


SCIENTIFIC REPORTS



OPEN

Topological motifs populate complex networks through grouped attachment

Jaejoon Choi^{1,2} & Doheon Lee^{1,3} 

Network motifs are topological subgraph patterns that recur with statistical significance in a network. Network motifs have been widely utilized to represent important topological features for analyzing the functional properties of complex networks. While recent studies have shown the importance of network motifs, existing network models are not capable of reproducing real-world topological properties of network motifs, such as the frequency of network motifs and relative graphlet frequency distances. Here, we propose a new network measure and a new network model to reconstruct real-world network topologies, by incorporating our Grouped Attachment algorithm to generate networks in which closely related nodes have similar edge connections. We applied the proposed model to real-world complex networks, and the resulting constructed networks more closely reflected real-world network motif properties than did the existing models that we tested: the Erdős–Rényi, small-world, scale-free, popularity-similarity-optimization, and nonuniform popularity-similarity-optimization models. Furthermore, we adapted the preferential attachment algorithm to our model to gain scale-free properties while preserving motif properties. Our findings show that grouped attachment is one possible mechanism to reproduce network motif recurrence in real-world complex networks.

Researchers have developed network models for real-world systems such as protein-protein interactions (PPIs), author collaborations, the World Wide Web (WWW), and social networks in order to analyze the relationship between the functions and structures in those real-world systems. Each real-world system has its own properties that can be described in terms of network measures such as network centralities, average path length, and degree distribution. The three classic models for describing real-world properties are the Erdős–Rényi (ER)¹, small-world (SW)², and scale-free (SF)³ models. Although several variations of these standard network models and other models have been proposed, these three models are still widely used in network analysis^{4–6}. Recently, two hyperbolic geometrical models have been developed: popularity-similarity-optimization (PSO)⁷, and nonuniform popularity-similarity-optimization (nPSO)^{8,9} models. These models have been proved to be able to reproduce real-world properties such as clustering, small-worldness, power-lawness, rich-clubness and community structure^{8,10,11}.

Network motifs are recurrent and statistically significant partial subgraphs or patterns¹², and graphlets are small connected non-isomorphic induced subgraphs¹³. Although these two concepts are defined slightly differently, they are commonly used interchangeably. Various studies on topological measures of networks have highlighted the importance of network motifs and graphlets in analyzing real-world networks properties, including scale-free, geometric, complex, or high-order networks^{13–19}. Some proposed topological measures of network motifs and graphlets include frequency of network motif¹⁴, graphlet degree distributions (GDD)¹⁵, and relative graphlet frequency distances (RGF-distances)¹³.

Here we suggest a new network measure which represents real-world topological properties and a new network model incorporating the grouped attachment (GA) to resemble real-world topological properties of network motifs. To validate the GA models, we show that the GA model networks have motif properties more similar to real-world networks than other tested conventional models.

¹Bio-Synergy Research Center, 291 Daehak-ro, Yuseong-gu, Daejeon, Republic of Korea. ²Department of Genetics, Harvard Medical School, 77 Avenue Louis Pasteur, Boston, Massachusetts, United States of America. ³Department of Bio and Brain Engineering, Korea Advanced Institute of Science and Technology (KAIST), 291 Daehak-ro, Yuseong-gu, Daejeon, Republic of Korea. Correspondence and requests for materials should be addressed to D.L. (email: dhlee@kaist.ac.kr)

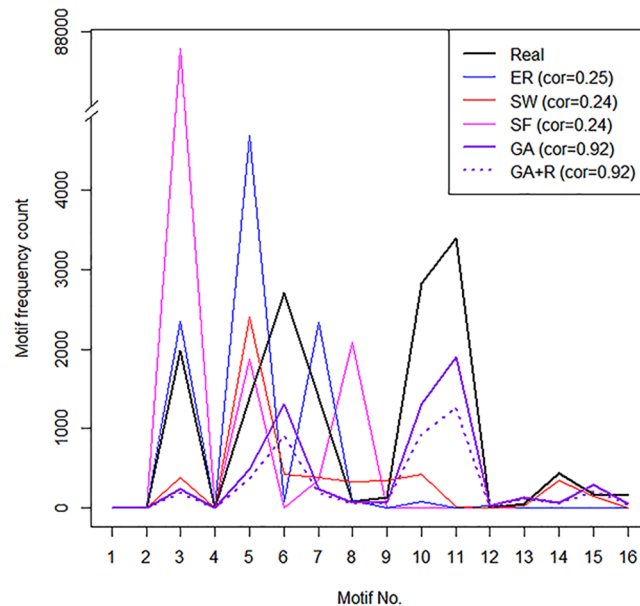


Figure 1. Motif property analysis result of canonical Wnt signaling pathway in the NCI/Nature database. Directed 3-node motif frequency of a real-world network and its corresponding model networks of existing network models (ER, SW and SF) and our models (GA and GA + R). The horizontal axis is directed 3-node motif number, and the vertical axis is motif frequency count. Pearson correlation coefficients between the real network and corresponding model networks are stated in legends. GA and GA + R models show higher similarity of motif frequencies to the real-world network compared to other network models. Legend: ER = Erdős-Rényi; SW = Small-world; SF = Scale-free; GA = Grouped attachment.

