Humanities & Social Sciences Communications

ARTICLE

https://doi.org/10.1057/s41599-023-01863-z OPEN

Check for updates

Inferring links in directed complex networks through feed forward loop motifs

Satyaki Roy[1]✉, Ahmad F. Al Musawi[2,3] & Preetam Ghosh[3]

Complex networks are mathematical abstractions of real-world systems using sets of nodes and edges representing the entities and their interactions. Prediction of unknown interactions in such networks is a problem of interest in biology, sociology, physics, engineering, etc. Most complex networks exhibit the recurrence of subnetworks, called network motifs. Within the realm of social science, link prediction (LP) models are employed to model opinions, trust, privacy, rumor spreading in social media, academic and corporate collaborations, liaisons among lawbreakers, and human mobility resulting in contagion. We present an LP metric based on a motif in directed complex networks, called feed-forward loop (FFL). Unlike nearest neighbor-based metrics and machine learning-based techniques that gauge the likelihood of a link based on node similarity, the proposed approach leverages a known dichotomy in the motif distribution of directed networks. Complex networks are sparse, causing most nodes and their associated links to have low motif participation. Yet, due to intrinsic network motif-richness, few links participate in many distinct motif substructures. Thus, the FFL-based metric combines the presence and absence of motifs as a signature to outperform baseline metrics on ten directed social and biological network datasets. We conclude with the future of the FFL-based metric in dynamic network inference as well as its use in designing combined metrics using network motifs of varying orders as features.

[1] Bioinformatics & Computational Science, Frederick National Laboratory for Cancer Research, Frederick, MD, USA. [2] Department of Information Technology, University of Thi Qar, Thi Qar, Iraq. [3] Department of Computer Science, Virginia Commonwealth University, Richmond, VA, USA. ✉email: satyaki.roy@nih.gov

Introduction

Complex networks are a ubiquitous tool for representation among entities in real-world systems (Strogatz, 2001). There are several types of networks including social, ecological, computer, biological, biochemical, technological, brain networks, climate, social networks, etc. Within biological networks, they may be classified based on the biomedical implications of the interaction among genes, proteins, drugs and target proteins, metabolites, neurons, organisms in a food chain, etc. (Crichton et al., 2018). An important tool in the effort to understand the possibility of known interactions between any two network entities is the prediction of the likelihood of them sharing a link. With the advent of deep machine learning, link prediction metrics can help find the similarity between nodes and their likelihood of sharing a link, by probing varying depths of their neighborhoods (Zhang and Chen, 2018; Mayo et al., 2015).

Several machine learning-based computational techniques are being leveraged to predict links in complex networks. Graph embedding techniques represent nodes as low-dimensional vectors while pre-serving the structural relationships within networks. Such techniques can mainly be classified into matrix decomposition, random walk, and deep learning. These techniques are being combined to develop ensemble learning models, while deep learning methods on graphs are being augmented with graph signal processing methods to achieve higher prediction accuracy (Chen et al., 2022; Cheung et al., 2020). In addition to the similarity in terms of local neighborhoods, recent link prediction models are encoding community or cluster-level closeness into the node embeddings (Saxena et al., 2022). An example of such clusters is the recurrent subgraphs, called network motifs that characterize large-scale natural and engineered complex networks (Milo et al., 2002; Stone et al., 2019). Apart from their over-representation in complex networks than expected by chance, the network motifs have been shown to act as signaling modules that control the response to external stimuli (Han et al., 2007) and render functional as well as structural robustness against perturbation (Paul and Radde, 2016; Roy et al., 2020; Dey et al., 2019; Ghosh et al., 2011). This has led to the characterization of the importance of nodes (and links) in complex networks in terms of their node (or link) motif participation (or centrality) (Koschützki et al., 2007; Wang et al., 2014). There is a growing body of literature on the role of motifs of three nodes (namely, feedback loop, feed-forward loop, etc.), four nodes (e.g., *bi fan*), and beyond (Milo et al., 2002; Ma'ayan et al., 2005).

Motifs are a promising tool to unravel the effect of drugs on cellular targets (Wu et al., 2016) and driver genes based on their change in motif profiles between normal and disease states in biological networks (Liu et al., 2014). It is also a pertinent resource for inferring relationships between entities in complex networks through link prediction metrics based on participation in common or shared motifs in undirected networks (Jia et al., 2017; Aghabozorgi and Khayyambashi, 2018). Entities in undirected networks exhibit the *transitive property* of relationships, resulting in the formation of triangles or closed triads as network motifs. For instance, in a social setting, if node 1 trusts node 2 and node 2 trusts node 3, then node 1 has reasons to trust node 3 (Jin and Zafarani, 2017). Thus, the notion of triadic closure in social network theory states that the existence of links (u, v) and (v, w) among three nodes u, v, w in an undirected network suggests that there is a tendency of forming the link (u, w) (Simmel, 1908). Triadic closure serves as the basis for commonness in the respective neighborhoods as a criterion for any pair of nodes to be connected. It can also be used to understand the evolution of networks over time (Liben-Nowell and Kleinberg, 2003).

Application of link prediction algorithms in social science. This tendency of forming triads (also called *common neighbor* property), whereby two individuals are likely to be friends if they have many common friends, is a key basis of our social groups (Wang et al., 2011). Figure 1a shows the common neighbor property in a social network, where Bill is more likely to be friends with Bob (than Mary) because the two have a common neighbor John. Similarly, social groups exhibit *homophily*, whereby people seek out individuals who are similar to them. These characteristics can be leveraged to predict how social ties are formed at a given time or evolve over time. Let us discuss the use of link prediction (LP) techniques in determining academic and corporate collaborations, liaisons among criminals or terrorists, opinions, trust, privacy, and rumor spreading in social media, and human mobility resulting in disease contagion.

Unknown or hidden social ties. LP algorithms have been applied to the data imputation problem in a social context, where unknown or missing information can be filled out with high confidence to make informed recommendations (Ouzienko and Zoran, 2014). These recommender systems deal with the prediction of citations of papers or co-authorships (Shibata et al., 2012; Benchettara et al., 2010a) or learning unknown associations among criminals in *criminal network analysis* (CNA) to find new leads into hidden liaisons, such as drug syndicates and terrorist networks. In the case of the latter, supervised machine learning and deep reinforcement learning-based LP models are considered to enable law enforcement agencies to deal with organized crime

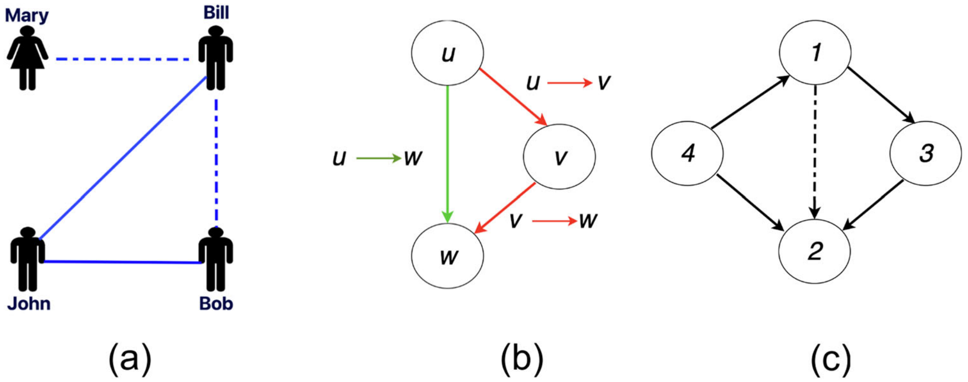

Fig. 1 Triads and motifs. a Common neighbor property of the social network, where Bill is more likely to be friends with Bob (than Mary) because they have a common neighbor John, **b** feed Forward Loop motif with three links, direct link $u \rightarrow v$ and indirect links $v \rightarrow w$, and $u \rightarrow w$; **c** directed link (1, 2) playing in $v \rightarrow w$ in the motif (4, 1, 2) and $u \rightarrow w$ in the motif (1, 3, 2).

syndicates and activities (Lim et al., 2019, 2021). LP models are a potential tool to find likely yet unfamiliar associations among organizations that may lead to *patent cooperation* (Chen et al., 2021).

These algorithms tap into behavioral studies which suggest that the interaction among individuals on online social platforms is influenced by their peers (Huo et al., 2018). Thus, its applicability has been studied in gauging interpersonal relationships (Esslimani and Brun, 2011) as well as *sentiments* in social relationships among individuals or organizations (Papadimitriou et al., 2012). Analysis of signed social networks from online social networking sites, where *trust* and *distrust* are denoted by + and – signs, is used to gauge attitudes, opinions, and trust (Leskovec et al., 2010a). Matrix factorization and latent factor-based LP models have incorporated the level of trust and behavioral properties into the recommenders (Yang et al., 2012; Jiang et al., 2020).

Evolution of social relationships. Social dynamics tend to change over time, making it imperative to encode temporal information into recommendation systems for academic and criminal networks. The community has looked at similar problems on CNA as well as academic citations and co-authorships in the context of varying times. LP models have been applied to databases of scientific literature to model the evolving motivation of authors to collaborate (Zhang et al., 2017). There is a suggestion that the combination of time-varying criminal networks with personalized information, like arrest warrants, judicial judgment, wiretap records, and police station proximity, can improve the prediction of future liaisons (Lim et al., 2019).

LP models applied to dynamic virtual as well as physical social networks leverage the history of past interactions to predict future associations (Jiang et al., 2020; Tsugawa and Hiroyuki, 2013). For example, combining LP models with communication history among individuals can help determine future *face-to-face interactions* in social settings, such as academic conferences. The prediction of evolving social dynamics in a multi-layered social network has been analyzed by gauging the information flow through trusted central nodes. This approach has been shown to minimize unreliability spawning from *rumor* spreading (Fan et al., 2019). Finally, these algorithms can be adapted to address the challenges in trust and *privacy*. Research has gone into the use of LP to minimize the trust deficit from the addition of new members (i.e., nodes) and associations (or links) in dynamic social groups (Bhagat et al., 2010).

