Applied Network Science

RESEARCH Open Access

Improving accuracy of expected frequency of uncertain roles based on efficient ensembling

Soshi Naito and Takayasu Fushimi[*]

*Correspondence:
takayasu.fushimi@gmail.com

School of Computer Science,
Tokyo University of Technology,
1404-1 Katakura-machi,
Hachioji-city, Tokyo 192–0982,
Japan

Abstract

This study tackles the problem of extracting the node roles in uncertain graphs based on network motifs. Uncertain graphs are useful for modeling information diffusion phenomena because the presence or absence of edges is stochastically determined. In such an uncertain graph, the node role also changes stochastically according to the presence or absence of edges, so approximate calculation using a huge number of samplings is common. However, the calculation load is very large, even for a small graph. We propose a method to extract uncertain node roles with high accuracy and high speed by ensembling a large number of sampled graphs and efficiently searching for all other transitionable roles. This method provides highly accurate results compared to simple sampling and ensembling methods that do not consider the transition to other roles. In our evaluation experiment, we use real-world graphs artificially assigned uniform and non-uniform edge existence probabilities. The results show that the proposed method outperforms an existing method previously reported by the authors, which is the basis of the proposed method, as well as another current method based on the state-of-the-art algorithm, in terms of efficiency and accuracy.

Introduction

In network science, counting the number of motifs, which are small subgraphs, is an important task for understanding the characteristics of a graph. This technique has been studied for many years, starting with the work of Milo et al. (2002), Wernicke (2005), Itzhack et al. (2007), Grochow and Kellis (2007), Ahmed et al. (2015), Pinar et al. (2017). There has also been considerable research on extending the concept of a motif. A motif-role that defines the role of each node based on structural equivalence in motifs has been proposed, and promising results obtained using it have been reported (Ohnishi et al. 2010; McDonnell et al. 2014). Figure 1 shows 13 motifs and 30 roles for a directed 3-node subgraph. In a directed graph, a connected subgraph consisting of three nodes is classified into one of 13 patterns based on the graph's isomorphism. Furthermore, a node of the directed graph is classified into one of 30 patterns based on the structural equivalence in the motif. Following existing research, this paper defines the node's role as the position of its appearance in subgraphs as shown in Fig. 1. Using motif-based roles, we can infer, for example, that the node appearing as Role 13 plays the role of transmitting information and that the node appearing as Role 24 plays the role of transmitting

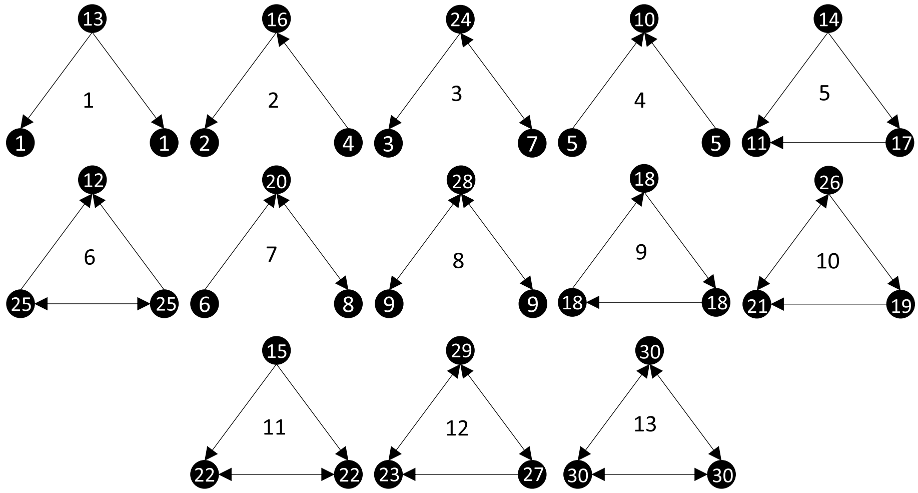


Fig. 1 13 motifs and 30 roles for directed 3-node subgraphs

received information to other nodes. Accordingly, extracting the motif-based role of each node can be applied, for example, in identifying important influencers in viral marketing.

Information diffusion over social networks can be treated as an uncertain graph, where the existence of edges between nodes is probabilistic. In the last few years, the study of uncertain graphs has attracted considerable attention in the field of network science. The counting of motifs and roles in uncertain graphs can facilitate a more detailed analysis of a given graph, and it is expected to be used in a wide range of fields such as marketing, urban planning, and protein analysis. For uncertain graphs with L uncertain edges, 2^L possible graphs need to be enumerated; in addition, the number of motifs (roles) for each of them needs to be counted, and the numbers with the weight of the occurrence probability of each possible graph need to be averaged. However, the number of possible graphs is very large, and even for small graphs it is difficult to compute the exact expectation. Therefore, in general, sampling-based approximations have been adopted.

The LINC algorithm of Ma et al. (2019) is a state-of-the-art technique for counting motifs in uncertain graphs. Instead of counting the number of motifs for all sample graphs from scratch, LINC focuses on the structural similarity between sample graphs, and it efficiently updates the number of motifs by considering only the difference edges between two sample graphs. In a situation of low uncertainty, that is, extremely high or low edge probability, LINC can compute the expected frequency more quickly than can naive sampling-based methods.

The aim of this study is to extract groups of nodes with similar motif-based roles, and promising results to this end have already been reported (Pržulj 2007; Guerrero et al. 2008; Ohnishi et al. 2010; McDonnell et al. 2014; Sarajlić et al. 2016). Therefore, this study follows these frameworks, which consist of three steps: counting motifs or roles and constructing feature vectors, calculating node similarity, and clustering nodes. In the context of an uncertain graph, we need to sample and ensemble either graphs, vectors, similarities, or clusters. It is important to evaluate the amounts of difference among the exact clustering results obtained by processing all possible graphs, depending on the steps at which sampling and ensembling are performed. In our previous study (Naito and Fushimi 2021), we proposed an efficient ensemble method, graph-ensemble, to ensemble possible graphs sampled from the given uncertain graph; it then generates a weighted graph we call an ensembled-graph, where the edge weight is the ratio of graphs with edge existence to the total number of sampled graphs; finally, the method counts the roles from this weighted graph by considering the edge weights. Experimental evaluations have compared the vector-ensemble and similarity-ensemble methods, both derived from the LINC algorithm, with the graph-ensemble method. The results show that the graph-ensemble method outputs similar results to the previous methods but much faster. On the other hand, the results of subsequent experiments show that the error of the graph-ensemble methods in presenting exact results is, to some extent, larger than the vector-ensemble and the similarity-ensemble methods. This is because the graph-ensemble method integrates sampled graphs and counts the roles from it; consequently, absent edges in some samples are eliminated by present edges in other samples and changes in the motif-roles cannot be considered. The vector-ensemble and similarity-ensemble methods count roles from each sampled graph, and thus changes in motif roles can be considered. Therefore, there is need for a method that is as efficient as the graph-ensemble method but also as effective as the vector-ensemble and similarity-ensemble methods. In this study, we improve the error by making it as small as that of the vector-ensemble method. This is done by considering the change in role due to the probabilistic absence of edges at the expense of a certain degree of speed that is possible with the graph-ensemble method.

As an extension of the conference version of this work (Naito and Fushimi 2021), we propose the extended graph-ensemble method, add graphs to illustrate our experiments, and compare and evaluate the proposed method along with existing methods from the viewpoint of error. Furthermore, the pseudo-code related to our method is added.

This paper is organized as follows: "Related work" section introduces related research. "Problem framework" section sets up the problem addressed in this study. "Existing methods" and "Proposed method: extended graph-ensemble method" sections describe the existing and proposed methods, and "Experimental evaluations" section presents evaluation experiments using each method. Finally, "Conclusion" section summarizes this study and mentions future work.

Related work

In this study, we consider the problem of extracting motif-based roles for uncertain graph nodes. Therefore, we briefly discuss related work in terms of network motifs, role extraction, and uncertain graphs.

Network motifs

Motif counting techniques have been studied for many years, starting with the pioneering work of Milo et al. (2002). Various algorithms in these techniques have been developed for different purposes (Wernicke 2005; Itzhack et al. 2007; Grochow and Kellis 2007; Ahmed et al. 2015; Pinar et al. 2017). Wernicke proposed a hash-based algorithm called ESU, which avoided the need for storing all subgraphs in a hash table and improved the efficiency of motif counting by not counting the same subgraph twice

(Wernicke 2005). Itzhack et al. proposed an efficient algorithm to traverse a breadth-first search tree with the target node as the root. It represents the existence of a link in a subgraph as a bit string, and it can efficiently identify motif patterns without checking the isomorphism of each subgraph (Itzhack et al. 2007). This study adopts the algorithm of Itzhack et al. for motif counting from sample graphs. Grochow and Kellis proposed an efficient algorithm for searching for a single motif (Grochow and Kellis 2007). This algorithm constructs a partial mapping from a particular graph to a target motif. In addition, the algorithm introduces a method called symmetric-break to avoid multiple counting of motifs, which greatly improves execution time. Ahmed et al. proposed a parallel algorithm for three- and four-node motifs that does not enumerate all motif instances but counts certain motifs, such as cliques and cycles, and uses the transition relations between motifs to compute all other motifs analytically (Ahmed et al. 2015). Pinar et al. proposed a divide-and-conquer algorithm that identifies the substructure of each found subgraph and divides it into smaller ones. Pinar et al. (2017). However, although it is a very efficient method, it cannot be applied to directed networks.

