Exploiting Longer Cycles for Link Prediction in Signed Networks

Kai-Yang Chiang*
UT Austin
kychiang@cs.utexas.edu

Nagarajan Natarajan*
UT Austin
naga86@cs.utexas.edu

Ambuj Tewari*
UT Austin
ambuj@cs.utexas.edu

Inderjit S. Dhillon
UT Austin
inderjit@cs.utexas.edu

ABSTRACT

We consider the problem of link prediction in signed networks. Such networks arise on the web in a variety of ways when users can implicitly or explicitly tag their relationship with other users as positive or negative. The signed links thus created reflect social attitudes of the users towards each other in terms of friendship or trust. Our first contribution is to show how any quantitative measure of social imbalance in a network can be used to derive a link prediction algorithm. Our framework allows us to reinterpret some existing algorithms as well as derive new ones. Second, we extend the approach of [6], by presenting a supervised machine learning based link prediction method that uses features derived from longer cycles in the network. The supervised method outperforms all previous approaches on 3 networks drawn from sources such as Epinions, Slashdot and Wikipedia. The supervised approach easily scales to these networks, the largest of which has 132k nodes and 841k edges. Most real-world networks have an overwhelmingly large proportion of positive edges and it is therefore easy to get a high overall accuracy at the cost of a high false positive rate. We see that our supervised method not only achieves good accuracy for sign prediction but is also especially effective in lowering the false positive rate.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous ; J.4 [**Computer Applications**]: Social and Behavioral Sciences—*Sociology*

General Terms

Algorithms, Experimentation

*Equal contribution to the work.

CIKM'11, October 24–28, 2011, Glasgow, Scotland, UK.

Keywords

Link & Graph Mining, Web & Social Knowledge Management

1. INTRODUCTION

Many research problems in the analysis of networks, social or otherwise, have been tackled by modeling the presence or absence of relationships between pairs of entities using a graph where a relationship is simply indicated by a positively weighted edge. Recently, real-world networks have appeared that do not fit this simple description. In e-commerce sites such as Ebay, users of the network develop *trust* and *distrust* in *agents* in the network; websites such as Amazon let members express their likes and dislikes toward the purchased products; online review and news websites such as Epinions and Slashdot allow readers to approve or denounce reviews and articles; certain nodes in a computer network may be detected as *byzantine* by a sub-network, and communication through a byzantine node is considered unreliable. In all such cases, it is helpful to think of the edges between entities as being *signed* either positively or negatively.

The study of signed networks dates back to the early 1950s [1] where dislikes and distrusts were modeled as negative weight edges in a graph. A formal model encompassing different types of interpersonal relationships was proposed and notion of *balance* in signed graphs was defined.

We note that the nature and complexity of many graph problems change once negative edges are introduced. For example, the shortest-path problem in the presence of cycles with negative edges is known to be NP-hard. Consequently, solutions to problems in unsigned social networks are usually not immediately applicable to signed networks. In particular, signed link prediction has connections to social balance theory while no such connection exists for the standard (unsigned) link prediction task.

The proliferation of online *signed* networks, and the recent advances in social network analysis, including the problem of link prediction in particular, naturally led researchers to the problem of predicting the *sign* of a link in signed networks[3, 6]. We follow this line of work and consider the problem of predicting signs in a signed network. Our proposed methods for the problem are motivated by the *general theory of social balance*[1]. We build on the work of Leskovec et al.[6], driven by the realization that higher order cycles in a signed graph yield a "measure of imbalance" suggested by the general theory of social balance. We show how these measures

can be successfully exploited for sign prediction. We also go beyond these simple measures of imbalance and propose a supervised machine learning approach that achieves state-of-the-art performance on datasets drawn from sources as diverse as Epinions, Slashdot and Wikipedia. Our largest network has more than 100k nodes and 800k edges. Our main contributions can be summarized as follows. First, we show how any quantitative measure of imbalance in a network can be used to derive a link prediction algorithm. We discuss different measures of imbalance that are derived from social balance theory and signed graph theory. In particular, we show that the *Katz* measure on a signed graph is quite naturally connected to a measure of imbalance. Second, we show that using measures of imbalance that depend on *higher order cycles* can improve the quality of sign prediction. The effect is more pronounced in the case of edges with zero *embeddedness*. The nodes of such edges, by definition, do not share any common neighbors. Third, our supervised learning approach to sign prediction uses features derived from longer cycles and achieves state-of-the-art performance. In particular, our method has smaller *false positive rate*, compared to the methods in [6]. Moreover, the performance improvement is consistent across diverse networks (in the sense of formation and composition of signed edges) such as Epinions, Slashdot and Wikipedia.

