Balance in Signed Bipartite Networks

Tyler Derr
Data Science and Engineering Lab
Michigan State University
derrtyle@msu.edu

Cassidy Johnson
University of the Pacific
c_johnson49@u.pacific.edu

Yi Chang
School of Artificial Intelligence
Jilin University, China
yichang@jlu.edu.cn

Jiliang Tang
Data Science and Engineering Lab
Michigan State University
tangjili@msu.edu

ABSTRACT

A large portion of today's big data can be represented as networks. However, not all networks are the same, and in fact, for many that have additional complexities to their structure, traditional general network analysis methods are no longer applicable. For example, signed networks contain both positive and negative links, and thus dedicated theories and algorithms have been developed. However, previous work mainly focuses on the unipartite setting where signed links connect any pair of nodes. Signed bipartite networks on the one hand, are commonly found, but have primarily been overlooked. Their complexities of having two node types where signed links can only form across the two sets introduce challenges that prevent most existing literature on unipartite signed and unsigned bipartite networks from being applied. On the other hand, balance theory, a key signed social theory, has been generally defined for cycles of any length and is being used in the form of triangles for numerous unipartite signed network tasks. However, in bipartite networks there are no triangles and furthermore there exist two types of nodes. Therefore, in this work, we conduct the first comprehensive analysis and validation of balance theory using the smallest cycle in signed bipartite networks - signed butterflies (i.e., cycles of length 4 containing the two node types). Then, to investigate the applicability of balance theory aiding signed bipartite network tasks, we develop multiple sign prediction methods that utilize balance theory in the form of signed butterflies. Our sign prediction experiment on three real-world signed bipartite networks demonstrates the effectiveness of using these signed butterflies for not only sign prediction, but paves the way for improvements in other signed bipartite network analysis tasks.

KEYWORDS

Signed Bipartite Networks; Balance Theory; Sign Prediction

1 INTRODUCTION

Much of the data being produced today can be represented in networks. However, many networks exist that fall outside the range of being able to apply the general network analysis methods to them due to their added complexities in structure. One such type that has become increasingly ubiquitous, especially with the growing popularity of online social media and e-commerce, are signed networks, which not only have positive links, but also allow for the construction of negative links. For example, negative links can be used to represent distrust, a mechanism to warn others of a potential "scammer" in online e-commerce, or in social networks they can represent the connections with our foes (or blocked users).

Previous work and theories have primarily focused on unipartite signed networks, which are networks that have a single node type and signed links are able to connect any two nodes in the network. However, a common form of signed networks that have primarily been overlooked – signed bipartite networks. These networks have two sets of nodes and links are only able to be formed between nodes of different types. Actually, signed bipartite networks appear across multiple domains. For example, in e-commerce, a signed bipartite network can be constructed between buyers and sellers in multi-vendor marketplaces when the users are asked to rate the other after each transaction and helpfulness ratings from users to reviews can be naturally denoted as a signed bipartite network. Another important application of signed bipartite networks is from the political science domain, more specifically, we observe that indeed the United States Congress is inherently a signed bipartite network formed from the representatives and the bills they have voted on (where the "Yea" and "Nay" votes can be represented as positive and negative links, respectively) [11, 21]. In addition, many online systems, such as Netflix and YouTube, adopt "thumbs-up" or "thumbs-down" rating systems that can also be formulated as signed bipartite networks.

Although there have been works focused on unsigned bipartite networks, these methods are lacking the capability to handle the further complexities of negative links. Similarly, methods developed for unipartite signed networks might not be applicable when having

the two node types or limiting the possible connections in the network. For example, a fundamental theory that explains the social phenomena of the link structure in signed network analysis is balance theory [6, 17]. It suggests that a cycle in signed networks with an even number of negative links is balanced, which is typically stated as "a friend of my friend is my friend" while an "enemy of my friend is my enemy". In unipartite signed networks balance theory has been extensively applied on signed triangles (i.e., the smallest undirected cycle) across various real-world networks to obtain better performance across modeling [9, 33, 42], measuring [5, 37, 41, 45], and mining applications [4, 7, 20, 23] . However, in signed bipartite networks it is fundamentally impossible to have any triangles while having the two different types of nodes. Therefore it is important to understand balance theory in signed bipartite networks and its possibility to enhance applications, due to the prevalence of signed bipartite networks. Thus, dedicated efforts are desired for signed bipartite networks in additional to unipartite signed networks and unsigned bipartite networks.

In this paper, we perform an initial investigation of balance theory in undirected signed bipartite networks. As aforementioned, balance theory has been utilized to advance numerous tasks in unipartite signed networks and sign prediction is the one being benefited most. Hence, we then investigate how to utilize balance theory to boost sign prediction in signed bipartite networks. This paves the way for using balance theory for other network analysis tasks in signed bipartite networks. The main contributions of the paper are summarized as the following:

- We conduct the first comprehensive analysis and validation of balance theory for signed bipartite networks;
- We leverage balance theory for the construction of multiple sign prediction methods.
- We perform experiments on three real-world signed bipartite network datasets to understand balance theory and sign prediction in signed bipartite networks.

The remainder of this paper is organized as follows. Our signed bipartite datasets are introduced along with our analysis and investigation of extending balance theory to signed bipartite networks in the form of signed butterflies in Section 2. Then, in Section 3 we present numerous sign prediction methods for signed bipartite networks based on the signed butterflies. Next we perform experiments for predicting the sign of missing links using our proposed methods in Section 4. Related work is presented in Section 5 and briefly discussed. Finally, conclusions and future work are given in Section 6.

2 BALANCE THEORY IN UNDIRECTED SIGNED BIPARTITE NETWORKS

In this section, we will introduce the datasets we have collected for this study. Thereafter we discuss balance theory from a general signed network perspective, then we validate its applicability in signed bipartite networks, and perform a preliminary analysis on our datasets; but first, we introduce the definitions and notations.

Consider an undirected signed bipartite network, $\mathcal{G} = (\mathcal{U}_B, \mathcal{U}_S, \mathcal{E}^+, \mathcal{E}^-)$, where $\mathcal{U}_B = \{b_1, b_2, \ldots, b_{n_B}\}$ and $\mathcal{U}_S = \{s_1, s_2, \ldots, s_{n_S}\}$ represent two mutually exclusive sets of homogeneous nodes with

Table 1: Notations.

Notations	Descriptions
$\mathbf{B}$	Undirected signed bidjacency matrix
$\mathbf{P}_B$	Adjacency matrix for the $\mathcal{U}_B$-projection network
$\mathbf{P}_S$	Adjacency matrix for the $\mathcal{U}_S$-projection network
$\mathbf{A}$	Adjacency matrix constructed from $\mathbf{B}, \mathbf{P}_B$, and $\mathbf{P}_S$
$\mathbf{U}$	Low-dimensional representation of nodes in $\mathcal{U}_B$
$\mathbf{V}$	Low-dimensional representation of nodes in $\mathcal{U}_S$

Table 2: Statistics on Signed Bipartite Networks.

	Bonanza	U.S. Senate	U.S. House
$n_B = |\mathcal{U}_B|$	7,919	1,056	1,281
$n_S = |\mathcal{U}_S|$	1,973	145	515
$|\mathcal{E}| = |\mathcal{E}^+| + |\mathcal{E}^-|$	36,543	27,083	114,378
% Links Positive	97.98%	55.31%	53.96%
% Links Negative	2.02%	44.69%	46.04%
Density of $\mathbf{B}$	2.339×10^{-3}	0.1769	0.1734

n_B and n_S representing the number of nodes for each set, respectively. $\mathcal{E}^+ \subset \mathcal{U}_B \times \mathcal{U}_S$ and $\mathcal{E}^- \subset \mathcal{U}_B \times \mathcal{U}_S$ represent the sets of positive and negative edges, respectively, between the two sets of nodes $\mathcal{U}_B$ and $\mathcal{U}_S$. We let $\mathcal{E} = \mathcal{E}^+ \cup \mathcal{E}^-$ be the set of all edges where $\mathcal{E}^+ \cap \mathcal{E}^- = \emptyset$, in other words, two nodes cannot have both a positive and negative edge between them. We use $\mathbf{B} \in \mathbb{R}^{n_B \times n_S}$ to represent the undirected signed bipartite biadjacency matrix of $\mathcal{G}$, where $\mathbf{B}_{ij} = 1, -1$, or 0, when there exists a positive, negative, or no link between b_i and s_j. We further summarize the major notations used throughout this work in Table 1.