Modeling human behavior. Another class of applications of LP lies in the modeling of *human behavior and contact* resulting in the transmission of infectious diseases, such as influenza and COVID-19. Combination of network clustering and LP can model the mixing pattern and spread dynamics in Korea (Kwon and Hang-Hyun, 2023) and design intervention and public policies that complement pharmaceutical measures (Antweiler et al., 2021). Finally, it is a useful tool to measure behavioral and *cultural perceptions* about intervention measures being adopted (Ma et al., 2022), particularly at a time when human sentiment can be gauged from freely available social media data (Wei et al., 2016).

Contributions of this work. In this work, we extend the idea of triadic closure in directed networks using network motifs. The feed forward loop (FFL) is a 3-node motif abundantly found in directed natural and engineered networks and therefore could be leveraged to understand the connectivity in directed complex networks (Gorochowski et al., 2018; Wang and Provan, 2009). We leverage the FFL motif to highlight a contradiction in directed complex network topologies we discuss hereafter. Complex networks have notably higher FFL presence than randomized networks (Roy et al., 2020); due to their motif-richness, a large number of FFL motifs are distributed among a few links, causing the average motif participation per link in complex networks to be typically higher than their randomized counterparts. Besides, a few hub nodes have a disproportionately high number of connections, making the complex networks scale-free in nature (Barabasi, 2009). Moreover, these networks exhibit the small-world property, wherein most nodes are not directly connected but it is possible to travel from one node to another in very few hops (Wang and Chen, 2003; Sampaio Filho et al., 2015). It follows that the well-connected hubs and their associated links are likely to participate in a high number of FFL motifs than the non-hubs. We explore whether the lack of connectivity of the majority of the (non-hub) nodes (termed network sparseness (Wang and Chen, 2003; Tang et al., 2019) and the unevenness in motif distribution among the nodes could lend a perspective to link prediction in directed networks.

With this intuition, we explore an FFL motif-based link prediction measure, which harnesses some interesting features of directed complex networks. Let us understand it using the example of a kind of directed biological network, called the transcriptional regulatory network (TRNs). FFL motif, as shown in Fig. 1a and discussed in the section "Feed forward loop motif", is an acyclic triangle. Our prior analysis (Roy et al., 2020, 2017, 2021) has shown that FFL is the most frequent triangular motif in TRNs, greatly outnumbering the cyclic triangle, called the feedback loops. We refer to a connected triplet of three nodes as a directed triad (or simply triad). (4,1,2) and (1,3,2) are examples of triads in Fig. 1b.

Since complex networks like TRNs are rich in FFL motifs, the likelihood of the existence of a link, say (1, 2) in Fig. 1c is high as it completes two FFL motifs (4, 1, 2) and (1, 3, 2). This property is analogous to existing common neighbor-based metrics (such as common neighbors, Jaccard, Adamic Adar, resource allocation, etc. (Ghorbanzadeh et al., 2021) in undirected networks. Additionally, the three directed links of an FFL denoted by an ordered triplet (u, v, w) (formed by links $u \rightarrow v$, $v \rightarrow w$, $u \rightarrow w$) have distinct designations. A link prediction metric should therefore consider the likelihood of the link given that it plays a designated role in an FFL motif. This is apparent in the case of the link (1, 2) that has designations $v \rightarrow w$ and $u \rightarrow w$ in FFLs (4, 1, 2) and (1, 3, 2), respectively. Thus, combining the evidence of (1, 2) in its network for both designations is necessary. Moreover, we have also reported that TRNs are extremely sparse, and the majority of nodes are likely to have no motif participation, making the absence of FFL motifs a possible signature as well. *Thus, we present an FFL-based link prediction metric (in the section "Feed forward loop-based link prediction and analysis") that learns the link existence likelihood given the presence and absence of FFL motif triads and combines the two likelihoods as the weighted sum over all possible FFL motif triads.*

Overall, the FFL-based approach leverages Bayes' theorem to estimate the possibility of connecting two nodes in terms of the number of triads and non-triads they participate in. We define a triad as a triplet of nodes with at least two directed links, whereas a *non-triad* is a triplet with at most one link. The proposed metric works in two stages (see Fig. 2). *First*, it estimates the probability of the existence of a link $c \equiv (x, y)$ in a directed network given a specific triad as the product of the probability that nodes x, y participate in the triad and the likelihood of a link existing in the network given that triad. *Second*, it analogously calculates the link probability of c given a specific non-triad as the product of the probability that nodes x, y participate in the non-triads and the likelihood that a link exists in the network given that non-triad.

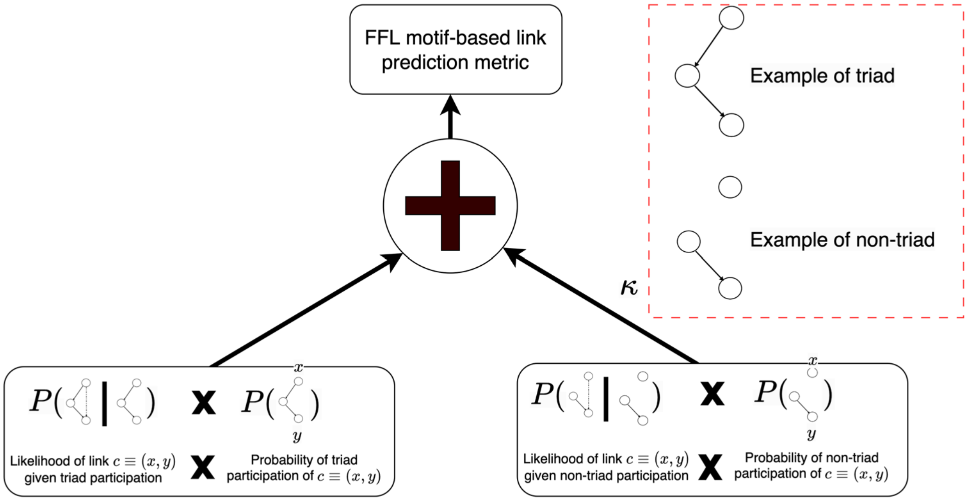
FFL motif-based link prediction metric
Example of triad
Example of non-triad
κ
x
y
Likelihood of link $c \equiv (x, y)$ given triad participation
X
Probability of triad participation of $c \equiv (x, y)$
Likelihood of link $c \equiv (x, y)$ given non-triad participation
Probability of non-triad participation of $c \equiv (x, y)$

Fig. 2 Motif-based link prediction. A schematic representation of the feed forward loop (FFL) motif-based approach, showing link prediction probability as the weighted sum of the probability of the existence of a link $c \equiv (x,y)$ given the existence of a triad and the probability of the existence of a link c given the existence of a non-triad.

Finally, it combines the two scores by a weighted sum determined by a real, non-zero parameter κ.

We conclude a discussion on the application of the motif-based approach in (1) inferring links evolving in dynamic natural networks (namely, social, and biological), engineered networks, and recommendation systems; and (2) building deep machine learning models that employ network motifs of varying orders to refine link prediction accuracy (refer to the section "Discussion").

Method

Directed graph. A graph is an ordered pair $G = (V, E)$ where V is a finite, non-empty set of objects called *vertices* (or nodes); and E is a (possibly empty) set of 2-subsets of V, called *edges* (Newman, 2003). A directed graph is a graph in which edges have directions. A directed edge $(u, v)\in E$ allows unidirectional information flow from vertex u to v and not necessarily from v to u. Some directed graphs may have weights associated with them; these weights can be discrete or continuous. *Graph density* of a directed network is measured as the ratio between the number of links to the maximum possible number of links, i.e., $\frac{|E|}{|V|\times(|V|-1)}$.

Feed forward loop motif. A feed forward loop (FFL) motif is a triplet of three nodes (u, v, w), forming an acyclic triangle (see Fig. 1b). FFL is one of the most abundant 3-node network motifs in complex networks, such as transcriptional networks. Nodes u, v, w are termed master regulator, intermediate regulator and target node, and there are three links, direct link $u \to v$ and indirect links $v \to w$, and $u \to w$.

The importance of a link (u, v) in a directed complex network G is often quantified by FFL motif-based centrality. Edge motif centrality is defined as the number of FFL motifs $m \equiv (u, v, w)$ it participates in, i.e.,

$$C(x, y; G) = \sum_{w\in V} 1[(u, v) \in E\&(u, w) \in E\,\&(u, w) \in E] \tag{1}$$

In Eq. (1), 1 is an indicator variable that assumes a value of 1 if a motif $m \equiv (u, v, w)$ exists in the form of directed links (u, v), (v, w), (u, w), and 0 otherwise. Edge motif centrality can be used to estimate the FFL motif distribution of G by plotting the frequency of links (u, v) that participate in k FFL motifs ($\forall 0 \le k \le \max_{e\in E}(C)$). Moreover, links participating in several FFLs can have distinct roles or designations—they may be $u \to v$, $v \to w$, $u \to w$ links of different FFLs. Figure 1c shows link (1, 2) participating in $v \to w$ in the motif (4, 1, 2) and $u \to w$ in the motif (1, 3, 2).

FFL motif density is defined as the ratio between the number of FFLs in G over all possible directed triplets of nodes u, v, w that can form FFLs and calculated as

$$D(G) = \frac{\sum_u \sum_v \sum_w 1[(u, v) \in E\,\&(u, w) \in E\&(u, w) \in E]}{|V|\times(|V|-1)\times(|V|-2)} \tag{2}$$

Feed forward loop-based link prediction and analysis. The FFL motif-based link prediction measures the probability of link prediction from node x to node y (i.e., $x \to y$) in terms of the FFL motifs in directed network G. For each triad (illustrated in Fig. 3), there is a closing link (denoted by c) that completes the triad. For example, $c = u \to w$ for triad 110. The score for the link (x, y), given by $S(x, y)$ is the combination of the probability of (x, y) being the closing link c, given the presence as well as the absence of each triad. The metric can be understood in three phases.

