers for peers: a practical algorithm for discovering hierarchy in weighted networks. *Data Mining and Knowledge Discovery* 31, 3 (2017), 702–738.
- [45] YANG, S.-H., SMOLA, A. J., LONG, B., ZHA, H., AND CHANG, Y. Friend or frenemy? predicting signed ties in social networks. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval* (New York, NY, USA, 2012), SIGIR '12, Association for Computing Machinery, pp. 555–564.
- [46] YANO, T., SMITH, N. A., AND WILKERSON, J. D. Textual predictors of bill survival in congressional committees. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (2012), pp. 793–802.
- [47] YOGATAMA, D., HEILMAN, M., O'CONNOR, B., DYER, C., ROUTLEDGE, B. R., AND SMITH, N. A. Predicting a scientific community's response to an article. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing* (Edinburgh, Scotland, UK., July 2011), Association for Computational Linguistics, pp. 594–604.
- [48] ZASLAVSKY, T. Signed graphs. *Discrete Applied Mathematics* 4, 1 (1982), 47–74.
- [49] 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.