2.1 Signed Bipartite Networks

We have collected three signed bipartite networks for this study. The first signed bipartite network is from the e-commerce website Bonanza[1]. Bonanza is similar to eBay[2] and Amazon Marketplace[3] in that users create an account for which they can buy or sell various goods. After a buyer purchases a product from a seller, both are able to provide a rating about the other along with a short comment. At the time of collection, Bonanza was using a rating scale of "Positive", "Neutral", and "Negative" to rate another user after a transaction. For representing the buyers and sellers, we use $\mathcal{U}_B$ and $\mathcal{U}_S$, respectively.

The next two datasets are representing the role call votes combined from the 1st to 10th United States Congress. More specifically, we collected two separate datasets[4]; one for the U.S. Senate and the other for the U.S. House of Representatives (which we will refer to as U.S. House). In each of these datasets we represent the bills that were voted by the set $\mathcal{U}_B$ and the senators or representatives by $\mathcal{U}_S$. If a congressperson voted "Yea" or "Nay" for the bill, then we represent these as positive or negative links between them, respectively, and leave the connection missing otherwise.

Note that for simplicity throughout the rest of this work we will refer to the nodes in $\mathcal{U}_B$ as "buyers" and those in $\mathcal{U}_S$ as "sellers". In Table 2 we report some basic statistics of our three collected datasets.

---

[1]http://www.bonanza.com
[2]http://www.ebay.com
[3]http://www.amazon.com
[4]https://www.govtrack.us/data/

Table 3: Signed Butterfly Statistics on Signed Bipartite Networks.

Signed Butterfly Isomorphism Classes	Bonanza				U.S. Senate				U.S. House of Representatives			
	Count	%	E%	s	Count	%	E%	s	Count	%	E%	s
(A) $(+, +, +, +)$	2554388	0.986	0.922	386	13404168	0.262	0.094	4142	227660420	0.244	0.085	17459
(B) $(+, -, -, +)$	3830	0.001	7.8e-04	40	5595440	0.110	0.122	-277	103731010	0.111	0.123	-1137
(C) $(+, +, -, -)$	726	2.8e-04	7.8e-04	-29	9404006	0.184	0.122	1349	173875858	0.186	0.123	5843
(D) $(+, -, +, -)$	456	1.7e-04	7.8e-04	-35	5537080	0.108	0.122	-302	101409932	0.109	0.123	-1368
(E) $(-, -, -, -)$	20	7.7e-06	1.7e-07	30	6815324	0.133	0.040	3414	137478104	0.147	0.045	15104
Balanced	2559420	**0.988**	0.924		40756018	**0.797**	0.500		744155324	**0.797**	0.500	
(F) $(+, +, +, -)$	30685	0.012	0.076	-390	6225745	0.122	0.302	-2811	109763190	0.118	0.289	-11565
(G) $(+, -, -, -)$	100	3.9e-05	3.2e-05	2	4118075	0.081	0.197	-2099	79053742	0.085	0.210	-9430
Unbalanced	30785	**0.012**	0.076		10343820	**0.203**	0.500		188816932	**0.203**	0.500	

We note that in the Bonanza dataset there is a significant imbalance between the number of positive and negative links as compared to the two U.S. Congress datasets. Although these datasets are representing vastly different real-world social structures, we next investigate balance theory [6, 17] to the signed bipartite network setting.

2.2 Signed Butterflies in Bipartite Networks

In signed networks one of the most fundamentally studied social theories is balance theory [6, 17], which discusses the settings in signed networks that are socially "balanced" (i.e., stable), and those that are more likely to change (to be balanced) due to the social tensions involved in maintaining "unbalanced" and seemingly unnatural connections. In recent signed network analysis works balance theory is usually investigated and then applied towards many tasks[28, 39, 44], but almost always in the form of triangles (or cycles of length 3) in a unipartite signed network. As seen in Figure 1, there are four possible configurations between the three nodes. We can further observe in Figure 1 that triangles (a) and (b) are balanced (due to having an even number of negative links), while (c) and (d) are unbalanced. Nevertheless, as previously mentioned, since there are no triangles in signed bipartite networks and they have two different node types, it is unknown whether balance theory is still applicable towards a bipartite setting.

In this subsection, we will therefore introduce how we plan to extend the usage of balance theory to the smallest signed cycles (i.e., butterflies) in undirected signed bipartite networks. Next we investigate and present our initial analysis of these signed butterflies in three real-world signed bipartite networks.

2.2.1 Signed Butterfly Isomorphism Classes. In unsigned bipartite networks, one commonly investigated structure is that of a "butterfly" [1, 36], which is a cycle of length 4. More formally, a butterfly is the simplest cohesive higher-order structure and also a complete biclique. Thus, this provides the most natural structure to investigate as a possible extension for balance theory in signed bipartite networks.

Just as there are different types of signed triangles, there are different types of signed butterflies. In Figure 2 we present the 7 non-isomorphic undirected signed butterflies. Note that there are five that adhere to balance theory while only two are categorized
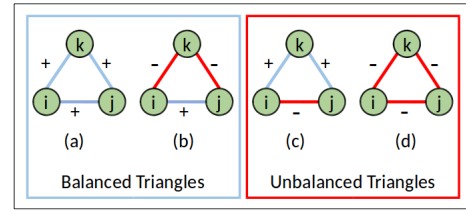
Figure 1: Undirected Signed Triangle Isomorphism Classes.
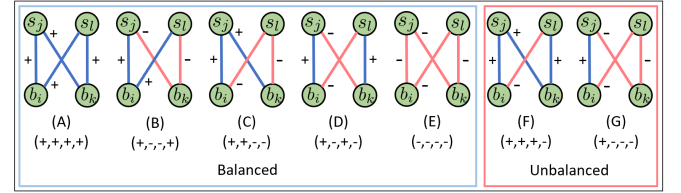
Figure 2: Undirected Signed Butterfly Isomorphism Classes.

as unbalanced. We use the notation $(*, *, *, *)$ to denote a signed butterfly isomorphism class that represents the links between the buyers and sellers (b_i, s_j, b_k, s_l) (in that order with the last sign connecting s_l and b_i). The simplest of types are $(+, +, +, +)$ and $(-, -, -, -)$, which denote the classes having all positive or all negative links, respectively, and both are balanced due to having an even number of negative links (and can be seen in Figures 2(A) and 2(E), respectively). We can interpret the $(+,+,+,+)$ class as the situations where two buyers have bought from the same two sellers and the sentiment amongst them across the four purchases was positive. Next, we have $(+, +, +, -)$ and $(+, -, -, -)$, which are the two unbalanced classes of signed butterflies (since they have an odd number of negative links). In Figure 2(F) we have the signed butterfly isomorphism class that encompasses all the signed butterflies with a single negative link. We can observe that no matter where this single negative link is placed, we always have one buyer with two positive links, one buyer with a positive and negative link, and similar structure for the two sellers. The isomorphism class $(+, -, -, -)$ can be seen as the complement (if defined as swapping link signs in a signed network) of the class $(+, +, +, -)$ and defined in a similar way, but with swapping the positive and negative links in the definition. This leaves the signed butterflies having

two positive and two negative links, of which we have three iso-morphism classes. In Figure 2(D) we see the class $(+, -, +, -)$ is used to represent signed butterflies where all buyers and sellers have one positive and one negative link in their cycle. When one of the buyers has two positive links, while the other buyer has two positive links, we observe in Figure 2(B) that both sellers have a single positive and single negative link, and define the isomorphism class of $(+, -, -, +)$. Finally, the last type of signed butterfly has both buyers connected positively to one seller, and negatively to the other, which we represent as the class $(+, +, -, -)$ shown in Figure 2(C).

