Understanding Missing Links in Bipartite Networks with MissBiN

Jian Zhao, Maoyuan Sun, Francine Chen, and Patrick Chiu

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 a visual analysis system, MissBiN, to involve analysts in the loop for making sense of link prediction results. MissBiN equips a novel method for link prediction in a bipartite network by leveraging the information of bi-cliques in the network. It also provides an interactive visualization for understanding the algorithm outputs. The design of MissBiN is based on three high-level analysis questions (what, why, and how) regarding missing links, which are distilled from the literature and expert interviews. We conducted quantitative experiments to assess the performance of the proposed link prediction algorithm, and interviewed two experts from different domains to demonstrate the effectiveness of MissBiN as a whole. We also provide a comprehensive usage scenario to illustrate the usefulness of the tool in an application of intelligence analysis.

Index Terms—Missing link prediction, bipartite network, bi-clique, interactive visualization, visual analytics.

1 Introduction

Many real-world systems can be modeled as *bipartite networks* (i.e., two-mode networks). That is, 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 based on roll call vote records [1], investigating gene-expression networks in bioinformatics [2], and identifying potential coalitions with entities co-occurrence networks from reports in intelligence analysis [3].

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 [4]. Link prediction is extremely useful in practice because real-world data is often noisy and incomplete. For example, as our knowledge on many biological networks is limited, applying link prediction can guide laboratory experiments, instead of blindly checking all possible protein interactions [5]. Also, link prediction can be employed for recommending products to users based on purchase networks in e-commerce [6], and suggesting friendships in social networks [4].

In practice, analysts need to leverage their domain knowledge to examine algorithmic results. They may ask: why is this link identified as missing with a high probability, does it make sense to have a link between these two nodes in the domain, and how will the network change by adding one or several detected missing links? This is because the algorithm output is usually a list of scores or probabilities for all potential missing links, which may be difficult to interpret. Moreover, these results can be inaccurate due to unseen flaws in algorithms. By leveraging human domain knowledge based on algorithmic outputs, analysts can improve the overall performance for real-world tasks. However, it remains challenging to effectively browse computational results and explore answers to analytical questions about missing links.

In order to address the above issues, we propose a visual analysis system, MissBiN, for detecting and examining missing links in bipartite networks, in extension to our prior work [7], [8]. First, this system contributes a novel link prediction approach that leverages the information of bi-cliques in a network (Section 4.3), which is partially inspired by structural hole theory in social science [9], [10]. The method can be applied, with any existing link prediction algorithms (e.g., neighbor-based techniques [11]), to both weighted and unweighted networks. Second, we develop an interactive visualization to present detected missing links, allowing for a better understanding of computed missing links and their impact on a bipartite network. The visualization enables analysts to compare an original network with one having specific links interactively added by analysts. This comparison is achieved via two of the most commonly used network analysis methods: metric-based (e.g., computing node betweenness [12]) and motif-based (e.g., detecting cliques [13]).

The design rationale of MissBiN is grounded by a set of three high-level analysis questions, which are obtained through a literature survey and semi-structured interviews with domain experts. To validate the link prediction in MissBiN, we conducted quantitative experiments on three real-world datasets. The results show that our approach outperforms baseline methods. In addition, we assessed the visual interface of MissBiN by carrying out interviews with experts from two different domains: management science and geographical science. The experts' feedback reveals the effectiveness of MissBiN, especially on the

- *Jian Zhao (correspondence author) is with University of Waterloo. E-mail: jianzhao@uwaterloo.ca.*
- *Maoyuan Sun is with Northern Illinois University. E-mail: smaoyuan@niu.edu.*
- *Francine Chen and Patrick Chiu are with FXPAL. Emails: {chen,chiu}@fxpal.com.*

Manuscript received April 19, 2005; revised August 26, 2015.

interactivity of investigating the detected missing links in their domain-specific datasets. Moreover, to concretely demonstrate the usefulness of MissBiN, we walk through a comprehensive usage scenario in intelligence analysis.

In summary, our contributions in this paper include:

1) A novel missing link prediction algorithm for bipartite graphs inspired by social science theories;
2) A visual analysis system, MissBiN, for exploring and understanding the predicted missing links; and
3) Evaluations of the algorithm performance based on real-world datasets as well as the whole system based on expert interviews.

2 Related Work

In this section, we review techniques for analyzing and visualizing bipartite networks and discuss algorithms for link prediction in both general and bipartite networks.

2.1 Bipartite Network Analysis

One of the key computational approaches for analyzing a general network (i.e., containing only one type of node) is to calculate the node centrality indices (e.g., betweenness and closeness), which characterizes the importance of a node [12], [14]. These metrics are also applicable for bipartite networks [15], [16]. In addition, any methods for general network analysis can be employed on a projected bipartite network (a transformation of a bipartite network to a general network by combining one type of nodes with links) [16], but some information may be lost.

Another branch of techniques is to identify special groups of nodes, such as motifs (e.g., chain, star, and clique), clusters, and communities [16]. Due to particular properties of bipartite networks, the motif-based analysis mainly focuses on extracting bi-cliques (e.g., using LCM [17] and MBEA [13]). Also, biclustering techniques (e.g., spectral co-clustering [18]) can be applied to simultaneously group two types of nodes, relaxing the criteria of bi-cliques.

Without losing generality, MissBiN supports visual analysis of bipartite networks based on the aforementioned two common approaches: metric-based and motif-based. Specifically, MissBiN allows analysts to interactively investigating the influence of particular missing links by comparing the results of these two types of analyses on networks with and without these links.

2.2 Bipartite Network Visualization

Similar to visualizing general networks, two main approaches to presenting bipartite networks are: node-link diagrams and matrices. Node-link diagrams emphasize entities (i.e., nodes) and are more commonly seen, but suffer from increased visual clutters for larger and denser networks [19]. On the other hand, matrices emphasize relationships (i.e., links) and are suitable for many network analysis tasks [20].

One example method, based on node-link diagrams, is Anchored Maps [21] that fixes the positions of nodes in one set. Jigsaw's List View [22] is another typical example, where different types of nodes are organized in different lists and links are applied to represent their connections. Similarly, Focus+Context lists are employed to show large bipartite networks [23], [24]. Another variation is to hide links and employ nested layouts in lists to reveal bipartite relations by duplicating nodes (e.g., ConTour [25]). Radial layout of nodes has also been applied. For example, AlertWheel [26] places nodes onto two concentric rings and utilizes the edge-bundling technique to display links. To further emphasize detected bi-cliques, BiSet [27] shows them as edge bundles between two lists of nodes and provides interactive features for directing users to potentially useful ones [28], [29], [30]. By relaxing the positions of nodes, BicOverlapper [31] displays a bipartite network with a node-link diagram and highlights sub-network motifs using boundaries.

For a second approach, bipartite networks are usually shown as a bi-adjacency matrix, where rows and columns represent two different types of nodes and links are revealed as matrix cells at corresponding locations. Example systems include BiVoc [32], Bicluster viewer [33], Expression Profiler [34], and BicOverlapper 2.0 [35]. An exception is BiDots [36] which organizes bi-cliques in rows to emphasize patterns and places nodes in columns, and allows for interactive repositioning of nodes.

Based on the above two approaches, hybrid visualization techniques have been proposed to combine the advantages of node-link diagrams and matrices. For example, based on NodeTrix [37], Furby [38] and Bixplorer [39] display each bi-clique as an individual matrix and connect them with links to show the entire network. Instead of visualizing the original network, Xu et al. [40] applied a similar method for the projected bipartite network. Matchmaker [41] and VisBricks [42] use extensive charts (e.g., heatmaps and parallel coordinates) to display bi-cliques and bundled links to indicate their relationships.

While the design of MissBiN has been inspired by the above systems, none of them has addressed the problems of detecting and visualizing missing links. More particularly, we employ a matrix-based design because links are the focus in our tasks and need to be emphasized visually.

2.3 Missing Link Prediction Algorithms

Common link prediction algorithms for networks roughly fall into two major categories: learning-based, and similarity-based. There are some comprehensive surveys such as [11], [43], [44], [45].

The learning-based methods usually treat link prediction as a binary classification problem and train a machine learning model to predict the class label (i.e., positive for potential linking) for each non-connected node pair. One typical approach is feature-based classification, which extracts features based on node attributes, topological structures, social theories, or combinations of them [46], [47], [48]. Another is based on probabilistic graph models including relational model [49] and entity-relationship model [50]. These techniques, although effective, are less general, often requiring some additional information (e.g., semantic node attributes) in addition to the observed network structure.

The similarity-based methods attempt to compute a similarity score for every non-connected pair of nodes

and rank all these potential links. Ways of computing the similarity metrics include random-walk based simulation, and neighbor-based measures, such as common neighbors, Jaccard coefficient, Adamic-Adar coefficient, and preferential attachment [11], [43]. Researchers have extended some of the similarity metrics to the bipartite network scenario [51]. Particularly, Xia et al. [52] proposed to increase link prediction accuracy via measuring structural holes in networks [10]. However, it cannot fit generally with any node similarity metric.

We move one step further by integrating important structural information in bipartite networks, on top of the link prediction scores generated by existing common approaches, to improve the performance. In this paper, we adopt the similarity-based approach (Section 4.2), because the information that the learning-based approach requires for training is usually dataset-specific (e.g., node attributes). The learned models may only perform well on networks with data features and typologies similar to the training set. To handle a more generic situation, we aim to perform link prediction only based on the topology of a network. Moreover, our approach (Section 4.3) can be used with any link prediction methods that produce a list of scores.