Likelihood of FFL motif given the existence of a triad. First, we measure the likelihood of (x, y) being the closing link c that completes an FFL motif Δ, given the existence of a triad t. The likelihood score is based on the Bayes rule: the product of the probability of (x, y) participating in a triad t and the likelihood of $c \equiv (x, y)$ given that it participates in t, i.e.,

$$P(\Delta; t, G) = P(c|t; G)\times P_{x\to y}(t; G) \tag{3}$$

In the above equation, the two terms are calculated as follows:

$$P_{x\to y}(t; G) = \frac{\text{Number of } t\left(x, y\right)\text{participating in } G}{\max\limits_{(x',y')} \text{ Number of } t\left(x', y'\right)\text{participating in } G} \tag{4}$$

$$P(c|t; G) = \frac{\text{Number of } \Delta\, t \text{ participating in} G}{\text{Number of } t \text{ in } G} \tag{5}$$

Likelihood of FFL motif given the absence of a triad. Applying the same logic, the probability of a closing link c given the absence of a triad t or equivalently, the existence of a non-triad (depicted in Fig. 3 and denoted by $\hat{t}$), is calculated as

$$P(c; \bar{t}, G) = P(c|\bar{t}; G)\times P_{x\to y}\left(\hat{t}; G\right) \tag{6}$$

Weighted sum of the two likelihoods. The effect of the existence (and absence) of a triad on a closing link $c \equiv (x,y)$ is weighed by a

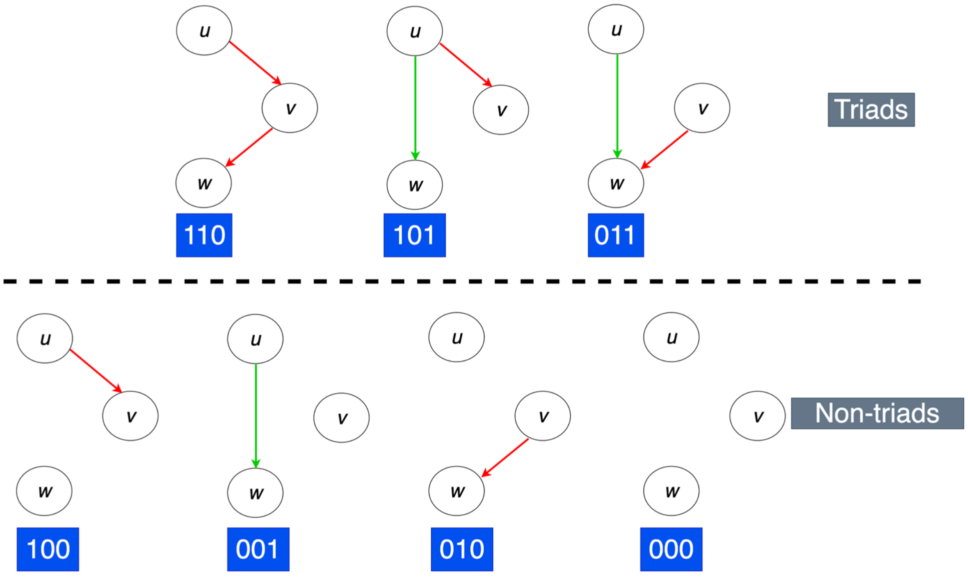

Fig. 3 FFL triads. Three triads and non-triads in directed complex networks and their representation in bitstrings of length 3: a value of 1 or 0 correspond to the presence or absence of the directed links (*u*, *v*), (*v*, *w*), (*u*, *w*) in a possible FFL motif *m* = (*u*, *v*, *w*) among three nodes *u*, *v*, *w* in a directed complex network.

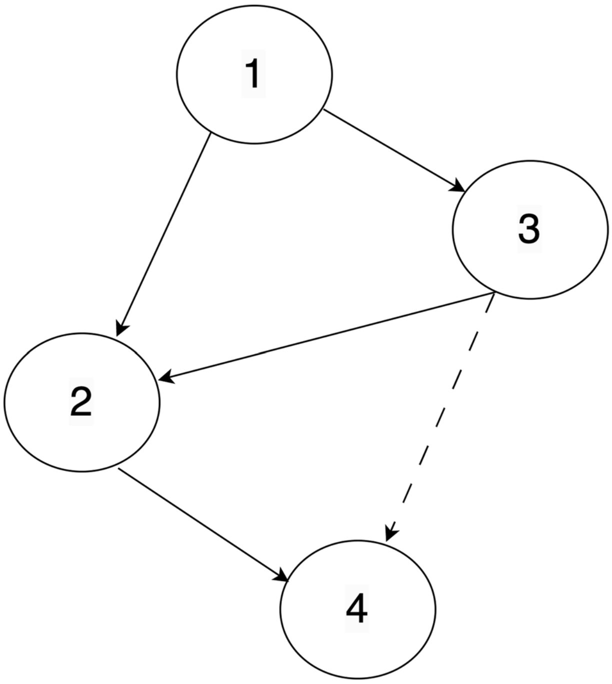

Fig. 4 Illustrative example of a prediction of link (3, 4) in a 4-node network, only considering the presence and absence of triad 110 where the links (*u*, *v*) and (*v*, *w*) are present.

learnable parameter κ (where $\kappa \geq 0$). Finally, combining Eqs. (3) and (6), the link prediction score is calculated over all possible triads $t \in T = [011{,}101{,}110]$, as

$$S(x,y) = \sum_{t \in T} P(\Delta; t, G) + \kappa \times P(c; \bar{t}, G) \tag{7}$$

One may consider three different weighing parameters $\kappa_{u\rightarrow v}$, $\kappa_{v\rightarrow w}$, and $\kappa_{u\rightarrow w}$ for triads $t = v \rightarrow w$, $u \rightarrow w$; $u \rightarrow v$, $u \rightarrow w$, and $u \rightarrow v$, $v \rightarrow w$, respectively.

Illustrative example: Let us consider an illustrative example of predicting the probability of link (3 4) in a directed network (refer to Fig. 4). To keep it simple, we will focus on triad 110 (shown in Fig. 3), where the links (*u*, *v*) and (*v*, *w*) are present. The triads of structure 110 available are (1, 2, 3), (1, 2, 4), (3, 1, 2), (3, 2, 4), (1, 3, 2)—the first four are open triads and the last one is a closed triad. As per Eqs. (4) and (5), the probability of (3, 4) given triad structure 110 is $P_{x\rightarrow y}(t; G) \times P(c|t; G) = \frac{1}{1} \times \frac{1}{5} = \frac{1}{5}$. Analogously, there are 19 non-triads (i.e., triplets of nodes with at most one link among them), enumerated as follows:

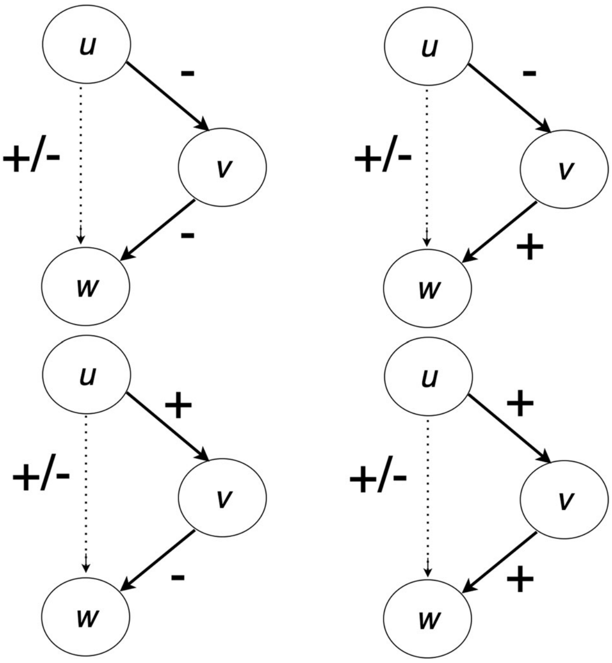

Fig. 5 Signed triads. Four combinations of triad 110 (from Fig. 2) in a directed complex network with signed links.

$$(1,3,4), (1,4,2), (1,4,3), (2,1,3), (2,1,4), (2,3,1), (2,3,4), (2,4,1), (2,4,3)$$

Table 1 Summary of the 50 subnetworks sampled from 10 datasets.

Network	|E|	FFL density	M-L ratio (sampled)	M-L ratio (ER)	κ
E. coli (Schaffter et al., 2011)	3758	9.64×10^{-7}	0.66	0.02	10.0
Yeast (Schaffter et al., 2011)	12,873	1.13×10^{-6}	0.43	0.03	1.0
Mouse (Han et al., 2017)	6490	1.57×10^{-6}	0.70	0.04	0.01
Human (Han et al., 2017)	8427	2.84×10^{-6}	1.31	0.06	0.01
Metabolic (Schellenberger et al., 2010)	5802	6.81×10^{-6}	2.98	0.10	0.01
Reco (Rossi and Ahmed, 2015).	17,359,346	2.90×10^{-6}	0.34	0.06	1.0
Citation (Rossi and Ahmed, 2015)	3,148,447	0.0002	8.77	1.05	1.0
Email (Yin et al., 2017)	25,571	0.0009	15.64	2.81	10.0
Wikipedia (Leskovec et al., 2010b)	103,689	9.38×10^{-5}	4.84	0.61	1.0
Twitter (Rossi and Ahmed, 2015)	834,798	3.09×10^{-7}	0.05	0.01	1.0

The table shows the average number of directed links (|*E*|); the ratio of the average number of FFL motifs to the maximum possible number of FFLs (FFL density); FFL *M*otif-to-*L*ink ratio in sampled networks and the Erdos–Renyi (ER) random graphs of the same graph densities; weighing parameter κ for FFL-based link prediction.

$$(3,1,4),(3,2,1),(3,4,1),(3,4,2),(4,1,2),(4,1,3),(4,2,1),(4,2,3),(4,3,1),(4,3,2).$$

Only 5 of these triads, namely, (1,4,2), (1,4,3), (2,1,4), (2,3,4), (3,4,2) have the (*u*, *w*) link. This makes the probability of (3, 4) given non-triad $P_{x\to y}(\bar{t};G)\times P(c|\bar{t};G)=\frac{1}{2}\times\frac{5}{19}=\frac{5}{38}$. The final probability score is $\frac{1}{5}+\kappa\times\frac{5}{38}$.