2.2.2 Signed Butterfly Analysis. In Table 3 we report our analysis after counting the number of signed butterflies for each isomorphism class as shown in Figure 2. We further calculated the percentage each isomorphism class takes up of the total signed butterfly count in each dataset (given in column "%"). Next, we analyzed the significance of these signed butterflies being found in signed bipartite networks and wanted to test whether they are overrepresented or underrepresented. Remember, balance theory would suggest that balanced isomorphism classes (A) through (E) should appear frequently while (F) and (G) (being unbalanced) should appear less frequently. To quantify this, extending the approach taken in [28], we calculate "$E\%$" as the expected percentage of total signed butterflies to fall into the given isomorphism class when randomly reassigning the positive and negative signs to the signed bipartite network. In other words, for example, "$E\%$" for the isomorphism class $(+, -, -, -)$ is calculated by $\binom{4}{1}\left((|\mathcal{E}^+|/|\mathcal{E}|) \times (|\mathcal{E}^-|/|\mathcal{E}|)^3\right)$, since there are 4 permutations of having a single positive link in a signed butterfly in class $(+, -, -, -)$ and the probability of each link appearing in a signed network with randomly assigned link signs would be the independent probabilities of having a single positive link (i.e., $|\mathcal{E}^+|/|\mathcal{E}|$) and three negative links (i.e., $|\mathcal{E}^-|/|\mathcal{E}|$). Finally, the value "$s$" is used to denote the number of standard deviations the actual count differs from our calculated expected number (based on "$E\%$") for each signed butterfly type and just as in [28], a positive (or negative) "s" value signifies appearing significantly more (or less) than expected.

We first observe that the large majority of signed butterflies in our three signed bipartite networks are indeed balanced. Furthermore, they are significantly more balanced than expected based on the link sign ratio in the given network (i.e., comparing columns "%" and $E\%$). The second observation is that all unbalanced signed butterflies across the three datasets are significantly underrepresented, except for the $(+, -, -, -)$ butterflies in Bonanza, where it shows a minimal over representation. Similarly, across all datasets the $(+, +, +, +)$ and $(-, -, -, -)$ signed butterflies are significantly overrepresented, further strengthening the applicability of balance theory in signed bipartite networks. However, the isomorphism classes involving two positive and two negative links appear to not always be found overrepresented. For example, the class where all buyers and sellers have one positive and one negative link, i.e., $(+, -, +, -)$, is less commonly found than expected across all three datasets.

In summary, our findings suggest that: 1) we can use signed butterflies to extend balance theory for signed bipartite networks; and 2) signed bipartite networks adhere to balance theory when

defined in terms of signed butterflies, thus making them applicable to advance numerous tasks in signed bipartite networks.

2.3 Signed Caterpillars in Bipartite Networks

In this subsection, we discuss the notion of "signed caterpillars", which we denote as paths of length 3 that are missing just one link to becoming a signed butterfly.

A signed caterpillar can take on one of eight different forms, since it is composed of three links being ether positive or negative. Note that all caterpillar types have the potential to be transformed into a signed butterfly (i.e., closed into a cycle of length 4) that is either balanced or unbalanced. If a signed caterpillar contains an even number of negative links, we refer this as a "balanced path" and balance theory would suggest a positive (or negative) link transforming it into a balanced (or unbalanced) signed butterfly. Similarly, we define an a signed caterpillar as an "unbalanced path" when having an odd number of negative links and balance theory would suggest a negative (or positive) link to close into a balanced (or unbalanced) signed butterfly.

3 SIGN PREDICTION FOR UNDIRECTED SIGNED BIPARTITE NETWORKS

With the aforementioned definitions and notations, we formally define the problem of sign prediction in undirected signed bipartite networks as the following:

Given an undirected signed bipartite network $\mathcal{G} = (\mathcal{U}_B, \mathcal{U}_S, \mathcal{E}^+, \mathcal{E}^-)$ represented as a biadjacency matrix $\mathbf{B} \in \mathbb{R}^{|\mathcal{U}_B| \times |\mathcal{U}_S|}$, we seek to predict the signs of no link pairs $(b_i, s_j) \in \{\mathcal{U}_B \times \mathcal{U}_S\} \backslash \{\mathcal{E}^+ \cup \mathcal{E}^-\}$.

Sign prediction in signed networks has been previously studied [7, 19, 27, 35, 49]. However, in the signed bipartite setting, many of these methods are no longer applicable, since there are no triangles. In Section 2.2, we validated that the large majority of signed butterflies in signed bipartite networks are balanced. Methods for predicting link signs in unipartite signed networks can be categorized into three main groups: 1) supervised methods; 2) low-rank approximation methods; and 3) propagation based methods. Therefore we develop a representative sign prediction method specific to signed bipartite networks from each group. More specifically, we propose: 1) a supervised classification method that uses signed caterpillars/butterflies; 2) extend a low-rank modeling method to ensure the predicted signs favor creating more balanced signed butterflies; and 3) a random walk based approach that integrates one-mode projection networks for $\mathcal{U}_B$ and $\mathcal{U}_S$ constructed using balance theory.

3.1 Signed Caterpillars Based Classifier

One common approach towards predicting links or link signs in both signed and unsigned networks is to frame the task in terms of a supervised classification problem [2, 7, 27, 32]. Here we extend the idea to the signed bipartite setting by formulating the problem of predicting the sign between a buyer b_i and a seller s_j by extracting features from either the individuals (i.e., their positive and negative degrees) or local neighborhood features based on balance theory (i.e., signed caterpillars).

To train our model we construct a training dataset consisting of known signed links (between a buyer and seller). Then, after having

a trained model, we can extrapolate what we learned from the training data to predict a positive or negative sign for an unknown buyer and seller pair. More specifically, we use a logistic regression model following the prediction on directed signed unipartite networks work in [27].

Feature Extraction. The two different sets of features we evaluate are either based on the two nodes degree distributions or information about how many signed caterpillars they are the two endpoints of (i.e., they would be the buyer and seller connection transforming the signed caterpillar to a balanced or unbalanced signed butterfly). Thus, the feature vector $\mathbf{x}_{ij}^d$ for the pair (b_i, s_j) includes the the positive and negative degrees for both b_i and s_j. In comparison, $\mathbf{x}_{ij}^{sc}$ contains the counts for each of the 8 possible signed caterpillars that have b_i and s_i as the endpoints. The expectation is that the features $\mathbf{x}_{ij}^{sc}$ will be more informative than those of $\mathbf{x}_{ij}^d$ because they would provide a vast amount of informaiton as to whether their link sign is likely to be positive or negative according to balance theory when considering the types of signed butterflies that would be constructed. This is in comparison to only using the degrees in a method similar in nature to a signed preferential attachment model with $\mathbf{x}^d$. We denote the supervised classifiers that use $\mathbf{x}_{ij}^d$ and $\mathbf{x}_{ij}^{sc}$ as SCd and SCsc, respectively.

3.2 Low-Rank Sign Prediction

In recent years the low-rank matrix factorization approaches have been gaining popularity for numerous applications involving link related network predictions [18, 34, 34, 38]. Although some of these works have focused on signed networks [14, 18, 38], none are structured to select link signs that would explicitly push towards more signed butterflies being balanced in signed bipartite networks. Thus, we first introduce a basic matrix factorization approach to model the signed bipartite network using the biadjacency matrix $\mathbf{B}$. Then we introduce how we can successfully modify this model through the inclusion of additional pairs of buyers and sellers derived from suggested implicit signed links that would construct the most balanced signed butterflies with the suggested link sign.

3.2.1 Basic Matrix Factorization Model. The set of existing edges in $\mathbf{B}$ are denoted in the set $\mathcal{E} = \{(b_i, s_j)|\mathbf{B} \neq 0\}$. In terms of the link sign prediction task we would like to discover two latent matrices $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_{n_B}] \in \mathbb{R}^{d \times n_B}$ and $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{n_S}] \in \mathbb{R}^{d \times n_S}$ of dimension d for the set of buyers and sellers, respectively, to solve the following optimization problem:

$$\min_{\mathbf{U}, \mathbf{V}} \sum_{(b_i, s_j) \in \mathcal{E}} \max\left(0, 1 - \mathbf{B}_{ij}(\mathbf{u}_i^\top \mathbf{v}_j)\right)^2 + \lambda\left(|\mathbf{U}|_F^2 + |\mathbf{V}|_F^2\right) \quad (1)$$

where $\mathbf{u}_i^\top \mathbf{v}_j$ is used to model the link sign between buyer b_i to seller s_j. Note that when the real link sign (i.e., $\mathbf{B}_{ij}$) and the predicted link sign (i.e., $\mathbf{u}_i^\top \mathbf{v}_j$) are of the same sign (i.e., both positive or both negative) then $\mathbf{B}_{ij}(\mathbf{u}_i^\top \mathbf{v}_j)$ is positive, and if over 1 then there is no loss. However, when the real and predicted values have differing signs then there is a higher loss value associated to drive the minimization during the training process. Following the work in [18] we use Stochastic Gradient Descent (SGD) to minimize the objective in Eq. (1).