Our experiments with real-world networks show that not all relationships formed in a social network conform to the intuitions underlying social balance theory. Indeed, we find that real-world networks are too complex to be described by a simple formalism. Given the complexity of these networks, longer cycles seem to contain in them more useful information for predicting signs of links.

2. RELATED WORK

Signed networks have received attention in the last decade in the context of clustering and link prediction. The problem of predicting edge signs in a social network was first considered by Guha et al.[3], albeit in a slightly different setting. They develop a trust propagation framework to predict the trust (or distrust) between pairs of nodes. First, a combined matrix is derived from the adjacency matrix which captures all one-step propagations (corresponding to cases $k = 2$ and $k = 3$ in our k-cycle method). The propagated trusts and distrusts are then computed as a linear combination of powers of the combined matrix. They apply variants of their method on the Epinions network and find that higher iterations of propagation tend to have a beneficial effect on the accuracy of prediction.

Kunegis et al.[5] study the spectral properties of signed networks and use kernels derived from a signed variant of the graph Laplacian for link prediction. They also consider power sum of the adjacency matrix (up to degree 4) for link prediction. However, they do not propose any supervised learning approach that learns the coefficients in their models.

Leskovec et al.[6] first considered an explicit formulation of the sign prediction problem. Their prediction methods are based on the theory of social balance and status[7]. The idea is that the sign of an edge (i, j) should minimize the number of *unbalanced triangles* involving the edge (i, j), i.e. triangles with an odd number of negative edges (see Section 4). They also propose a supervised machine-learning formulation of the problem, and show that the features derived from different types of triangles, combined with first-order features of a node like the number of incoming positive edges, make better predictors of the edge sign. The prediction model is also shown to generalize fairly well across different online social networks. Their model, however, does not look beyond such local structures as triangles. These local features lead to impressive accuracies in the networks that they consider. Nevertheless, it is natural to ask if the acccuracy can be further boosted by considering higher order features.

3. PRELIMINARIES

For us, a graph will mean a signed graph unless otherwise stated. This formally means that $G = (V, E, \Sigma)$ where $V = \{1, 2, \ldots, |V|\}$ is a finite set of *nodes* or *vertices* of the graph. The set E consists of *edges* of the form $\{i, j\}$ or (i, j), for $i, j \in V$, depending on whether the graph is *undirected* or *directed* (we assume there are no self-loops). The third component of a graph is a mapping $\Sigma : E \to \{+1, -1\}$ giving a *sign* to each edge. We will assume that our graphs are connected (weakly connected if directed). We will use the terms 'network' and 'graph' interchangeably. An undirected graph has an associated *adjacency matrix* $A \in \{-1, 0, +1\}^{|V|\times|V|}$. For undirected graphs, A is symmetric, i.e. $A = A^T$, while for directed graphs it will not be symmetric in general. It will also be convenient to define the positive part A^+ and negative part A^- as: $A^+ := \max(A, 0)$, $A^- := \min(A, 0)$ where max/min are applied entry-wise. With these definitions, we have $A = A^+ + A^-$.

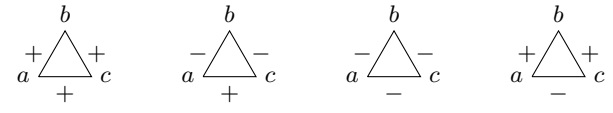
A *path* (of length or order k) in an undirected graph is a sequence $i_1, i_2, \ldots, i_{k+1}$ of vertices such that $\{i_j, i_{j+1}\}$ is an edge for $1 \le j \le k$. A *simple path* is a path with no repeated vertices. A *cycle* is a path with $i_1 = i_{k+1}$. A *simple cycle* is a cycle with no repeated vertices (except the first one). Similar definitions yield paths, simple paths, cycle and simple cycles in directed graphs. Denote the set of all order k cycles and simple cycles of a graph G by $C_k(G)$ and $SC_k(G)$ respectively. The *indicator* $\mathbf{1}[P]$ is 1 if predicate P is true and 0 otherwise. We let power take precedence over subscript, i.e. $A^k_{i,j}$ denotes $(A^k)_{i,j}$.