Labeled directed links. The idea of FFL motif triad completion can be extended to directed complex graphs with labeled links. An example of a link label can be the sign (i.e., + or −) showing the up-regulation or down-regulation of a gene by a transcription factor in transcriptional networks (see the description of *network datasets* in section "Results"). In signed networks, the number of possible triads will be higher. Figure 5 (top) shows the 4 combinations within the 110 triads. We show the 12 triadic combinations for motif $m \equiv (u,v,w)$ in Fig. 5 (bottom).

The link prediction score for unlabeled network (Eq. (7)) can be extended for signed networks using the set of 12 triadic combinations *T* (Table 1), as

$$S(x \to^{\gamma} y) = \sum_{t\in T} P(\Delta^{\gamma};t,\ G) + \kappa\times P(\Delta^{\gamma};t',\ G) \tag{8}$$

In Eq. (8), $P(\Delta^{\gamma};\, t,G)$ is the probability that triad *t* is closed by $c \equiv (x \to^{\gamma} y)$ forming a signed FFL motif Δ^{γ} and $P(c^{\gamma};\bar{t},G)$ is the probability of $P(\Delta^{\gamma};t',G)$ forming the third link in non-triad t'.

Association rule-mining. Machine learning (ML) models are trained on input (or training) data to make predictions on unseen data (Jordan and Mitchell, 2015). Association rule mining is a class of ML algorithms that learns the relationships among entities from patterns existing in relational databases, such as transactional or medical data (Zhao and Bhowmick, 2003). Given a set of items $\mathcal{I}=\{i_1,i_2,\cdots\}$ and set of transactions $\mathcal{D}=\{d_1,d_2,\cdots\}$, an itemset *x* is defined as a combination of items, i.e., $x \in \mathcal{I}$. Metrics such as support, confidence, and lift are employed to quantify the importance of rules. We utilize support defined as the frequency of appearance of an itemset *x* in the transaction, calculated as

$$\mathrm{Supp}(x) = \frac{d \in \mathcal{D} : x \subset d}{|D|} \tag{9}$$

The support for each itemset is normalized to a score in the range 0 and 1, by dividing by the maximum support.

For any triplet $u, v, w \in V$, we employ the measures of *accuracy* (see section "Accuracy") and *precision* (see section "Precision") (Osisanwo et al., 2017) to estimate the likelihood of the existence of a link $u \to w$, given the existence of $u \to v$ and $v \to w$. Both these measures are defined on the basis of true positive (TP), true negative (TN), false positive (FP), and false negative (FN), as we discuss hereafter.

Accuracy. Expected accuracy in link prediction is given by

$$A = \frac{TP + TN}{TP + TN + FP + FN} \tag{10}$$

We measure expected accuracy using Eq. (11) (defined hereafter). We define a bitstring of length 3, say $[b_{u,v}, b_{v,w}, b_{u,w}]$, where a value of 1 or 0 corresponds to the presence or absence of the directed links $u \to v$, $v \to w$, $u \to w$ in a possible FFL motif $m = (u, v, w)$ among 3 nodes *u*, *v*, *w* in a directed complex network. For example, 001 refers to the triad lacking in $u \to v$, $v \to w$ and the presence of $u \to w$ (illustrated in blue in Fig. 3). Finally, $\eta(001)$ denotes the number of triplets (among all possible triplets *u*, *v*, *w*) that satisfy the configuration of triad 001.

$$A = \frac{\eta(000)+\eta(010)+\eta(100)+\eta(111)}{(\eta(000)+\eta(001))+(\eta(010)+\eta(011))+(\eta(100)+\eta(101))+(\eta(110)+\eta(111))} \tag{11}$$

Precision. It is the fraction of all positive predictions that are truly positive. It is calculated as

$$P = \frac{TP}{TP + FP} \tag{12}$$

In the context of triplet analysis, we measure *expected precision* as

$$P = \frac{\eta(111)}{\eta(111)+\eta(001)+\eta(101)+\eta(011)} \tag{13}$$

As discussed above, we employ accuracy and precision to find support for triad completion: the likelihood of the presence of link $\boldsymbol{u} \to \boldsymbol{w}$, if and only if both $\boldsymbol{u} \to \boldsymbol{v}$ and $\boldsymbol{v} \to \boldsymbol{w}$ are present in a directed complex network (i.e., logical conjunction). Specifically, high accuracy and low precision would indicate that the contribution of the TN towards the accuracy is higher than that of TP.

Subgraph sampling. We sample subgraphs from the directed networks. In Fig. 6, we input TRN *G* and the order (i.e., the required number of nodes in sampled subgraph) *r*, of the required subgraph. We initialize an empty graph *H* and add a well-connected node $u \in V(G)$ as the first node. Subsequently, we iteratively add nodes to *H* by randomly selecting a new node, say *v*, that belongs to the neighborhood of $u \in V(H)$ in *G* and include the directed links between the newly added and existing nodes *u* and *v* from *G*. This is required to ensure a single

```
Algorithm 1 Sample - TRN
Input. Directed graph $G$, $r$
Output. Directed graph $H$
Convert $G$ to undirected graph $G'$
$H \leftarrow \emptyset$
Select random seed $u \in V$ with probability $\frac{deg(u)}{\sum_{v \in V} deg(v)}$
$V(H) \leftarrow V(H) \cup \{u\}$
while $|H| < r$ do
    $v \leftarrow$ random neighbor of $u \in V(G')$ $(u \notin V(H))$, say $v$, from $G'$
    Add $V(H) \leftarrow V(H) \cup \{v\}$
    for $u \in V(H)$ do
        if $(u, v) \in E(G)$ then
            Add a new edge $E(H) \leftarrow E(H) \cup (u, v)$
        end if
        if $(v, u) \in E(G)$ then
            Add a new edge $E(H) \leftarrow E(H) \cup (v, u)$
        end if
    end for
end while
```

Fig. 6 Subgraph sampling algorithm.

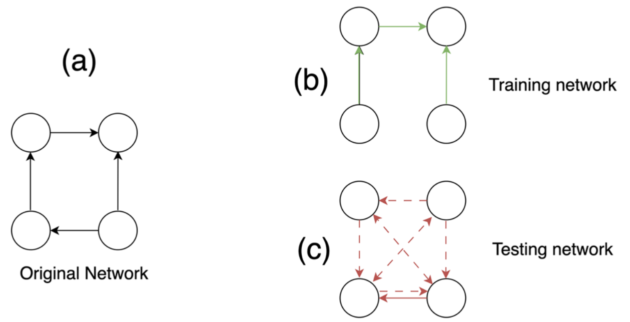

Fig. 7 Training and testing networks. Split of **a** directed complex network dataset into **b** training network with training links E_{train} shown as green dotted lines and **c** testing network with training links colored brown, comprising remaining links $E{-}E_{\text{train}}$ (solid lines) and nonexistent links (dotted lines).

connected component in the generated subgraph. This process terminates when the required subgraph order r is realized.

AUC-ROC. Receiver operator characteristic (ROC) is a binary classification measure that plots true positive rate $(= \frac{\text{TP}}{\text{TP+FN}})$ against false positive rate $(= \frac{\text{FN}}{\text{TP+FN}})$ at different threshold values. Area under the curve (AUC) gives the classifier's ability to distinguish between classes. $0.5 < \text{AUC} \leq 1$ indicates that the classifier distinguishes the positive class values from the negative class values since the classifier is able to detect more numbers of TP and TN than FN and FP (Davis and Goadrich, 2006).

Results

We carry out experiments on 10 directed complex network datasets (enlisted below). For each, we sample 50 subgraphs and record the mean (and standard deviation in) the AUC-ROC. The summary statistics of the sampled subnetworks are shown in Table 1.

Training and testing split. We divide each directed complex network dataset into training and testing datasets. Given a network $G(V, E)$ shown in Fig. 7a, the set of directed training links E_{train} is a subset of the complete set of edges, i.e., $E_{\text{train}} \subset E$. The network $G(V, E_{\text{train}})$ (Fig. 7b) is used to train the link prediction models, while the model is tested as a combination of (1) remaining links $E{-}E_{\text{train}}$ labeled 1 and (2) nonexistent links labeled 0, as illustrated in Fig. 7c.

Network Datasets. Here are the descriptions of the 10 network datasets.