This allows us to then utilize the learned low-dimensional representations for each buyer and seller to predict the sign of unknown buyer and seller pairs. However, although this model is effectively learning a representation that can accurately predict the existing links, it does not explicitly control whether the signs of non-existing links are actually going to predict link signs that adhere to balance theory (i.e., having more signed butterflies balanced than unbalanced). Therefore we denote this method simply as MF. Next, we will present an extension to this basic framework to further ensure more signed butterflies between the missing links are balanced.

3.2.2 Matrix Factorization with Balance Theory. As previously discussed, the aforementioned basic matrix factorization approach given in Eq. (1) does not explicitly enforce the non-existing link signs to favor balanced relationships. Instead it can only focus on learning low-dimensional representations for each buyer and seller such that the model minimizes the error on predicting the existing link signs. The approach we have selected is to further encourage the model learning link signs for buyer and seller pairs that currently do not exist in the signed bipartite network, but would convert many signed caterpillars into balanced signed butterflies if they were to exist.

The first step is calculating whether balance theory would suggest a positive or negative link for each buyer and seller pair (b_i, s_j), that currently do not have a link between them, based on the types of signed caterpillars they're jointly involved in and the endpoints of.

THEOREM 3.1. *Given a signed undirected biadjacency matrix $\mathbf{B}$, then the matrix $\hat{\mathbf{S}} = \mathbf{B}\mathbf{B}^T\mathbf{B} \odot \overline{\mathbf{B}}$ is such that $sign(\hat{\mathbf{S}}_{ij})$ suggests the sign of a non-existent link in $\mathbf{B}$ that would result in a net gain of $|\hat{\mathbf{S}}|$ additional balanced signed butterflies created (after subtracting the number of potential unbalanced signed butterflies created simultaneously) if the suggested signed link were to be added between b_i and s_j, where we define $\overline{\mathbf{B}}$ as $\overline{\mathbf{B}}_{ij} = 0$ if $\mathbf{B}_{ij} \neq 0$ and $\overline{\mathbf{B}}_{ij} = 1$ when $\mathbf{B}_{ij} = 0$.*

PROOF. If we let $\mathbf{A} = \begin{bmatrix} \mathbf{0} & \mathbf{B} \\ \mathbf{B}^T & \mathbf{0} \end{bmatrix}$ be the adjacency matrix in $\mathbb{R}^{|\mathcal{U}| \times |\mathcal{U}|}$. We can observe that $\mathbf{A}^3 = \begin{bmatrix} \mathbf{0} & \mathbf{B}\mathbf{B}^T\mathbf{B} \\ \mathbf{B}^T\mathbf{B}\mathbf{B}^T & \mathbf{0} \end{bmatrix}$. We note that in [12] it has been shown $\mathbf{A}^l = \mathbf{M}_B^l - \mathbf{M}_U^l$, where $\mathbf{M}_B^l, \mathbf{M}_U^l \in \mathbb{R}^{|\mathcal{U}| \times |\mathcal{U}|}$ store the number of balanced and unbalanced paths of length l, respectively, between all pairs of nodes in a signed network represented as $\mathbf{A}$. Thus, since $\mathbf{A}_{ij}^3 = \left[\mathbf{B}\mathbf{B}^T\mathbf{B}\right]_{ij}$ for some buyer b_i and seller s_j, we observe that this represents the number of of balanced paths of length 3 subtracted by the number of unbalanced paths of length 3. By definition of a signed caterpillar, if one is a balanced path, then it would suggest a positive link to close to be a balanced signed butterfly, but if it was formed by an unbalanced path it would require the closing link to be negative to form a balanced butterfly. Therefore, it follows that $sign\left(\left[\mathbf{B}\mathbf{B}^T\mathbf{B}\right]\right) = sign\left(\mathbf{M}_B^l - \mathbf{M}_U^l\right)$ indeed represents the sign that would promote the creation of more balanced signed butterflies, and similarly for the net gain of balanced butterflies being formed equaling the absolute value of their difference (i.e, $|\mathbf{M}_B^l - \mathbf{M}_U^l|$). It is then easy to extend to only the buyer and seller pairs b_i and s_j in $\left[\mathbf{B}\mathbf{B}^T\mathbf{B}\right] \odot \overline{\mathbf{B}}$ after taking the element-wise product with $\overline{\mathbf{B}}$ that zeros out the pairs that have an existing link. □

Note that $\hat{\mathbf{S}}$ can also be calculated (sometimes more efficiently) using the following:

$$\hat{\mathbf{S}}_{ij} = \begin{cases} \left[\mathbf{B}\mathbf{B}^\top \mathbf{B} \right]_{ij} & \text{if } \mathbf{B}_{ij} = 0 \\ 0 & \text{otherwise} \end{cases}$$

to avoid using the potentially very dense matrix $\overline{\mathbf{B}}$ for sparse signed bipartite networks.

Using Theorem 3.1 we can construct additional sets $\mathcal{E}_i^+$ and $\mathcal{E}_i^-$ of implicit positive and negative links, respectively, suggested by balance theory that would create the highest net gain of balanced signed butterflies in the signed bipartite network. We define these sets as follows:

$$\hat{\mathcal{E}}_i^+ = \{(b_i, s_j) | \hat{\mathbf{S}}_{ij} > 0 \text{ and } \hat{\mathbf{S}}_{ij} \in top_k(\hat{\mathbf{S}})\}$$
$$\hat{\mathcal{E}}_i^- = \{(b_i, s_j) | \hat{\mathbf{S}}_{ij} < 0 \text{ and } \hat{\mathbf{S}}_{ij} \in bottom_k(\hat{\mathbf{S}})\} \quad (2)$$

where $top_k(\hat{\mathbf{S}})$ and $bottom_k(\hat{\mathbf{S}})$ are used to denote the k largest and smallest values, respectively, in $\hat{\mathbf{S}}$.

We formulate our object that incorporates balance theory as follows:

$$\min_{\mathbf{U}, \mathbf{V}} \sum_{(b_i, s_j) \in \mathcal{E}} \max \left(0, 1 - \mathbf{B}_{ij}(\mathbf{u}_i^\top \mathbf{v}_j)\right)^2 + \lambda \left(|\mathbf{U}|_F^2 + |\mathbf{V}|_F^2\right)$$

$$+ \alpha \sum_{(b_i, s_j) \in \hat{\mathcal{E}}_i^+} \max \left(0, 1 - \hat{\mathbf{S}}_{ij}(\mathbf{u}_i^\top \mathbf{v}_j)\right)^2$$

$$+ \beta \sum_{(b_i, s_j) \in \hat{\mathcal{E}}_i^-} \max \left(0, 1 - \hat{\mathbf{S}}_{ij}(\mathbf{u}_i^\top \mathbf{v}_j)\right)^2 \quad (3)$$

where α and β are used to control the level at which we incorporate the modeling of signed butterflies through the inclusion of the implicit positive and negative links, respectively. We again note that these implicit positive and negative links are implied by balance theory by using $\hat{\mathbf{S}}$, which effectively counts for each node pair (b_i, s_j) what the net gain of total balanced signed butterflies would be once including the link with the suggested sign (according to the majority count of signed caterpillars being of balanced or unbalanced paths of length 3). We denote this matrix factorization method using balance theory as MFwBT.