3 System Design and Overview

Here we introduce the design of MissBiN, a visual analysis system for detecting and understanding missing links in bipartite networks. Our main goal is to allow analysts to better investigate the characteristics of a network by combining automatic link prediction with interactive exploration. Thus, they may better understand the meaning of missing links by integrating their domain knowledge with algorithmic results.

As few works have been done in the visual analysis of missing links in networks, to design MissBiN, we aim to answer the following *What*, *Why*, and *How* questions about missing links. This is inspired by the questions asked by Brehmer et al. in their work about forming abstract visualization tasks [53]. Further, we conducted two 30-minute semi-structured interviews with our domain experts from management science and geographical science to verify and consolidate these questions. The first expert is a full professor at the management school of a university; and the second expert is a postdoctoral researcher at the computer science department of a university. More detailed background of the experts will be described in Section 7. The analysis questions are:

Q1. *What are the missing links?* The tool should provide an effective and robust method of discovering missing links in the network. Also, it should offer an overview of the scores of detected missing links, and alternative results from different link prediction algorithms, to combine the benefits of different approaches.

Q2. *Why is a link missing?* The tool should support human in the loop for investigating potential meanings behind a missing link by providing information such as the data context (e.g., node neighbors). This is because automatic algorithms may generate results that are not meaningful for particular domains.

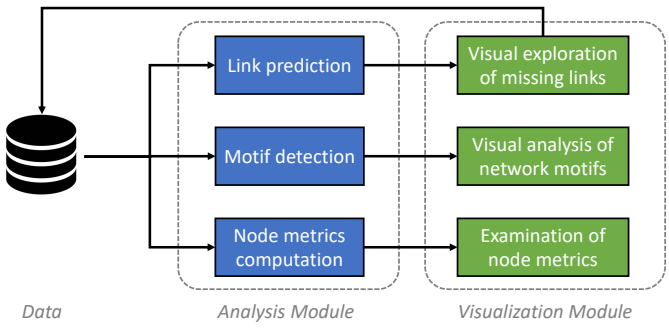


Fig. 1. The MissBiN system architecture.

Q3. *How does a missing link impact?* The tool should allow for evaluating the influence of particular missing links to better understand the characteristics of the network and the detected missing links. Therefore, an analyst is able to wisely utilize the valuable parts of the results.

In summary, these questions are the motivating factors on studying missing data in a systematic way, gradually deepening the analysis. The *What* question is to *detect* missing links by employing the algorithm and then some visual representation. The *Why* question is to *verify* missing links by leveraging an analyst's domain knowledge through visual interfaces. The *How* question is to *mitigate* missing links by testing the links' effects as if they exist in the network, with a combination of human judgment and algorithm outputs. Note that there exist more specific analytical questions. We use these three high-level questions to motivate our visualization design in a similar way to Brehmer et al.'s approach [53].

Following these questions and rationales, we design and develop MissBiN, which includes an analysis module and a visualization module (Figure 1). The analysis module supports missing link prediction in bipartite networks and two of the most common ways of observing networks, including node metrics and sub-network motifs. The novel link prediction method leverages the structural information of bi-cliques in the networks, which can be integrated with most link prediction algorithms [43]. The visualization module shows outputs from the analysis module and supports analysts in exploring the data with user interactions. An analyst can visually investigate computed missing links, and further examine the potential impact of missing links by comparing analytical results of the original network to those of the network with some hypothetically-added links.

.

4 Missing Link Prediction of MissBiN

To answer the *What* question (**Q1**) in MissBiN, we propose a novel bi-clique oriented approach for link prediction by augmenting existing similarity-based algorithms with topological information of bi-cliques.

4.1 Problem Definition

A bipartite network can be defined as $G = \langle X, Y, E \rangle$, where X and Y are two non-overlapping sets of nodes and E is

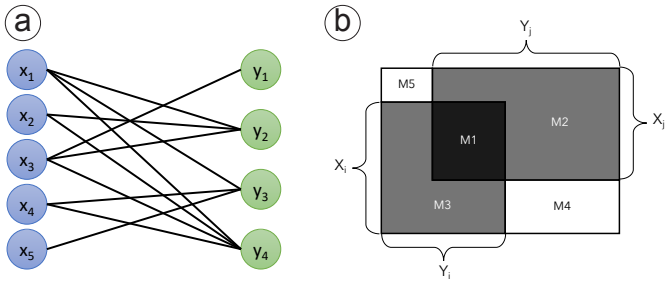


Fig. 2. a) An example bipartite network, where the two groups of nodes are: $X = \{x_i\}$ and $Y = \{y_i\}$. b) An illustration of two overlapped bi-cliques, $C_i = \langle X_i, Y_i, E_i \rangle$ and $C_j = \langle X_j, Y_j, E_j \rangle$, each of them shown as a rectangle (bi-adjacency matrix), where $E_i = X_i \times Y_i$ and $E_j = X_j \times Y_j$. Therefore, $M_1 = E_i \cap E_j$ represents the intersection; $M_2 = E_j - E_i$ and $M_3 = E_i - E_j$ represent their differences; and $M_4 = (X_i - X_j) \times (Y_j - Y_i)$ and $M_5 = (X_j - X_i) \times (Y_i - Y_j)$ represent the links needed to form a larger bi-clique.

the set of links that only exist between X and Y, i.e., $e = \langle x, y \rangle \in E$ 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. Thus, a link prediction problem is to identify which links are likely missing in the set $U - E$.

4.2 Status Quo Similarity-Based Link Prediction

This type of method first computes the similarity score of every non-connected pair of nodes in the set $U - E$. Then, it can generate a ranked list of missing links with decreasing scores for prediction.

One way to compute the similarity between pairs of nodes is via a random walk. We adopt a specific approach called *random walk with restart (RWR)* [54]. Consider a random walker starting from node x. At each step of the walk, it iteratively moves to a random neighbor of the current node with probability α and restarts the walk at node x with probability $1 - \alpha$. 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 other set. As the above measure is not symmetric, the final similarity score of each pair of nodes is: $s(x, y) = p_{xy} + p_{yx}$.

Another way of measuring similarity is based on comparing neighborhoods of two nodes. The assumption is that the more similar the topology of two neighborhoods is, the more likely the link connecting the two nodes is missing [11], [43]. This can be applied to both general networks and bipartite networks. However, for a bipartite network, neighbors of two possible connected nodes, $x \in X$ and $y \in Y$, must be from different sets of nodes. Following the ideas in [51], [52], 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 the two-hop links of a node x. That is, $\gamma(x) = \{\langle x_i, y \rangle \in E : x_i \neq x, y \in \Gamma(x)\}$. For example, in Figure 2a, $\Gamma(x_4) = \{y_3, y_4\}$ and $\gamma(x_4) = \{\langle x_1, y_3 \rangle, \langle x_5, y_3 \rangle, \langle x_1, y_4 \rangle, \langle x_2, y_4 \rangle, \langle x_3, y_4 \rangle\}$. Based on this definition, a number of similarity metrics can be applied to compare the neighbor context of two nodes, $\gamma(x)$ and $\gamma(y)$:

- *common neighbors*: $s(x, y) = |\gamma(x) \cap \gamma(y)|$;
- *Jaccard coefficient*: $s(x, y) = \frac{|\gamma(x) \cap \gamma(y)|}{|\gamma(x) \cup \gamma(y)|}$;
- *Adamic-Adar coefficient*: $s(x, y) = \sum_{\langle m,n \rangle \in \gamma(x) \cap \gamma(y)} \frac{1}{\log |\Gamma(m)| \cdot |\Gamma(n)|}$;
- and *preferential attachment*: $s(x, y) = |\gamma(x)| \cdot |\gamma(y)|$.

4.3 Novel Enhancement with Bi-Clique Information

Based on the above algorithms, we propose a novel approach that integrates one important type of structure in bipartite networks, *bi-cliques* (complete bipartite graphs). Formally, a bi-clique is defined as a sub-network, $G' = \langle X', Y', E' \rangle$, where $X' \subseteq X, Y' \subseteq Y$, and $E' \subseteq E$, and there exists a link $e = \langle x, y \rangle \in E'$ between every pair of nodes, $x \in X$ and $y \in Y$. Many algorithms have been proposed to efficiently detect all bi-cliques in a network, and in this paper, we adopt the MBEA algorithm [13].

Our intuition is that missing links between 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 [9], [10]. That is, a person tends to know what other people in the same community know; and to increases the overall information flow in the entire network, it is beneficial to add (missing) connections that link people from different communities. The literature also indicates that the structural hole theory is useful in bipartite network analysis [52], but their method is less general (see Section 2).

As is shown in Figure 2b, considering two bi-cliques as two communities that have some nodes in common; each missing link between the non-overlapping nodes (i.e., M_4 and M_5) 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 community 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 them, and thus each of the missing links carries less value.