Results

Network motifs in network models. While recent studies have shown the importance of analyzing network motifs and graphlets^{20–25}, current network models are not capable of reproducing real-world topological properties of network motifs. To show the incapability of the network models, we examined network motif frequencies of the canonical Wnt signaling pathway (see Supplementary Fig. S1) in the NCI (National cancer institute)/Nature database. For the given real network, we generated corresponding model networks by the network models (ER, SW and SF). For each network model, 100 random networks are generated with input parameters optimized from the real network (See method session for detailed description of network generation), and directed 3-node motif frequency distributions of the networks are examined.

As we show in Fig. 1, none of previous network models (ER, SW and SF) reproduced a motif frequency distribution of the real-world network (black line) compared to our proposed model (GA and GA + R) networks (purple lines). GA model networks have significantly high correlations (correlation coefficients = 0.92) with the real network compared to previous network models. This result shows that previous network models have quite different network motifs from the real network, and our proposed model networks have higher network motif similarity to the real network compared to other network models.

Co-neighborhood of a graph. To begin to address a new network model which can reproduce real-world topological properties of network motifs, we focus on the concept of common neighbors index²⁶ (Fig. 2). Common neighbors index represents the likelihood that two nodes interact increases if overlap of their first-node-neighbors (adjacent nodes) increases. Several studies^{27–32} claimed that nodes in the same community or cluster have high similarities and common neighbors index are highly related to community or cluster structures in networks. Nodes in the same cluster can be clustered based on various criteria such as vertex connectivity or neighborhood similarity²⁸, and nodes in a community structure show high similarities²⁹. We assumed that neighborhood similarities of related nodes could be a key solution to reproduce real-world topological properties of network motifs. Therefore, we suggest a new network measure, co-neighborhood, which shows neighborhood similarity of nodes in the network.

Let $G = (V, E)$ be a graph with node set V and edge set E . We defined co-neighborhood of a graph as an average value of Jaccard's indices^{33,34} of edges in the graph:

$$cn(G) = \frac{1}{|E|} \sum_{e \in E} JC(u, v), \quad (1)$$

where two nodes $u, v \in V$ are connected by an edge $e \in E$, and $JC(u, v)$ is a Jaccard's index of node u and v . Jaccard's index is normalized common neighbors index^{33,34}. (See method section for the equation of Jaccard's index).

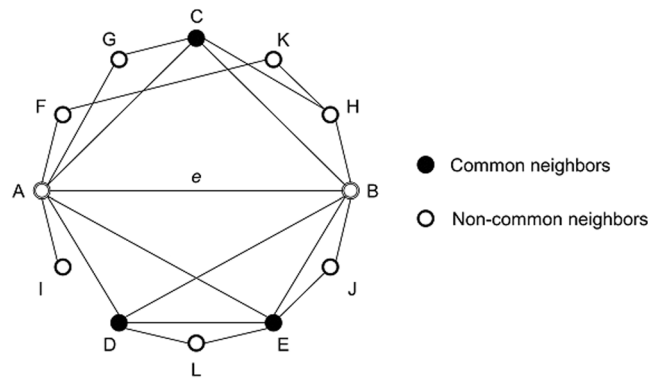


Figure 2. Common neighbors and Jaccard's index of an edge. The figure shows an example of common neighbors. Among all neighbor nodes (C, D, E, F, G, H, I, J, K and L) of edge e (A-B), nodes which are adjacent to both of A and B are co-neighbor nodes (C, D and E). Jaccard's index of edge e (A-B) is $3/10 = 0.3$.

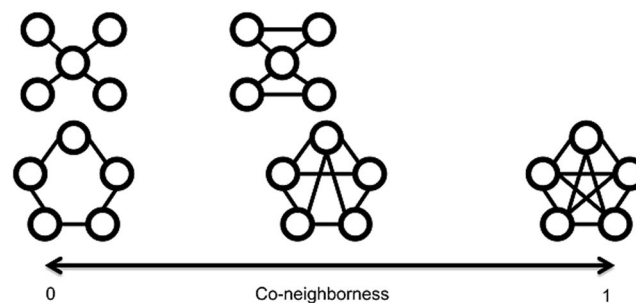


Figure 3. Co-neighborhood of a graph. Example graphs are illustrated with their co-neighborhood. A graph with no triangle subgraph (fully connected 3-node subgraph) has co-neighborhood as 0. A fully connected graph has co-neighborhood as 1.

The co-neighborhood of a graph has range from 0 to 1. If the co-neighborhood is close to 0, few common neighbors of an edge (two adjacent nodes) exist. If the co-neighborhood is close to 1, most adjacent nodes of an edge (two adjacent nodes) are common neighbors. A graph with no 'triangle subgraph' (fully connected 3-node subgraph) has co-neighborhood as 0. A fully connected graph has co-neighborhood as 1 (Fig. 3).

There are several existing network measures (see Supplementary Note) which are related to co-neighborhood such as graph density or average/global clustering coefficient³⁵. These measures can have high correlation with co-neighborhood, but have different values as co-neighborhood has a distinctive definition (see Supplementary Table S3).