3.3 Random Walk Based Sign Prediction

Typical propagation based methods, such as the random walk with restart [40] have seen many variants and been applied to solve link prediction and ranking related tasks in unsigned unipartite networks. However, signed bipartite networks pose multiple challenges that prevent them from directly using the typical methods. One such problem is that bipartite networks do not have a stationary distribution and thus do not converge [31]. One way of handling this problem in unsigned bipartite networks is considered a "lazy" random walk, where the walker will probabilistically stay at the same node. We will later use this method as a comparison against our proposed random walk based method. Furthermore as seen in previous sign prediction methods for unipartite signed networks, balance theory is the key component towards obtaining higher performance when predicting the sign of unknown links. Thus, due to our analysis of the signed butterflies, indeed signed bipartite networks are showing high levels of balance and therefore
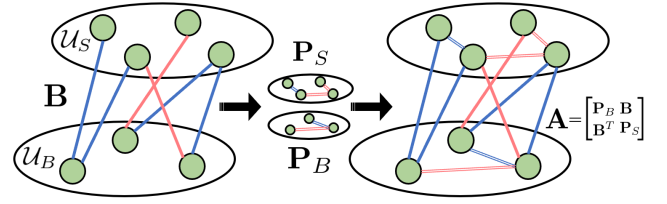

Figure 3: High-level intuition of how we construct A.

we should also be using balance theory to guide the random walk based method for signed bipartite networks towards a solution having more balanced relations.

Here we present a random walk based approach that integrates the $\mathcal{U}_B$ and $\mathcal{U}_S$ one-mode projection adjacency matrices, which are constructed using balance theory, to aid in handling the issues faced with the bipartite setting, and develop a signed random walk based approach to not only allow a proper transition matrix, but to furthermore have the random walker be promoting balance theory. The first step will be the construction of a signed adjacency matrix $\mathbf{A}$ based on balance theory, followed by defining a signed transition matrix that can further promote and propagate balanced relations throughout the network.

3.3.1 Constructing the one-mode adjacency matrices. In unsigned bipartite network analysis one-mode projections are typically used for both analysis and aiding to solve various tasks [16, 50, 51]. They are constructed by creating a projection network that creates implicit connections between nodes of the same type. In terms of our definitions, two one-mode projection networks can be performed, one that connects the buyers in $\mathcal{U}_B$ together amongst themselves and the other for the sellers in $\mathcal{U}_S$ by constructing seller to seller links; these relations can be represented in the adjacency matrices $\mathbf{P}_B \in \mathbb{R}^{|\mathcal{U}_B| \times |\mathcal{U}_B|}$ and $\mathbf{P}_S \in \mathbb{R}^{|\mathcal{U}_S| \times |\mathcal{U}_S|}$. A visual example can be seen in Figure 3 when going from $\mathbf{B}$ to $\mathbf{P}_S$ and $\mathbf{P}_B$ from left to right through the first arrow.

We note that there is not just one way to discover these implicit connections between pairs of users in the same set, and in fact there are many possible methods for one-mode projections [50, 51]. It has also been studied that using different methods to construct the projection networks can cause drastic changes to the usability and performance [47]. In wanting to carefully construct these projection networks, we choose to utilize balance theory in the form of signed triangles. Next we will discuss the formation of the adjacency matrix $\mathbf{P}_B$, and a similar process can be followed for constructing $\mathbf{P}_S$ (although we only discuss $\mathbf{P}_B$ here).

Based on the ideas of common neighbor similarity in unsigned networks, we will possibly connect two buyers b_i and b_j if they have at least one seller in common they are linked to. Let the number of common sellers that b_i and b_j agree upon (in terms of link sign) be denoted as ns_{ij}^A. Similarly let ns_{ij}^D denote the number of sellers these two buyers disagree on in terms of link sign. Then we define $\mathbf{P}_{Bij} = \mathbf{P}_{Bji} = ns_{ij}^A - ns_{ij}^D$, which we can see is taking the number of sellers they agree upon in terms of signed connections (i.e., both either negatively or positively connected to that seller) and subtracting the number of sellers they disagree on (i.e., the

sellers where one buyer has a positive link while the second buyer has a negative connection with that seller). We can now further see the connection between $\mathbf{P}_{Bij}$ and the common neighbor similarity method. It is easy to verify that out of all the triangles formed between b_i, b_j, and the sellers s_k they are commonly linked to, that using the links $\mathbf{B}_{ik}$, $\mathbf{B}_{jk}$ and $\mathbf{P}_{Bij}$, we see that the majority will adhere to balance theory. This is by design since if $na_{ij}^A > na_{ij}^D$ then $sign(\mathbf{P}_{Bij})$ is positive and closing the ns_{ij}^A triangles to be balanced, while the lesser number of ns_{ij}^D will close to be unbalanced. Note that a similar argument can be given when $ns_{ij}^A < ns_{ij}^D$ and if $ns_{ij}^A = ns_{ij}^D$ then $\mathbf{P}_{Bij} = 0$ and no signed triangles are formed. Ultimately, we construct a parameterized version as follows:

$$\mathbf{P}_{Bij} = \begin{cases} 0 & \delta_n < ns_{ij}^A - ns_{ij}^D < \delta_p \\ ns_{ij}^A - ns_{ij}^D & \text{otherwise} \end{cases} \quad (4)$$

where δ_p and δ_n are used to define thresholds for the necessary magnitude of $ns_{ij}^A - ns_{ij}^D$ to have a non-zero value in $\mathbf{P}_{Bij}$. This allows us to ignore adding smaller values (e.g., 2), since in some settings having such a small value might not be very significant and thus we might not want to construct a link between b_i and b_j. Note that for simplicity we allow δ_p and δ_n to be shared for constructing both $\mathbf{P}_B$ and $\mathbf{P}_S$.

3.3.2 Performing the random walk. Now having the two projection adjacency matrices $\mathbf{P}_B$ and $\mathbf{P}_S$, we can use them to construct an adjacency matrix $\mathbf{A} \in \mathbb{R}^{|\mathcal{U}| \times |\mathcal{U}|}$, which will be the unibipartite signed network we perform our random walk on, where $\mathcal{U} = \{b_1, \ldots b_{n_B}, s_1, \ldots s_{n_S}\}$. In Figure 3 we show the high-level intuition of how to construct $\mathbf{A}$. First we denote $\hat{\mathbf{B}}$ as the row normalized biadjacency matrix where $\hat{\mathbf{B}}_{ij} = \mathbf{B}_{ij}/\sum_k |\mathbf{B}_{ik}|$. We similarly construct row normalized adjacency matrices $\hat{\mathbf{P}}_B$ and $\hat{\mathbf{P}}_S$. Now we can formulate $\mathbf{A}$ as follows:

$$\mathbf{A} = \begin{bmatrix} \hat{\mathbf{P}}_B & \omega\hat{\mathbf{B}} \\ \omega\hat{\mathbf{B}}^T & \hat{\mathbf{P}}_S \end{bmatrix} \quad (5)$$

where ω is a parameter that can be used to bias the random walker to favor the real links in our signed bipartite network as compared to the implicit links we obtained through the $\mathcal{U}_B$ and $\mathcal{U}_S$ one-mode projection networks. Next we construct a similar row normalized adjacency matrix $\hat{\mathbf{A}}$ where $\hat{\mathbf{A}}_{ij} = \mathbf{A}_{ij}/\sum_k |\mathbf{A}_{ik}|$.

Finally, we utilize $\hat{\mathbf{A}}$ in a random walk propagation model where we define $\mathbf{Y}$ to be the matrix holding the inferred link signs as follows:

$$\mathbf{Y}_{ij} = \sum_k \hat{\mathbf{A}}_{ik} \mathbf{Y}_{kj} \quad (6)$$

Next we describe how the above adheres to balance theory in terms of triangles of the adjacency matrix $\mathbf{A}$. This is because for some k if $\hat{\mathbf{A}}_{ik}\mathbf{Y}_{kj} > 0$ it increases $\mathbf{Y}_{ij}$ ensuring it to be positive, which would be a triangle consisting of either three positives, or two negatives and a positive. Similarly when $\hat{\mathbf{A}}_{ik}\mathbf{Y}_{kj} < 0$ we are decreasing $\mathbf{Y}_{ij}$ and encouraging it to be negative, thus also following balance theory. The closed form solution that includes the restart capability,

with probability $(1 - c)$, is given to be the following:

$$\mathbf{Y} = (1 - c)(\mathbf{I} - c\hat{\mathbf{A}})^{-1} \quad (7)$$

Note that each signed butterfly involving b_i, s_j, b_k, and s_l in the original network $\mathbf{B}$ now consists of up to $\binom{4}{3}$ triangles in $\hat{\mathbf{A}}$. Thus, when we are encouraging balanced triangles here in $\mathbf{Y}$ this correlates to having balanced signed butterflies in the upper right corner of $\mathbf{Y}$, which is where we obtain the link sign predictions (i.e., when predicting the sign between b_i and s_j we have use $\hat{\mathbf{B}}_{ij'}$ where $j' = (n_B + j)$. We denote this method as Signed Bipartite Random Walk (SBRW).