Following this intuition, we develop an algorithm to re-rank the missing link list generated by the above similarity-based methods. As is shown in Algorithm 1, the algorithm computes weights, w_e for all missing links based on the bi-cliques in the network. The weight of a link is the sum of all the values calculated when processing each pair of bi-cliques (line 4-13), in which the value is determined by the size of the two bi-cliques and their overlap (line 9). Intuitively, as is shown in Figure 2b, the value computed in each iteration (line 9) corresponds to $\frac{|M_1|}{|M_4|+|M_5|}$ where $|\cdot|$ represents the cardinality, i.e., the number of links in the bi-clique or the "area" of the matrix. We add this value to every missing links in M_4 and M_5 for the current pair of bi-cliques (line 10-12). We only compute and accumulate w_e for the bi-clique pairs that have an overlap ratio o (line 5) larger than a threshold (line 6-8), where o corresponds to $\frac{|M_1|}{\sum_{i=1}^{5} M_t}$. This filters out bi-clique pairs with no or small overlap, which cause marginal effect on w_e but need more computation time. Then, we normalize the weights and the similarity scores by their maximum values, and generate a new ranked list with the new scores with $s'(x, y) = w(x, y) \cdot s(x, y)$.

We also tested other ways of computing w_e, such as letting $w = \frac{|M_1|}{\sum_{t=2}^{5} M_t}$ (line 9). We finally decided to use the form in Algorithm 1 based on our experiments. One thing to note is that the above re-ranking approach can

Algorithm 1: Missing link ranking.

Input : A list of bi-cliques, $L = \{C_i = \langle X_i, Y_i, E_i \rangle\}$, detected in a bipartite network $G = \langle X, Y, E \rangle$.
Output: Weights, w, for all non-observed (missing) links in G.

foreach $e \in X \times Y - E$ **do**
  $w_e \leftarrow 0$;
end
foreach bi-clique pair (C_i, C_j) from L **do**
  $o \leftarrow \frac{|X_i \cap X_j| \cdot |Y_i \cap Y_j|}{|X_i \cup X_j| \cdot |Y_i \cup Y_j|}$;
  if $o <$ *threshold* **then**
    continue;
  end
  $w = \frac{|X_i \cap X_j| \cdot |Y_i \cap Y_j|}{|X_i - X_j| \cdot |Y_j - Y_i| + |X_j - X_i| \cdot |Y_i - Y_j|}$;
  foreach $e \in \{\langle x, y \rangle; x \in X_i - X_j, y \in Y_j - Y_i\} \cup \{\langle x, y \rangle; x \in X_j - X_i, y \in Y_i - Y_j\}$ **do**
    $w_e \leftarrow w_e + w$;
  end
end

be employed to the probabilities generated by any link prediction algorithm.

5 Evaluation of Missing Link Prediction

To validate the accuracy of the proposed link prediction approach in MissBiN, we conducted quantitative experiments with three bipartite networks. In this section, we first describe these datasets, and then discuss the experimental design and results.

5.1 Experimental Datasets

The first dataset is a weighted person-place network extracted from the Atlantic Storm corpus [55], which contains 111 fictional intelligence reports. We extracted person and place entities from the reports, and computed relationship weights between entities based on the numbers of co-occurrences and normalized word distances in the text (the higher the weight is, the stronger the relationship that two entities have). Because real-world dataset is usually noisy, we followed the method in [52], and we further extracted the core network by removing nodes with less than three neighbors, resulting in a bipartite network with 207 person nodes, 165 place nodes, and 1,718 links.

The second dataset is a weighted user-conversation bipartite network detected from Slack communication messages of a group within an IT company. We divided chat message logs in one month into multiple conversations based on the time intervals between messages. Then, we identified the users in these conversations, and constructed a bipartite network connecting users and conversations. The weight of a link is calculated based on the number of words that a user-contributed to a conversation. Again, we extracted the core network for our experiments, leading to 41 users, 61 conversations, and 258 links.

The third network is built from the IEEE VIS publication corpus [56] that contains meta-data of the papers published from 1990 to 2015. We constructed an unweighted bipartite network between authors and papers using this information. Similarly, we filtered out nodes with less than three neighbors and obtained a final network of 442 authors, 1,160 papers, and 2,140 links.

5.2 Experimental Setup

Because there is no ground truth for missing links, we followed a scheme commonly used in the literature (e.g., [51]) to design our experiments. That is, we randomly removed a certain number of links from an original network, applied the link prediction algorithms on this new network, and measured the performance by comparing the predicted links with the removed (actually missing) links (i.e., the ground truth).

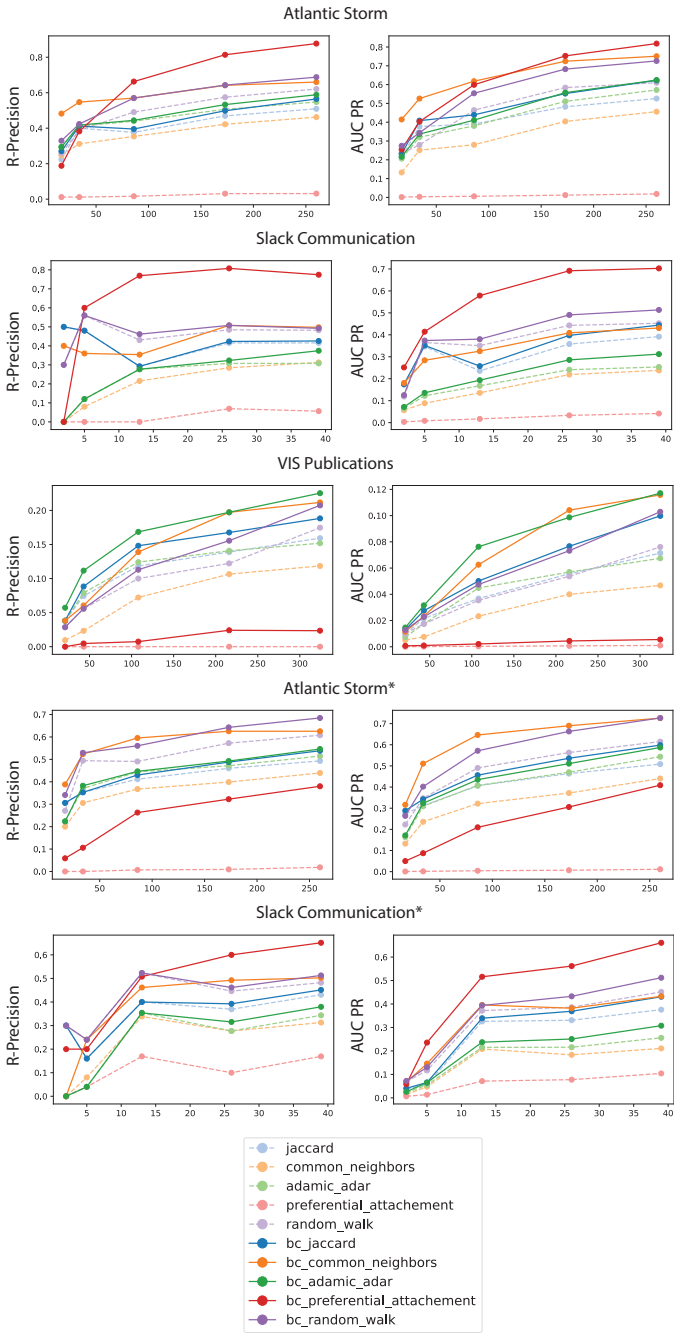


Fig. 3. Experimental results of our bi-clique (bc) oriented methods (Section 4.3) and the baselines (Section 4.2), i.e., bc_[...] v.s. [...], on R-Precision or AUC PR with the five networks, including three unweighted networks and two weighted networks (denoted with *). The x-axis is the number of links removed from the original dataset in order to construct the input network. The links are removed at 1%, 2%, 5%, 10%, and 15% of the original dataset.

We evaluated our bi-clique oriented approach with re-ranking of the results from five existing link prediction algorithms, including common neighbors, Jaccard coefficient, Adamic-Adar coefficient, preferential attachment, and random walk methods [11], [43]. For each algorithm, to test its performance under different situations, we randomly removed 1%, 2%, 5%, 10%, and 15% of links from the list of all links of an input network. For each of these conditions, we performed the experiment five times in order to reduce sampling bias. As two of our datasets are weighted bipartite networks (i.e., Atlantic Storm and Slack communication), we binarized these two networks and conducted our experiments on five different datasets, including three unweighted networks and two weighted ones. For the weighted networks, we extended the algorithms to their weighted versions, in particular, computing weighted sums based on links using the equations in Section 4. Algorithm 1 was not modified because it does not consider link weights, although weighted areas of $M1$, $M2$, and $M3$ can be used, which is left for future work.

For implementation, we set 0.05 as the threshold of Algorithm 1. The bi-clique detection method, MBEA [13], requires a minimum bi-clique size as its parameter, where we used 3 for both types of nodes. Thus, the input to Algorithm 1 contained bi-cliques larger than 3×3. For the random walk [54], we chose 101 as the number of iterations.

5.3 Results

We used two metrics from the field of information retrieval to measure the performance of the algorithms: *R-Precision* and *AUC PR* [57]. As the algorithms generate a ranked list of missing links with scores, the R-Precision is the ratio of the number of relevant items retrieved to the rank when the rank equals the number of relevant items in the collection. For example, if we remove n links from the network and the algorithm retrieves m of those links in its top-n results, the R-Precision is $\frac{m}{n}$. The AUC PR is the area under the precision-recall curve. Using the same example, consider that the algorithm retrieves m correct links in its top-k results. The curve is formed by iteratively computing the precision ($\frac{m}{k}$) and the recall ($\frac{m}{n}$) with k ranging from 1 to the entire rank list length.