3.1 Problem set-up

We adopt the framework [3, 6] of predicting the sign of a single edge that has been suppressed, using the rest of the network. Formally, given a graph $G = (V, E, \Sigma)$, and a test edge $e_{\text{test}} \in E$, we want to predict $\Sigma(e_{\text{test}})$, using only the edges in $E - \{e_{\text{test}}\}$. When using a supervised machine learning approach it is convenient to think of $\{e' \in E - \{e_{\text{test}}\} : \Sigma(e') = +1\}$ and $\{e' \in E - \{e_{\text{test}}\} : \Sigma(e') = -1\}$ as the set of positive and negative examples respectively.

3.2 Basics of Social Balance Theory

We briefly review the rudiments of the theory of social balance. This theory rests on the premise that certain configurations of positive ('is a friend of' or 'trusts') and negative ('is an enemy of' or 'distrusts') edges between individuals are socially more plausible than others. A more detailed treatment can be found in [2, Chapter 5]. For example, in the case of three individuals a, b and c, the left two configurations below are more likely than the right two:



on the basis of sayings such as 'a friend of a friend is a friend' and 'an enemy of an enemy is a friend'. Accordingly, the first two triangles are called *balanced* while the latter two are said to be *unbalanced*. Formally, a complete undirected signed graph is called *balanced* iff all triangles in it are balanced. Of course, real world networks are not complete by any means. Hence, the notion of balance can be extended by using the idea of filling in missing entries. An undirected signed graph is called *balanced* iff it is possible to add signed edges to make it a balanced and complete graph. It turns out that this 'local' definition based on looking at triangles only is equivalent to a 'global' definition where we say that a graph is balanced if and only if its vertices can be divided into two mutually exclusive and exhaustive sets X and Y (with one of them possibly empty) such that all edges within X and within Y are positive while all edges with one end in X and the other in Y are negative.

The following theorem relates balance to the existence of simple cycles with an odd number of negative edges.

THEOREM 1. ([4, 1]) *A signed graph is balanced iff there are no simple cycles with odd number of negative edges.*

Motivated by this theorem, we call a cycle (simple or otherwise) *balanced* if it has an even number of negative edges and *unbalanced* otherwise.

4. METHODS BASED ON MEASURES OF SOCIAL IMBALANCE

The main idea developed in this section is that any quantitative measure of social imbalance in a graph can be used to design a link prediction algorithm. The sign prediction for a given edge is the one that minimizes the social imbalance in the resulting graph (once the signed edge has been added to the network).

In a complete graph, perfect balance, by definition, implies the absence of any unbalanced triangles. This motivates a simple measure of imbalance, namely the total number of unbalanced triangles in a graph. Thus, we define,

$$\mu_{\text{tri}}(G) := \sum_{\tilde{\sigma} \in SC_3(G)} \mathbf{1}\left[\tilde{\sigma} \text{ is unbalanced}\right] . \qquad (1)$$

A definition essentially similar to the one above appears in the recent work of van de Rijt [10, p. 103] who observes that the equivalence between $\mu_{\text{tri}}(G) = 0$ and G being balanced holds only for complete graphs.

For an incomplete graph, imbalance might manifest itself only if we look at longer simple cycles. Accordingly, we define a higher-order analogue of (1),

$$\mu_k^s(G) := \sum_{i=3}^{k} \beta_i \sum_{\tilde{\sigma} \in SC_i(G)} \mathbf{1}\left[\tilde{\sigma} \text{ is unbalanced}\right] . \qquad (2)$$

where $k \geq 3$ and β_i's are coefficients weighting the relative contributions of unbalanced simple cycles of different lengths. If we choose a decaying choice of β_i, like $\beta_i = \beta^i$ for some $\beta \in (0,1)$, then we can even define an infinite-order version $\mu_\infty^s(G)$ by setting $k = \infty$ above. It is clear that $\mu_\infty(\cdot)$ is a genuine measure of imbalance in the sense formalized by the following theorem which follows directly from Theorem 1).