For comparison, if we set the two one-mode projection matrices to the identity matrix (i.e., $\mathbf{P}_B = \mathbf{P}_S = \mathbf{I}$) and set $\omega = 1$ then Eq. 6 becomes the equation for a lazy random walk method, which we denote as LazyRW.

4 EXPERIMENT

In this section, we empirically evaluate our proposed sign prediction methods for signed bipartite networks that harness balance theory. We seek to answer the following: (1) Does the extended balance theory to signed butterflies in the bipartite setting provide an increase in performance for sign prediction? and (2) How do the proposed methods work/compare? To address these questions we perform experiments to measure the performance for each of the proposed sign prediction methods across three real-world signed bipartite networks. To better understand our methods and the contribution of balance theory, we also follow-up with a parameter sensitivity analysis for the major parameters of our methods.

4.1 Experimental Settings

Here we discuss the settings used for our experiments on sign prediction in signed bipartite networks. As previously discussed in Section 2.1 we have collected three signed bipartite networks for this study, namely, Bonanza, U.S. Senate, and U.S. House. For our sign prediction experiments we have randomly selected 10% of the links as test, utilized a random 5% for validation purposes of tuning the hyperparameters of our models, and the remaining 85% as training for each of our datasets. More specifically, each method is only given access to the signed bipartite network induced from the training links, then, for each edge in the testing set, we compare the ground truth link sign with the link sign the specific method suggests for that undirected pair. For evaluation we use both F1 and Area Under the receiver operating characteristic Curve (AUC), since the positive and negative links are unbalanced especially in the Bonanza dataset. To the best of our knowledge this is the first study of predicting link signs in signed bipartite networks; hence other existing methods either for unipartite signed or unsigned bipartite networks are likely not applicable. The main investigation is two-fold. First, we want to test the applicability of balance theory (based on signed butterflies) to aid in sign prediction. Second, we want to provide insights to guide practical usage of sign predictors with different types of signed bipartite networks. Thus, we only provide a comparison against the methods we have presented in this work. Our code and data are available at https://github.com/DSE-MSU/signed-bipartite-networks.

Table 4: Link Sign Prediction Results in terms of (AUC,F1).

| Sign Prediction Method	Bonanza	U.S. Senate	U.S. House
SCd	(0.553 , 0.959)	(0.638 , 0.654)	(0.625 , 0.635)
SCsc	(0.664 , 0.674)	(0.812 , 0.823)	(0.827 , 0.837)
MF	(0.593 , 0.903)	(0.792 , 0.812)	(0.831 , 0.846)
MFwBT	(0.608 , 0.905)	(0.814 , 0.827)	(0.834 , 0.848)
LazyRW	(0.547 , 0.979)	(0.808 , 0.821)	(0.815 , 0.827)
SBRW	(0.582 , 0.949)	(0.836 , 0.849)	(0.846 , 0.858)

4.2 Comparison Results

The results across our three signed bipartite networks in terms of AUC and F1 can be found in Table 4 and the first observation we make is that there is not one proposed method that outperforms the others across all the datasets.

The second observation we make is that the three methods SCsc, MFwBT, and SBRW, which receive aid in prediction from balance theory when defined using signed butterflies, always perform better than their respective baseline method (i.e., SCd, MF, LazyRW) that only use generic signed network information in terms of AUC and only in two cases the F1 is worse. In the Bonanza dataset we have the SCd and LazyRW outperforming SCsc and SBRW, respectively, in terms of F1 (although performing worse in AUC). The reason for this is the heavy imbalance between the positive and negative links in this dataset, more specifically, almost 98% of the links are positive, which is generally a setting where the AUC measurement is preferred to understand the performance better. Therefore we can see that to better detect the few negative links comes at the sacrifice of misclassifying some of the positive links, which is why the F1 of SCsc and SBRW is less than SCd and LazyRW, but comes with a significant increase in AUC. In general we observe that in fact the usage of signed butterflies for sign prediction in signed bipartite networks provides a very significant improvement in almost all cases. This fact suggests that we can give a positive answer to our first question – the usage of balance theory in the form of signed butterflies for sign prediction in signed bipartite networks indeed provides an empirically verifiable improvement.

In the U.S. Senate and U.S. House datasets, for the methods constructed based on intuitions of how to correctly ensure more balanced signed butterflies are being created when predicting missing link signs (i.e, SCsc, MFwBT, and SBRW), we see the low-rank model outperforms the the supervised classifier approach, while the random walk method performs the best (for both AUC and F1).

However, unlike the two U.S. Congress datasets, in the Bonanza dataset we actually observe the complete opposite behavior (in terms of AUC) for the ranking of methods that utilize the signed butterfly based balance theory. We hypothesize this is due to the heavy class imbalance between the positive and negative links. With this imbalance the SBRW method might be unable to directly handle this setting as the parameters only focus on separating real/implicit and balance/unbalance through ω and δ_p/δ_n. Futhermore, if most negative links are involved in balance relationships then actually

(a) MFwBT (AUC)   (b) MFwBT (F1)
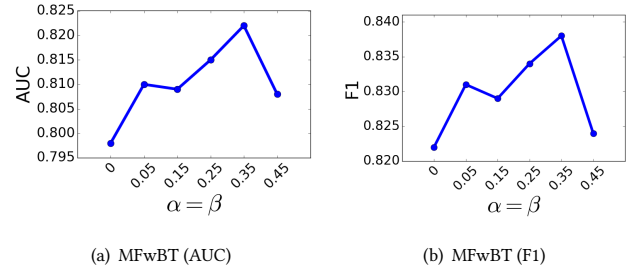
Figure 4: Parameter Sensitivity on α and β in MFwBT on the U.S. Senate dataset.

this would cause even more positive links to be constructed in the two one-mode projection matrices (since two negatives would result in a positive link being created). In comparison, MFwBT is able to more accurately control the ratio of positive to negative implicit links being used in the training procedure (through selecting the size of both $\hat{\mathcal{E}}_i^+$ and $\hat{\mathcal{E}}_i^-$) when extracting them from investigating which links would cause the most signed caterpillars to turn into balance signed butterflies. Also, we note that in our study we fixed $\alpha = \beta$, but this mechanism would further allow MFwBT to balance the contribution of implicit positive and negative links towards learning the most effective representations. Finally, although we see a drastic improvement in terms of AUC for the SCsc method, we also observe this comes at great cost to the F1 measure, and thus this method is just discovering a trade-off of predicting more negative links. This is because we have tuned our logistic regression model to use weights on each training example inversely proportional to the frequency of that link type.

4.3 Parameter Analysis

Among our three proposed sign prediction methods, the low-rank modeling with balance theory (MFwBT) and random walk (SBRW) methods contain interesting hyperparameters from the perspective of wanting to further understand balance theory in signed bipartite networks.

In our MFwBT method, we discussed that we can control the number of suggested implicit positive and negative links from signed buttefies being included in $\mathcal{E}_i^+$ and $\mathcal{E}_i^-$, respectively. We performed a grid search for both the size of $\mathcal{E}_i^+$ and $\mathcal{E}_i^-$, in the set $\{0, 1000, 10000\}$. We discovered that the best setting when considering across the three datasets was having $|\mathcal{E}_i^+| = 1000$ and $|\mathcal{E}_i^-| = 10000$, which we suspect is due to the class imbalance and having more explicit positive links than negative links. Furthermore, the values of α and β were used to control the contribution of training on both positive and negative links suggested based on signed butterflies (i.e., links in $\mathcal{E}_i^+$ and $\mathcal{E}_i^-$), respectively. For simplicity of our analysis we set $\alpha = \beta$ and report the performance on our validation set for the U.S. Senate dataset in Figure 4. We observe that updating the node representations using suggested signed links (that were selected since they would close the most signed caterpillars into balanced signed butterflies) provides an improvement over not taking balance theory into account (which is when $\alpha = \beta = 0$), but care should be taken to not put too much focus on these implicit links. We observe similar findings in our other datasets.