Role extraction

Extracting node roles from a network is an important research topic. Role extraction methods are largely divided into two types, graph-based and feature-based methods (Rossi and Ahmed 2015). Graph-based methods, such as concept and extraction algorithms of regular equivalence (Everett and Borgatti 1994) and structural equivalence (Lorrain and White 1971) have been proposed. These concepts focus on local structures such as relationships among neighboring nodes similar to network motifs, but extracting exactly equivalent nodes is costly. More recently, by relaxing the concept of equivalence, many feature-based role discovery techniques have been proposed (Henderson et al. 2011, 2012; Rossi et al. 2012, 2013; Gilpin et al. 2013).

Feature-based methods transform the graph representation into a feature representation, so in that sense, our method belongs to this category. Some studies defined the motif-based roles (a.k.a orbits) and graphlet degree vector for each node, whose element is the number of roles (Pržulj 2007; Guerrero et al. 2008; McDonnell et al. 2014). Przulj constructed a vector of 73 kinds of orbits obtained from 2- to 5-node graphlets and attempted to quantify the similarity among graphs or nodes (Pržulj 2007). McDonnell et al. proposed a transformation matrix from motif-frequency vector to role-frequency vector to efficiently compute the number of roles for each node or the whole graph (McDonnell et al. 2014). Our study also defines the feature vector of each node based on the number of roles of each node, but we count the number of roles based on Itzhack's algorithm, not McDonnell's one.

Furthermore, some methods calculated the similarity between the vectors and clustered the nodes into groups. Ohnishi et al. analyzed an inter-firm network using motif-roles and found economically meaningful clusters of nodes (Ohnishi et al. 2010). Sarajlic et al. discovered the core–broker–periphery structure from world trade networks and predicted the economic attributes of each country node (Sarajlić et al. 2016). Following the promising results of the above studies, our role extraction framework consists of counting roles, constructing feature vectors, calculating node-similarity, and clustering nodes.

Uncertain graphs

Research on uncertain graphs has been pursued in a wide range of contexts. One important task is the extension of existing graph analysis methods, including node centrality, clustering, embedding, and motif counting, to uncertain graphs.

Pfeiffer et al. extended certain representative structural indices for the deterministic graph, i.e., shortest path length, clustering coefficient, and betweenness centrality ranking, to uncertain graphs by introducing the expected value of each index for the occurrence probability of each possible graph (Pfeiffer and Neville 2011). Such a notion and sampling-based approximation have been widely used in subsequent research on uncertain graphs, including the work in this study.

Ceccarello et al. developed a node clustering method for uncertain graphs and reduced the basic problem to *k*-center and *k*-median problems (Ceccarello et al. 2017). In this method, the distances between nodes are defined by the inverse of the connection probability among them, which is efficiently and accurately estimated by the Monte Carlo sampling method.

Hu et al. proposed an embedding method for uncertain graphs, which constructs a matrix of expected proximities of all node pairs in an uncertain graph and reduces the number of the matrix dimensionality via a matrix factorization technique to obtain low-dimensional vectors for the nodes [10]. This method uses the Jaccard coefficient for the set of adjacent nodes when calculating the expected proximity between nodes, that is, it calculates the similarity between nodes based on the local structure. Similarly, our method constructs vectors based on the expected number of motif roles, which represents the local structure. The procedure is reversed because the purpose of Hu's method of obtaining a low-dimensional vector from the similarity between nodes and that of our method of obtaining a similarity matrix from a low-dimensional vector is different.

Motif counting for uncertain graphs has not yet been thoroughly studied. The following are some of the major studies on the subject. Tran et al. proposed a method to compute an unbiased estimator of the number of motifs from noisy and incomplete data, but the method assumes that all edges have uniform joint probabilities and does not apply to non-uniform probabilities (Tran et al. 2013).

Ma et al. proposed two sampling-based algorithms to obtain basic statistics such as the mean, variance, and probability distribution of motif counts (Ma et al. 2019). The first is a simple sampling method, called PGS, which samples a large number of possible graphs from uncertain graphs and counts the instances of a single motif from each sample graph. However, the method requires a sufficient number of samples to accurately estimate the average number of motifs based on Hoeffding's inequality. The second, more efficient method, called LINC, uses the structural similarity between sample graphs to update the frequency of motifs by examining only edge differences between consecutive samples. It outputs the same results as PGS but runs much faster when the same samples are used. In this work, we consider the LINC algorithm a state-of-the-art technique and propose a more efficient ensemble algorithm than those equipped with a role counting routine by LINC.

Table 1 Notation

Notation	Description
$G = (V, E)$	Deterministic graph or backbone graph
V, E	Sets of nodes and edges
$N = \|V\|, L = \|E\|$	Numbers of nodes and edges
$g = (U, F)$	Motif, or connected subgraph
$\bar{d} = L/N$	Average degree of each node
$\mathbf{R}, \mathcal{R}, \bar{\mathbf{R}}$	R-dimensional vectors of N nodes, i.e., $N \times R$ matrix
R	Number of role patterns
$\mathbf{C}, \mathcal{C}, \bar{\mathbf{C}}$	$N \times N$ similarity matrix
$\mathbf{H}, \mathcal{H}, \bar{\mathbf{H}}$	$N \times K$ affiliation matrix
K	Number of role clusters
$\Gamma(v)$	Set of adjacent nodes of node v
$\mathcal{G} = (G, \mathbf{p})$	Uncertain graph
$p(e), p$	Edge-existence probability
$G_s = (V, E_s)$	Sampled graph of an uncertain graph
S	Number of samples
$\Pr[G]$	Occurrence probability of graph G
$\delta()$	Kronecker delta function
$\mathcal{D}_{s,s'}$	Set of edges that appear in G_s but not in $G_{s'}$, and vice versa.

Problem framework

This study deals with the problem of extracting node groups with similar motif-roles in uncertain graphs. For convenience of explanation, the case of role extraction based on a motif composed of three nodes and directed edges is described here; however, the method can be applied to role extraction based on a small k node motif that is not limited to $k = 3$, regardless of whether it is directed or undirected. Table 1 summarizes the nomenclature used in this paper. As for R, C, H, the calligraphic font of the capital letter represents the role vectors, the similarity matrix, and the affiliation matrix of an uncertain graph; the bold font with subscript represents those of a deterministic graph sampled from an uncertain graph; the bold font with over-bar shows those of the ensembled (averaged) version.

Motif-role extraction

First, we formulate the problem of extracting a motif-role from the deterministic graph, $G = (V, E)$. Here, V is a set of nodes, E is a set of edges, $N = |V|$ is the number of nodes, and $L = |E|$ is the number of edges. A motif is a graph with a few nodes and edges among them, and it is considered a building block of a large graph. For a set of nodes $U \subset V$ and edges among them $F = (U \times U) \cap E$, we define $g = (U, F)$ as a motif when g is a connected graph. In this study, we focus on the directed 3-node motif, i.e., $|U| = 3$ and $|F| \leq 6$. The number of patterns of edge-existence states between all pairs of U is $2^{|F|} = 2^6$. Among these, 54 patterns are connected ones, and by coordinating them according to the graph-isomorphism, the number of subgraph patterns is 13, i.e., the number of patterns of a directed 3-node motif is 13 as shown in Fig. 1.

Role was first defined as the structural equivalence in the graph, and role discovery as any process that divides nodes into classes of structurally equivalent nodes (Lorrain and

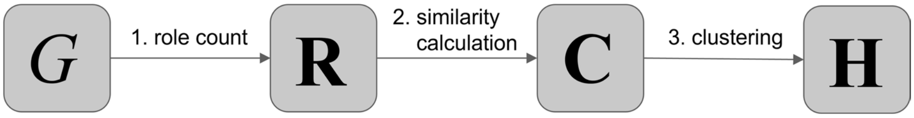


Fig. 2 Procedure of role extraction

White 1971). Relaxing this definition, in this study, according to the study by McDonnell et al. (2014), the role is defined based on the structural equivalence in the motif. In the directed 3-node motif, there are 30 types of roles as shown in Fig. 1. In this study, role extraction is accomplished by the following three steps: 1) constructing the role vector for each node, 2) calculating the similarity between role vectors for all node pairs, and 3) extracting node groups based on similarity (see Fig. 2). In the construction of the role vector for each node in step 1, the numbers of roles R are counted for each node v, and the appearance frequency is arranged in the R dimension vector $\mathbf{r}_v$. The ith element in $\mathbf{r}_v$ represents the number of times the node v appears as role i. In the case of the directed 3-node motif, the number of role types is $R = 30$ as shown in Fig. 1. The matrix in which the role vectors of all N nodes are arranged is expressed as $\mathbf{R} = [\mathbf{r}_1, \ldots, \mathbf{r}_N]^T$, where $\mathbf{r}^T$ represents the transpose of $\mathbf{r}$. In step 2, the cosine similarity between role vectors $c_{u,v} = \frac{\mathbf{r}_u^T \mathbf{r}_v}{||\mathbf{r}_u||||\mathbf{r}_v||}$ is used to calculate the similarity of all node pairs. Let the similarity of all $N \times N$ node pairs be the similarity matrix $\mathbf{C} = [c_{u,v}]_{u \in V, v \in V}$. In step 3, all nodes are classified into K clusters by the greedy method of k-medoids clustering (Nemhauser et al. 1978), which outputs the affiliation matrix $\mathbf{H} = [h_{u,k}]_{u \in V, K=1}^{K}$, where $h_{u,k} = 1$ if node u belongs to cluster k, otherwise $h_{u,k} = 0$. In this way, the role extraction process outputs K clusters, each of which consists of nodes with similar role vectors.