THEOREM 2. *Fix a (possibly incomplete) graph G. Let $\beta_i > 0$ be any sequence such that $\mu_\infty^s(G)$ is well-defined. Then, $\mu_\infty(G) > 0$ iff G is unbalanced.*

This suggests that we could use $\mu_\infty(\cdot)$ as a measure of imbalance to derive link prediction algorithms. However, enumerating simple cycles of a graph is a hard problem. In particular, if we could count simple cycles of length n in a graph with n vertices in polynomial time, we would solve the NP-complete Hamiltonian cycle problem. To get around this computational issue, we slightly change the definition of $\mu_k(\cdot)$ to the following.

$$\mu_k(G) := \sum_{i=3}^{k} \beta_i \sum_{\sigma \in C_i(G)} \mathbf{1}\left[\sigma \text{ is unbalanced}\right] . \qquad (3)$$

As before, we allow $k = \infty$ provided the β_i's decay sufficiently rapidly. The only difference between these definitions and the previous one is that here we sum over *all* cycles, not just simple ones. However, we still get a valid notion of imbalance as stated by the following result (proof will be given in a longer version of the paper).

THEOREM 3. *Fix a (possibly incomplete) graph G. Let $\beta_i > 0$ be any sequence such that $\mu_\infty(G)$ is well-defined. Then, $\mu_\infty(G) > 0$ iff G is unbalanced.*

The basic idea of using a measure of imbalance for predicting the sign of a given query link i, j, such that $i \neq j$ and $\{i, j\} \notin E$ is as follows. Given a graph G and query $\{i, j\}$ for $i, j \in V, i \neq j$, we construct two graphs: $G^{+(i,j)}$ and $G^{-(i,j)}$. These are obtained from G by augmenting its edge-set with $\{i, j\}$ and attaching a $+1$ and -1 sign to it respectively. Given a measure of imbalance, $\mu(\cdot)$, the predicted sign of $\{i, j\}$ is then simply:

$$\text{sign}\left(\mu\left(G^{-(i,j)}\right) - \mu\left(G^{+(i,j)}\right)\right) . \qquad (4)$$

Note that, to be able to this quickly, we should use a $\mu(\cdot)$ for which the quantity (4) is efficiently computable. We now consider the measures mentioned in the previous subsection to ensure that this is indeed the case for them.

Somewhat surprisingly, for $\mu(\cdot) = \mu_3(\cdot)$, the prediction (4) simply amounts to computing the (i, j) entry in the matrix A^2 where A is the (signed) adjacency matrix of G. In fact, a more general result is true (proof will be given in a longer version of the paper).

THEOREM 4. *Let $G = (V, E, \Sigma)$ be an undirected signed graph and let $i \neq j$ be such that $\{i, j\} \notin E$. Let $G^{+(i,j)}$ and $G^{-(i,j)}$ be the augmented graphs as defined above. Then, for any $k \geq 2$,*

$$\sum_{\sigma \in C_k\left(G^{-(i,j)}\right)} \mathbf{1}\left[\sigma\right] - \sum_{\sigma \in C_k\left(G^{+(i,j)}\right)} \mathbf{1}\left[\sigma\right] = A_{i,j}^{k-1}$$

where $A \in \{-1, 0, +1\}^{|V| \times |V|}$ is the adjacency matrix of G.

Using Theorem 4, it is easy to see that

$$\text{sign}\left(\mu_k\left(G^{-(i,j)}\right) - \mu_k\left(G^{+(i,j)}\right)\right) = \text{sign}\left(\sum_{t=3}^{k} \beta_t A_{i,j}^{t-1}\right)$$

and that the above is true even for $k = \infty$. In the special case $\beta_k = \beta^{k-1}$ with $\beta < 1/\|A\|_2$, we can sum the above infinite series to get the Katz prediction rule for edge sign prediction: $\text{sign}\left(\left((I - \beta A)^{-1} - I - \beta A\right)_{i,j}\right)$. Katz has been successfully used as a link prediction method for *unsigned* networks [8] but here we see it reappearing for link

prediction in *signed* networks from a social balance point of view. We find this connection between Katz and social balance intriguing and believe, to the best of our knowledge, that it has not been made before.