Figure 3 displays the results of our experiments with the three datasets, in which the columns represent the Atlantic Storm, Slack communication, and IEEE VIS publication networks respectively, and the first two rows indicate their unweighted versions and the last two rows are the weighted ones. The performance metrics (the y-axis, R-Precision or AUC PR) were computed in each run with the input network built by removing a certain percentage of the links (the x-axis). Table 1 further shows the average performance of each condition of the experimental results in numbers.

From Figure 3, we can observe that the bi-clique oriented methods enhance the baselines in all the conditions with different levels of improvement on both R-Precision and AUC PR. Some of the performance gain is substantial, where the maximum improvement appears with preferential attachment (PA) for the unweighted Atlantic Storm dataset (0.564 for R-Precision and 0.557 for AUC PR).

For the unweighted Atlantic Storm and Slack communication networks, the preferential attachment based method performs the best. For the VIS publications network, the overall performance is worse than those for other datasets, which might be because the network is sparse. The best performing method is based on Adamic-Adar coefficient (AA), but the largest improvement appears with common neighbors (CN). For weighted networks, the best performers are common neighbors (CN) and preferential attachment (PA) with the Atlantic Storm and Slack communication datasets respectively. Further, from Figure 3, the performance of all the algorithms are generally better as more links are removed from the original datasets. This may result from that the task is harder when there are few correct missing links but a lot of possible connected links; while, as more and more links are removed, the performance might drop because of less information remained in the networks. However, future studies need to be conducted to verify this.

Moreover, for a random prediction on a bipartite network $G = \langle X, Y, E \rangle$ with a fraction f links removed for the experiment, the probability of selecting a correct link is $p = \frac{fE}{XY-(1-f)E}$. For selecting fE links, the probability is approximately $fE \cdot p$, and thus the R-Precision is $\frac{fE \cdot p}{fE} = p$. Using this equation, the average R-Precision for this random algorithm with the same experimental settings can be obtained: 0.003, 0.007, 0.0003 for Atlantic Storm, Slack communication, and VIS publications networks, respectively. We can see that our approach is orders of magnitude better, even for the VIS publications network where it performs the worst.

6 Visual Interface of MissBiN

Algorithms are not always perfect, and the computed missing links may not always be meaningful. Real-world scenarios are far more complicated, and it is difficult to consider every nuance in all domains for the algorithm design. Thus, it is critical to involve analysts in the loop to examine algorithmic results, which couples the flexibility of humans with the scalability of machines.

To this end, we design a visual interface as part of MissBiN to help analysts to better make sense of missing links identified by the aforementioned methods in bipartite networks. This visualization module consists of five interactively-coordinated views: a Network View and a Link List View to support the exploration of missing links, a Motifs Overview and a Detail View to offer the analysis of motifs, and a Metrics View to display node-based metrics (Figure 4). These views display outputs from the analysis module in visual forms to allow analysts to effectively answer the *What*, *Why*, and *How* questions about missing links. In general, we aim to design a simple visualization that can be easily-understood and self-explanatory.

6.1 Visual Exploration of Missing Links

MissBiN supports the visual exploration of predicted missing links through two views. First, the Network View (Figure 4a) 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

TABLE 1
More details of the experimental results shown in Figure 3.

		Atlantic Storm			Slack communication			VIS publications			Atlantic Storm*			Slack communication*		
R-Precision	JA	.395	.428	.032	.421	.424	.004	.105	.126	.020	.405	.423	.018	.332	.341	.009
	CN	.359	.580	.221	.179	.424	.245	.066	.129	**.063**	.342	**.552**	**.209**	.202	.339	.138
	AA	.440	.455	.016	.202	.219	.016	.107	**.152**	.045	.405	.418	.013	.203	.218	.015
	PA	.021	**.585**	**.564**	.025	**.590**	**.565**	.000	.012	.012	.007	.226	.219	.096	**.432**	**.336**
	RW	.467	.531	.064	.451	.464	.013	.096	.112	.016	.487	**.552**	.065	.398	.407	.009
AUC PR	JA	.398	.451	.053	.301	.325	.024	.039	.053	.014	.393	.444	.051	.225	.249	.024
	CN	.305	.607	.302	.148	.326	.178	.025	.063	**.039**	.300	**.578**	**.277**	.133	.285	.152
	AA	.398	.429	.031	.170	.200	.030	.039	**.068**	.029	.379	.406	.027	.153	.177	.025
	PA	.008	**.566**	**.557**	.021	**.528**	**.507**	.000	.003	.002	.005	.212	.207	.055	**.406**	**.352**
	RW	.435	.516	.082	.346	.377	.031	.039	.052	.013	.448	.525	.078	.279	.308	.029

* denotes weighted networks. JA: Jaccard coefficient; CN: common neighbors; AA: adamic-ardar coefficient; PA: preferential attachment; RW: random walk. For each condition (i.e., in a table cell), the three numbers indicate (1) the average metric of the baseline, (2) the average metric of the proposed method, and (3) the improvement, over the five runs (on removing different numbers of links from the original dataset). The highest performance and improvement are highlighted in bold for each metric in each dataset.

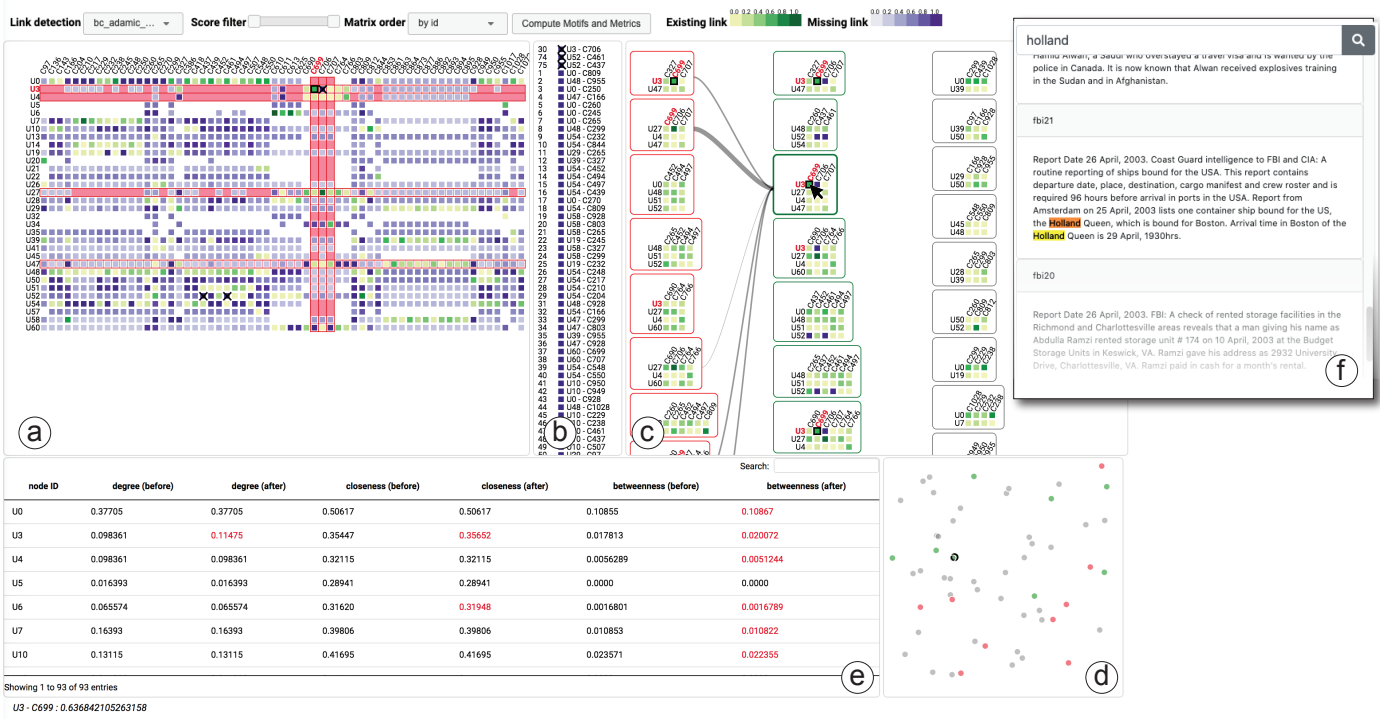


Fig. 4. An analyst is investigating the predicted links in a person-conversation bipartite network using MissBiN which consists of: a) a Network View, b) a Link List View, c) a Motifs Detail View, d) a Motifs Overview, and e) a Metrics View. f) A Document View is added later for showing context in a specific case of exploring intelligence reports (Section 8). The existing links in the network are shown in a yellow-green colorscale, where the intensity reveals the weight of a link. The predicted missing links are displayed in a white-purple colorscale, where a darker color reflects a higher score determined by the link prediction algorithm. Hovering over a bi-clique in b) highlights the corresponding rows and columns in a) in red.

