MissBiN: Visual Analysis of Missing Links in Bipartite Networks

Jian Zhao∗  Maoyuan Sun†  Francine Chen∗  Patrick Chiu†
∗FX Palo Alto Laboratory  †Northern Illinois University

ABSTRACT
The analysis of bipartite networks is critical in a variety of application domains, such as exploring entity co-occurrences in intelligence analysis and investigating gene expression in bio-informatics. One important task is missing link prediction, which infers the existence of unseen links based on currently observed ones. In this paper, we propose MissBiN that involves analysts in the loop for making sense of link prediction results. MissBiN combines a novel method for link prediction and an interactive visualization for examining and understanding the algorithm outputs. Further, we conducted quantitative experiments to assess the performance of the proposed link prediction algorithm and a case study to evaluate the overall effectiveness of MissBiN.

Keywords: Link prediction, bipartite network, visual analytics.

1 INTRODUCTION
Many real-world systems can be modeled as bipartite networks (i.e., two-mode networks). There are two types of nodes in a network and links only exist between different node types. Bipartite relationship analysis has been applied in a variety of application domains, such as studying political leanings with voter-vote networks [5] and investigating gene-expression networks in bioinformatics [23].

One important network analysis task is link prediction (i.e., detecting missing links), which infers the existence of implicit relationships between nodes based on currently observed links [21]. Link prediction is extremely useful in practice because real-world data is often noisy or incomplete; for example, guiding biological experiments in checking protein interaction networks [4].

In practice, analysts need to leverage their domain knowledge to examine algorithmic results. This is because the algorithm output is usually a list of scores or probabilities, which may be difficult to interpret. Moreover, these results can be inaccurate. They may ask: why is this link identified as missing with a high score, does it make sense to have a link between these two nodes, and how will the network change by adding one or several missing links? However, there still is a lack of effective tools to allow analysts to browse the results and explore answers to these questions.

To address these issues, we propose MissBiN, a visual analysis tool for detecting and examining missing links in bipartite networks. First, we contribute a novel link prediction approach that leverages the information of bi-cliques, inspired by structural hole theory in social science [3]. The method can be applied with any existing algorithm (e.g., neighbor-based techniques [24]), to both weighted and unweighted networks. Second, we develop an interactive visualization to present detected missing links, allowing for a better understanding of the missing links and their impact. The visualization enables analysts to compare networks with and without specific links interactively added by analysts. To evaluate MissBiN, we conducted quantitative experiments on real-world datasets and an in-depth case study of analyzing entity relationships in documents.
of nodes and uses link bundling. Radial layout of nodes has also been applied [7]. Second, bipartite networks can be shown as a bi-adjacency matrix, where rows and columns represent two different types of nodes. Example systems include BiVoc [10], Bicluster viewer [12], Expression Profiler [16], and BicOverlapper 2.0 [26]. An exception is BiDots [41] which organizes bi-cliques in rows to emphasize patterns and places nodes in columns. Hybrid techniques have been proposed to combine the advantages of node-link diagrams and matrices. For example, based on NodeTrix [13], Furby [30] and Bixplorer [8] display each bi-clique as an individual matrix and connect them with links. Xu et al. [38] applied a similar method for the projected bipartite network. Further, Matchmaker [20] and VisBricks [19] augment this visual representation with extensive charts (e.g., heatmaps and parallel coordinates) to display bi-cliques.

While some visual design of MissBiN is inspired by the above systems, none of them has addressed the problems of detecting and visualizing missing links. Particularly, we employ a matrix-based design because links are the focus in our analysis scenarios.

3 MISSING LINK PREDICTION OF MISSBIN

3.1 Problem Definition and Node Similarity

A bipartite network can be formally defined as $G = (X, Y, E)$, where $X$ and $Y$ are two non-overlapping sets of nodes and $E$ is the set of links that only exist between $X$ and $Y$, i.e., $e = (x, y) \in E$ where $x \in X$ and $y \in Y$. For a bipartite network, the number of all possible links is $|X| \cdot |Y|$ and we denote this set of links as $U$.

A link prediction problem is to identify which links are likely missing in the set $U - E$. As discussed in Section 2.2, similarity-based methods first compute the similarity score of every non-connected pair of nodes, and then generate a ranked list of missing links with decreasing scores for prediction. These types of methods are more generic because they purely rely on the network topology and are not biased to this visual representation with extensive charts. Thus, we adopt this for development and experiments, but our approach can be used with any link prediction methods that generate link probabilities.