Uncertain graph

This study targets uncertain graphs, in which the existence of edges between nodes is probabilistically determined. The uncertain graph $\mathcal{G} = (G, p)$ is defined by the backbone graph $G = (V, E)$, consisting of the node set V and the edge set E, and the existence probability of each edge $p : E \to (0, 1]$. Since the uncertain graph can be expressed as a set of its possible graphs, it is expressed as $\mathcal{G} = \{G_i = (V, E_i); E_i \subseteq E\}$. Assuming that the number of uncertain edges is L, the number of possible graphs in the uncertain graph is $2^L = |\mathcal{G}|$. Following the related study, the occurrence probability $\Pr[G_i]$ for each possible graph G_i is calculated based on independent Bernoulli trials for all edges:

$$\Pr[G_i] = \prod_{e \in E_i} p(e) \prod_{e \in E \setminus E_i} (1 - p(e)).$$

Motif-role extraction in uncertain graph

Next, we formulate the problem of extracting the motif-role from the uncertain graph $\mathcal{G}$. To solve the above role extraction problem exactly for uncertain graphs, it is necessary to perform the three previously listed steps for all possible graphs $\mathcal{G} = \{G_i = (V, E_i); E_i \subseteq E\}$ and ensemble the results in consideration of the occurrence probability $\Pr[G_i]$ of each possible graph G_i as follows:

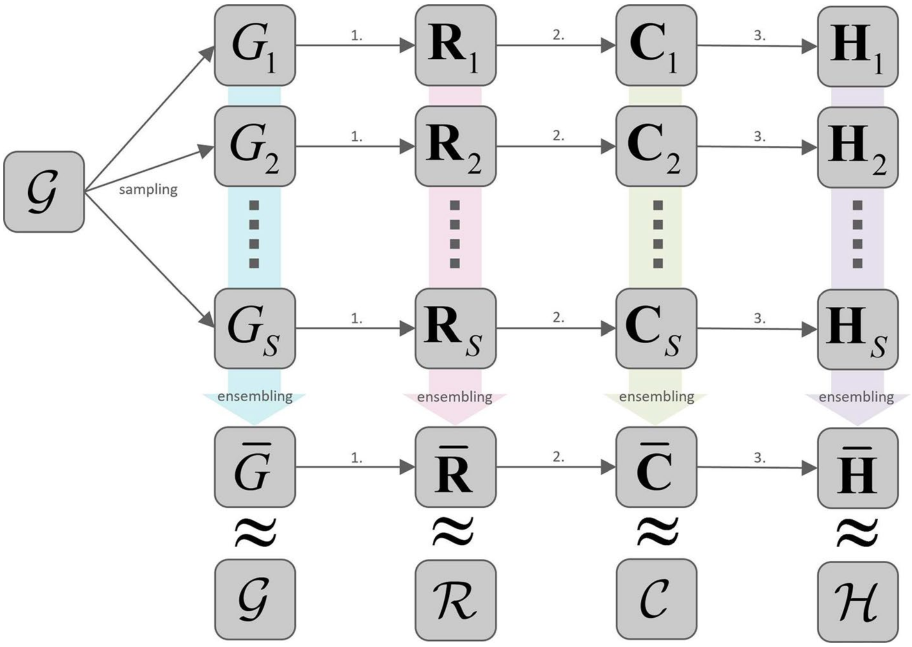


Fig. 3 Four ensemble methods

$$\mathcal{H} = \underset{G \in \mathcal{G}}{\Phi} (\mathbf{H}_G;\ \Pr[G]).$$

Here, Φ is an operator of the ensemble, and it indicates that the clustering result $\mathbf{H}_G$ of each possible graph G is ensembled in consideration of the weight of the occurrence probability $\Pr[G]$. To obtain an exact ensemble result for an uncertain graph with L uncertain edges, sampling and ensembling are required for the number of possible graphs 2^L; this process is difficult to implement even for a small graph. Therefore, approximation by sampling is generally adopted.

Existing methods

This section describes four ensemble methods that sample possible graphs from an uncertain graph and output clustering results. As shown in Fig. 3, four ensemble methods use S possible graphs, $\{G_1, \ldots, G_S\}$, sampled from the given uncertain graph $\mathcal{G}$. The graph-ensemble method ensembles sampled graphs, generates a weighted graph $\bar{G}$, and counts motif-roles from the weighted graph. The vector-ensemble method ensembles role vectors $\{\mathbf{R}_1, \ldots, \mathbf{R}_S\}$ obtained from each sampled graph and generates an averaged role matrix (vectors) $\bar{\mathbf{R}}$. The similarity-ensemble method ensembles similarity matrices $\{\mathbf{C}_1, \ldots, \mathbf{C}_S\}$ calculated from each role matrix (vectors) and generates an averaged similarity matrix $\bar{\mathbf{C}}$. The cluster-ensemble method ensembles affiliation matrices $\{\mathbf{H}_1, \ldots, \mathbf{H}_S\}$ obtained from each similarity matrix and generates an ensembled affiliation matrix $\bar{\mathbf{H}}$, which is a clustering result. When sampling many graphs, i.e., $S \simeq 2^L$, the ensembled results, $\bar{G}$, $\bar{\mathbf{R}}$, $\bar{\mathbf{C}}$ and $\bar{\mathbf{H}}$ become close to the true results $\mathcal{G}$, $\mathcal{R}$, $\mathcal{C}$ and $\mathcal{H}$. The details of these existing methods are described in the following subsections.

Graph-ensemble method

First, we explain the graph ensemble method proposed in our previous study (Naito and Fushimi 2021). The procedure for outputting the similarity matrix $\bar{C}$ in the graph ensemble method (hereinafter, the GE method) is shown in Algorithm 1. In the GE method, ensembling is performed on a group of sample graphs $\{G_1, \ldots, G_S\}$, $G_s = (V, E_s)$, $E_s \subseteq E$ to generate an ensembled graph $\bar{G}$ (see Algorithm 2):

$$\mathcal{G} = \underset{G\in\mathcal{G}}{\Phi}(G;\ \Pr[G]) \simeq \overset{S}{\underset{s=1}{\Phi}}(G_s;\ 1/S) = \bar{G}.$$

Here, $\bar{G} = (V, \bar{E}, \bar{p})$ is a weighted graph with weights $\bar{p}(e) = \sum_{s=1}^{S} \delta(e \in E_s)/S$, which means the sample probability of edge e appearing in S sample graphs, and $\delta(cond)$ is a Boolean function that returns 1 if the condition *cond* is True and 0 if it is False.

Algorithm 1 Graph-Ensemble method: GE$(\mathcal{G}, S)$

```
Input: $\mathcal{G} = (G, p)$,  $G = (V, E)$,  $S$
Output: $\bar{\mathbf{C}}$
Initialize: $\bar{\mathbf{R}} = [\bar{r}_{u,i}]_{u \in V, 1 \leq i \leq 30} \leftarrow \mathbf{0}$
$\bar{G} \leftarrow$ emsemble_graphs$(\mathcal{G}, S)$
$\Psi \leftarrow$ search_connected_triples$(\bar{G})$
for $G^{(m)} \in \Psi$ do
    $G^{(m)} = (V^{(m)}, E^{(m)}, \mathbf{p}^{(m)})$
    $\bar{\mathbf{R}} \leftarrow$ count_roles$(G^{(m)}, \bar{\mathbf{R}})$
end for
$\bar{\mathbf{C}} \leftarrow$ cosine_similarity$(\bar{\mathbf{R}})$
```

Algorithm 2 Construction of ensembled graph: ensemble_graphs$(\mathcal{G}, S)$

```
Input: $\mathcal{G} = (G, p)$,  $G = (V, E)$,  $S$
Output: $\bar{G} = (V, \bar{E}, \bar{p})$
Initialize: $\forall e \in E$,  $\bar{p}(e) \leftarrow 0$
Initialize: $\bar{E} \leftarrow \emptyset$
for $s = 1 : S$ do                                   ▷ Sample and ensemble graphs
    $G_s = (V, E_s) \leftarrow$ sample_graph$(\mathcal{G})$
    for $e \in E_s$ do
        $\bar{p}(e)$ += $1/S$
    end for
    $\bar{E} \leftarrow \bar{E} \cup \{e\}$
end for
```