5. SUPERVISED METHOD BASED ON LONGER CYCLES

The methods derived from the measures of imbalance in the previous section rely on social balance theory for link prediction in signed networks. However, real world networks may not conform to the prediction of social balance theory or may do so only to a certain extent. To deal with this situation, we use measures such as Katz to derive *features* that can then be fed to a supervised machine learning algorithm along with the signs of the known edges in the network. We draw upon research in unsigned link prediction where the Katz measure has been empirically demonstrated to produce competitive results [8]. However, recent research [9] shows that learning the weights β_i based on supervised machine learning approaches tends to increase link prediction accuracy. It is thus natural to expect that the relative weights for cycles of various lengths may be better estimated by taking into account the evidence in the training data corresponding to the *given network*.

We pose the problem of predicting the sign of an observed link as a standard (binary) classification problem in machine learning, using positive and negative examples. Our training set consists of pairs $(e, \Sigma(e))$ where e ranges over edges whose signs are given to us and $\Sigma(e)$'s are the given signs. Given the dataset, we wish to learn a classifier that can predict the sign of a given test edge e_{test} that was not part of the training set.

We now extend the supervised learning approach of [6] by introducing features derived from *longer cycles*. In the process, we obtain supervised variants of the cycle-based sign prediction methods introduced in the previous section.

5.1 Features from Longer Cycles

Let us now describe the features of a *directed* edge $e = (i, j)$. Note that, unlike methods presented in Section 4, we here consider *directed* signed networks. Social balance theory has mostly been concerned with undirected network and hence the methods in Section 4 deal with undirected networks only. Here, we are weakening our reliance on social balance theory and can therefore naturally deal with directed graphs as well.

To motivate our longer cycle based features, let us first recall the feature construction used in Leskovec et al [6]. Fix an edge $e = (i, j)$. Consider an arbitary common neighbor (in an undirected sense) k of i and j. The link between i and k can be in 4 possible configurations: $i \xrightarrow{+} k$, $i \xleftarrow{+} k$ $i \xrightarrow{-} k$, or $i \xleftarrow{-} k$. Similarly, there are 4 possible configurations for the link between k and j. Thus, we can get a total of 16 features for the edge e by considering the number of common neighbors k in each of the $4 \times 4 = 16$ configurations.

This corresponds to a supervised variant of k-cycle method for $k = 3$. In terms of matrix powers, these sixteen features are nothing but the (i, j) entry in the sixteen matrices: $(A^{b_1})^{t_1} \cdot (A^{b_2})^{t_3}$ where $b_1, b_2 \in \{\pm\}$ and $t_i \in \{T, 1\}$. A criticism against using only these triangle-based features is that there could be many people in the social network who do not share friends. In fact, this is the case in most of the networks that are used in [6]. The reason their method is able to predict well on such pairs is that they additionally use 7 "degree-type" features like in-degree and out-degree (and their signed variants). Thus, the prediction for edge with zero emdeddedness (embeddedness refers to the number of common neighbors of the vertices of an edge) relies completely on such degree based features. These degree features tend to introduce a bias in learning. For example, a node that is predisposed to make positive relationships, biases the classifier to predict positive relationships.

Table 1: Network Statistics

	Epinions	Slashdot	Wikipedia
No. of nodes	131,828	82,144	7,065
No. of edges	840,799	549,202	103,561
Fraction of + edges	0.8529	0.7740	0.7884
Fraction of − edges	0.1471	0.2260	0.2116
Normalized MOI-3	0.0950	0.1335	0.2165

This criticism thus necessitates incorporating features from higher-order cycles. Generalizing the construction for $k = 3$ case, for the edge (i, j), the features can be obtained as the (i, j) entries in the 4^{k-1} matrices

$$\left(A^{b_1}\right)^{t_1} \cdot \left(A^{b_2}\right)^{t_2} \cdots \left(A^{b_{k-1}}\right)^{t_{k-1}} , \qquad (5)$$

with $b_i \in \{\pm\}, t_i \in \{T, 1\}$.

Note that the number of features is exponential in k, and therefore it is not feasible to obtain features from arbitrarily long cycles. We use supervised higher order cycle (HOC) methods for $k \leq 5$ in the experiments.