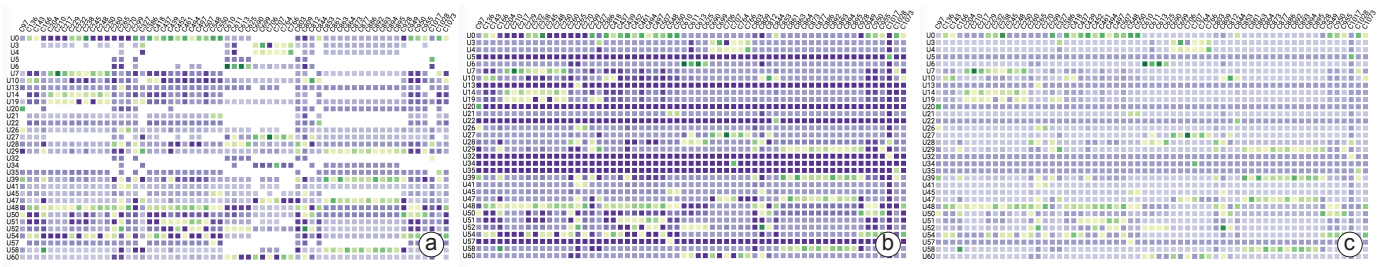


Fig. 5. The Network View of the bi-clique oriented link predictions of a person-conversation bipartite network based on (a) the adamic-adar coefficient, (b) the preferential attachment, and (c) the random walk with restart algorithms.

columns. We use this matrix-based design because it is more effective to visualize dense networks [19], [20], where in our case we need to show a large number of links, including both the original and predicted missing links. This helps to reveal *What* the missing links are (**Q1**).

Second, the Link List View (Figure 4b) shows the computed missing links linearly by probability or score, where each link is visualized in a similar way to that in the Network View. Additional information such as the rank and the connecting nodes of the link is provided. This Link List View works together with the Network View, allowing an analyst to better explore the link prediction from different perspectives (**Q1**).

A number of user interactions are offered. In the Network View, an analyst can reorder the rows and columns of the matrix with certain criteria such as the node label, the average prediction score, and the total number of detected missing links. Hovering over a link or a node highlights the corresponding row and/or column, and displays some detailed information, such as the prediction score. Similar interactions are offered in the Link List View. An analyst can also filter the matrix based on the prediction score, for example, to reveal the most probable missing links suggested by the algorithm. Such user interactions help analysts understand the missing links by providing the data context of a node from different perspectives (**Q2**). However, other interactions such as interactive legend [58] can be integrated to enhance the visual exploration experience.

Moreover, different link prediction algorithms can be applied and viewed in the visualization. For example, from Figure 5, we can observe that different algorithms may generate significantly different predictions. Thus, an analyst can use the visualization to combine the advantages of various methods in real-world applications (**Q2**). However, a side-by-side visualization of graphs [59] needs to be developed to facilitate this, which is left for future work.

These two views offer an overview of the structure of a bipartite network and the performance of the missing link prediction (**Q1**). Having this image in their mind, analysts can further utilize their domain knowledge to investigate the detected missing links and understand the meaning behind them (**Q2**). Specifically, an analyst can explore link prediction results and hypothetically add certain missing links to examine their influence with visual analysis of motifs and metrics described in the following. The added links are marked as black crosses on the matrix and also displayed at the top of the list (Figure 4ab). A group of links can be added at once by selecting them from the matrix.

6.2 Visual Analysis of Network Motifs

Motif (i.e., a sub-network context of closely related nodes) analysis is one major approach to understanding the topology of a network. In bipartite networks, a bi-clique is one of the most important structural patterns. Other motifs (e.g., trees and chains) are less meaningful and not utilized much. MissBiN provides a Detail View (Figure 4c) and an Overview (Figure 4d) for browsing the motifs at different scales (**Q2**). In the Motifs Detail View, bi-cliques are displayed as small multiples of matrices in similar visual encodings to the Network View (Figure 4a), highlighting the most important motifs. Essentially, a bi-clique is a portion of the bi-adjacency matrix of the entire network. In addition, the Motifs Overview displays all the bi-cliques as dots in a two-dimensional space based on MDS projection [60]. The distance between two bi-cliques is measured with the sum of the two Jaccard distances of the corresponding node sets of the bi-cliques (i.e., $1 - \frac{X_i \cap X_j}{X_i \cup X_j} + 1 - \frac{Y_i \cap Y_j}{Y_i \cup Y_j}$).

These two views not only support the visual exploration of all bi-cliques detected in the network, but also the investigation of the impact, if certain missing links are added (**Q3**). The Motifs Detail View supports comparing the two sets of bi-cliques detected in the networks with and without added links by an analyst. MissBiN organizes the bi-cliques of the new network (with added links) in three columns: *removed* bi-cliques (those in the original set but not appear in the new set), *newly-added* ones (those appear in the new set but not the original set), and *unchanged* ones (those appear in both sets), in borders of red, green, and gray, respectively (Figure 4c). Specifically, bi-cliques can be added when the new links make nodes connected to form a new bi-clique; and bi-cliques can be removed when the new links make smaller bi-cliques merge into a bigger one. In each column, the default order of bi-cliques is by size, which can be changed to other sorting criteria. Similarly, the Motifs Overview encodes these bi-cliques in the three different colors (Figure 4d).

We further compute the similarity between the added and removed bi-cliques using the Jaccard distance to support a better understanding of the structural changes and the impact of missing links (**Q3**). In the Motifs Detail View, when an analyst hovers over a bi-clique, this information is shown as gray ribbons connecting the related bi-cliques, with the thickness mapped to their pairwise similarity value (Figure 4c). Moreover, corresponding rows and columns of the hovered-over bi-clique in both the Overview and the Detail View are highlighted in the Network View to offer more context. Although in MissBiN we currently only allow for one type of motif (i.e., bi-clique), other forms of motif analysis can be easily supported with a similar visualization.

6.3 Examination of Node Metrics

Computing node-metrics is a common way of getting to learn a big picture of the characteristics of a network in social sciences and other domains (**Q2**). The Metrics View in MissBiN (Figure 4e) supports this type of analysis by presenting a number of metrics in a tabular view: degree, closeness, and betweenness centralities of before and after adding certain missing links, etc. Changes of metric values are highlighted in red, revealing the effect of added links (**Q3**). This table is also interactively linked with other views. For example, hovering over a row emphasizes the corresponding node in the Network View. Since there might be a large number of nodes (rows), a search function is provided, and hovering over a node in other views automatically navigates to that row in the table.

7 Initial Expert Feedback

To evaluate the visual interface of MissBiN, we conducted qualitative interviews with the two experts whom we

talked to during the design of the system. As mentioned before, they have different technical backgrounds including management science and geographical science. During the interview, we first reviewed the background of this research and introduced the features of the tool. We then loaded the datasets that the experts were interested in, and let them conduct free-form visual analysis with MissBiN. We provided help if the experts were confused about the tool features or had questions about anything. The think-aloud protocol was employed. We also observed their interactions with the tool and took notes when necessary. After the free-form analysis, we conducted follow-up interviews with the experts to collect their further feedback. Each interview lasted for about an hour.

In general, both experts appreciated MissBiN in many aspects, particularly the interactivity of missing links investigation, and they were willing to use the tool in their research in the future. In the following, we discuss the experts' feedback in more detail.

7.1 Interview Study I

The first expert (E1), a management school professor, has extensive research experience in analyzing the social dynamics of people in large organizations and online communities with quantitative methods. For his interview, he explored an employee-conversation network extracted from chat logs of an enterprise communication tool. That is, the network contains two types of nodes including employee and conversation, and each link reveals that an employee participated in a conversation.

Examining overall communication. E1 wanted to investigate patterns revealed in the social relationships of employees in the company, as shown in Figure 4. He advocated that link prediction in such a network is critical, and it implies that some employees could beneficially participate in certain conversations (**Q1**). As there were many missing links with high probabilities, E1 commented that "*These links can be used for recommendation of people to connect and chat channels to join, and indicators for the social health of the entire team.*"

Identifying missing social interaction. E1 also appreciated the Motifs Detail View (Figure 4c) where he could investigate the effect of missing links from the motifs analysis perspective and had a lower-level view of the network topology (**Q3**). He commented that "*I would like to see this applied to the analysis of general network patterns (motifs), such as stars and loops.*" Another aspect in this view that E1 liked was the support of comparing detected motifs. He mentioned that "*It seems this [missing] link is critical because it dramatically changes the cliques in the network.*" He further hovered over specific bi-cliques, mentioning that "*This clique is mainly overlapped with these old ones, so not much new information is gained.*" Thus, E1 was able to investigate whether the effect of the selected missing links is significant or not through comparing the bi-cliques. Moreover, E1 said that the Motifs Overview was interesting and he never thought about viewing network patterns in this way. After some exploration, he added "*Look here! This green dot (an added motif) is far away from anybody including the red dots (removed motifs). It means this new pattern is quite unique and the links I just clicked are critical for this group of nodes.*"

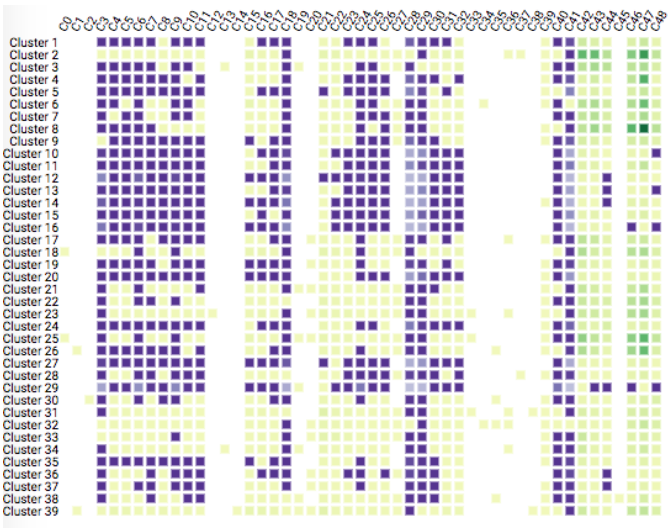


Fig. 6. Visualization of the bipartite network between crime types and region clusters in Washington DC in 2017.