1. *Citations network* comprises authors as nodes and directed edges represent the citation of authors by other authors (Rossi and Ahmed, 2015). It has 28,093 nodes and 3,148,447 links.
2. *Email network* is a network of members of a European institution and directed edges indicate that emails are sent from one member to another (Yin et al., 2017). It contains 1005 nodes and 25,571 links.
3. *Recommendation network* from a Czech dating site, where nodes are users and directed edges represent ratings given by users to other users (Rossi and Ahmed, 2015). It contains 220,970 nodes and 17,359,346 links.
4. *Wikipedia network* comprises a network of Wikipedia 7115 users and 103,689 directed links existing from voter users to users competing to become system administrator (Leskovec et al., 2010b).
5. *Twitter network* is a network of 465,017 social network users and 834,798 directed links pointing from the follower to the followed user node (Rossi and Ahmed, 2015).
6. *Transcriptional regulatory network* (TRN) comprises a network of proteins, called Transcription Factors (TFs), and target genes. Directed links exist from one TF to a TF/gene whose expression it regulates. We consider 4 TRNs:
   - *Escherichia coli* (or *E. coli*) bacteria with 1565 nodes and 3758 links, *Saccharomyces cerevisiae* (or yeast) (Schaffter et al. 2011) with 4441 nodes and 12,873 links
   - Human (with 2862 nodes and 8427 links) and mouse (Han et al. 2017) (with 2456 nodes and 6490 links). For *E. coli* TRN, the links are labeled with positive (+) or negative (−) signs, indicating that the TF is up- or down-regulating a given TF/gene.
7. *Metabolic network* represents the metabolic reactions of *E. coli*, where a node is a metabolite, and each directed link points from the input to the product of a reaction (Schellenberger et al., 2010). It has 1039 nodes and 5802 links.
8. *Dynamic social network of Dutch school friendships* comprising a series of snapshots of the friendships among

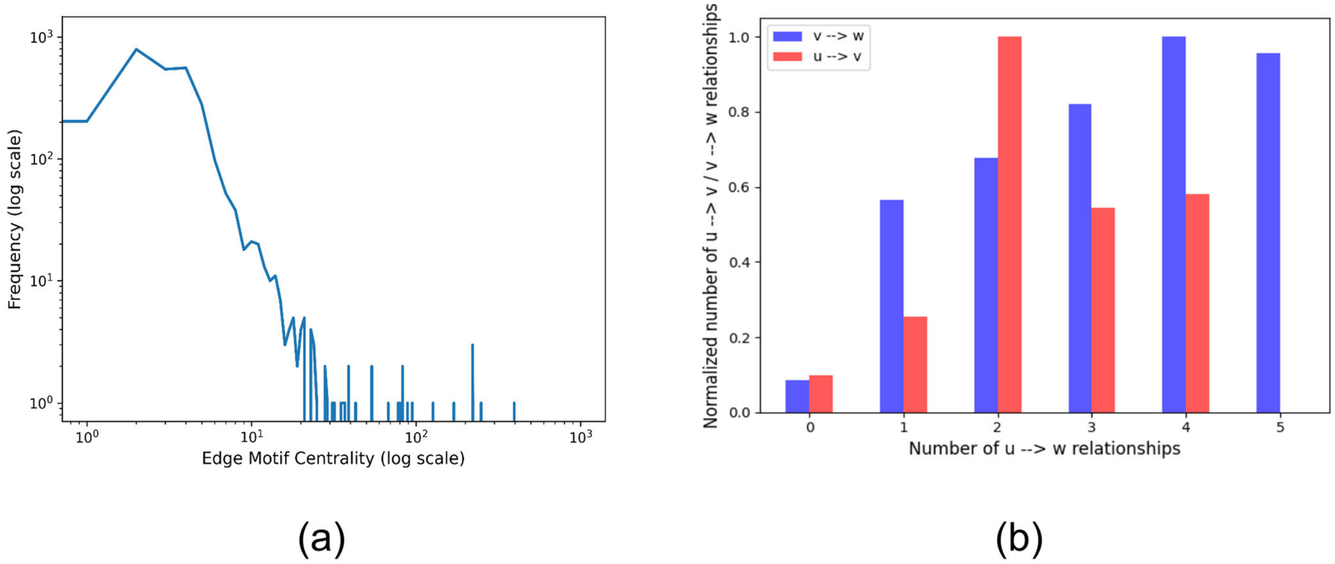

Fig. 8 Motif participation of directed links in *E. coli* transcriptional networks. a Log–log plot of the frequency of links for a given edge motif centrality, showing the decline in the frequency of edges with centrality > 3. **b** Links with high $u \rightarrow w$ participation have high $u \rightarrow v$, $v \rightarrow w$ participation (normalized by maximum $u \rightarrow v$, $v \rightarrow w$ participation among all links).

freshmen at secondary school in The Netherlands, in 2003–2004 (Snijders et al., 2010).

Baseline metrics. We consider the following baseline metrics (all of which, except triadic similarity and DeepWalk, are taken from (Ghorbanzadeh et al., 2021)). For any node $u \in V$, the notations Γ_u, Γ^{i}_u and Γ^{o}_u denote the sets of all neighbors, in-neighbors, and out-neighbors of u.

1. *Jaccard similarity* is a normalized common neighbor-based metric in an undirected network and is calculated as

$$JC(x,\ y) = \frac{\Gamma_u \cap \Gamma_v}{\Gamma_u \cup \Gamma_v} \tag{14}$$

2. *Common neighbor incoming* is a metric based on the common in-neighbors in a directed network and is calculated as

$$CNI\left(x, y\right) = \Gamma^i_u \cap \Gamma^i_v \tag{15}$$

3. *Common neighbor outgoing* is a metric based on the common out-neighbors in a directed network and is calculated as:

$$CNO\left(x, y\right) = \Gamma^{\mathrm{o}}_u \cap \Gamma^{\mathrm{o}}_v \tag{16}$$

4. *Common neighbors incoming and outgoing* combines the CNI and CNO scores by adding them up, i.e.,

$$CNIO\left(x, y\right) = CNI\left(x, y\right) + CNO\left(x, y\right) \tag{17}$$

5. *Adamic Adar* is another common neighbor-based similarity measure that assigns higher importance to a common neighbor with fewer neighbors. It is calculated as:

$$\mathrm{AA}\left(x, y\right) = \sum_{z \in \Gamma_u \cap \Gamma_v} \frac{1}{\log \Gamma_z} \tag{18}$$

6. *Resource allocation* works on a very similar principle, as

$$\mathrm{RA}(x,\ y) = \sum_{z \in \Gamma_u \cap \Gamma_v} \frac{1}{\Gamma_z} \tag{19}$$

7. *Motif triads* measure the likelihood of link existence among nodes x and y in an undirected network based on the number of 13 triadic motifs ($\phi(x,\ y,\ z)$) they share with a common neighbor z (Aghabozorgi and Khayyambashi, 2018). It is calculated as:

$$\mathrm{MS}(x,\ y) = \frac{\sum_{z \in \Gamma_u \cap \Gamma_v} \phi(x,\ y,\ z) \times \frac{1}{13}}{\left|\Gamma_u \cap \Gamma_v\right|} \tag{20}$$

8. *DeepWalk* is a neural network-based approach that learns the latent representation among the nodes in an undirected network while preserving the topological relationships among the nodes (Perozzi et al., 2014). The similarity among nodes u,$v \in V$ can be estimated by the dot product of their latent vectors z_u, z_v, i.e., similarity$(u,\ v) = z^{\mathrm{T}}_u.z_v$.

Motif participation of directed links. Any FFL motif (u,v,w) constitutes three links $u \rightarrow v$, $v \rightarrow w$, $u \rightarrow w$. The number of such FFL motifs a directed link participates in is given by its edge motif centrality (Eq. (1)). As discussed in the section "Introduction", complex networks are characterized by an over-representation of network motifs, i.e., there are more network motifs than expected by chance. Table 1 shows that the average FFL motif per link in the 50 sampled subnetwork datasets is notably higher than those of Erdos–Renyi (ER) random graphs of the same graph densities. At the same time, these networks have very low graph density; hence, the majority of FFL motifs are concentrated among a few directed links, while a vast majority of links do not participate in FFLs. To illustrate this, we show in Table 1 that the ratio between the average number of FFL motifs and the maximum possible number of FFLs (termed FFL *density*) is very low.

We highlight the two features of directed complex networks using the *E. coli* transcriptional regulatory network. Figure 8a is a log–log plot of the frequency of links (on the y-axis) for a given edge motif centrality (on the y-axis), showing the decline in the frequency of edges with edge motif centrality > 3. This reaffirms that the majority of the links have low motif participation. It is worth noting that the few links with high-edge motif centrality typically participate in multiple roles (i.e., $u \rightarrow v$, $v \rightarrow w$, $u \rightarrow w$) of several FFLs. To show this, we record the average number of role $u \rightarrow v$, $v \rightarrow w$ participation given the $u \rightarrow w$ participation of directed links in *E. coli*. Figure 8b depicts that links with high $u \rightarrow w$ participation tend to have high $u \rightarrow v$, $v \rightarrow w$ participation (normalized by maximum $u \rightarrow v$, $v \rightarrow w$ participation among all links). *Overall, this suggests that the presence and the absence of*

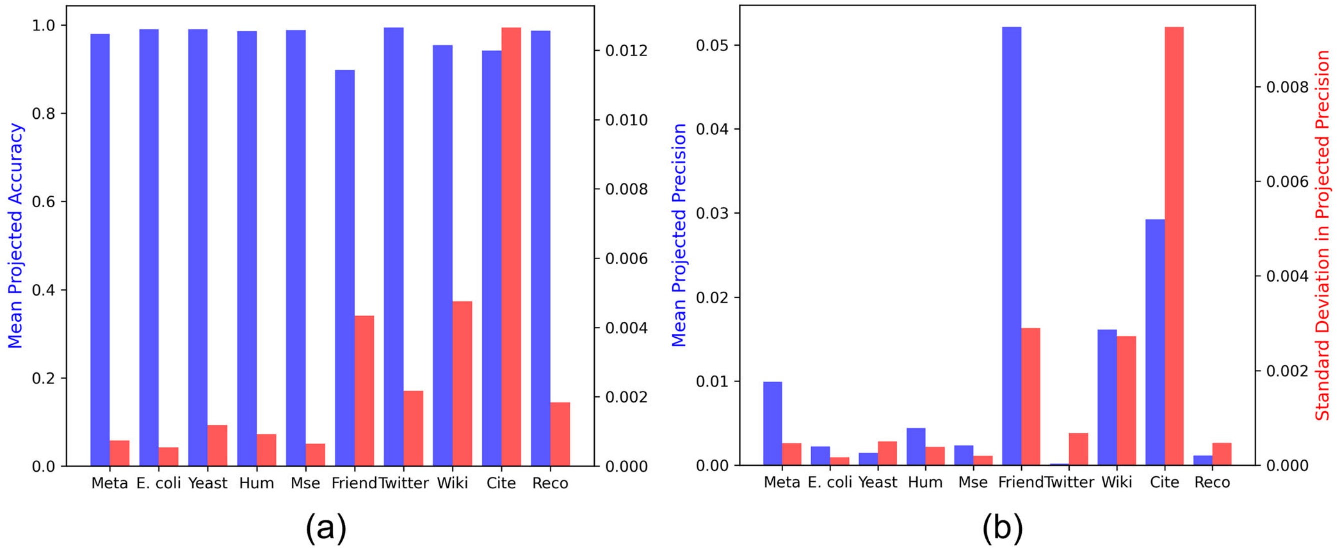

Fig. 9 FFL motif completion. Mean (and standard deviation in the) in the prediction success of $u \rightarrow w$ link given the existence of $u \rightarrow v$ and $v \rightarrow w$ across all possible ordered triplets of nodes in the complex network datasets in terms of **a** accuracy and **b** precision.