Grouped attachment. To reflect the real-world co-neighborhood property, we consider the following grouped attachment procedure (Fig. 4). We get three values as input parameters: (1) n as a node count, (2) p as an edge probability ($0 < p < 1$) and (3) q as a groupness probability ($p < q < 1$, see Supplementary Note). Starting from a single node graph G_0 , at every repeat we create and add a highly interconnected graph F , which is generated by an edge extension model (see Supplementary Note) with an input probability q . Then, we create edges that connect the nodes in the graph F to the nodes in the graph G . Among the nodes in the graph G , we select nodes with a probability p/q to be connected, and for every selected nodes, we connect to the nodes in the graph F with a probability q ($p < q < 1$). We repeat the procedures until the graph G has n nodes.

By following the procedures above, we can generate edges connecting the graph G to the graph F with a probability p in total. Compared to a random selection of edges with a probability p , the procedures guarantee high neighborhood similarities of nodes in the graph F , when q has a high value. If q has a low value, edges are created almost similar as random selection with probability p , which leads to generate a graph similar to the ER model.

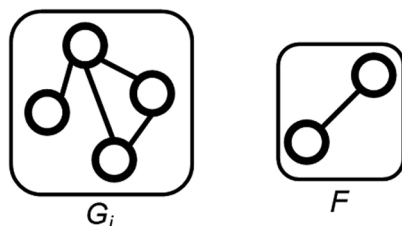
The grouped attachment has similarities and differences with the preferential attachment (SF model)³. Indicated by their names, both models have growing characteristics which is implemented through attachments. Both models start with a small number of nodes (one node for preferential attachment), and at every repeat a new group of nodes (one node for preferential attachment) are added with edges that link the new nodes to the nodes already present in the system. However, grouped attachment does not preferentially attach nodes, which means nodes are not connected depending on degree of nodes. Instead, grouped attachment adds group of nodes at each repeat, while preferential attachment adds a single node at each repeat.

Furthermore, we implemented GA with revised p model (GA + R model) to adjust a total edge density to be p . As represented in the grouped attachment explanation, the procedures only guarantee a density of edges

Step 1. Start node

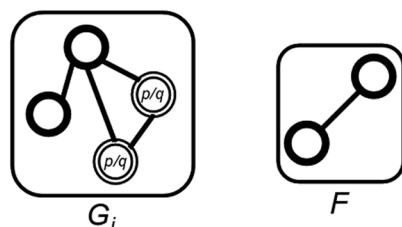


Step 2. Subgraph F generation

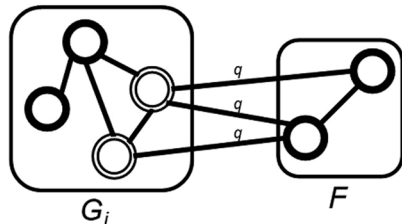


Step 3. Connecting edge generation

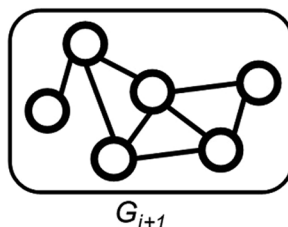
Step 3-1. Target node selection



Step 3-2. Edge generation



Step 4. Graph G update



Step 5. Repeat Step 2-Step 4

Figure 4. Network generation procedures of grouped attachment model. The GA model gets three input parameters; n : node count, p : edge probability, and q : groupness probability. Like the preferential attachment model, attachments are processed repeatedly.

connecting the existing graph with the added graph (edges between the graph G and the graph F) to be p . On the other hand, a density of edges of an added graph (the Graph F) is independent of p . As our model is designated to have a total edge density as p (like the ER model), we adjusted a density of connecting edges (edges between the graph G and the graph F) to be p' , which guarantees a total edge density to be p (Fig. 5). We deduced p' by calculating an edge density of the added region (edges in the graph F and connecting region between the graph G and F) for each repeat (see Supplementary Note for p' calculation).

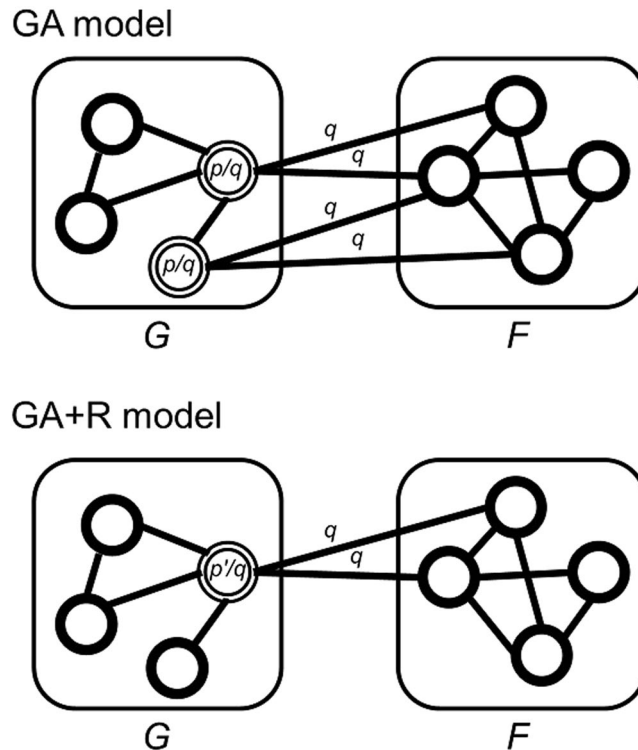


Figure 5. Edge generation procedures of GA and GA + R model. The figure shows an example of edge generation procedures (Step 3 of Fig. 4) of both models. Both GA and GA + R models are supposed to have edge density as p , while edge density of graph F is not guaranteed to be p . To adjust total edge density to be p , the edge generation procedure of GA + R model selects target nodes with probability of p'/q , instead of p/q , which change the edge density of connecting edges (between graph G and graph F) from $p (= p/q \times q)$ to $p' (= p'/q \times q)$.