The number of features can quickly become unmanageable, and computationally infeasible, as soon as k is beyond 5. While dimensionality of the feature space may be the primary concern, the combinatorial nature of the features also raises the following intuitive concern: the interpretability of features rendered by high-order cycles, say when $k = 6$, composed of different signs and directions, is a challenge. For example, it is intuitively hard to appreciate the difference between two walks $i \xrightarrow{+} k_1 \xrightarrow{+} k_2 \xrightarrow{-} k_3 \xrightarrow{+} k_4 \xrightarrow{+} j$ and $i \xrightarrow{+} k_1 \xrightarrow{+} k_2 \xleftarrow{-} k_3 \xrightarrow{+} k_4 \xrightarrow{+} j$.

With this realization, one way to reduce the number of features yet retain the information in longer cycles, is to consider the underlying undirected graph, ignoring the directions. In particular, the kth order features will be from the matrices $A^{b_1} \cdot A^{b_2} \ldots \cdot A^{b_{k-1}}$ with $b_i \in \{\pm\}$. Since we are considering the undirected graph, we ensure that the features are symmetric by summing features of the form $A^{b_1}A^{b_2}$ and $A^{b_2}A^{b_1}$. Thus the number of k-th order features to compute is reduced to $O(2^k)$ from $O(4^k)$. Though the number of features is still exponential in k, the construction of features becomes much easier for small values of k.

We use a simple logistic regression where the imbalance of an edge is modeled as a linear combination of the features, which are imbalances in cycles of various lengths and characteristics themselves. Let $\Phi : V \times V \to \mathbb{R}^p$ denote the feature map. We have, $P(\Sigma(u, v) = +1) = 1/(1 + \exp\left(-w_0 - \sum_{i=1}^{p} w_i \Phi_i(u, v)\right))$. The prediction for edge (u, v) is given by $\text{sign}(\langle w, \Phi(u, v)\rangle)$.

6. EXPERIMENTS

We consider three online social networks — Epinions, Slashdot and Wikipedia[6] (downloaded from `snap.stanford.`

Table 2: Accuracy of HOC Methods

	Epinions	Slashdot	Wikipedia
HOC-3	0.9014	0.8303	0.8424
HOC-5	0.9080	0.8469	0.8605

Table 3: False Positive Rate of HOC Methods

	Epinions	Slashdot	Wikipedia
HOC-3	0.4756	0.5575	0.5488
HOC-5	0.4441	0.5070	0.4817

edu). All the networks have explicit sign labels on the links. Refer to Table 1 for the statistics of the networks. Note that MOI-3 is normalized by the total number of triangles. Refer to [6] for description of the networks. We have 2 families of methods: one based on measures of imbalance (MOI) from Section 4 and the other based on the supervised machine learning approach involving higher order cycles (HOC) described in Section 5. Both families depend on a parameter $k \geq 3$ that denotes the order of the cycles that the method is based on. For MOI, we consider k up to 10 and for HOC we consider $k = 3, 4, 5$. Note that the set of features used by HOC-$(k+1)$ is a strict superset of the features used by HOC-k. We also remind the reader that MOI-3 and HOC-3 are the methods considered in [6].

We evaluate and compare MOI methods using a *leave-one-out* type methodology: each edge in the network is successively removed and the method tries to predict the sign of that edge using the rest of the network. For HOC methods, we resort to 10-*fold cross-validation*. We (randomly) created 10 disjoint test folds each consisting of 10% of the total number of edges in the network. For each test fold, the remaining 90% of the edges serve as the training set. For a given test fold, the feature extraction and logistic model training happens on a graph with the test edges removed. We report accuracies and false-positive rates by averaging them over the 10 folds.

6.1 Results

Our experiments on the three online social networks show that higher order cycles benefit the accuracy of sign prediction and lower the false positive rate. Furthermore, the results are consistent across the three diverse networks. Figure 2 shows the accuracy of MOI based methods. Note that the accuracy is shown for edges with embeddedness under certain threshold. Firstly, we see that accuracy is non-decreasing in embeddedness threshold. Next, it is clear that higher-order methods perform significantly better than MOI-3 (triangles) method. Finally, the performance boost is large for edges with low embeddedness. This is expected as edges of low embeddedness by definition do not have many common neighbors for their end-points, and higher-order cycles have relatively better information for such edges than others. We also observe from our experiments that beyond $k = 5$, the performance gain is not very significant.