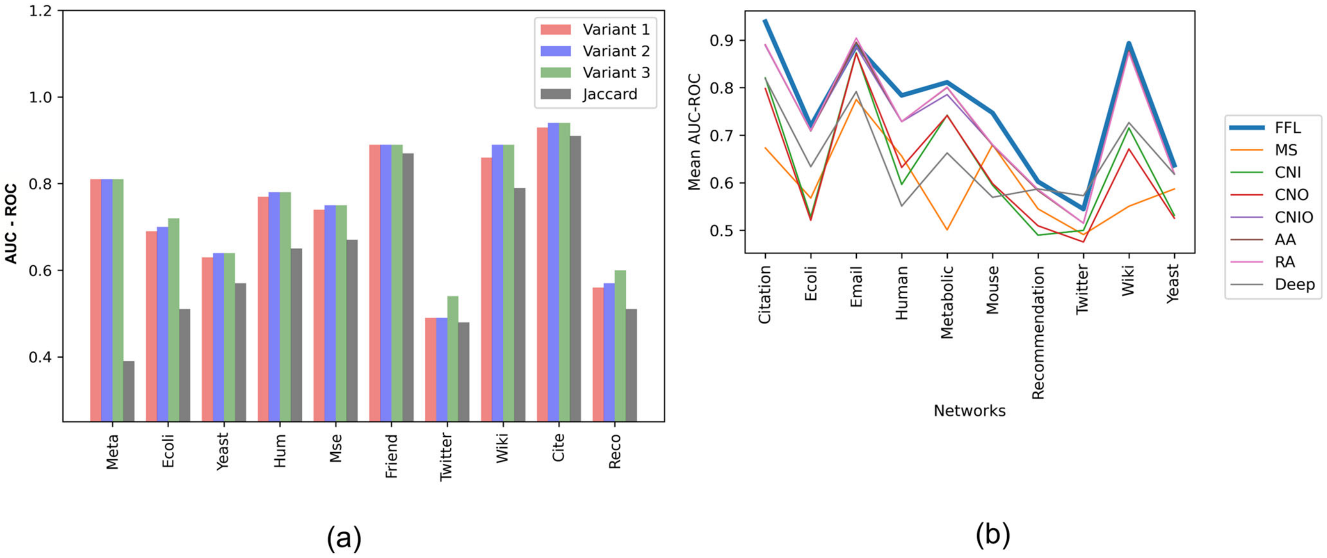

Fig. 10 Comparison with baseline algorithms. AUC-ROC scores of **a** three variants of FFL-based scoring mechanisms and Jaccard. The three variant scores are based on the existence of triads, the sum of existence and absence of triads, and the weighted sum of existence and absence of triads; **b** variant 3 of FFL scoring against baseline algorithms: MS, CNI, CNO, CNIO, AA, RA, and DeepWalk (see baseline metrics in section "Results").

FFL motifs can both be useful measures of link prediction in directed complex networks.

FFL motif completion. The FFL motif triad completion rule enforces the conjunctive logic – it tests the likelihood that the $u \rightarrow w$ link of an FFL motif (u,v,w) exists if and only if both $u \rightarrow v$ and $v \rightarrow w$ are present (refer section "Association rule-mining" for details). We apply the rule to all possible ordered triplets of nodes and record the accuracy scores (Eq. (11)). Figure 9a shows that the mean accuracy scores for the 10 network datasets vary between 89.7% and 99.4%, showing the high tendency of links to complete FFL motif triads.

The high accuracy can be a function of high true positive (TP) as well as true negative (TN) (see section "Accuracy"). To get a clearer picture of which out of TP and TN contribute more to the accuracy, we estimate the *expected precision* (fraction of positive predictions that are TPs; refer to Eq. (13)) for the stated FFL triad completion rule. Figure 9b shows that the mean precision ranges between 0.018% and 5.21% across the 10 networks, suggesting that the high accuracy stems more from the FPs than TPs.

Note that the common neighbor-based link prediction metrics (Jaccard, Adamic Adar, Resource Allocation, etc.) account for both TP and TN by predicting a high likelihood of triad completion if two nodes share common neighbors. However, they fail to account for the fact that the majority of the links in directed complex networks do not participate in any motifs. We show in Table 1 that the sampled subnetworks from the 10 datasets exhibit a low FFL motif density, explaining the need for incorporating both the presence as well as the absence of motif triads into the prediction score.

Comparison with baselines. We evaluate the link prediction accuracy of the following three variants of the proposed scoring metric: a score based on the (**v1**) existence of triads (Eqs. 3), (v2)

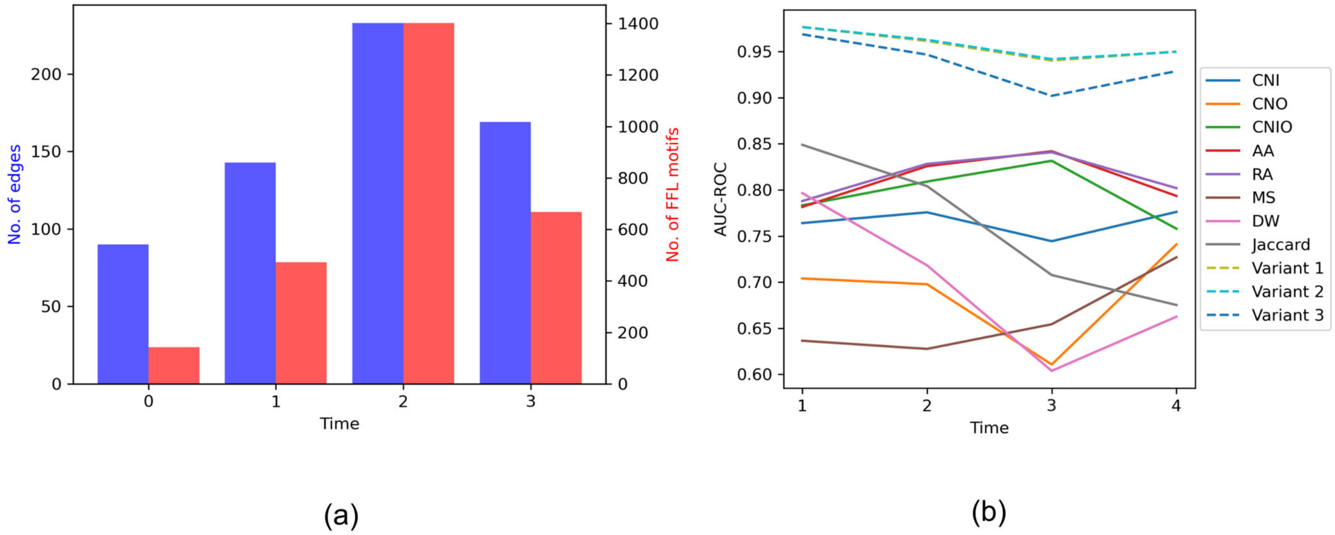

Fig. 11 Dynamic network analysis. a Number of nodes, links, and FFL motifs in four snapshots of a dynamic social network (Wei et al., 2016). **b** Comparison of the mean AUC-ROC scores across 50 runs (on the dynamic network) of three variants of FFL-based scoring mechanisms and Jaccard against those of baseline algorithms: MS, CNI, CNO, CNIO, AA, RA, DeepWalk.

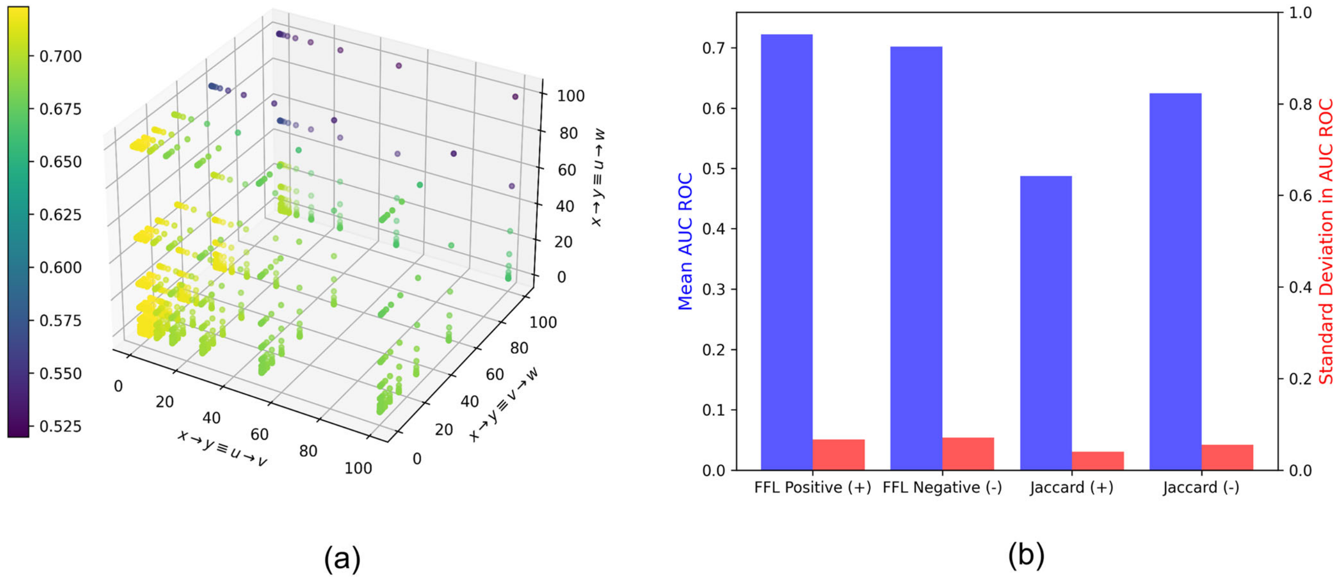

Fig. 12 Variations in labels and weighing parameter for 50 subnetworks of *E. coli* TRN. a Mean AUC-ROC scores for $\kappa_{u\rightarrow v}$, $\kappa_{v\rightarrow w}$, $\kappa_{u\rightarrow w} \in [0.01, 100]$; **b** comparison of the mean (and standard deviation in) AUC-ROC for positive and negative labeled directed links of FFL-based against Jaccard metric.

sum of existence and absence of triads (combining Eqs. (3) and (6), and (v3) weighted sum of existence and absence of triads (Eq. 5). The value of the weighing parameter (κ) is chosen such that the AUC-ROC scores are maximized. We compare the AUC-ROC scores of the 3 variants against the Jaccard similarity baseline for 50 sampled subnetworks of the 10 network datasets. Figure 10a shows that all three variants of the FFL-based metric significantly outperform Jaccard. Furthermore, $v1 \leq v2 \leq v3$ for all the datasets. Clearly, incorporating the absence of triads is an important feature that improves prediction accuracy (hence, $v1 \leq v2$). Controlling the contribution of existence and absence of triads using a weighing parameter κ further improves performance (hence, $v2 \leq v3$). The values of κ for different subnetworks are summarized in Table 1.