7.2 Interview Study II

The second expert (E2), a computer science postdoctoral researcher, conducts research between geographical information analysis and data visualization. He is interested in developing geospatial analytical models and visual tools for supporting decision making. The dataset used in his interview was a bipartite network of crimes and locations in Washington DC in 2017. The crimes are categorized by the type of offense (e.g., robbery), method (e.g., knife), and time of the day (e.g., evening), resulting in 49 different crime incident categories. Moreover, the locations are grouped into 40 different clusters.

Investigating overall crime status. E2 was curious about the correlations between different crimes and locations. From the Network View of the dataset (Figure 6), he was able to identify that some crimes are very rare, such as burglary with guns during a day (C12), and some region clusters have a much higher number of incidents such as Cluster 2 (Columbia Heights, Mt. Pleasant, Pleasant Plains, Park View) and Cluster 8 (Downtown, Chinatown, Penn Quarters, Mt. Vernon, N. Capitol St.) [61]. Moreover, the high-frequency crimes that occurred in these clusters were theft related (C42-C48) which appeared in almost every cluster. Based on the E2's observation of the correlations, the missing link prediction indicates potential crimes might happen in certain clusters, and those with high probabilities were in the aforementioned clusters (**Q1**). E2 commented that "*I like the ability of examining the prediction of multiple algorithms at the same time, so I can do a more comprehensive analysis and combine the results.*" But he suggested that it would be better to facilitate the algorithm comparison with a side-by-side visualization, where currently he could only open two windows in the browser to do so.

Discovering flaws in data processing. From his exploration of the predicted missing links, E2 was surprised that many of the link probabilities are relatively high, as indicated by the many dark purple squares in Figure 6. Based on his domain knowledge, E2 said that it cannot be all true because of the geographical, demographic, and

economical features of different neighborhoods in the city. By digging into some of the local regions of the network with MissBiN, E2 suspected that the above results might due to the preprocessing of the dataset that classified the crime incidents (**Q2**). He explained that "*This coarse categorization of the incidents results in less distinguishable connections in the network which may confuse the [link prediction] algorithm.*" He further suggested that adding features with geographical bias in this domain may generate better predictions.

8 Usage Scenario

To demonstrate the usefulness of MissBiN, in this section, we walk through a scenario of intelligence analysis with *The Sign of the Crescent* [55] 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 extracted 284 unique entities and 495 relationships based on co-occurrence of the entities in the same report. For this scenario, we analyze a person-location bipartite network containing 49 persons and 104 locations as well as 328 connections between them, which is the largest and most important bipartite network in this dataset. Examining missing links in such an application domain can be greatly helpful, because the information is often incomplete in intelligence analysis. In order to facilitate the analysis, we developed a simple Document View on top of the MissBiN interface, which displays all the related reports in the dataset based on user-selected entities or relationships. The Document View also offers a basic search function using keywords matching.

Suppose that Michelle is an intelligence officer, and she is assigned to identify suspicious persons and activities from these reports. She launches MissBiN and loads the person-location bipartite network. The system automatically runs the missing link prediction and displays the results. The key steps of her analysis are shown in Figure 7.

First, Michelle 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. Galab - Afghanistan* has the highest probability (Figure 7a). Thus, Michelle 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 remained bi-cliques of the network. Of the eight newly formed bi-cliques, a person named *H. Pakes* appears the most frequently, which is in six bi-cliques (Figure 7b).

Navigating back to the Network View, she finds that *H. Pakes* connects with many locations with high probability. Michelle further sorts the nodes decreasingly based on average missing link probability, and confirms this observation since *H. Pakes* is ranked the second. But the first person node has too few connections with the locations, which seems an isolated node. Therefore, 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, who has been a member of the terrorist organization *Al Qaeda*. The reports also reveal that *H. Pakes* was involved in shipping explosive materials from *Holland* to the US.

After, Michelle shifts her focus to the location node *The Netherlands* because *H. Pakes* has a fake Dutch passport. She then identifies two more high-probability missing links connecting *The Netherlands* with *A. Ramaz* and *F. Goba*, ranked the fourth and the sixth respectively, as seen from the Link List View (Figure 7c). Michelle adds these two missing links and again re-computes the motifs and node metrics with MissBiN. This time the system generates 17 new bi-cliques. With some exploration, Michelle identifies that *Charlottesville* and *Virginia* are two frequent US locations appearing in these bi-cliques (Figure 7d). Moreover, she finds that the betweenness of *H. Pakes* increases significantly after adding the two missing links, confirming that *H. Pakes* is a very important person and the added links are critical.

Further reading in the reports related to the nodes in the new bi-cliques reveals that *M. Galab* (appeared in the missing link with highest probability, *M. Galab - Afghanistan*, as mentioned earlier) is a participant of a terrorist organization named *HAMAS*. Moreover, *M. Galab* holds a valid student visa at the *University of Virginia* for several years. Hence, Michelle decides to investigate the location *Charlottesville*, and from the Network View she finds that it has a high weight connection to *Y. Mosed*, which is second highest except the link to *M. Galab* (Figure 7e). From the reports, *Y. Mosed* is also a member of *HAMAS* and has a valid student visa at the same university.

From the Link List View, Michelle discovers that a particular person named *F. Goba* who appears frequently in missing links with high probability. 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, and then she hypothesizes that *F. Coba*, *M. Galab*, and *Y. Mosed* followed the orders of *A. Ramaz*, planed to attack a train to *New York City* with a bomb made with the explosive materials shipped by *H. Pakes*.

Next, Michelle 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. Thus, Michelle starts to use the Motifs Overview to explore them (Figure 7f). One interesting pattern appears—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, Michelle finds two unseen names: *B. Dhaliwal* and *C. Webster*. She further investigates these two people from the reports and learns that *B. Dhaliwal* is also a fake name, who served *Taliban*. More investigation can be continued to find out if *B. Dhaliwal* is related to any planed terrorist activities and the identified train attack.

9 Discussion

While the previous studies have indicated the effectiveness of MissBiN, there still exist limitations in the system as well as the evaluation.

First, the missing link prediction process may be less scalable for very large networks. The time complexity of the standard link prediction algorithms is $O(|X||Y|)$ for a bipartite graph $G = \langle X, Y, E \rangle$. On top of this, we adjust

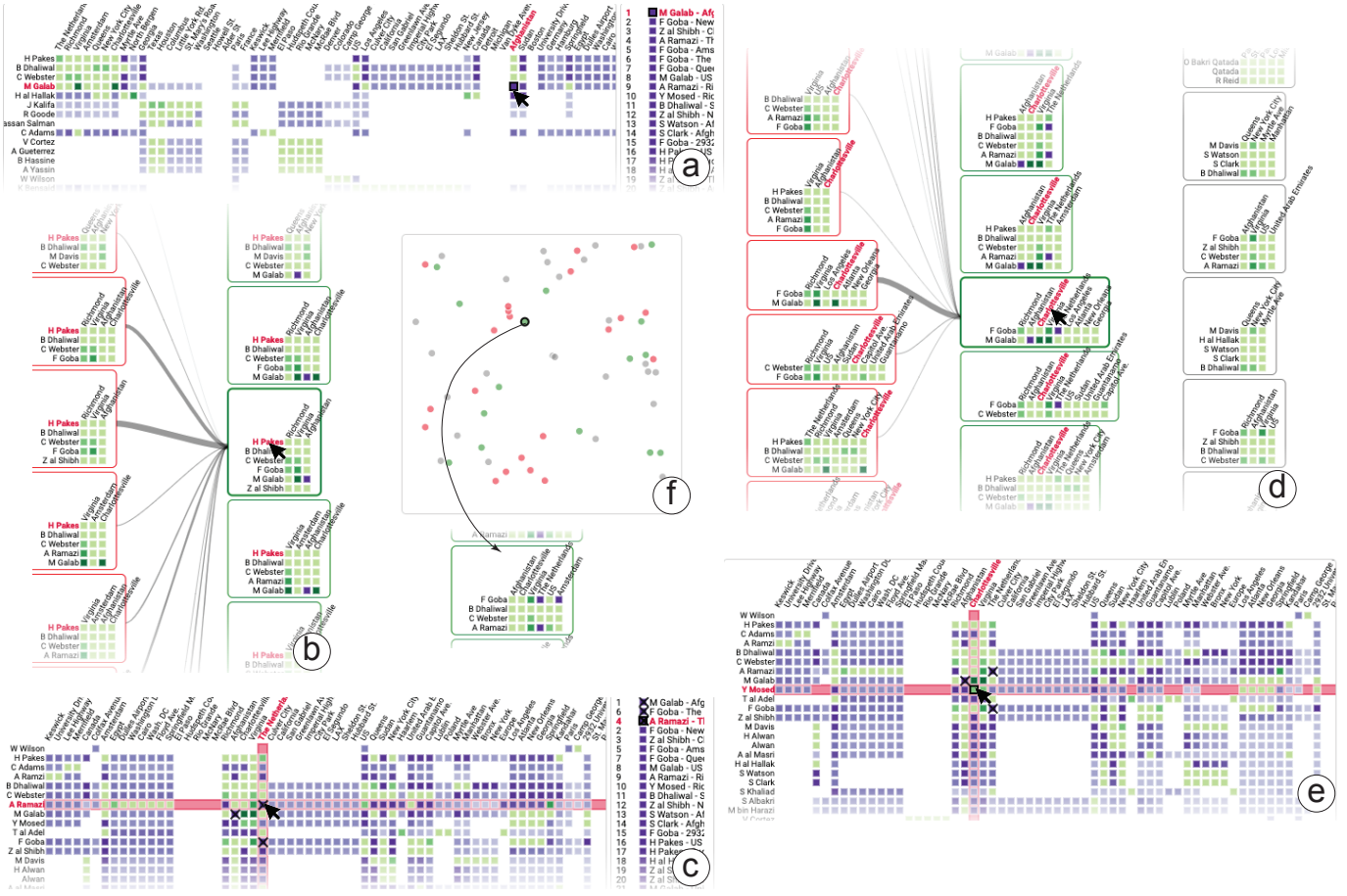


Fig. 7. An officer is analyzing the intelligence reports of *The Sign of the Crescent* dataset using MissBiN.