One way to compute the node similarity is via a random walk, and here we employ random walk with restart (RWR) [33]. To accommodate bipartite networks, we let the random walker run for an odd number of iterations, such that it always stops at nodes in the set of two neighborhoods are, the more likely the random walk scores from node $x$ to node $y$ and the opposite.

Another way of measuring similarity is based on comparing the neighborhoods of two nodes. The intuition is that the more similar the topology of two neighborhoods are, the more likely the link connecting the two nodes is missing. For a bipartite network, neighbors of two possible connected nodes must be different. Following the ideas in [6, 37], we define the one-hop neighbors of a node $x$ in a bipartite network as $\Gamma(x)$, and we further define $\gamma(x)$ as the set of two-hop links of a node $x$. That is, $\gamma(x) = \{(x, y) \in E : x \neq y \in \Gamma(x)\}$. For example, in Figure 1a, $\Gamma(x_4) = \{y_1, y_4\}$ and $\gamma(x_4) = \{(x_1, y_1), (x_3, y_3), (x_4, y_4)\}$. A number of similarity metrics can be applied to compare the neighborhood context of two nodes, $\gamma(x)$ and $\gamma(y)$, such as:

- common neighbors: $s_{xy} = |\gamma(x) \cap \gamma(y)|$;
- Jaccard-Adar coefficient: $s_{xy} = \frac{|\gamma(x) \cap \gamma(y)|}{|\gamma(x) \cup \gamma(y)|}$;
- Adamic-Adar coefficient: $s_{xy} = \frac{1}{\sum_{e \in \gamma(x) \cap \gamma(y)} \log |\Gamma(e)|}$.

and preferential attachment: $s_{xy} = \frac{|\gamma(x)| \cdot |\gamma(y)|}{\log |\Gamma(x)| \cdot |\Gamma(y)|}$.  

3.2 Enhancement based on Bi-Clique Information

Based on the above methods, we propose to integrate a key structure in bipartite networks, maximal bi-cliques, for link prediction. Formally, a maximal bi-clique is defined as a sub-network, $G' = (X', Y', E')$, where $X' \subseteq X$, $Y' \subseteq Y$, and $E' \subseteq E$, and there exists a link $e = (x, y) \in E'$ between every pair of nodes, $x \in X'$ and $y \in Y'$.

Figure 1: a) An example bipartite network $G = (X,Y,E)$. b) An illustration of two overlapped bi-cliques, $C_1 = (X,Y,E_1)$ and $C_2 = (X,Y,E_2)$; each of them is shown as a bi-adjacency matrix.

Algorithm 1: Finding link weighting based on bi-cliques.

Data: A list of bi-cliques, $L = \{C_i = (X_i,Y_i,E_i)\}_i$, detected in a bipartite network $G = (X,Y,E)$.

Result: Scores, $w_i$, for all non-observed (missing) links in $G$.

For each $e \in U - E$, $U = \{\langle x,y \rangle : x \in X, y \in Y\}$:

1. For each bi-clique pair $(C_i, C_j)$ from $L$ do:
   a) $w_{ij} \leftarrow 0$;
   b) If $\sigma < 0$ then continue;
   c) For each $e = \langle x,y \rangle \in (X_i - X_j) \cap (X_j - X_i) \cap (Y_i - Y_j) \cap (Y_j - Y_i)$ do:
      i) $w_{ij} \leftarrow w_{ij} + \sum_{i \in \gamma(x)} \sum_{j \in \gamma(y)} |\gamma(x)| \cdot |\gamma(y)|$.

Our intuition is that missing links connecting nodes from different bi-cliques to form a larger bi-clique should carry more weight. This is inspired by the structural hole theory in social science [3]. That is, a person tends to know what other people in the same community know, and the most beneficial connections come from the weak ties that link people from different communities. Also, the literature indicates that it is useful in bipartite network analysis [37]. However, their method cannot fit generally with any node similarity metric.

As is shown in Figure 1b, consider two bi-cliques as two communities that have some nodes in common; each missing link between the non-overlapping nodes (i.e., $M_4$) from the two communities contributes to the formation of a bigger community that benefits all the nodes. If the two communities have many nodes in common, each of the few missing links that can be added carries more value, as a bigger bi-clique can be formed fairly easily. On the other hand, if the two communities have less in common, then more links need to be added to merge the two bi-cliques into a larger one, and each of the missing links carries less value.