Next, for an ensembled graph $\bar{G}$, we search for connected-triples based on the algorithm of Itzhack et al. (2007) (Algorithm 3). In Algorithm 3, $\Gamma(u) = \{v;\ (u, v) \in \bar{E} \wedge (v, u) \in \bar{E}\}$ at Line 6 stands for a set of adjacent nodes of node u, and $\bar{\Gamma}(u)$ at Line 7 is a set of nodes searched for in the for-loop at Line 6. That is, $\Gamma(u) \setminus \bar{\Gamma}(u)$ at Line 8 represents a set of adjacent nodes of node u that are not searched for at Line 6. Then, in the searched for connected-triples $G^{(m)}$, the role of each node is identified and counted in consideration of the weight $\bar{p}(e)$. By aligning the number of roles for each node and regarding it as a vector, we construct $(N \times R)$ role vectors (matrix) $\bar{\mathbf{R}}$ (Algorithm 4). In detail, (1) we represent the presence or absence of 6 edges between the 3 nodes u, v, w of a connected-triple $G^{(m)}$ by the 6-bit bit string $\mathbf{b}_u$ via the motif2bits function; (2) we obtain the role number $i \leftarrow \mathrm{Rcode}(\mathbf{b}_u)$ from the dictionary Rcode, which is a correspondence table between the bit string and the role number; (3) we add an occurrence probability $\Pr[G^{(m)}]$ to the ith element of the role vector of node u, $\bar{r}_{u,i}$, where $\Pr[G^{(m)}] \leftarrow \prod_{e \in E^{(m)}} p(e) \prod_{e \in E \setminus E^{(m)}} (1 - p(e))$ is the occurrence probability of connected-triple $G^{(m)}$ calculated based on the presence/absence of 6 edges and their probabilities of existence. For directed 3-node motifs, by bit-shifting the bit string $\mathbf{b}_u$ focused on node u, the bit strings $\mathbf{b}_v, \mathbf{b}_w$ focused on the other 2 nodes v, w can be obtained. For motifs with more than 3 nodes, this is not a simple bit shift, but the bit string can be obtained in a similar manner.

Algorithm 3 Search connected triples: search_connected_triples(G)

```
Input: $G = (V, E, p)$
Output: $\Psi$
Initialize: $m \leftarrow 0$                                   ▷ Index for connected triples
Initialize: $\Psi \leftarrow \mathrm{list}()$                          ▷ Empty list
for $u \in V$ do
    for $v \in \Gamma(u)$ do
        $\bar{\Gamma}(u) \leftarrow \{v\}$
        for $w \in \big(\Gamma(u) \setminus \bar{\Gamma}(u)\big) \cup \Gamma(v)$ do
            $V^{(m)} \leftarrow \{u, v, w\}$
            $E^{(m)} \leftarrow V^{(m)} \times V^{(m)},\ \mathbf{p}^{(m)} = [p(e)]_{e \in E^{(m)}}$
            $G^{(m)} = (V^{(m)}, E^{(m)}, \mathbf{p}^{(m)})$
            $\Psi[m] \leftarrow G^{(m)}$
            $m \mathrel{+}= 1$
        end for
    end for
end for
```

Algorithm 4 Count and update of #roles: count_roles($G^{(m)}, \bar{\mathbf{R}}$)

```
Input: $G^{(m)} = (V^{(m)}, E^{(m)}, \mathbf{p}^{(m)}), \bar{\mathbf{R}} = [\bar{r}_{u,i}]_{u \in V, 1 \le i \le 30}$
Output: $\bar{\mathbf{R}}$
$V^{(m)} \rightarrow \{u, v, w\}$
$\mathbf{b}_u \leftarrow \mathrm{motif2bits}(G^{(m)})$
$\mathbf{b}_v \leftarrow (\mathbf{b}_u \ll 2 | \mathbf{b}_u \gg 4) \& ((1 \ll 6) - 1)$
$\mathbf{b}_w \leftarrow (\mathbf{b}_v \ll 2 | \mathbf{b}_v \gg 4) \& ((1 \ll 6) - 1)$
$(i, j, k) \leftarrow (\mathrm{Rcode}[\mathbf{b}_u], \mathrm{Rcode}[\mathbf{b}_v], \mathrm{Rcode}[\mathbf{b}_w])$        ▷ Get role code
$\Pr[G^{(m)}] \leftarrow \prod_{e \in E^{(m)}} p(e) \prod_{e \in E \setminus E^{(m)}} (1 - p(e))$
$\bar{r}_{u,i} \mathrel{+}= \Pr[G^{(m)}]$                                 ▷ Increment #roles
$\bar{r}_{v,j} \mathrel{+}= \Pr[G^{(m)}]$
$\bar{r}_{w,k} \mathrel{+}= \Pr[G^{(m)}]$
```

After constructing the role vectors of each node, $\bar{\mathbf{H}}$ is output by classifying each node into clusters based on the matrix $\bar{\mathbf{C}}$, whose elements are the similarity between the role vectors. In this method, the ensembled graph $\bar{G}$ is obtained by ensembling S graphs with L edges. Let p be the average edge existence probability, where the expected number of edges in each sample graph is pL; accordingly, an ensembled graph can be obtained with $O(SpL)$. For one ensembled graph, the connected three nodes are searched for according to the algorithm of Itzhack et al., and the number of roles of all N nodes is counted. Therefore, as with the computational complexity of the 3-node motif count, when the average degree is $\bar{d}$, the ensemble role vectors $\bar{\mathbf{R}}$ is obtained with a computational complexity of $O(N\bar{d}^2)$).

Vector-ensemble method

The role-vector-ensemble method (hereinafter, VE method) generates an ensembled role vector $\bar{\mathbf{R}}$ by averaging the role vector $\{\mathbf{R}_1, \ldots, \mathbf{R}_S\}$ obtained from the sample graph G_s:

$$\mathcal{R} = \underset{G \in \mathcal{G}}{\Phi}(\mathbf{R}_G;\ \Pr[G]) \simeq \frac{1}{S}\sum_{s=1}^{S} \mathbf{R}_s = \bar{\mathbf{R}}.$$

Then, the cosine similarity $\bar{\mathbf{C}}$ is calculated from the obtained ensembled role vector $\bar{\mathbf{R}}$. Each node is divided into 1 of K clusters based on the similarity matrix, and $\bar{\mathbf{H}}$ is output. When constructing the role vector $\mathbf{R}_s$ from each sample graph G_s, the LINC algorithm (Ma et al. 2019), which is the state-of-the-art technique, is used. The LINC algorithm focuses on the difference $D_{s,s'} = (E_s \setminus E_{s'}) \cup (E_{s'} \setminus E_s)$ between the edge sets E_s and $E_{s'}$ in the two sample graphs G_s and $G_{s'}$, and only the number of appearances of the roles related to edge $e \in D_{s,s'}$ whose existence/absence state has changed is updated. Let p be the average edge appearance probability. The expected value of the number of edges that change state is $2L(p - p^2)$; hence, it is effective when the uncertainty is small, such as when $p = 0.1$ or $p = 0.9$. In this way, S role vectors (matrices) $\{\mathbf{R}_1, \ldots, \mathbf{R}_S\}$, each of which is an $(N \times R)$ matrix, are efficiently calculated and averaged to obtain an ensembled role vector $\bar{\mathbf{R}}$. Therefore, if $\bar{m}$ is the average number of motif instances including each edge, the ensembled role vector $\bar{\mathbf{R}}$ is obtained with a computational complexity $O(S(L(p - p^2)\bar{m} + NR))$ by the VE method.

Similarity-ensemble method

In the similarity-ensemble method (hereinafter, SE method), the average of similarity matrices $\{\mathbf{C}_1, \ldots, \mathbf{C}_S\}$ calculated from role vectors $\{\mathbf{R}_1, \ldots, \mathbf{R}_S\}$ is calculated, and the ensembled similarity matrix $\bar{\mathbf{C}}$ is generated:

$$\mathcal{C} = \underset{G \in \mathcal{G}}{\Phi}(\mathbf{C}_G;\ \Pr[G]) \simeq \frac{1}{S}\sum_{s=1}^{S} \mathbf{C}_s = \bar{\mathbf{C}}.$$

Then, based on the ensembled similarity matrix $\bar{\mathbf{C}}$, all of the nodes divided into clusters and $\bar{\mathbf{H}}$ are outputted. In the SE method, the number of roles in sample graphs is counted and updated based on the LINC algorithm, as in the VE method. In this way, S similarity matrices $\{\mathbf{C}_1, \ldots, \mathbf{C}_S\}$, each of which is an $(N \times N)$ matrix, are calculated and then averaged. Therefore, the dominant computational complexity of the SE method to obtain the ensembled similarity matrix $\bar{\mathbf{C}}$ is $O(SN^2)$.[1]

Cluster-ensemble method

The cluster-ensemble method ensembles the clustering results $\{\mathbf{H}_1, \ldots, \mathbf{H}_S\}$ and produces the membership matrix $\bar{\mathbf{H}}$:

[1] The number of roles is counted based on the LINC algorithm; however, the amount of calculation required for the ensemble of the similarity matrix is dominant.