We compare the accuracy of FFL-based scoring (variant 3) against other baseline metrics, namely, MS, CNI, CNO, CNIO, AA, RA, and DeepWalk. For DeepWalk, we have considered random walks of length 10, and each node is embedded as vectors of length 16. Figure 10b illustrates that FFL outperforms the baselines for almost all complex networks because it (1) predicts link existence based on the existence and absence of triads and (2) strikes a balance between the two through the weighing parameter (κ). Resource allocation (RA), Adamic Adar (AA), and common neighbor incoming and outgoing (CNIO) come closest to the FFL-based metric for email, Wikipedia, and citation subnetworks, because these subnetworks are relatively dense and have a higher propensity for FFL triad completion, as reported in Table 1. (RA and AA seem indistinguishable as they have the same AUC-ROC for most networks.)

We explore how the motif-based approach performs when applied to different snapshots of time-varying social networks (discussed under *dynamic network* at the beginning of the "Results" section). We record the number of links and FFL motifs in each of the four timepoints of the 26-node dynamic network (see Fig. 11a) and show that the mean AUC score across 50 runs of 3 variants of the motif-based approach, particularly **v2**, is higher than that of the baseline measures (see Fig. 11b).

Weighing parameter and variations in labels. We discuss in section "Weighted sum of the two likelihood" that the parameter κ accounts for the presence of a link given the absence of a triad t (refer to Fig. 2 for all possible triads $t = v \rightarrow w$, $u \rightarrow w$; $u \rightarrow v$, $u \rightarrow w$, and $u \rightarrow v$, $v \rightarrow w$). For any FFL motif m ≡ (u, v, w), the weighing parameter k can be further classified into $\kappa_u \rightarrow v$, $\kappa_v \rightarrow w$, and $\kappa_u \rightarrow w$ corresponding to triad closing link $c \equiv (x,y) \equiv (u, v),(v, w),(u, w)$, respectively, given the absence of triad t (given by $\bar{t}$). The goal is to examine whether the use of specific weighing parameters instead of a generic one κ could better control the contribution of different triadic patterns toward the directed link (x, y) to be predicted. To study this, we define i as 15 evenly spaced real numbers in range $[-2, 2]$ and generate each of $\kappa_{u\rightarrow v}$, $\kappa_{v\rightarrow w}$ $\kappa_{u\rightarrow w} = 10^i$ $\forall i$. The resultant κ parameters range between [0.01, 100], where more values are close to 0.01 than they are to 100. This is helpful since the highest accuracy scores are achieved for low values (<10) of κ (see Table 1). Figure 12a depicts that the mean AUC-ROC over 50 sampled *E. coli* TRN subnetworks, where the best scores are once again achieved for lower κ values. The best mean AUC-ROC scores of ≈0.75 ± 0.03 are achieved for $\kappa_{u\rightarrow v}$, $\kappa_{v\rightarrow w}$, $\kappa_{u\rightarrow w} = 0.14$, 1.93, 0.01, which outperforms the earlier best FFL accuracy (≈0.72 ± 0.04) with a common κ. This suggests that including parameters that control different FFL motif substructures can improve prediction accuracy. Needless to say, these parameter values need fine-tuning, as they may vary across network types.

The FFL-based metric can effectively predict link labels. In the case of *E. coli* transcriptional network (TRN), the links are signed positive (+) or negative (−). We employ the signed variant of the FFL metric (Eq. (8)) that employs 12 signed FFL triad variants to predict the signs of *E. coli* TRN and compare the AUC-ROC scores for both signs against the Jaccard metric. Figure 12b shows that for both signs, the proposed FFL-based link prediction metric outperforms Jaccard.

Discussion

This metric opens up several research directions. *First*, there is a growing interest in the area of inference models for dynamic social and biological networks with the goal of learning the changing relationship among genes, proteins, metabolites, etc. (Oates and Mukherjee, 2012; Nalluri et al., 2017). In the domain of genomics, the advent of high-throughput data acquisition tools has facilitated access to finer-resolution datasets, such as single-cell data (in addition to bulk tissue data). In the existing study, the FFL-based metric predicts the link existence score based on the likelihood of link existence given the presence (and absence) of directed triads in the entire network. Since motifs are abundantly present in biological networks, given any time-series data of fine resolution, machine learning models can be trained on the history of motif participation of nodes of interest (namely, cells, proteins, genes, etc.) to predict future connections. We have demonstrated in the section "Comparison with baselines" that the proposed approach, when applied to snapshots of dynamic networks, outperforms baselines. We are currently extending this approach to temporal links (i.e., links connecting two different timepoints) in spatiotemporal networks as well as predict future connections in other evolving directed complex networks. Some application domains are (1) management (i.e., related to security, performance, and troubleshooting) of engineered networks through the identification of beneficial and harmful links (Liao and Striegel, 2012) and (2) improving recommendations in e-commerce and academic networks (Benchettara et al., 2010b).

Second, using the signed *E. coli* transcriptional network as a case study, we have demonstrated that the FFL-based metric can be useful to predict labels of directed links. It is apparent that there will be considerable computational complexity as the number of possible labels (and consequently directed triads) increases. To handle this, it will be necessary to identify triads that do not contribute to the overall link existence score. This process of selecting informative triads can be mapped to a feature selection problem and these triads can vary greatly across complex networks. Third, we would like to generalize the proposed metric to other network motifs. While two- and three-node network motifs are more widely studied (Jiang et al., 2006), such an analysis promises to reveal higher-order motifs of interest in specific complex network types. Finally, network motifs of varying orders can be used as features to build and train deep machine-learning models that predict the existence of links in large (and dynamic) networks.

Conclusions

In this work, we presented a metric that leverages the feed forward loop (FFL) motifs to predict links in directed complex networks. Existing common neighbor-based link prediction metrics rely solely on the triadic-richness of complex networks and gauge the likelihood of link existence based on the number of shared triads between a pair of nodes. We propose to account for the sparse topologies of directed complex networks that make the absence of triads an equally important criterion for link prediction. Thus, the proposed FFL-based metric measures score as the weighted sum of the likelihood of a link given the presence and absence of directed triads. Our experimental analysis of 10 real complex network datasets (biological, social, communication, etc.) shows that the proposed metric exhibits a significantly higher AUC-ROC than standard baseline approaches.

Data availability

The code and datasets generated during and/or analyzed during the current study are available in the GitHub repository, https://github.com/satunr/MOTIF_LP.

Received: 10 October 2022; Accepted: 19 June 2023;
Published online: 29 June 2023

References

Aghabozorgi F, Khayyambashi M (2018) A new similarity measure for link prediction based on local structures in social networks. Physica A 501:12–23

Antweiler D et al. (2021) Towards the detection and visual analysis of COVID-19 infection clusters. In: EuroVis workshop on visual analytics, European Association for Computer Graphics

Barabasi A (2009) Scale-free networks: a decade and beyond. Science 325(5939):412–413

Benchettara N, Kanawati R, Rouveirol C (2010b) Supervised machine learning applied to link prediction in bipartite social networks. In: 2010 international conference on advances in social networks analysis and mining. Institute of Electrical and Electronics Engineers (IEEE), pp. 326–330

Benchettara N, Rushed K, Céline C (2010a) A supervised machine learning link prediction approach for academic collaboration recommendation. In: Proceedings of the fourth ACM conference on Recommender systems. Association for Computing Machinery (ACM)

Bhagat S et al. (2010) Privacy in dynamic social networks. In: Proceedings of the 19th international conference on World wide web. Association for Computing Machinery (ACM)

Chen W, Hui Q, Kuo C (2021) Partner selection in China interorganizational patent cooperation network based on link prediction approaches. Sustainability 13.2:1003

Chen Y, Hsiao C, Wu C (2022) An ensemble model for link prediction based on graph embedding. Decision Support Syst 157:113753

Cheung M et al. (2020) Graph signal processing and deep learning: convolution, pooling, and topology. IEEE Signal Process Mag 37(6):139–149

Crichton G et al. (2018) Neural networks for link prediction in realistic biomedical graphs: a multi-dimensional evaluation of graph embedding-based approaches. BMC Bioinform 19(1):1–11

Davis J, Goadrich M (2006) The relationship between precision-recall and roc curves. In: Proceedings of the 23rd international conference on machine learning. pp. 233–240

Dey A et al. (2019) What network motifs tell us about resilience and reliability of complex networks. Proc Natl Acad Sci USA 116(39):19368–19373

Esslimani I, Brun A (2011) Densifying a behavioral recommender system by social networks link prediction methods. Soc Netw Anal Min 1:159–172