Figure 1 shows the distribution of edge embeddedness in the data sets. Observe that a significant fraction of the edges have low embeddedness in all the networks. Thus, for a good fraction of edges, we observe a large increase in accuracy of higher-order MOI based methods, in all the data sets.

The results for the supervised HOC methods are shown in Tables 2 & 3 and Figure 3. In all the data sets, there is a small improvement in accuracy by using higher order cycles (HOC-5), as shown in Table 2. The false positive rate, however, reveals a more interesting phenomenon in Table 3. Indeed, higher order methods (such as HOC-5) significantly reduce the false positive rate as compared to that of HOC-3. However Figure 3 shows that, unlike MOI based methods, edge embeddedness does not seem to affect the decrease in false positive rate for HOC methods. We see this trend across all the data sets.

7. CONCLUSION

We see that longer cycles significantly benefit sign prediction, and do so consistently across many real-world networks. We presented a framework to obtain a link prediction algorithm, using any quantitative measure of imbalance. Higher order cycles came as a natural generalization of local triangles, and furthermore, the generalization is well-founded by the general theory of social balance. Finally, we observe that the edges appearing in real-world signed networks do not necessarily conform to the intuitions underlying the social balance theory, and longer cycles contain more useful information for predicting edge signs.

8. ACKNOWLEDGMENTS

This research was supported by NSF grant CCF-0916309 and DOD Army grant W911NF-10-1-0529.

9. REFERENCES

[1] D. Cartwright and F. Harary. Structure balance: A generalization of Heider's theory. *Psychological Review*, 63(5):277–293, 1956.

[2] D. Easley and J. Kleinberg. *Networks, Crowds, and Markets*. Cambridge University Press, 2010.

[3] R. Guha, R. Kumar, P. Raghavan, and A. Tomkins. Propagation of trust and distrust. In *WWW*, pages 403–412, 2004.

[4] F. Harary. On the notion of balance of a signed graph. *Michigan Mathematical Journal*, 2(2):143–146, 1953.

[5] J. Kunegis, S. Schmidt, A. Lommatzsch, J. Lerner, E. W. D. Luca, and S. Albayrak. Spectral analysis of signed graphs for clustering, prediction and visualization. In *SDM*, pages 559–570, 2010.

[6] J. Leskovec, D. Huttenlocher, and J. Kleinberg. Predicting positive and negative links in online social networks. In *WWW*, pages 641–650, 2010.

[7] J. Leskovec, D. Huttenlocher, and J. Kleinberg. Signed networks in social media. In *CHI*, pages 1361–1370, 2010.

[8] D. Liben-Nowell and J. Kleinberg. The link-prediction problem for social networks. *Journal of the American Society for Information Science and Technology*, 58(7):1019–1031, 2007.

[9] Z. Lu, B. Savas, W. Tang, and I. S. Dhillon. Supervised link prediction using multiple sources. In *ICDM*, pages 923–928, 2010.

[10] A. van de Rijt. The micro-macro link for the theory of structural balance. *Journal of Mathematical Sociology*, 35(1):94–113, 2011.