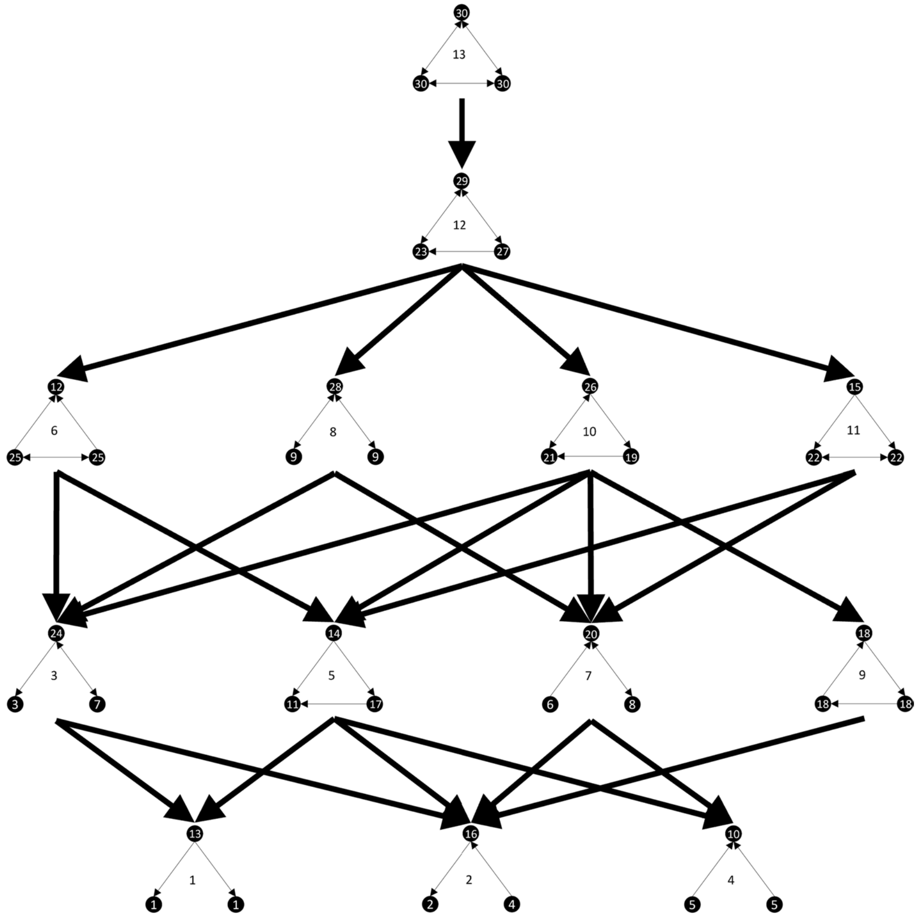


Fig. 4 Transition of roles in directed 3-node motifs

$$\mathcal{H} = \underset{G\in\mathcal{G}}{\Phi}(\mathbf{H}_G;\ \Pr[G]) \simeq \underset{s=1}{\Phi}{}^{S}(\mathbf{H}_s;\ 1/S) = \bar{\mathbf{H}}.$$

Unlike ensembling for supervised classification results, ensembling unsupervised clustering results is a challenging task because the correspondence relationship between obtained clusters is not clear and its degree has to be considered. Therefore, since no ensemble method for clustering results has been established yet, we do not discuss the issue in this article.

Proposed method: extended graph-ensemble method

This study proposes the extended graph-ensemble method (hereinafter, Ext-GE method) to calculate the expected value of the role frequency of each node under the assumption that motif stochastically collapses and shifts to another role. As shown in Fig. 4, the hierarchy can be defined for each role according to the number of edges in the corresponding motif. From Fig. 4, Motif 13 and Role 30 with 6 edges are at the top level, and Motifs 1, 2, 4 and Roles 1, 2, 4, 5, 10, 13, 16 with 2 edges are at the bottom level. This hierarchical diagram shows which motif changes to which by erasing one edge. For example, if any one of the six edges in Motif 13 disappears, it becomes Motif 12; of the four edges in Motif 11, if any one of the edges outgoing from Role 15 node disappears, it becomes Motif 7, and if any one of the bidirectional edges between Role 22 nodes disappears, it becomes Motif 5. When the edge is absent stochastically, the upper role changes to the lower role, and the frequency of appearance of the lower role increases. Therefore, when counting the role frequency of each node for $\bar{G} = (V, \bar{E}, \bar{p})$ ensembled with S sample graphs G_s, $1 \leq s \leq S$, the subordinate motif of the corresponding motif is searched for, and the number of roles included in that motif is also counted at the same time (Algorithm 5). In the while-loop at Line 6 to 18 in Algorithm 5, the motif of the connected triples searched for in the ensemble graph and its subordinate motifs and roles are also considered.

In detail, at Line 4, to search for all of the lower motifs without duplication and without omission, we express the edge-existence state as a bit string **b** via the motif2bits function, as do the above-mentioned methods GE, VE, and SE. By repeatedly performing the bit AND operation at Line 8 and the subtraction at Line 17 for $\mathbf{b}_u$, in which all 6 bits are initialized with 1 at Line 5, the subordinate motifs and roles are searched for efficiently and comprehensively. The while-loop repeats at most 64 times in the case of a directed 3-node motif represented by 6 bits. The other parts are the same as the count_roles function in Algorithm 4. The bits2motif function at Line 11 is the inverse function of motif2bits at Line 4, which returns the graph structure whose edge states, i.e., presence or absence, is expressed as $\mathbf{b}_u$, $\mathbf{b}_v$, $\mathbf{b}_w$. Algorithm 6 shows the whole picture of the Ext-GE method. The only difference between this method and the GE method (Algorithm 1) is whether the transition to the lower roles should be considered when calculating the expected value of the role number at Line 8.

Algorithm 5 Count and update of #roles and lower #roles: count_lower_roles($G^{(m)}, \bar{\mathbf{R}}$)

```
 1: Input: $G^{(m)} = (V^{(m)}, E^{(m)}, \mathbf{p}^{(m)}), \bar{\mathbf{R}} = [\bar{r}_{u,i}]_{u \in V, 1 \leq i \leq 30}$
 2: Output: $\bar{\mathbf{R}}$
 3: $V^{(m)} \to \{u, v, w\}$
 4: $\mathbf{b} \leftarrow$ motif2bits($G^{(m)}$)
 5: $\mathbf{b}_u \leftarrow (1 \ll 6) - 1$
 6: while True do
 7:     if $\mathbf{b}_u < 0$ then break end if
 8:     $\mathbf{b}_u$ &= $\mathbf{b}$
 9:     $\mathbf{b}_v \leftarrow (\mathbf{b}_u \ll 2 | \mathbf{b}_u \gg 4) \& ((1 \ll 6) - 1)$
10:     $\mathbf{b}_w \leftarrow (\mathbf{b}_v \ll 2 | \mathbf{b}_v \gg 4) \& ((1 \ll 6) - 1)$
11:     $G^{(n)} \leftarrow$ bits2motif($\mathbf{b}_u, \mathbf{b}_v, \mathbf{b}_w$)
12:     $(i, j, k) \leftarrow$ (Rcode[$\mathbf{b}_u$], Rcode[$\mathbf{b}_v$], Rcode[$\mathbf{b}_w$])        ▷ Get role code
13:     $\Pr[G^{(n)}] \leftarrow \prod_{e \in E^{(n)}} p(e) \prod_{e \in E \setminus E^{(n)}} (1 - p(e))$
14:     $\bar{r}_{u,i}$ += $\Pr[G^{(n)}]$        ▷ Increment #roles
15:     $\bar{r}_{v,j}$ += $\Pr[G^{(n)}]$
16:     $\bar{r}_{w,k}$ += $\Pr[G^{(n)}]$
17:     $\mathbf{b}_u$ −= 1
18: end while
```

Table 2 Basic statistics of datasets

Dataset	#nodes *N*	#edges *M*
Gnutella (peer-to-peer file sharing) (Leskovec and Krevl 2014)	10,876	39,994
Blog (trackback among weblogs) [1]	12,047	53,315
Enron (email communication) (Klimt and Yang 2004)	19,603	210,950
Hepth (citation in arXiv) (Leskovec and Krevl 2014)	27,400	352,504
Celegans (neural network of neurons and synapses) (Marinka Zitnik Rok Sosič and Leskovec 2018)	131	764

```
Algorithm 6 Extended Graph-Ensemble method: Ext-GE($\mathcal{G}, S$)
Input: $\mathcal{G} = (G, p),\ \ G = (V, E),\ \ \ S$
Output: $\bar{\mathbf{C}}$
Initialize: $\bar{\mathbf{R}} = [\bar{r}_{u,i}]_{u \in V, 1 \leq i \leq 30} \leftarrow \mathbf{0}$
$\bar{G} \leftarrow$ emsemble_graphs($\mathcal{G}, S$)
$\Psi \leftarrow$ search_connected_triples($\bar{G}$)
for $G^{(m)} \in \Psi$ do
    $G^{(m)} = (V^{(m)}, E^{(m)}, \mathbf{p}^{(m)})$
    $\bar{\mathbf{R}} \leftarrow$ count_lower_roles($G^{(m)}, \bar{\mathbf{R}}$)
end for
$\bar{\mathbf{C}} \leftarrow$ cosine_similarity($\bar{\mathbf{R}}$)
```