the prediction scores with bi-cliques, and the worst-case time complexity is $O(2^n n)$ with the MBEA algorithm [13], where $n = \max(|X|, |Y|)$. Further development of the bi-clique detection, such as approximation-based methods, will increase the performance.

Second, while the visual interface of MissBiN offers two main approaches to analyzing networks: metric-based and motif-based, other views need to be developed to support domain-specific tasks. For example, a map of geographical locations is required by E2 to better understand the context of data, and a Document View is developed for the intelligence analysis scenario. However, we strive to design a general visual analysis framework for understanding missing links, and customized views can be easily integrated for different tasks and applications. Also, to facilitate the comparison of different algorithm results, a juxtaposition view can be easily developed based on the original Network View. Moreover, we focus on analyzing nodes in the Metrics View, and metrics on edges can be added easily to support more comprehensive analyses.

Third, while the matrix design in MissBiN is effective in presenting larger and denser networks compared to the node-link diagram [20], it is still an open challenge to visualize extremely large networks. Multi-scale visualizations equipped with aggregation, pre-computation, and focus+context techniques [62], [63], [64] could be used to enhance the scalability of the Network and Motif Views.

Fourth, although we interviewed experts from different domains and provide a comprehensive usage scenario, the evaluation of MissBiN can be enhanced. While the effectiveness of the algorithm has been verified with quantitative experiments, more evaluation such as deployment studies and experiments needs to be conducted to investigate how the tool can be used in real-world settings as well as in other domains.

Although having limitations, MissBiN possesses several key advantages, especially its generalizability for analyzing missing links in any bipartite networks. Both the algorithm and the visual interface do not depend on any domain- or dataset-specific features. On the algorithm side, while we adopt similarity-based methods as the basis, learning-based methods can also be used before the re-weighting with detected bi-cliques. While the structural hole theory that inspired us is discovered in social science, it has become an important foundation in many other fields such as economics and computer science [9]. On the visual interface side, the views are general for displaying both weighted and unweighted networks. However, for unweighted networks, the node-link diagram, which is easier to understand, could be applied to the Motifs Detail View. Further, these views are designed for not depending on any specific meta-data attributes of the nodes or links in the networks. Additional domain-specific views showing this information can be easily integrated to support more complex tasks.

10 Conclusion

We have presented MissBiN, a visual analysis tool for exploring and understanding missing links in bipartite networks. MissBiN offers a novel approach for missing link prediction by using the information of bi-cliques in networks. It can be integrated with a variety of link prediction algorithms. Moreover, MissBiN provides an interactive visualization to present computed missing links and support the investigation of their meaning and influence in the network by comparing networks with and without selected missing links. In order to evaluate MissBiN, quantitative experiments, expert interviews, and a use case were conducted for assessing both the algorithm and the visualization. Results indicated that our algorithm outperforms the corresponding baselines, and MissBiN, as an integrated system consisting of the algorithm and interactive visualization, is useful for understanding missing information in different applications.

Acknowledgments

This research is supported in part by the NSERC Discovery Grant and NSF Grant IIS-1850036 and IIS-2002082. Part of the work was completed while the authors were at FXPAL.

References

[1] C. Carrubba, M. Gabel, and S. Hug, "Legislative voting behavior, seen and unseen: A theory of roll-call vote selection," *Legislative Studies Quarterly*, vol. 33, no. 4, pp. 543–572, 2008.

[2] S. C. Madeira and A. L. Oliveira, "Biclustering algorithms for biological data analysis: A survey," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 1, no. 1, pp. 24–45, 2004.

[3] H. Wu, J. Vreeken, N. Tatti, and N. Ramakrishnan, "Uncovering the plot: detecting surprising coalitions of entities in multi-relational schemas," *Data Mining and Knowledge Discovery*, vol. 28, no. 5-6, pp. 1398–1428, 2014.

[4] D. Liben-Nowell and J. Kleinberg, "The link prediction problem for social networks," in *Proceedings of the Twelfth International Conference on Information and Knowledge Management*, 2003, pp. 556–559.

[5] C. V. Cannistraci, G. Alanis-Lobato, and T. Ravasi, "From link-prediction in brain connectomes and protein interactomes to the local-community-paradigm in complex networks," *Scientific Reports*, vol. 3, no. 1, 2013.

[6] Z. Huang, X. Li, and H. Chen, "Link prediction approach to collaborative filtering," in *Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries*, 2005, pp. 141–142.

[7] J. Zhao, F. Chen, and P. Chiu, "A generic visualization framework for understanding missing links in bipartite networks," in *SIGGRAPH Asia 2018 Posters*, 2018.

[8] J. Zhao, M. Sun, F. Chen, and P. Chiu, "Missbin: Visual analysis of missing links in bipartite networks," in *IEEE Visualization Conference (VIS)*, 2019, pp. 71–75.

[9] M. S. Granovetter, "The strength of weak ties," *American Journal of Sociology*, vol. 78, no. 6, pp. 1360–1380, 1973.

[10] R. S. Burt, *Structural Holes: the Social Structure of Competition*. Harvard University Press, 1995.

[11] V. Martínez, F. Berzal, and J.-C. Cubero, "A survey of link prediction in complex networks," *ACM Computing Surveys*, vol. 49, no. 4, pp. 1–33, 2016.

[12] C. Prell, *Social Network Analysis: History, Theory and Methodology*. SAGE Publications, 2011.

[13] Y. Zhang, C. A. Phillips, G. L. Rogers, E. J. Baker, E. J. Chesler, and M. A. Langston, "On finding bicliques in bipartite graphs: a novel algorithm and its application to the integration of diverse biological data types," *BMC Bioinformatics*, vol. 15, no. 1, p. 110, 2014.

[14] P. Bonacich, "Power and centrality: A family of measures," *American Journal of Sociology*, vol. 92, no. 5, pp. 1170–1182, 1987.

[15] S. P. Borgatti, "Two-mode concepts in social network analysis," in *Computational Complexity*. Springer, 2012, pp. 2912–2924.

[16] A. S. Asratian, T. M. J. Denley, and R. Häggkvist, *Bipartite Graphs and their Applications*. Cambridge University Press, 1998.

[17] T. Uno, T. Asai, Y. Uchida, and H. Arimura, "An efficient algorithm for enumerating closed patterns in transaction databases," in *International Conference on Discovery Science*, 2004, pp. 16–31.

[18] Y. Kluger, "Spectral biclustering of microarray data: Coclustering genes and conditions," *Genome Research*, vol. 13, no. 4, pp. 703–716, 2003.

[19] R. Keller, C. M. Eckert, and P. J. Clarkson, "Matrices or node-link diagrams: Which visual representation is better for visualising connectivity models?" *Information Visualization*, vol. 5, no. 1, pp. 62–76, 2006.

[20] M. Ghoniem, J.-D. Fekete, and P. Castagliola, "On the readability of graphs using node-link and matrix-based representations: A controlled experiment and statistical analysis," *Information Visualization*, vol. 4, no. 2, pp. 114–135, 2005.

[21] K. Misue, "Anchored maps: Visualization techniques for drawing bipartite graphs," in *Human-Computer Interaction. Interaction Platforms and Techniques*. Heidelberg, 2007, vol. 4551, pp. 106–114.

[22] J. Stasko, C. Görg, and Z. Liu, "Jigsaw: Supporting investigative analysis through interactive visualization," *Information visualization*, vol. 7, no. 2, pp. 118–132, 2008.

[23] H.-J. Schulz, M. John, A. Unger, and H. Schumann, "Visual Analysis of Bipartite Biological Networks," in *Eurographics Workshop on Visual Computing for Biomedicine*, C. Botha, G. Kindlmann, W. Niessen, and B. Preim, Eds., 2008.

[24] J. Abello, S. G. Kobourov, and R. Yusufov, "Visualizing large graphs with compound-fisheye views and treemaps," in *Graph Drawing*. Springer, 2005, pp. 431–441.

[25] C. Partl, A. Lex, M. Streit, H. Strobelt, A.-M. Wassermann, H. Pfister, and D. Schmalstieg, "Contour: Data-driven exploration of multi-relational datasets for drug discovery," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 12, pp. 1883–1892, 2014.