(a) SBRW (AUC) (b) SBRW (F1)
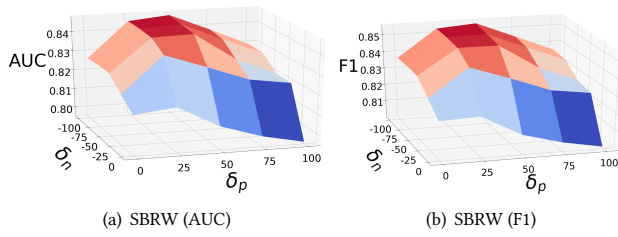
Figure 5: Parameter Sensitivity on δ_p and δ_n in SBRW on the U.S. House dataset.

For our SBRW method, there are two main sets of parameters ω and the threshold pair δ_p with δ_n. We varied ω at a large granularity in the set $\{1, 2, 3, 5\}$, and observed there was not as much significant difference as found in varying δ_p and δ_n, thus we selected the best on average across the three datasets of $\omega = 2$ and fixed this value to investigate the impact of δ_p and δ_n on the performance for predicting the missing link signs.

In Figure 5, for the U.S. House dataset, we varied both δ_p in the set $\{0,25,50,75,100\}$ and similarly for δ_n in the set $\{0,-25,-50, -75, -100\}$. Note that although we saw similar trends across the three datasets, the specific magnitude of δ_n and δ_p we needed to tune separately for each dataset due to the average magnitude of ns_{ij}^A and ns_{ij}^D (i.e., number of common sellers b_i and b_j agree or disagree on, respectively) for constructing $\mathbf{P}_B$ and similarly for $\mathbf{P}_S$, although we fixed δ_p and δ_n for constructing both one-mode projection matrices. We observe in terms of both AUC and F1 from Figure 5 that indeed using these two thresholds to avoid implicit links that do not have a significant amount of information (i.e., low magnitude of $|ns_{ij}^A - ns_{ij}^D|$) provides great improvement to our method. It appears that implicit positive links that have low support are helpful to include. However, it seems better to avoid inferred negative links, which we obtained based on balance theory in the form of signed triangles between two buyers and a seller (similarly for the case of two sellers and a buyer). Although including a few is helpful, ones that have low amount of balance theory support from the network are better left out of the propagation process.

5 RELATED WORK

Our work is mostly related to signed network analysis [26, 28, 39], which has primarily been studied in the unipartite setting, and also to unsigned bipartite methodologies [1, 24, 51].

There have been numerous works that focus on sign prediction under the unipartite setting [7, 13, 19, 27, 35, 49]. In [7] a supervised classifier was presented exploiting balance theory through cycles of length 3 and greater to predict signs. However, unlike our work, they were focused in the directed unipartite signed network setting. In comparison, we perform a comprehensive analysis where we first extend balance theory to signed bipartite networks and then use our proposed signed butterflies for sign prediction in the setting of having two types of nodes where links are only between differing types. Signed network embedding [10, 43, 44, 48] is a related area of research that seeks to learn a representation for each node in a signed network that can then later be used for a plethora of tasks including sign prediction, node classification, and visualization.

However, they are not specifically designed for bipartite networks and those that utilize balance theory primarily harnessed the social theory by leveraging triangles, which would not be applicable here. Another related domain is that of applying signed network analysis to political networks, such as congressional vote analysis [11, 21, 29].

Although here we perform sign prediction, our work is closely related to link prediction in unsigned bipartite networks [3, 24, 25, 30]. One such approach is a local method similar to common-neighbors, but given for bipartite networks in [8]. In [25] numerous methods are provided for predicting links in bipartite networks, such as graph kernels (e.g., Exponential kernel [22]) and a preferential attachment based approach to predict missing links. In [16] they use the one-mode projection matrix to construct a candidate set and only predict links if nodes are among the discovered set of similar nodes. Most recently, a bipartite network embedding algorithm was presented in [15] that achieves state-of-the-art performance on predicting links in unsigned bipartite networks, by using random walks to extract implicit links to boost performance beyond only using explicit links, which is somewhat similar to the ideas of MFwBT.

6 CONCLUSION

In conclusion, signed bipartite networks are a specific type of networks that have become increasingly ubiquitous, but yet fall between the cracks since their added complexities coming from both the negative links and bipartite setting have left both methods and theories lacking the capability to correctly handle them. Meanwhile, balance theory, a key signed social theory, has shown to provide vast improvements in modeling, measuring, and mining tasks related to signed networks when utilized in the form of triangles. Thus, we provided an initial investigation of balance theory in signed bipartite networks that: (1) extends the definition in the form of signed butterfly isomorphism classes; (2) validated that indeed balanced signed butterflies are found significantly more often as compared to unbalanced in signed bipartite networks; (3) leveraged balance theory for the construction of multiple sign prediction methods; and (4) performed experiments on three real-world signed bipartite networks to provide insight into both balance theory and sign prediction in signed bipartite networks.

Our future work will first consist of gaining an even better understanding of the dynamics of signed bipartite networks and how social theories such as balance theory affect their construction/evolution. Thereafter, we plan to utilize signed butterflies for other network analysis tasks in the signed bipartite setting such as network embedding [10, 44] and tie strength prediction [46]. We also plan to further pursue the usefulness of the signed bipartite network formulation of the US Congress for in-depth analysis and prediction tasks such as "swing votes", which are votes coming from a representatives that is voting against what their party suggests.

REFERENCES

[1] Sinan G Aksoy, Tamara G Kolda, and Ali Pinar. 2017. Measuring and modeling bipartite graphs with community structure. *Journal of Complex Networks* 5, 4 (2017), 581–603.

[2] Mohammad Al Hasan, Vineet Chaoji, Saeed Salem, and Mohammed Zaki. 2006. Link prediction using supervised learning. In *SDM06: workshop on link analysis, counter-terrorism and security.*

[3] Oussama Allali, Clémence Magnien, and Matthieu Latapy. 2013. Internal link prediction: A new approach for predicting links in bipartite graphs. *Intelligent Data Analysis* 17, 1 (2013), 5–25.

[4] Pranay Anchuri and Malik Magdon-Ismail. 2012. Communities and balance in signed networks: A spectral approach. In *Advances in Social Networks Analysis and Mining (ASONAM), 2012 IEEE/ACM International Conference on.* IEEE, 235–242.

[5] Phillip Bonacich and Paulette Lloyd. 2004. Calculating status with negative relations. *Social Networks* 26, 4 (2004), 331–338.

[6] Dorwin Cartwright and Frank Harary. 1956. Structural balance: a generalization of Heider's theory. *Psychological review* 63, 5 (1956), 277.

[7] Kai-Yang Chiang, Nagarajan Natarajan, Ambuj Tewari, and Inderjit S Dhillon. 2011. Exploiting longer cycles for link prediction in signed networks. In *Proceedings of the 20th ACM international conference on Information and knowledge management.* ACM, 1157–1162.

[8] Simone Daminelli, Josephine Maria Thomas, Claudio Durán, and Carlo Vittorio Cannistraci. 2015. Common neighbours and the local-community-paradigm for topological link prediction in bipartite networks. *New Journal of Physics* 17, 11 (2015), 113037.

[9] Tyler Derr, Charu Aggarwal, and Jiliang Tang. 2018. Signed Network Modeling Based on Structural Balance Theory. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management.* ACM, 557–566.

[10] Tyler Derr, Yao Ma, and Jiliang Tang. 2018. Signed graph convolutional networks. In *2018 IEEE International Conference on Data Mining (ICDM).* IEEE, 929–934.

[11] Tyler Derr and Jiliang Tang. 2018. Congressional Vote Analysis Using Signed Networks. In *2018 IEEE International Conference on Data Mining Workshops (ICDMW).* IEEE, 1501–1502.

[12] Tyler Derr, Chenxing Wang, Suhang Wang, and Jiliang Tang. 2018. Relevance Measurements in Online Signed Social Networks. In *Proceedings of the 14th International Workshop on Mining and Learning with Graphs (MLG).*

[13] Tyler Derr, Zhiwei Wang, and Jiliang Tang. 2018. Opinions Power Opinions: Joint Link and Interaction Polarity Prediction in Signed Networks. In *Advances in Social Networks Analysis and Mining (ASONAM), 2018 IEEE/ACM International Conference on.* IEEE.