Topological property of GA model networks. To validate our models, we computed RGF-distances²⁵ between a canonical Wnt signaling pathway in the NCI/Nature database (the pathway used in the introduction session) and its corresponding model networks of existing network models (ER, SW and SF) and our models (GA and GA + R) (see Supplementary Fig. S7). Creation processes of corresponding model networks of existing models are described in the Method session. RGF-distance compares the frequencies of the graphlets in two networks. To find the optimized q values for our models, we applied various q values ranging from 0.1 to 0.9 with an interval 0.1 (see Supplementary Note). Furthermore, we performed the same experiments to various types of undirected networks from Network Repository³⁶, co-authorship network of scientists, airport network among cities, and real world road network (see Supplementary Note).

As described ahead, Fig. 1 shows directed 3-node motif frequency distributions of a canonical Wnt signaling pathway and its corresponding model networks of the existing network models (ER, SW and SF) and our models (GA and GA + R) with optimized q values (0.8 and 0.8, respectively). None of the existing network models reproduced a motif frequency distribution of the real network. On the other hand, GA models show similar shapes (high frequency at motif No. 3, 5, 6, 7, 10 and 11) of directed 3-node motif frequency distributions with the real network (purple lines in Fig. 1).

Supplementary Table S1 shows RGF-distances between canonical Wnt signaling pathway and its corresponding model networks of the existing network models (ER, SW and SF) and our models (GA and GA + R). Having low RGF-distances can be interpreted as they have more similar graphlet frequencies to the real-world network. GA models with optimized q values (0.8 and 0.8, respectively) show better results (lower values) in RGF-distances compared to the existing network models.

According to the results of motif frequency distributions and RGF-distances, our models showed better performances compared to the existing network models. The importance of the results is implied by the similar patterns (high frequency at motif No. 3, 5, 6, 7, 10 and 11) of 3-node motif frequency distribution of GA models with the real network, while the existing network models showed different aspects. GA models had better performances compared to the existing network models not only on the RGF-distances, but also on the aspects of 3-node motif frequency distribution. These results indicate that GA models reproduce motif properties of the real network better than other models.

Table 1 shows RGF-distances and co-neighborhood values between various real-world networks (canonical Wnt signaling pathway, co-authorship of scientists, airport network among cities, and real-world road network) and their corresponding model networks of the existing network models (ER, SW, SF, and PSO/nPSO) and GA models. For PSO/nPSO models, the best performed results are shown among various input parameter settings