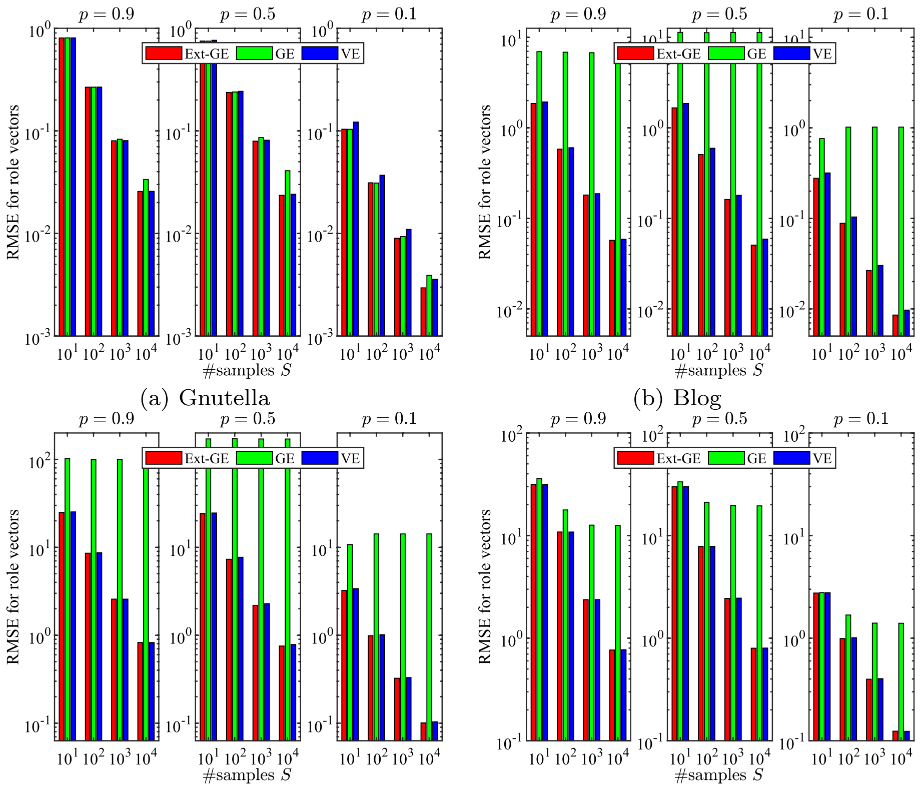
Experimental evaluations

In this study, we tackle the problem of counting the number of motif-derived roles for the nodes of the uncertainty graph and extracting node groups with similar role vectors. To confirm how the approximation of role counts by our methods affects the similarity between role vectors, and the final clustering result, we evaluate how accurately our method can output them against the true results.

Dataset and settings

In our experimental evaluations, role extraction based on the directed 3-node motif is performed on the following four directed graphs observed in the real world, and the effectiveness and efficiency of the proposed method is confirmed. The graph sizes are shown in Table 2. For these graphs, we set a uniform edge existence probability $p(e) = p \in [0.1, 0.2, \ldots, 0.9]$. For the last graph, we set a non-uniform probability $p(e) \sim \mathrm{Beta}(\alpha, \beta),\ \ (\alpha, \beta) \in \{(1.5, 5.0), (2.5, 2.5), (5.0, 1.5)\}$. The number of samples is set to $S \in \{10^1, 10^2, 10^3, 10^4\}$, and the number of clusters is set to $K = 10$. By varying the number of samples and clusters, we evaluated the variation in the error of the results and the execution time with the numbers of samples and clusters. Because we obtained similar results, only the results for $K = 10$ are presented in this study.

In our experiments, the true value for the expected number of roles $\mathcal{R}$ is calculated based on a previous work (Todor et al. 2015), which calculates the expected number of motifs for the uncertain graph; the true similarity matrix $\mathcal{C}$ is calculated from $\mathcal{R}$, and the true clustering result $\mathcal{H}$ is computed from $\mathcal{C}$. As an error measure for role vectors and a



(a) Gnutella

(b) Blog

(c) Enron

(d) Hepth

Fig. 5 RMSE of ensembled role vectors $\bar{\mathbf{R}}$ against true role vectors $\mathcal{R}$

similarity matrix, we employed root mean squared error (hereinafter, RMSE). For the true role vectors $\mathcal{R} = [r^*_{v,i}]_{v \in V, 1 \leq i \leq 30}$ and the approximated one $\bar{\mathbf{R}} = [\bar{r}_{v,i}]_{v \in V, 1 \leq i \leq 30}$,

$$\mathrm{RMSE} = \sqrt{\frac{1}{NR} \sum_{v \in V} \sum_{i=1}^{R} (r^*_{v,i} - \bar{r}_{v,i})^2}.$$

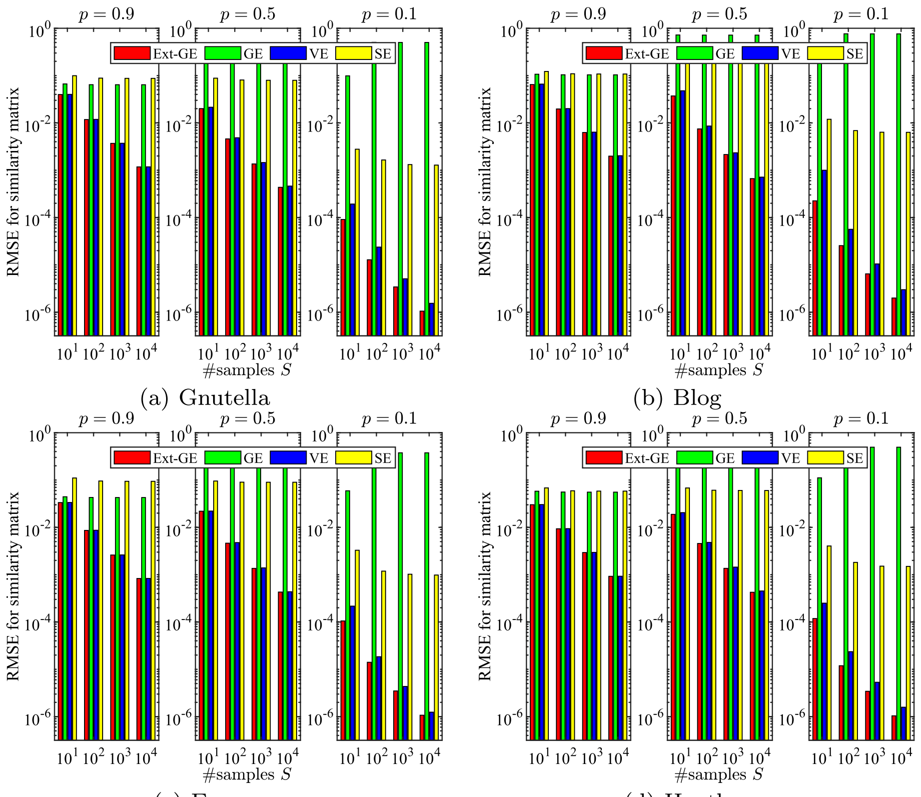
For the true similarity matrix $\mathcal{C} = [c^*_{u,v}]_{u \in V, v \in V}$ and the approximated one $\bar{\mathbf{C}} = [\bar{c}_{u,v}]_{u \in V, v \in V}$,

$$\mathrm{RMSE} = \sqrt{\frac{1}{N(N-1)/2} \sum_{u \in V} \sum_{v \in V, u < v} (c^*_{u,v} - \bar{c}_{u,v})^2}.$$

As s similarity measure for the true clustering result in $\mathcal{H}$ and the approximated one $\bar{\mathbf{H}}$, we employed normalized mutual information (hereinafter, NMI) (Kvålseth 2017).

Error evaluation for role vectors

First, we evaluated our method in terms of the error of the role vectors. Figure 5 illustrates the RMSE in a logarithmic scale with respect to the number of samples S. For almost all of the networks we used, we could make the following observations. As the size of the graph increases, the absolute number of appearance roles increases; therefore, the error value tends to increase. On the contrary, the lower the edge-existence



(a) Gnutella

(b) Blog

(c) Enron

(d) Hepth

Fig. 6 RMSE of ensembled similarity matrix $\bar{\mathbf{C}}$ against true similarity matrix $\mathcal{C}$