[14] Rana Forsati, Iman Barjasteh, Farzan Masrour, Abdol-Hossein Esfahanian, and Hayder Radha. 2015. Pushtrust: An efficient recommendation algorithm by leveraging trust and distrust relations. In *Proceedings of the 9th ACM Conference on Recommender Systems.* ACM, 51–58.

[15] Ming Gao, Leihui Chen, Xiangnan He, and Aoying Zhou. 2018. BiNE: Bipartite Network Embedding. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR '18).* ACM, New York, NY, USA, 715–724.

[16] Man Gao, Ling Chen, Bin Li, Yun Li, Wei Liu, and Yong-cheng Xu. 2017. Projection-based link prediction in a bipartite network. *Information Sciences* 376 (2017), 158–171.

[17] Fritz Heider. 1946. Attitudes and cognitive organization. *The Journal of psychology* 21, 1 (1946), 107–112.

[18] Cho-Jui Hsieh, Kai-Yang Chiang, and Inderjit S Dhillon. 2012. Low rank modeling of signed networks. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining.* ACM, 507–515.

[19] Amin Javari and Mahdi Jalili. 2014. Cluster-Based Collaborative Filtering for Sign Prediction in Social Networks with Positive and Negative Links. *ACM Transactions on Intelligent Systems and Technology (TIST)* 5, 2 (2014), 24.

[20] Jinhong Jung, Woojeong Jin, Lee Sael, and U. Kang. 2016. Personalized Ranking in Signed Networks Using Signed Random Walk with Restart. In *IEEE 16th International Conference on Data Mining, ICDM 2016, December 12-15, 2016, Barcelona, Spain.* 973–978. https://doi.org/10.1109/ICDM.2016.0122

[21] Hamid Karimi, Tyler Derr, Aaron Brookhouse, and Jiliang Tang. 2019. Multi-Factor Congressional Vote Prediction. In *Advances in Social Networks Analysis and Mining (ASONAM), 2019 IEEE/ACM International Conference on.* IEEE.

[22] Risi Imre Kondor and John Lafferty. 2002. Diffusion kernels on graphs and other discrete input spaces. In *Proceedings of the 19th International Conference on Machine Learning.* Morgan Kaufmann, 315–322.

[23] Srijan Kumar, Francesca Spezzano, VS Subrahmanian, and Christos Faloutsos. 2016. Edge Weight Prediction in Weighted Signed Networks.. In *ICDM.* 221–230.

[24] Jérôme Kunegis. 2015. Exploiting the structure of bipartite graphs for algebraic and spectral graph theory applications. *Internet Mathematics* 11, 3 (2015), 201–321.

[25] Jérôme Kunegis, Ernesto W De Luca, and Sahin Albayrak. 2010. The link prediction problem in bipartite networks. In *International Conference on Information*

[26] Jérôme Kunegis, Andreas Lommatzsch, and Christian Bauckhage. 2009. The slashdot zoo: mining a social network with negative edges. In *Proceedings of the 18th international conference on World wide web.* ACM, 741–750.

[27] Jure Leskovec, Daniel Huttenlocher, and Jon Kleinberg. 2010. Predicting positive and negative links in online social networks. In *Proceedings of the 19th international conference on World wide web.* ACM, 641–650.

[28] Jure Leskovec, Daniel Huttenlocher, and Jon Kleinberg. 2010. Signed networks in social media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.* ACM, 1361–1370.

[29] Mario Levorato and Yuri Frota. 2016. Brazilian Congress structural balance analysis. *arXiv preprint arXiv:1609.00767* (2016).

[30] Xin Li and Hsinchun Chen. 2013. Recommendation as link prediction in bipartite graphs: A graph kernel-based machine learning approach. *Decision Support Systems* 54, 2 (2013), 880–890.

[31] László Lovász et al. 1993. Random walks on graphs: A survey. *Combinatorics, Paul erdos is eighty* 2, 1 (1993), 1–46.

[32] Zhengdong Lu, Berkant Savas, Wei Tang, and Inderjit S Dhillon. 2010. Supervised link prediction using multiple sources. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on.* IEEE, 923–928.

[33] Mark Ludwig and Peter Abell. 2007. An evolutionary model of social networks. *European Physical Journal B–Condensed Matter* 58, 1 (2007).

[34] Aditya Krishna Menon and Charles Elkan. 2011. Link prediction via matrix factorization. In *Joint european conference on machine learning and knowledge discovery in databases.* Springer, 437–452.

[35] Athanasios Papaoikonomou, Magdalini Kardara, Konstantinos Tserpes, and Dora Varvarigou. 2014. Edge Sign Prediction in Social Networks via Frequent Subgraph Discovery. *IEEE Internet Computing* (2014).

[36] Seyed-Vahid Sanei-Mehri, Ahmet Erdem Sariyuce, and Srikanta Tirthapura. 2018. Butterfly Counting in Bipartite Networks. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining.* ACM, 2150–2159.

[37] Moshen Shahriari and Mahdi Jalili. 2014. Ranking nodes in signed social networks. *Social Network Analysis and Mining* 4, 1 (2014), 172.

[38] Jiliang Tang, Charu Aggarwal, and Huan Liu. 2016. Recommendations in signed social networks. In *Proceedings of the 25th International Conference on World Wide Web.* International World Wide Web Conferences Steering Committee, 31–40.

[39] Jiliang Tang, Yi Chang, Charu Aggarwal, and Huan Liu. 2016. A survey of signed network mining in social media. *ACM Computing Surveys (CSUR)* 49, 3 (2016), 42.

[40] Hanghang Tong, Christos Faloutsos, and Jia-yu Pan. 2006. Fast Random Walk with Restart and Its Applications. In *Data Mining, 2006. ICDM'06. Sixth International Conference on.* IEEE, 613–622.

[41] Vincent Traag, Yurii Nesterov, and Paul Van Dooren. 2010. Exponential Ranking: Taking into Account Negative Links. *Social Informatics* (2010), 192–202.

[42] Vida Vukašinović, Jurij Šilc, and Risth Škrekovski. 2014. Modeling acquaintance networks based on balance theory. *International Journal of Applied Mathematics and Computer Science* 24, 3 (2014), 683–696.

[43] Hongwei Wang, Fuzheng Zhang, Min Hou, Xing Xie, Minyi Guo, and Qi Liu. 2018. SHINE: signed heterogeneous information network embedding for sentiment link prediction. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining.* ACM, 592–600.

[44] Suhang Wang, Jiliang Tang, Charu Aggarwal, Yi Chang, and Huan Liu. 2017. Signed network embedding in social media. In *Proceedings of the 2017 SIAM International Conference on Data Mining.* SIAM, 327–335.

[45] Zhaoming Wu, Charu C Aggarwal, and Jimeng Sun. 2016. The troll-trust model for ranking in signed networks. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining.* ACM, 447–456.

[46] Rongjing Xiang, Jennifer Neville, and Monica Rogati. 2010. Modeling relationship strength in online social networks. In *Proceedings of the 19th international conference on World wide web.* ACM, 981–990.

[47] Muhammed A Yildirim and Michele Coscia. 2014. Using random walks to generate associations between objects. *PloS one* 9, 8 (2014), e104813.

[48] Shuhan Yuan, Xintao Wu, and Yang Xiang. 2017. SNE: signed network embedding. In *Pacific-Asia conference on knowledge discovery and data mining.* Springer, 183–195.

[49] Tongda Zhang, Haomiao Jiang, Zhouxiao Bao, and Yingfeng Zhang. 2013. Characterization and edge sign prediction in signed networks. *Journal of Industrial and Intelligent Information Vol* 1, 1 (2013).

[50] Tao Zhou, Jie Ren, Matúš Medo, and Yi-Cheng Zhang. 2007. Bipartite network projection and personal recommendation. *Physical Review E* 76, 4 (2007), 046115.

[51] Katharina Anna Zweig and Michael Kaufmann. 2011. A systematic approach to the one-mode projection of bipartite graphs. *Social Network Analysis and Mining* 1, 3 (2011), 187–218.

Processing and Management of Uncertainty in Knowledge-based Systems. Springer, 380–389.