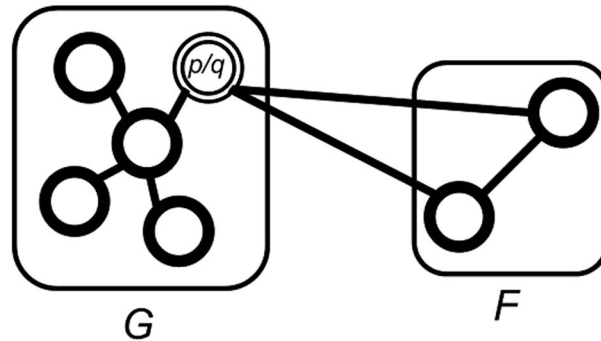
	Canonical_wnt	Ca_netscience	Inf_USAir	Inf_euroroad
Real-world network				
Co-neighborhood	0.19	0.30	0.26	0.01
Assortativity coefficient γ	-0.24	-0.08	-0.21	0.13
Corresponding model (existing) networks				
ER model				
RGF-distance	107.26 ± 4.51	112.89 ± 3.32	109.97 ± 3.74	75.26 ± 6.00
Co-neighborhood	0.01 ± 0.00	0.00 ± 0.00	0.02 ± 0.00	0.00 ± 0.00
Assortativity coefficient γ	0.00 ± 0.03	-0.01 ± 0.03	0.00 ± 0.02	-0.01 ± 0.03
SW model				
RGF-distance	46.32 ± 2.79	72.90 ± 1.27	21.01 ± 0.40	102.11 ± 2.63
Co-neighborhood	0.24 ± 0.01	0.17 ± 0.01	0.32 ± 0.01	0.00 ± 0.00
Assortativity coefficient γ	0.00 ± 0.04	0.00 ± 0.04	0.00 ± 0.02	0.01 ± 0.02
SF model				
RGF-distance	56.47 ± 4.07	118.56 ± 5.71	50.92 ± 1.55	131.14 ± 5.67
Co-neighborhood	0.02 ± 0.00	0.02 ± 0.01	0.04 ± 0.00	0.00 ± 0.00
Assortativity coefficient γ	-0.85 ± 0.05	-0.39 ± 0.06	-0.81 ± 0.03	-0.54 ± 0.11
nPSO model				
Optimized T value and C value	0.1, 8	0.1, 8	0.1, 8	0.1, 8
RGF-distance	28.64 ± 1.28	68.92 ± 1.75	34.61 ± 1.13	98.73 ± 0.16
Co-neighborhood	0.13 ± 0.00	0.08 ± 1.75	0.19 ± 0.01	0.00 ± 0.00
Assortativity coefficient γ	-0.10 ± 0.02	-0.19 ± 0.03	-0.09 ± 0.01	-0.24 ± 0.03
Corresponding model (GA) networks				
GA model				
Optimized q value	0.8	0.9	0.9	0.2
RGF-distance	22.97 ± 1.43	16.72 ± 2.85	33.30 ± 3.31	31.94 ± 10.14
Co-neighborhood	0.14 ± 0.01	0.19 ± 0.01	0.14 ± 0.01	0.02 ± 0.00
Assortativity coefficient γ	0.00 ± 0.04	-0.05 ± 0.05	-0.07 ± 0.03	0.04 ± 0.03
GA + R model				
Optimized q value	0.8	0.9	0.9	0.2
RGF-distance	25.91 ± 2.28	13.72 ± 2.66	32.33 ± 2.22	27.29 ± 17.86
Co-neighborhood	0.14 ± 0.01	0.20 ± 0.01	0.15 ± 0.01	0.01 ± 0.00
Assortativity coefficient γ	0.05 ± 0.06	0.02 ± 0.09	-0.02 ± 0.04	0.04 ± 0.03
GA + P model				
Optimized q value	0.9	0.9	0.9	0.1
RGF-distance	20.87 ± 2.73	52.99 ± 8.31	40.75 ± 2.29	87.74 ± 15.66
Co-neighborhood	0.17 ± 0.01	0.18 ± 0.01	0.15 ± 0.01	0.01 ± 0.00
Assortativity coefficient γ	-0.31 ± 0.03	-0.17 ± 0.03	-0.39 ± 0.02	-0.18 ± 0.02
GA + RP model				
Optimized q value	0.6	0.9	0.9	0.1
RGF-distance	29.23 ± 4.88	43.73 ± 15.83	43.86 ± 2.80	67.83 ± 11.02
Co-neighborhood	0.09 ± 0.01	0.19 ± 0.01	0.16 ± 0.01	0.01 ± 0.00
Assortativity coefficient γ	-0.32 ± 0.04	-0.15 ± 0.02	-0.34 ± 0.03	-0.16 ± 0.02

Table 1. RGF-distance, co-neighborhood and assortativity analysis results of various real-world networks. The table shows RGF-distance, co-neighborhood, assortativity coefficient γ values of four real-world networks and their corresponding model networks. For models, which require optimization, also stated the optimized parameter(s). Average values and standard deviations are stated together as the experiments are performed 10 times (100 times for canonical_wnt) per every condition, and averaged the results. Low RGF-distance represents high motif similarity to real-world networks. Co-neighborhood is our proposed measure, which is highly related to common neighbors and community/cluster structures. Low assortativity coefficient γ values are related to scale-free property. Best performed RGF-distances, relatively high co-neighborhood values, and relatively low assortativity coefficient γ values are stated in bold.

(T = 0.1, 0.5, and 0.9; C = 0, 4, and 8; See method session for details). Except for the Inf_USAir network, GA models outperformed in RGF-distances. For the Inf_USAir network, GA models might generate better results for q value over 0.9, which was not included in the experiment.

For three real-world networks (Canonical_wnt, Ca_netscience, and Inf_USAir) which have high co-neighborhood values (0.19, 0.30, and 0.26, respectively), GA models showed relatively high co-neighborhood values compared to ER and SF model. SW and PSO/nPSO models also showed relatively high co-neighborhood

GA model



GA+P model

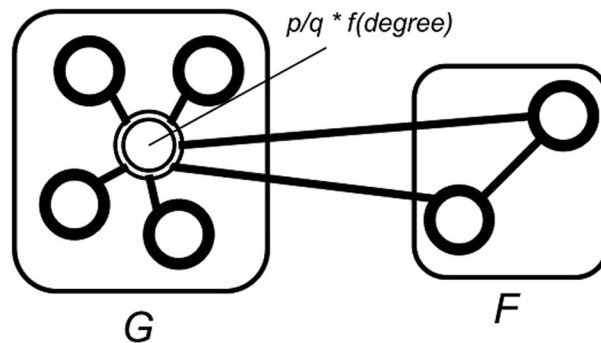


Figure 6. Edge generation procedures of GA and the preferential attachment adapted model (GA + P). The figure shows an example of edge generation procedures (Step 3 of Fig. 4) of both models. High degree nodes have higher probability to be selected in attachment procedure of GA + P model. GA + P model shows a scale-free property while maintaining a motif property of our model.

values, because SW and PSO/nPSO model generates networks with high clustering coefficients, and clusters increase co-neighborhood values of the network. Inf_euroroad network has a low co-neighborhood value (0.01) and all model networks showed low co-neighborhood values. As optimized q values for GA models of Inf_euroroad are also low, there seems to be a correlation between co-neighborhood and optimized q values, which might be a good following research topic. In general, corresponding model networks of similar co-neighborhood values with real-world networks showed low RGF-distances, which implies that co-neighborhood can be a good topological measure of network motifs.