Fan T et al. (2019) Information spread link prediction through multi-layer of social network based on trusted central nodes. Peer-to-Peer Netw Appl 12:1028–1040

Ghorbanzadeh H et al. (2021) A hybrid method of link prediction in directed graphs. Expert Syst Appl 165:113896

Ghosh P, Mayo M, Chaitankar V, Habib T, Perkins E, Das S (2011) Principles of genomic robustness inspire fault-tolerant WSN topologies: a network science-based case study. In: 2011 IEEE international conference on Pervasive Computing and Communications Workshops (PERCOM Workshops). Institute of Electrical and Electronics Engineers (IEEE), pp. 160–165

Gorochowski T, Grierson C, Di Bernardo M (2018) Organization of feed-forward loop motifs reveals architectural principles in natural and engineered networks Sci Adv 4(3):eaap9751

Han H et al. (2017) Trrust v2: an expanded reference database of human and mouse transcriptional regulatory interactions. Nucleic Acids Res 46(D1):D380–D386

Han Z et al. (2007) Signal transduction network motifs and biological memory. J Theor Biol 246(4):755–761

Huo Z, Xiao H, Xia H (2018) Link prediction with personalized social influence. In: Proceedings of the AAAI conference on artificial intelligence, vol. 32 Association for the Advancement of Artificial Intelligence (AAAI), (1)

Jia S, Gao L, Gao Y, Nastos J, Wen X, Zhang X, Wang H (2017) Exploring triad-rich substructures by graph-theoretic characterizations in complex networks. Physica A 468:53–69

Jiang B et al. (2020) Social recommendation in heterogeneous evolving relation network. In: Computational Science–ICCS 2020: 20th International Conference, Amsterdam, The Netherlands, June 3–5, 2020, Proceedings, Part I 20. Springer International Publishing

Jiang R, Tu Z, Chen T, Sun F (2006) Network motif identification in stochastic networks. Proc Natl Acad Sci USA 103(25):9404–9409

Jin S, Zafarani R (2017) Emotions in social networks: distributions, patterns, and models. In: Proceedings of the 2017 ACM on conference on information and knowledge management. Association for Computing Machinery (ACM), pp. 1907–1916

Jordan M, Mitchell T (2015) Machine learning: trends, perspectives, and prospects. Science 349(6245):255–260

Koschützki D, Schwöbbermeyer H, Schreiber F (2007) Ranking of network elements based on functional substructures. J Theor Biol 248(3):471–479

Kwon O, Hang-Hyun J (2023) Clustering and link prediction for mesoscopic COVID-19 transmission networks in Republic of Korea. Chaos 33.1:013107

Leskovec J, Huttenlocher D, Kleinberg J (2010b) Signed networks in social media. In: Proceedings of the SIGCHI conference on human factors in computing systems. pp. 1361–1370

Leskovec J, Daniel H, Jon K (2010a) Predicting positive and negative links in online social networks. In: Proceedings of the 19th international conference on World wide web

Liao Q, Striegel A (2012) Intelligent network management using graph differential anomaly visualization. In: 2012 IEEE network operations and management symposium. Institute of Electrical and Electronics Engineers (IEEE), pp. 1008–1014

Liben-Nowell D, Kleinberg J (2003) The link prediction problem for social networks. In Proceedings of the 12th international conference on Information and knowledge management. Association for Computing Machinery (ACM), pp. 556–559

Lim M et al. (2019) Situation-aware deep reinforcement learning link prediction model for evolving criminal networks. IEEE Access 8:16550–16559

Lim M et al. (2021) Performance optimization of criminal network hidden link prediction model with deep reinforcement learning. J King Saud Univ-Comput Inf Sci 33.10:1202–1210

Liu G, Li DZ, Jiang CS, Wang W (2014) Transduction motif analysis of gastric cancer based on a human signaling network. Braz J Med Biol Res 47:369–375

Ma N et al. (2022) Link prediction in supernetwork: risk perception of emergencies. J Inf Sci 48.3:374–392

Ma'ayan A et al. (2005) Formation of regulatory patterns during signal propagation in a mammalian cellular network. Science 309(5737):1078–1083

Mayo M, Abdelzaher A, Ghosh P (2015) Long-range degree correlations in complex networks. Comput Soc Netw 2(4):1–13

Milo R et al. (2002) Network motifs: simple building blocks of complex networks. Science 298(5594):824–827

Nalluri J, Barh D, Azevedo V, Ghosh P (2017) mirsig: a consensus-based network inference methodology to identify pan-cancer mirna–mirna interaction signatures. Sci Rep 7(39684):1–14

Newman M (2003) The structure and function of complex networks. SIAM Rev 45(2):167–256

Oates C, Mukherjee S (2012) Network inference and biological dynamics. Ann Appl Stat 6(3):1209

Osisanwo F et al. (2017) Supervised machine learning algorithms: classification and comparison. Int J Comput Trends Technol 48(3):128–138

Ouzienko V, Zoran O (2014) Imputation of missing links and attributes in longitudinal social surveys. Mach Learn 95:329–356

Papadimitriou A, Panagiotis S, Yannis M (2012) Fast and accurate link prediction in social networking systems. J Syst Softw 85.9:2119–2132

Paul D, Radde N (2016) Robustness and filtering properties of ubiquitous signaling network motifs. IFAC-PapersOnLine 49(26):120–127

Perozzi B, Al-Rfou R, Skiena S (2014) DeepWalk: Online learning of social representations. In: Proceedings of the 20th ACM SIGKDD international conference on knowledge discovery and data mining. pp. 701–710

Rossi R, Ahmed N (2015) The network data repository with interactive graph analytics and visualization. AAAI

Roy S, Ghosh P, Barua D, Das SK (2020) Motifs enable communication efficiency and fault-tolerance in transcriptional networks. Sci Rep 10(1):1–15

Roy S, Raj M, Ghosh P, Das S (2017) Role of motifs in topological robustness of gene regulatory networks. In: 2017 IEEE International Conference on Communications (ICC). Institute of Electrical and Electronics Engineers (IEEE), pp. 1–6

Roy S, Roy P, Nadendla V, Das SK (2021) Influence spread control in complex networks via removal of feed forward loops. In: 2021 International Conference on Computer Communications and Networks (ICCCN). IEEE, pp. 1–9

Sampaio Filho C et al. (2015) Mandala networks: ultra-small-world and highly sparse graphs. Sci Rep 5(1):1–6

Saxena A, Fletcher G, Pechenizkiy M (2022) Nodesim: node similarity-based network embedding for diverse link prediction. EPJ Data Sci 11(1):24

Schaffter T, Marbach D, Floreano D (2011) Genenetweaver: in silico benchmark generation and performance profiling of network inference methods. Bioinformatics 27(16):2263–2270

Schellenberger J, Park J, Conrad T, Palsson B (2010) Bigg: a biochemical genetic and genomic knowledgebase of large-scale metabolic reconstructions. BMC Bioinform 11(1):1–10

Shibata N, Yuya K, Ichiro S (2012) Link prediction in citation networks. J Am Soc Inf Sci Technol 63.1:78–85

Simmel G (1908) Soziologie. Duncker & Humblot, Leipzig

Snijders T, Gerhard V, Steglich C (2010) Introduction to stochastic actor-based models for network dynamics. Soc Netw 32.1:44–60

Stone L, Simberloff D, Artzy-Randrup Y (2019) Network motifs and their origins. PLoS Comput Biol 15(4):e1006749

Strogatz S (2001) Exploring complex networks. Nature 410(6825):268–276

Tang W, Constantino P, Daoutidis P (2019) Optimal sparse network topology under sparse control in laplacian networks. IFAC-PapersOnLine 52(20):273–278

Tsugawa S, Hiroyuki O (2013) Effectiveness of link prediction for face-to-face behavioral networks. PLoS ONE 8(12):e81727

Wang D et al. (2011) Human mobility, social ties, and link prediction. In: Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining. Association for Computing Machinery (ACM)

Wang J, Provan G (2009) On motifs and functional modules in complex networks. In: 2009 IEEE Toronto International Conference Science and Technology for Humanity (TIC-STH). Institute of Electrical and Electronics Engineers (IEEE), pp. 78–82

Wang P, Lü J, Yu X (2014) Identification of important nodes in directed biological networks: a network motif approach. PLoS ONE 9(8):e106132

Wang X, Chen G (2003) Complex networks: small-world, scale-free and beyond. IEEE Circuits Syst Mag 3(1):6–20

Wei J et al. (2016) Link prediction analysis of internet public opinion transfer from the individual perspective. Data Anal Knowl Discov 32.1:55–64

Wu F, Ma C, Tan C (2016) Network motifs modulate draggability of cellular targets. Sci Rep 6(1):1–11

Yang S et al. (2012) Fpriend or frenemy? Predicting signed ties in social networks. In: Proceedings of the 35th international ACM SIGIR conference on research and development in information retrieval. Association for Computing Machinery (ACM)

Yin H et al. (2017) Local higher-order graph clustering. In: Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining. Association for Computing Machinery (ACM), pp. 555–564

Zhang J et al. (2017) Uncovering mechanisms of co-authorship evolution by multirelations-based link prediction. Inf Process Manag 53.1:42–51

Zhang M, Chen Y (2018) Link prediction based on graph neural networks. Adv Neural Inf Process Syst 31

Zhao Q, Bhowmick S (2003) Association rule mining: a survey. Nanyang Technological University, Singapore, p. 135

Acknowledgements

This work was supported by NSF CBET 1802588.

Author contributions

SR conceptualized the work. SR and PG developed the methodology. SR and AM performed data curation, analysis, and visualization. All three authors participated in the validation and writing.

Funding

Open Access funding provided by the National Institutes of Health (NIH).

Competing interests

The authors declare no competing interests.

Ethical approval

This article does not contain any studies with human participants performed by any of the authors.

Informed consent

This article does not contain any studies with human participants performed by any of the authors.

Additional information

Correspondence and requests for materials should be addressed to Satyaki Roy.

Reprints and permission information is available at http://www.nature.com/reprints

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