Following this intuition, we develop an algorithm to re-rank the missing link list with scores $s(x,y)$ generated by the above similarity-based methods. As shown in Algorithm 1, it computes scores, $w_{ij}$, for all missing links (e.g., $M_4$ in Figure 1b) based on detected bi-cliques. When processing each pair of bi-cliques, $w_{ij}$ is updated with a value determined by the size of two bi-cliques and their overlap. Intuitively, in Figure 1b, the value computed in each iteration corresponds to the area of the intersection $M_4$ divided by the area of the missing part $M_4$. We only consider bi-clique pairs with overlap larger than a ratio $\theta$ to avoid overly biasing towards non-related bi-cliques. Then, we normalize the scores by the maximum value and generate a new ranked list of predicted missing links with the new scores: $s'(x,y) = w(x,y) \cdot s(x,y)$.

4 VISUAL INTERFACE OF MISSBIN

Algorithms are not always perfect in real-world usage, so some detected missing links may not be meaningful. We design a visual interface of MissBiN (Figure 2) allowing analysts to better make sense of predicted missing links.

4.1 Visual Exploration of Missing Links

MissBiN supports the visual exploration of predicted missing links through two views. First, the Network View (Figure 2a) displays...
the bi-adjacency matrix of a bipartite network, where the row and the column represent two different types of nodes respectively. The links are represented as squares in the intersections of rows and columns. The existing links in the network are shown in a yellow-green colorscale, reflecting the link weight. The predicted links are displayed in a white-purple colorscale with darker color indicating higher score. Second, the Link List View (Figure 2b) shows the computed missing links linearly by prediction score. A number of user interactions are offered, such as reordering rows and columns of the matrix as well as filtering based on the score.

These two views offer an overview of the bipartite network and the performance of the missing link prediction. We choose the matrix-based design for the Network View because our focus is links; matrices utilize much visual space for encoding links and are more scalable for denser networks [9, 17]. Moreover, the color channel cannot precisely reflect the scores and ranks of the missing links, we design the Link List View to augment the Network View for exploring the link prediction. An analyst can further explore the results and hypothetically add certain missing links to examine their influence with visual analysis of motifs and metrics (as described later). The added links are marked as black crosses on the matrix and also displayed at the top of the list (Figure 2a & b).

4.2 Visual Analysis of Network Motifs and Node Metrics

Motif analysis is one key approach to understanding the topology of a network. MissBiN provides a Detail View (Figure 2c) and an Overview (Figure 2d) for browsing the motifs at different scales. In the Motifs Detail View, bi-cliques are displayed as small multiples of matrices in similar visual encodings to the Network View because our focus is links; matrices utilize much visual space for encoding links and are more scalable for denser networks [9, 17]. Moreover, the color channel cannot precisely reflect the scores and ranks of the missing links, we design the Link List View to augment the Network View for exploring the link prediction. An analyst can further explore the results and hypothetically add certain missing links to examine their influence with visual analysis of motifs and metrics (as described later). The added links are marked as black crosses on the matrix and also displayed at the top of the list (Figure 2a & b).

5 Evaluation of Missing Link Prediction

5.1 Experimental Datasets

We used two real-world datasets to conduct experiments for testing our algorithm. The first is a person-place network extracted from the Atlantic Storm corpus containing 111 intelligence reports [15]. We identified 207 person nodes, 165 place nodes, and 1,718 links between them, from these reports. The second dataset is a weighted user-conversation bipartite network detected from Slack communication messages of a group within an IT company. The weight of a link is based on the number of words that a user contributed to a conversation. The resulting bipartite network contains 41 users, 61 conversations, and 258 links.
We evaluated our bi-clique oriented link prediction approach by the Crescent [15] dataset. It contains 41 fictional reports regarding three (RW) [24, 35]. For each algorithm, we randomly removed 1%, 2%, 5%, 10%, and 15% of links. For each condition, we performed the experiment five times to reduce sampling bias.

<table>
<thead>
<tr>
<th></th>
<th>Atlantic Storm</th>
<th>Slack communication</th>
</tr>
</thead>
<tbody>
<tr>
<td>JA</td>
<td>0.395</td>
<td>0.332</td>
</tr>
<tr>
<td>CN</td>
<td>0.359</td>
<td>0.202</td>
</tr>
<tr>
<td>AA</td>
<td>0.440</td>
<td>0.203</td>
</tr>
<tr>
<td>PA</td>
<td>0.021</td>
<td>0.096</td>
</tr>
<tr>
<td>RW</td>
<td>0.467</td>
<td>0.398</td>
</tr>
</tbody>
</table>

Table 1: For each condition (i.e., in a table cell), the three numbers indicate: the average metric of the baseline, the average metric of the proposed method, and the improvement over the five runs.