From these results, we can insist that GA models generate networks which have high topological similarities with real-world networks in the manner of RGF-distances. As well, co-neighborhood showed its potential to be a representing topological measure of network motifs.

Preferential GA models (GA + P and GA + RP models). As the scale-free model (SF model) has been widely analyzed and represented as a proper network model in various types of networks^{3,37,38}, we applied the preferential attachment procedures to our models. The preferential GA model (GA + P model) is a combined model which preferentially attaches nodes when connecting the added graph (Graph F) with the existing graph (Graph G) (Fig. 6). During the attachment procedure (step 3 in Fig. 4), we select $|V_G| \cdot p/q$ nodes depending on the distribution $deg(V_G)^\alpha + a$, where $deg(V_G)$ indicates a degree (in-degree for a directed graph) distribution of nodes in the Graph G , α indicates power of preferential attachment, and a indicates initial attractiveness of the nodes³. This procedure guarantees higher connection probabilities on nodes of higher degrees, while preserving a motif property of our model. Furthermore, we also implemented GA + RP model, which adapted preferential attachment to GA + R model.

To show preferential attachment procedures are well-implemented in preferential models (GA + P and GA + RP models), we computed RGF-distances²⁵, power law exponents³, and assortativity coefficients³⁹ of the networks. Scale-free networks have degree distributions following power law with exponents in the range between two and three³. As some of network generation models get power law exponents as one of their input parameters, we also measured assortativity coefficients as indirect measurements of scale-free property⁴⁰. Power law exponents and assortativity coefficients are measured from two real-world networks (Canonical_Wnt and Inf_USAir) and their corresponding model networks (Fig. 7). Networks generated by SF, PSO/nPSO, GA + P,

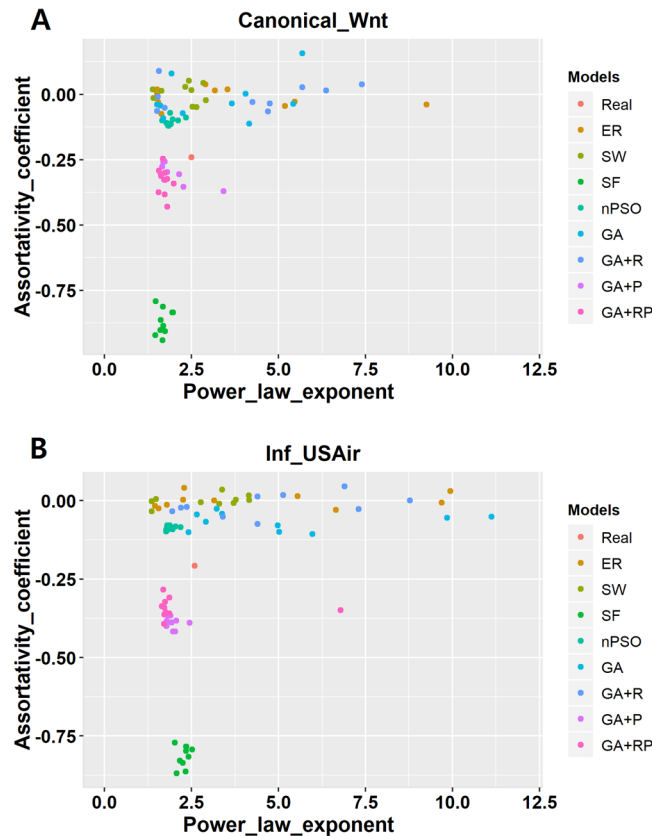


Figure 7. Assortativity coefficients and power law exponents of two real-world networks and their corresponding model networks. Two real-world networks, (A) Canonical_Wnt and (B) Inf_USAir, are selected as both of them have power law exponents between two and three, which represents scale-free property. Assortativity coefficients and power law exponents are measured from those networks and their corresponding model networks. For PSO/nPSO and GA models, we used the optimized input parameters stated in the Table 1. Most networks with low negative assortativity coefficients have power law exponents around two and three, which represents scale-free property. There exist couple of outliers in GA + P and GA + RP model networks.

and GA + RP have low assortativity coefficients and power law exponents around two to three, which represents those networks are scale-free and confirms that low negative assortativity coefficients are related to scale-free properties. In Table 1, networks generated by SF and PSO/nPSO models show low negative values of assortativity coefficient r , while networks generated by ER and SW models show r values close to 0. This result also supports that scale-free properties are represented with low negative values of r . As networks generated by preferential models (GA + P and GA + RP models) have power law exponents around two to three and show relatively low negative values of r compared to networks generated by non-preferential models (GA and GA + R models), we can claim that networks generated by preferential models have scale-free properties.