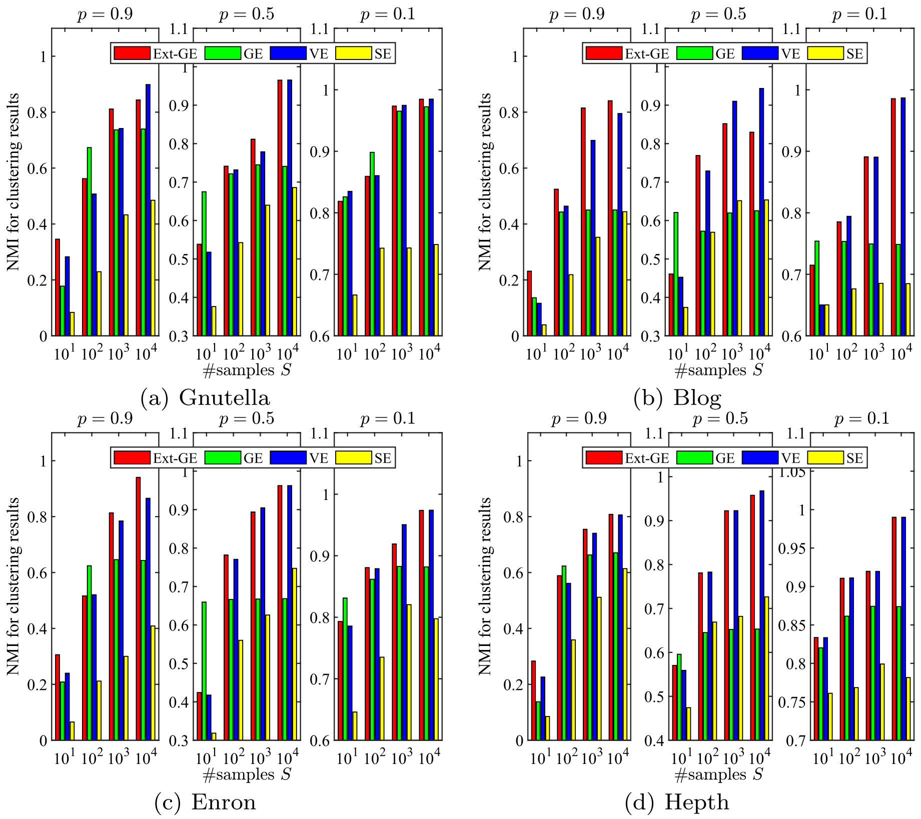
probability, the smaller is the absolute amount of the number of appearance roles; therefore, the error value tends to be smaller. As the number of samples increases, the error decreases with some exceptions, for example, the GE method for the Blog and Enron networks. Furthermore, Ext-GE achieves smaller errors than, or errors almost equal to, VE.

Error evaluation for similarity matrix

Next, we discuss the RMSE of similarity matrices. Figure 6 depicts the RMSE in a logarithmic scale to the number of samples S. For almost all of the networks we used, we made the following observations. As the size of the graph increases, the absolute number of appearance roles increases; therefore, the error value tends to increase. As the number of samples increases, the errors of Ext-GE and VE decrease, while those of GE and SE do not decrease. Furthermore, Ext-GE achieves smaller errors than VE.

Similarity evaluation for clustering results

Next, we confirm the effectiveness of our method of focusing on the similarity to the true clustering results. Figure 7 shows the NMI for the number of samples S. From these figures, we can make the following observations. In almost all cases, when the edge-existence probability is small and the number of samples is large, all methods produce more similar results to the true results (considering the difference in the axes' ranges).



(a) Gnutella

(b) Blog

(c) Enron

(d) Hepth

Fig. 7 NMI of clustering results $\bar{\mathbf{H}}$ against those by true similarity matrix $\mathcal{H}$

SE outputs the worst results; GE sometimes outputs good results depending on the networks; Ext-GE and VE stably output better results than the other methods independent of the networks, probability, and number of samples.

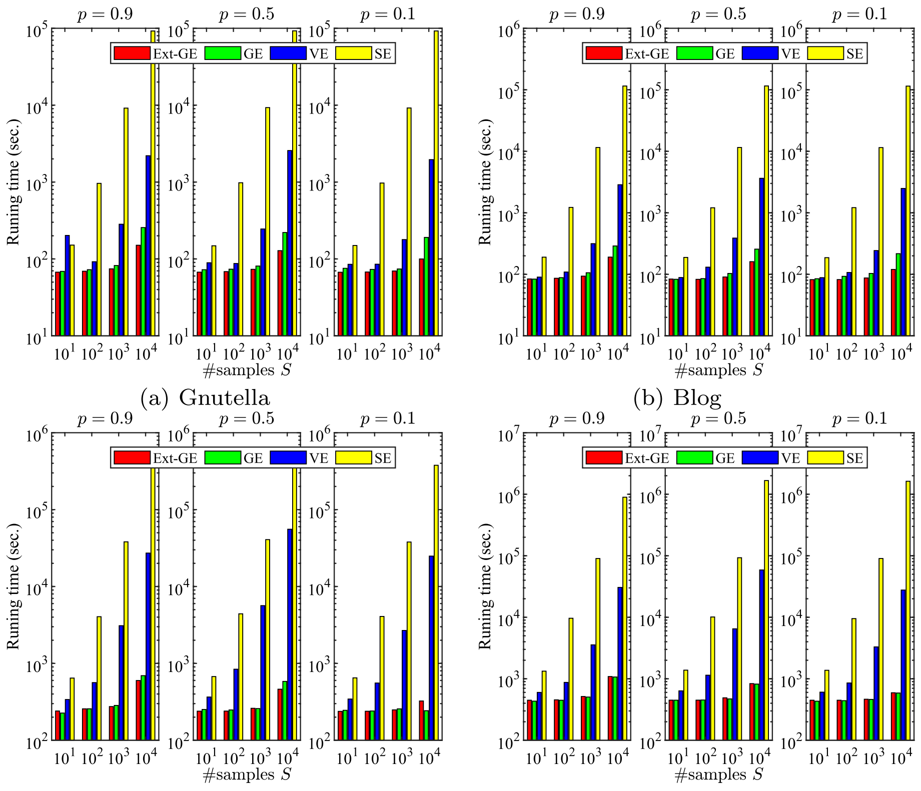
The error of the role vectors affects the error of similarity matrices and the final clustering results; therefore, more accurate role vectors are required.

Efficiency

Next, we evaluate our method in terms of computational efficiency. Figure 8 indicates the running time up to the outputs of the ensembled similarity matrix $\bar{\mathbf{C}}$ from the given uncertain graph, in a logarithmic scale with respect to the number of samples S. From these figures, for all of the networks, our Ext-GE is much faster than VE and SE, which are derived from the state-of-the-art LINC algorithm, especially when the number of samples is large.

Non-uniform setting

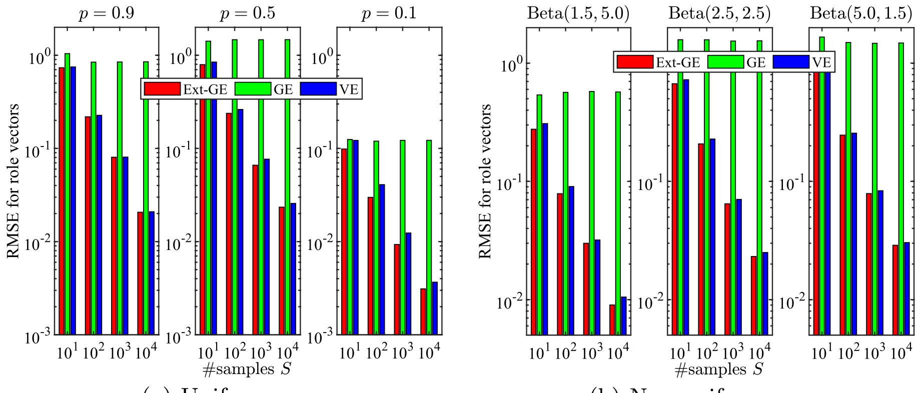
Finally, to confirm the difference between edge-existence probabilities, we compared the results for Celegans under the settings of uniform and non-uniform edge-existence probability. As a non-uniform setting, we set a non-uniform probability according to the beta distribution, $p(e) \sim \mathrm{Beta}(\alpha, \beta), \ \ (\alpha, \beta) \in \{(1.5, 5.0), (2.5, 2.5), (5.0, 1.5)\}$. The mean



(a) Gnutella (b) Blog

(c) Enron (d) Hepth

Fig. 8 Running time up to outputting the ensembled similarity matrix



(a) Uniform (b) Non-uniform

Fig. 9 RMSE of ensembled role vectors $\bar{\mathbf{R}}$ against true role vectors $\mathcal{R}$

value of the random numbers can be calculated as $\alpha/(\alpha + \beta)$, so they are about 0.23, 0.5, and 0.77.

Figures 9 and 10 show the RMSE of the role vectors and the NMI of the clusters with respect to the number of samples S. From the results, we can observe that there is no remarkable difference between uniform and non-uniform settings, i.e., our Ext-GE method achieves much smaller RMSE values and much higher NMI values than the

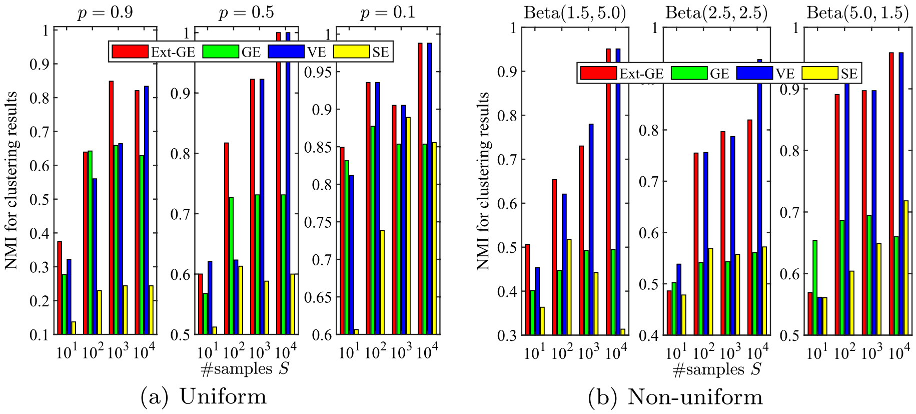


Fig. 10 NMI of clustering results $\bar{\mathbf{H}}$ against those by true similarity matrix $\mathcal{H}$

existing GE method, and it is comparable to the VE method, in any beta distribution. Although not shown here, there is a similar tendency in the RMSE of the similarity matrices. Furthermore, the computational costs of these methods do not depend on the probability values; in fact, the running times were confirmed to be almost the same.