[26] M. Dumas, J. Robert, and M. J. Mcguffin, "Alertwheel: radial bipartite graph visualization applied to intrusion detection system alerts," *IEEE Network*, vol. 26, no. 6, pp. 12–18, 2012.

[27] M. Sun, P. Mi, C. North, and N. Ramakrishnan, "Biset: Semantic edge bundling with biclusters for sensemaking," *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, no. 1, pp. 310–319, 2016.

[28] H. Wu, M. Sun, P. Mi, N. Tatti, C. North, and N. Ramakrishnan, "Interactive discovery of coordinated relationship chains with maximum entropy models," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 12, no. 1, p. 7, 2018.

[29] M. Sun, J. Zhao, H. Wu, K. Luther, C. North, and N. Ramakrishnan, "The effect of edge bundling and seriation on sensemaking of biclusters in bipartite graphs," *IEEE Transactions on Visualization and Computer Graphics*, 2018.

[30] M. Sun, D. Koop, J. Zhao, C. North, and N. Ramakrishnan, "Interactive bicluster aggregation in bipartite graphs," in *IEEE Visualization Conference (VIS)*, 2019.

[31] R. Santamaría, R. Therón, and L. Quintales, "Bicoverlapper: A tool for bicluster visualization," *Bioinformatics*, vol. 24, no. 9, p. 1212, 2008.

[32] G. A. Grothaus, A. Mufti, and T. Murali, "Automatic layout and visualization of biclusters," *Algorithms for Molecular Biology*, vol. 1, no. 1, pp. 1–15, 2006.

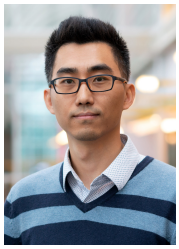
[33] J. Heinrich, R. Seifert, M. Burch, and D. Weiskopf, "Bicluster viewer: A visualization tool for analyzing gene expression data," in *Proceedings of the International Symposium Advances in Visual Computing*, 2011, pp. 641–652.

[34] M. Kapushesky, P. Kemmeren, A. C. Culhane, S. Durinck, J. Ihmels, C. Korner, M. Kull, A. Torrente, U. Sarkans, J. Vilo, and A. Brazma, "Expression profiler: next generation–an online platform for analysis of microarray data," *Nucleic Acids Research*, vol. 32, pp. W465–W470, 2004.

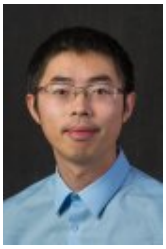
[35] R. Santamaria, R. Theron, and L. Quintales, "BicOverlapper 2.0: visual analysis for gene expression," *Bioinformatics*, vol. 30, no. 12, pp. 1785–1786, 2014.

[36] J. Zhao, M. Sun, F. Chen, and P. Chiu, "BiDots: Visual exploration of weighted biclusters," *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, no. 1, pp. 195–204, 2018.

[37] N. Henry, J.-D. Fekete, and M. McGuffin, "NodeTrix: a hybrid visualization of social networks," *IEEE Transactions on Visualization and Computer Graphics*, vol. 13, no. 6, pp. 1302–1309, 2007.
[38] M. Streit, S. Gratzl, M. Gillhofer, A. Mayr, A. Mitterecker, and S. Hochreiter, "Furby: fuzzy force-directed bicluster visualization," *BMC Bioinformatics*, vol. 15, no. Suppl 6, p. S4, 2014.
[39] P. Fiaux, M. Sun, L. Bradel, C. North, N. Ramakrishnan, and A. Endert, "Bixplorer: Visual analytics with biclusters," *Computer*, no. 8, pp. 90–94, 2013.
[40] P. Xu, N. Cao, H. Qu, and J. Stasko, "Interactive visual co-cluster analysis of bipartite graphs," in *Proceedings of IEEE Pacific Visualization Symposium*, 2016.
[41] A. Lex, M. Streit, C. Partl, K. Kashofer, and D. Schmalstieg, "Comparative analysis of multidimensional, quantitative data," *IEEE Transactions on Visualization and Computer Graphics*, vol. 16, no. 6, pp. 1027–1035, 2010.
[42] A. Lex, H.-J. Schulz, M. Streit, C. Partl, and D. Schmalstieg, "VisBricks: Multiform visualization of large, inhomogeneous data," *IEEE Transactions on Visualization and Computer Graphics*, vol. 17, no. 12, pp. 2291–2300, 2011.
[43] P. Wang, B. Xu, Y. Wu, and X. Zhou, "Link prediction in social networks: the state-of-the-art," *Science China Information Sciences*, vol. 58, no. 1, pp. 1–38, 2015.
[44] M. A. Hasan and M. J. Zaki, "A survey of link prediction in social networks," in *Social Network Data Analytics*. Springer, 2011, pp. 243–275.
[45] L. Lü and T. Zhou, "Link prediction in complex networks: A survey," *Physica A: Statistical Mechanics and its Applications*, vol. 390, no. 6, pp. 1150–1170, 2011.
[46] R. N. Lichtenwalter and N. V. Chawla, "Vertex collocation profiles: Subgraph counting for link analysis and prediction," in *Proceedings of the 21st International Conference on World Wide Web*, 2012, pp. 1019–1028.
[47] S. Scellato, A. Noulas, and C. Mascolo, "Exploiting place features in link prediction on location-based social networks," in *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2011, pp. 1046–1054.
[48] T. Wohlfarth and R. Ichise, "Semantic and event-based approach for link prediction," in *Practical Aspects of Knowledge Management*, T. Yamaguchi, Ed., 2008, pp. 50–61.
[49] K. Yu and W. Chu, "Gaussian process models for link analysis and transfer learning," in *Advances in Neural Information Processing Systems*, J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, Eds. Curran Associates, Inc., 2008, pp. 1657–1664.
[50] D. Heckerman, C. Meek, and D. Koller, *Probabilistic Entity-Relationship Models, PRMs and Plate Models*. MIT Press, 2004, pp. 201–239.
[51] Y.-J. Chang and H.-Y. Kao, "Link prediction in a bipartite network using wikipedia revision information," in *Proceedings of the Conference on Technologies and Applications of Artificial Intelligence*, 2012.
[52] S. Xia, B. Dai, E.-P. Lim, Y. Zhang, and C. Xing, "Link prediction for bipartite social networks: The role of structural holes," in *Proceedings of International Conference on Advances in Social Networks Analysis and Mining*, 2012.
[53] M. Brehmer and T. Munzner, "A multi-level typology of abstract visualization tasks," *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 12, pp. 2376–2385, 2013.
[54] H. Tong, C. Faloutsos, and J. yu Pan, "Fast random walk with restart and its applications," in *Proceedings of International Conference on Data Mining*, 2006.
[55] F. Hughes and D. Schum, "Discovery, proof, choice: the art and science of the process of intelligence analysis-preparing for the future of intelligence analysis," *Washington, DC: Joint Military Intelligence College*, 2003.
[56] P. Isenberg, F. Heimerl, S. Koch, T. Isenberg, P. Xu, C. D. Stolper, M. M. Sedlmair, J. Chen, T. Möller, and J. Stasko, http://www.vispubdata.org/site/vispubdata/, 2018.
[57] J. Davis and M. Goadrich, "The relationship between precision-recall and roc curves," in *Proceedings of the 23rd International Conference on Machine Learning*, 2006, pp. 233–240.
[58] N. H. Riche, B. Lee, and C. Plaisant, "Understanding interactive legends: a comparative evaluation with standard widgets," *Computer Graphics Forum*, vol. 29, no. 3, pp. 1193–1202, 2010.
[59] F. Beck, M. Burch, S. Diehl, and D. Weiskopf, "A taxonomy and survey of dynamic graph visualization," *Computer Graphics Forum*, vol. 36, no. 1, pp. 133–159, 2016.
[60] W. S. Torgerson, "Multidimensional scaling: I. theory and method," *Psychometrika*, vol. 17, no. 4, pp. 401–419, 1952.
[61] Washington DC Neighborhood Clusters, https://www.neighborhoodinfodc.org/nclusters/nclusters.html, 2018.
[62] Y. Hu and L. Shi, "Visualizing large graphs," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 7, no. 2, pp. 115–136, 2015.
[63] L. Lins, J. T. Klosowski, and C. Scheidegger, "Nanocubes for real-time exploration of spatiotemporal datasets," *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 12, pp. 2456–2465, 2013.
[64] J. Abello, J. Korn, and M. Kreuseler, "Navigating giga-graphs," in *Proceedings of the Working Conference on Advanced Visual Interfaces*, 2002.



Jian Zhao is an assistant professor with the School of Computer Science, University of Waterloo. His research interests include Information Visualization, Human-Computer Interaction, Visual Analytics and Data Science. His work contributes to the development of advanced interactive visualizations that promote the interplay of human, machine, and data.



Maoyuan Sun is an assistant professor with the Department of Computer Science, Northern Illinois University. His research falls in areas of visual analytics, information visualization, and human computer interaction, with applied domains in intelligence analysis, business intelligence, cyber security, and STEM education.

Fraincine Chen is a principal research scientist at FXPAL. Her research interests are in applied machine learning, with a focus around developing methods for extracting, organizing and helping users to make better use of information in different types of media, including text, images, video, audio and imaged text.

Patrick Chiu is a principal research scientist at FXPAL. His current research interests include multimedia applications and content analysis, human-computer interaction, and ubiquitous computing.