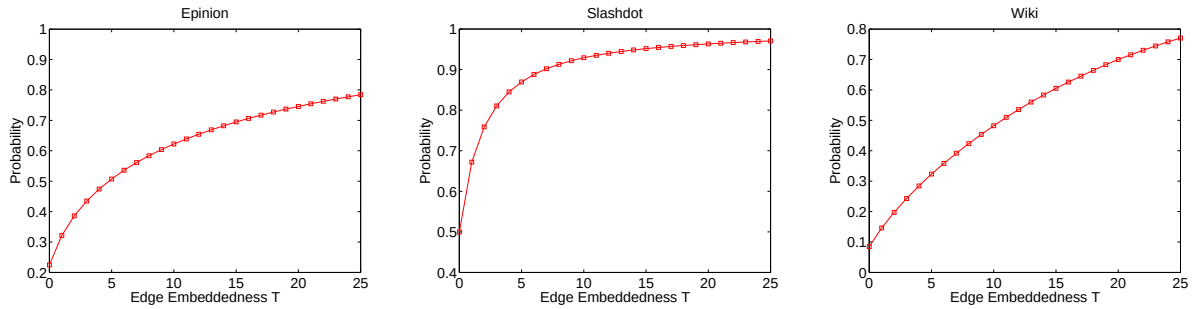


Figure 1: Cumulative Probability Distribution of Edge Embeddedness. These plots show the fraction of edges with embeddedness *no more* than T for various thresholds T. We see that these networks have a significant fraction of low embeddedness edges. For example, the fraction of edges with *zero* embeddedness (edges whose end-points do not share any common neighbor) is about $20\%, 50\%$ and 10% for Epinions, Slashdot and Wikipedia respectively.

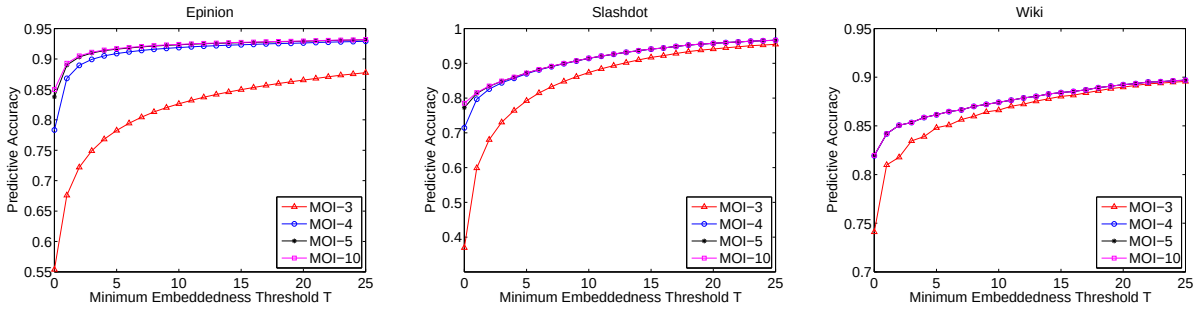


Figure 2: Accuracy of Measures of Imbalance (MOI) Based Methods for $k = 3, 4, 5, 10$. These plots show the accuracy of MOI-k methods for edges with embeddedness *at least* T for various thresholds T. We see that the difference in the performance of MOI-3 and higher order methods is larger when edges with lower embeddedness are considered. We also see that the improvement obtained by going beyond order 5 is not very significant.

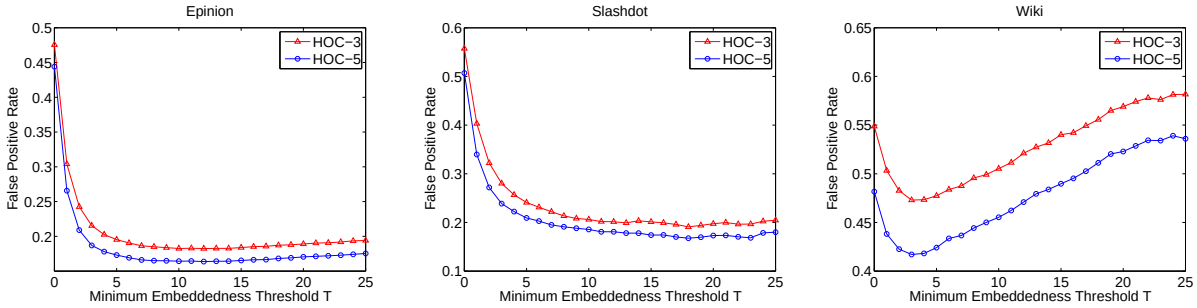


Figure 3: False Positive Rates of Higher Order Cycle (HOC) Methods for $k = 3, 5$. These plots show the false positive rate of HOC-k methods for edges with embeddedness *at least* T for various thresholds T. We see that considering higher order cycles has the benefit of significantly reducing false-positives while simultaneously achieving slightly better overall accuracy (see Table 2). However, unlike what we see for MOI methods, here the improvement does not seem to depend strongly on edge embeddedness. The false positive rates for HOC-4 are very similar to that of HOC-5 and hence are not shown.