Furthermore, it is notable that RGF-distances are quite similar between the non-preferential models (GA and GA + R models) and the preferential models (GA + P and GA + RP models) in canonical Wnt signaling pathway and airport network among cities, while preferential models had poor RGF-distances compared to non-preferential models in other networks (co-authorship network of scientists and real world road network). As real-world networks of canonical Wnt signaling pathway and airport network among cities showed relatively low negative values (-0.24 and -0.21 , respectively) of r , we can assume that those two networks have scale-free properties. Then, we can conclude that the preferential models showed good performances of motif properties in real-world networks with scale-free properties.

According to these results, we can claim that preferential models (GA + P and GA + RP) gained scale-free properties while preserving real-world motif properties. Also, it can be implied that preferential attachment procedures are well-implemented in preferential models (GA + P and GA + RP).

Discussion

In summary, we suggested a new network measure, co-neighborhood, and a new network model, grouped attachment, to represent real-world network topologies. We showed that some of real-world networks have high co-neighborhood, and reproducing the co-neighborhood can generate real-world topologies. As the preferential attachment is suggested to reproduce scale-free properties of real-world networks³, we suggested the grouped attachment to reproduce co-neighborhood of real-world networks. By applying the grouped attachment to random network generation, we have developed a new network model which has higher similarities of motif properties

with real-world networks. While existing network models could not reproduce motif frequency distribution of real-world networks, our proposed model showed higher similarities of motif frequencies with real-world networks quantitatively. Furthermore, we applied preferential attachment procedures to our model, to gain scale-free properties while preserving real-world motif properties.

Nevertheless, existing network models have their capability to reproduce some of real-world properties. While our models outperformed on reproducing network motif properties, existing models have their own specialties on different real-world properties; SF model for scale-free properties; SW model for small world properties; PSO/nPSO models for various properties stated in the introduction session. It would be responsible for users to choose appropriate models for a given task.

As co-neighborhood adopted the concept of Jaccard's index, graphs with high co-neighborhood would have pairs of adjacent nodes which are located closely in a hidden geometric space¹⁰. Furthermore, if you apply co-neighborhood (Jaccard's index) concept to the community/cluster structure, you can interpret it as "A pair of nodes in the community/cluster structure would likely to have similar common neighbors, so that they have high interconnections in the community/cluster and few connections to nodes out of the community/cluster". In this interpretation, we have focused on 'few connections to nodes out of the community/cluster'. We thought that not only the fewness of connections is important, but also those few connections should be (likely to be) connected to the same nodes (out of the community/cluster), not randomly. This idea is well implemented in our grouped attachment model. We assumed the highly interconnected graph F represents a community/cluster. When they attach to existing graph G , they make connections to specifically selected nodes (not to random nodes). These implementations led our model to have reproducibility of neighborhood similarities.

In-depth analysis of co-neighborhood and optimization of q values can be a good candidate following research topics. Our findings show that real-world complex networks are populated by topological motifs and the proposed model reproduces real-world topological properties. These findings can be applied to various network topology studies such as community detection⁴¹ and link prediction^{42,43}. Some of existing methods of both tasks, community detection and link prediction, are dependent on network models which reproduce real-world topologies. They use network models for network structure estimation, and infer results based on them. As our proposed GA models uniquely reproduces real-world motif properties, implementing our models might be a key solution to the tasks.

Methods

Common neighbor index and Jaccard index. Let u and v are network nodes, and $\Gamma(u)$ and $|\Gamma(u)|$ refer to the set of neighbors of u and the cardinality of the set, respectively. Common neighbor index²⁶ is defined as

$$CN(u, v) = |\Gamma(u) \cap \Gamma(v)|, \quad (2)$$

and Jaccard index^{33,34} is defined as

$$JC(u, v) = \frac{|\Gamma(u) \cap \Gamma(v)|}{|\Gamma(u) \cup \Gamma(v)|} = \frac{CN(u, v)}{|\Gamma(u) \cup \Gamma(v)|}. \quad (3)$$

Random network generation of existing models. All random networks of existing models (ER, SW, and SF) are generated through 'igraph' R package (package version 1.0.1)⁴⁴. To generate input parameters of network models, we measure $|N|$ (number of nodes), $|E|$ (number of edges), p (edge density), α (exponent of the fitted power-law distribution of degree), and a (minimum value from which the power-law distribution of degree was fitted) from a given real-world network.

For ER network generation, we adapted Erdős-Rényi model by utilizing `erdos.renyi.game()` function in 'igraph' package. We set the number of nodes to be $|N|$, and the edge probability to be p .

For SW network generation, we adapted Watts-Strogatz model by utilizing `watts.strogatz.game()` function in 'igraph' package. We set the dimension of the starting lattice to be 1, the size of the lattice along each dimension to be $|N|$, the neighborhood within which the vertices of the lattice will be connected to be $|E|/|N|$, and the rewiring probability to be 0.05.

For SF network generation, we adapted Barabasi-Albert (preferential attachment) model by utilizing `barabasi.game()` function in 'igraph' package. We set the number of vertices to be $|N|$, the power of the preferential attachment to be α , the number of edges to add in each time step to be $|E|/|N|$, and the attractiveness of the vertices with no adjacent edges to be a .

For PSO/nPSO network generation, we adapted nPSO model in the corresponding manuscript²⁷. We set the number of nodes to be $|N|$, the half of average degree to be $|E|/|N|$, the exponent of the power-law node degree distribution to be α . The random networks of PSO/nPSO model were generated with three different temperature values ($T = 0.1, 0.5, \text{ and } 0.9$), and three different numbers of communities ($C = 0, 4, \text{ and } 8$); $C = 0$ corresponds to the PSO model, while $C = 4, 8$ corresponds to the nPSO model.