Trajectories of information propagation in social media can be modeled as an uncertain graph with non-uniform edge-existence probabilities. Such an uncertain graph is observed as many instances where edges stochastically appear and disappear, and thus its true structure and true edge-existence probabilities cannot actually be known. Our method reflects this fact and ensembles many observed (sampled) graphs and outputs accurate results close to those obtained from the true structure. Therefore, our method is applicable to real-world uncertain graphs, and it is promising for accurately identifying important nodes in a viral marketing strategy.

Conclusion

In this study, for the task of motif-role extraction from an uncertain graph, we proposed an efficient and effective method, called the extended-graph ensemble method. It involves counting node roles defined by the position in motifs, calculating the similarity between nodes based on role vectors, and dividing all of the nodes into clusters with similar role vectors. This method ensembles sampled graphs and counts roles by considering the transition to lower-layer roles due to stochastically occurring edge-disappearance.

In experiments using real-world networks with added uniform and non-uniform edge probabilities, we confirmed the effectiveness and efficiency of our proposed method. The proposed method, the extended-graph-ensemble method, outputs results with smaller errors from the true values and works more quickly than the existing methods, including our previously proposed method, the graph-ensemble method, and the vector-ensemble and similarity-ensemble methods, which are both derived from LINC, the state-of-the-art technique. Accordingly, we conclude that the extended graph ensemble method is the most suitable for the problem addressed in this study.

Future tasks include motif-role extraction that is not limited to 3-node motifs, determination of the appropriate number of samples using Hoeffding's inequality, and more detailed analysis of the clustering results.

Acknowledgements
Not applicable.

Author contributions
SI performed the research and wrote the article. TF contributed to designing the proposed method and part of experimental evaluations. Both authors read and approved the final manuscript.

Funding
The second author is grateful for the financial support from JSPS Grant-in-Aid for Scientific Research (C) (Nos. 20K11940, 22K12279).

Availability of data and materials
The C++, Python codes and raw datasets used and analysed during the current study will be available at https://github.com/fuppo27?tab=repositories.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 27 February 2022 Accepted: 4 August 2022
Published online: 12 August 2022

References

Ahmed NK, Neville J, Rossi RA, Duffield N (2015) Efficient graphlet counting for large networks. In: 2015 IEEE international conference on data mining, pp 1–10. https://doi.org/10.1109/ICDM.2015.141
Ceccarello M, Fantozzi C, Pietracaprina A, Pucci G, Vandin F (2017) Clustering uncertain graphs. Proc VLDB Endow 11(4):472–484
Everett M, Borgatti S (1994) Regular equivalence: general theory. J Math Sociol 19(1):29–52
Gilpin S, Eliassi-Rad T, Davidson I (2013) Guided learning for role discovery (glrd): Framework, algorithms, and applications. In: Proceedings of the 19th ACM SIGKDD international conference on knowledge discovery and data mining. ACM, New York, pp 113–121
Grochow JA, Kellis M (2007) Network motif discovery using subgraph enumeration and symmetry-breaking. In: Proceedings of the 11th annual international conference on research in computational molecular biology, RECOMB'07. Springer, Berlin, pp 92–106
Guerrero C, Milenković T, Pržulj N, Kaiser P, Huang L (2008) Characterization of the proteasome interaction network using a QTAX-based tag-team strategy and protein interaction network analysis. Proc Natl Acad Sci 105(36):13333–13338. https://doi.org/10.1073/pnas.0801870105
Henderson K, Gallagher B, Eliassi-Rad T, Tong H, Basu S, Akoglu L, Koutra D, Faloutsos C, Li L (2012) Rolx: structural role extraction & mining in large graphs. In: Proceedings of the 18th ACM SIGKDD international conference on knowledge discovery and data mining. ACM, pp 1231–1239
Henderson K, Gallagher B, Li L, Akoglu L, Eliassi-Rad T, Tong H, Faloutsos C (2011) It's who you know: graph mining using recursive structural features. In: Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining. ACM, New York, pp 663–671
Hu J, Cheng R, Huang Z, Fang Y, Luo S (2017) On embedding uncertain graphs. In: ACM on conference on information and knowledge management, vol 123, pp 157–166
https://github.com/fuppo27/graph_dataset
Itzhack R, Mogilevski Y, Louzoun Y (2007) An optimal algorithm for counting network motifs. Phys A Stat Mech Appl 381:482–490. https://doi.org/10.1016/j.physa.2007.02.102
Klimt B, Yang Y (2004) The enron corpus: a new dataset for email classification research. In: Boulicaut JF, Esposito F, Giannotti F, Pedreschi D (eds) Machine learning: ECML 2004. Springer, Berlin, pp 217–226
Kvålseth TO (2017) On normalized mutual information: measure derivations and properties. Entropy. https://doi.org/10.3390/e19110631
Leskovec J, Krevl A (2014) SNAP datasets: Stanford large network dataset collection. http://snap.stanford.edu/data
Lorrain F, White H (1971) Structural equivalence of individuals in social networks. J Math Sociol 1(1):49–80
Ma C, Cheng R, Lakshmanan LVS, Grubenmann T, Fang Y, Li X (2019) Linc: a motif counting algorithm for uncertain graphs. Proc VLDB Endow 13(2):155–168. https://doi.org/10.14778/3364324.3364330
Marinka Zitnik Rok Sosič SM, Leskovec J (2018) BioSNAP datasets: Stanford biomedical network dataset collection. http://snap.stanford.edu/biodata
McDonnell MD, Yaveroglu ON, Schmerl BA, Iannella N, Ward LM (2014) Motif-role-fingerprints: the building-blocks of motifs, clustering-coefficients and transitivities in directed networks. PLoS ONE 9(12):1–25. https://doi.org/10.1371/journal.pone.0114503

Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U (2002) Network motifs: simple building blocks of complex networks. Science 298(5594):824–827. https://doi.org/10.1126/science.298.5594.824

Naito S, Fushimi T (2021) Motif-role extraction in uncertain graph based on efficient ensembles. In: Proceedings of the 10th international conference on complex networks and their applications, pp 501–513

Nemhauser GL, Wolsey LA, Fisher ML (1978) An analysis of approximations for maximizing submodular set functions. Math Program 14:265–294

Ohnishi T, Takayasu H, Takayasu M (2010) Network motifs in an inter-firm network. J Econ Interact Coord 5(2):171–180

Pfeiffer JJ, Neville J (2011) Methods to determine node centrality and clustering in graphs with uncertain structure. In: Proceedings of the fifth international conference on weblogs and social media. The AAAI Press, pp 590–593

Pinar A, Seshadhri C, Vishal V (2017) Escape: efficiently counting all 5-vertex subgraphs. In: Proceedings of the 26th international conference on World Wide Web, WWW '17. Republic and Canton of Geneva, CHE, pp 1431–1440. https://doi.org/10.1145/3038912.3052597

Pržulj N (2007) Biological network comparison using graphlet degree distribution. Bioinformatics 23(2):177–183. https://doi.org/10.1093/bioinformatics/btl301

Rossi RA, Gallagher B, Neville J, Henderson K (2012) Role-dynamics: fast mining of large dynamic networks. In: Proceedings of the 21st international conference companion on World Wide Webv, pp 997–1006

Rossi RA, Gallagher B, Neville J, Henderson K (2013) Modeling dynamic behavior in large evolving graphs. In: Proceedings of the sixth ACM international conference on web search and data mining. ACM, pp 667–676

Rossi RA, Ahmed NK (2015) Role discovery in networks. IEEE Trans Knowl Data Eng 27(4):1112–1131

Sarajlić A, Malod-Dognin N, Yaveroǧlu ON, Pržulj N (2016) Graphlet-based characterization of directed networks. Nature 123:89. https://doi.org/10.1038/srep35098

Todor A, Dobra A, Kahveci T (2015) Counting motifs in probabilistic biological networks. In: Proceedings of the 6th ACM conference on bioinformatics, computational biology and health informatics, BCB '15. Association for Computing Machinery, New York, pp 116–125. https://doi.org/10.1145/2808719.2808731

Tran N, Choi KP, Zhang L (2013) Counting motifs in the human interactome. Nat Commun 4:2241. https://doi.org/10.1038/ncomms3241

Wernicke S (2005) A faster algorithm for detecting network motifs. In: Proceedings of the 5th international conference on algorithms in bioinformatics, WABI'05. Springer, Berlin, pp 165–177

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.