5.2 Experimental Setup and Results

Because there is no ground truth for missing links, we followed a commonly used method [37]: randomly remove a certain number of links from an original network, apply the link prediction algorithms on this new network, and measure the performance by comparing the predicted links with the removed links (i.e., the ground truth). We evaluated our bi-clique oriented link prediction approach by comparing it with five corresponding baseline algorithms, including common neighbors (CN), Jaccard coefficient (JA), Adamic-Adar coefficient (AA), preferential attachment (PA), and random walk (RW) [24, 35]. For each algorithm, we randomly removed 1%, 2%, 5%, 10%, and 15% of links. For each condition, we performed the experiment five times to reduce sampling bias.

We used two metrics to measure the performance of the algorithms: R-Precision (the ratio between all the relevant items and the ranked number of relevant items in the collection), and AUC PR (the area under the precision-recall curve). Table 1 shows the average performance of each condition. Based on the results, we can observe that the bi-clique oriented methods enhance the baselines in all the conditions with different levels of improvement on both metrics. Some of the performance gain is substantial, where the maximum improvement appears with the PA condition of the Atlantic Storm dataset.

6 Case Study

To demonstrate the usefulness of MissBiN, in this section, we walk through a case study of investigative analysis with The Sign of the Crescent [15] dataset. It contains 41 fictional reports regarding three coordinated terrorist plots in the US, of which 24 are relevant to the plot. Using name-entity detection techniques, we analyze a person-location bipartite network containing 49 persons and 104 locations as well as 328 connections between them. Examining missing links in such an analysis can be greatly helpful, because the information is often incomplete. We developed a simple Document View on top of the MissBiN interface, which displays all the related reports based on user-selected entities or relationships. Our analyst’s task is to identify suspicious persons and activities from these reports. She launches MissBiN, which visualizes the person-location bipartite network (Figure 2). First, she adjusts the threshold to only show predicted missing links with probability higher than 0.7, because there are too many purple squares in the Network View, which is a bit overwhelming. Among all the missing links, M. Gialob-Afghanistan has the highest probability (Figure 2b). She thus adds this potential link by clicking it and re-computes the motifs and node metrics. The Motifs Detail View then displays the removed, added, and unchanged bi-cliques of the network. Of the eight newly formed bi-cliques, a person named H. Pakes appears the most, which is in six bi-cliques (Figure 2c). Navigating back to the Network View, she finds that H. Pakes connects with many locations with high probability. She further sorts the nodes in decreasing order based on average missing link probability, and confirms this observation since H. Pakes is ranked the second (Figure 3a). The first person node has too few connections with the locations, which seems an isolated node. Thus, she focuses on H. Pakes and reads a few reports regarding him. She then finds that H. Pakes is a person carrying a forged Dutch passport and his actual given name is Abu al Masri, a member of the terrorist organization Al Qaeda. The reports reveal that H. Pakes was involved in shipping explosive materials from Holland to the US.

From the Link List View, she discovers that a particular person named F. Goba appears frequently in missing links with high probability (Figure 3b). For example, its connections to New York City, Amsterdam, and Queens are ranked second, fifth, and seventh, respectively. She reads the reports related to these locations and connects them with the information obtained before. She finds that F. Goba was involved in attacking a train to New York City with a bomb made with the explosive materials shipped by H. Pakes. Next, she adds the missing links between F. Goba and New York City, Amsterdam, and Queens, which results in a lot of newly-formed and removed bi-cliques. Then, she uses the Motifs Overview to explore them (Figure 3c). One interesting pattern catches her eyes: a cluster of four red dots close to a green dot, indicating that these (red) bi-cliques are quite similar and may be merged together into one new (green) bi-clique after adding a few missing links. By exploring the new bi-clique in the Motifs Detail View (Figure 3d), she finds two names: B. Dhalival and C. Webster. She further investigates these two people from the reports and learns that B. Dhalival is also a fake name, who carries an Indian passport, and his actual name is S. Alokri, a Pakistani who served Taliban in Afghanistan. More investigation of the reports reveals that B. Dhalival was likely related to the previously identified train attack.

7 Conclusion and Future Work

We have presented MissBiN, a visual analysis tool for exploring and understanding missing links in bipartite networks. It offers a novel approach for missing link prediction by using the information of bi-cliques in networks. Moreover, MissBiN provides an interactive visualization to present computed missing links and support the investigation of their meaning and influence by comparing networks with and without selected missing links. Quantitative experimental results and a case study indicate promising applications of MissBiN in practice. However, more evaluation such as deployment studies needs to be conducted to investigate its usage in real-world settings and in other domains. Also, we plan to extend the scalability of both the algorithm and visualization of MissBiN. We want to employ faster, but approximated, bi-clique detection algorithms [40], and enhance the matrix view for large networks with multi-scale [34] and focus+context techniques [14].