Using each model, we generated 100 random networks (100 repeats) for directed networks and 10 random networks (10 repeats) for undirected networks. All experiments are performed to the networks and the results are averaged.

Counting motif frequencies. Counting network motif frequencies have been processed differently depending on directedness of the given network. For directed networks, we utilized `graph.motifs()` function in 'igraph' R package⁴⁴. For undirected networks, we utilized `countMotif()` function in 'NeMo' R package (package version 1.0.1)⁴⁵.

RGF-distance calculation. RGF-distance compares the frequencies of the appearance of all 3–5-node graphlets in two networks¹³. Between two graphs G and H , RGF-distance is defined as

$$D(G, H) = \sum_{i=1}^{29} |F_i(G) - F_i(H)|, \quad (4)$$

where

$$F_i(G) = -\log\left(\frac{N_i(G)}{T(G)}\right). \quad (5)$$

$N_i(G)$ is the number of graphlets (motif frequency count) of type i ($i \in \{1, \dots, 29\}$) for graphlet size from 3 to 5 in a network G , and

$$T(G) = \sum_{i=1}^{29} N_i(G) \quad (6)$$

is the total number of graphlets of G . Graphlet types can be referred to motif numbers in network motifs. In our experiments, we computed RGF-distances between real-world networks and their corresponding model networks.

Power law exponent measurement. Power law exponent k is a measure of scale-free property. Degree distribution of a scale-free network follows power law with exponent $2 < k < 3$. Power law exponent k of network is measured through `power.law.fit()` function in ‘igraph’ R package (package version 1.0.1)⁴⁴.

Assortativity coefficient measurement. Assortativity coefficient r is a measure of the likelihood for nodes to connect to other nodes with similar degrees³⁹, and is related to a scale-free metric⁴⁰. Assortativity coefficient ranges between -1 and 1 . When $r = 1$, the network is completely assortative. When $r = 0$, the network is non-assortative. When $r = -1$, the network is completely disassortative. Assortativity coefficient r of network is measured through `assortativity.degree()` function in ‘igraph’ R package (package version 1.0.1)⁴⁴.

Code availability. The R script implementing the GA models and the co-neighborhood is available at https://github.com/bisl-kaist/GA_model.

Data availability. The Canonical Wnt signaling pathway data analyzed during the current study are available in the NCI/Nature database by ‘import network from web services’ in Cytoscape⁴⁶. The Co-authorship of scientists⁴⁷, the airport network among cities⁴⁸, and the real-world road network data⁴⁹ analyzed during the current study are available in the Network Repository, <http://networkrepository.com/>³⁶. Detailed descriptions of four networks are stated in Supplementary Note.

References

- Erdős, P. & Rényi, A. On random graphs, I. *Publicationes Mathematicae (Debrecen)* **6**, 290–297 (1959).
- Watts, D. J. & Strogatz, S. H. Collective dynamics of ‘small-world’ networks. *nature* **393**, 440–442 (1998).
- Barabási, A.-L. & Albert, R. Emergence of scaling in random networks. *science* **286**, 509–512 (1999).
- Berry, D. & Widder, S. Deciphering microbial interactions and detecting keystone species with co-occurrence networks. *Frontiers in microbiology* **5**, 219 (2014).
- Sporns, O. & Bullmore, E. T. From connections to function: the mouse brain connectome atlas. *Cell* **157**, 773–775 (2014).
- Costa, L. D. F. *et al.* Analyzing and modeling real-world phenomena with complex networks: a survey of applications. *Advances in Physics* **60**, 329–412 (2011).
- Papadopoulos, F., Kitsak, M., Serrano, M. Á., Boguná, M. & Krioukov, D. Popularity versus similarity in growing networks. *Nature* **489**, 537 (2012).
- Muscoloni, A. & Cannistraci, C. V. Leveraging the nonuniform PSO network model as a benchmark for performance evaluation in community detection and link prediction. *New Journal of Physics* (2018).
- Muscoloni, A. & Cannistraci, C. V. A nonuniform popularity-similarity optimization (nPSO) model to efficiently generate realistic complex networks with communities. *New Journal of Physics* **20**, 052002 (2018).
- Muscoloni, A., Thomas, J. M., Ciucci, S., Bianconi, G. & Cannistraci, C. V. Machine learning meets complex networks via coalescent embedding in the hyperbolic space. *Nature Communications* **8**, 1615 (2017).
- Muscoloni, A. & Cannistraci, C. V. Rich-clubness test: how to determine whether a complex network has or doesn’t have a rich-club? *arXiv preprint arXiv 1704.03526* (2017).
- Shen-Orr, S. S., Milo, R., Mangan, S. & Alon, U. Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nature genetics* **31**, 64–68 (2002).
- Pržulj, N., Corneil, D. G. & Jurisica, I. Modeling interactome: scale-free or geometric? *Bioinformatics* **20**, 3508–3515 (2004).
- Schreiber, F. & Schwobbermeyer, H. Frequency concepts and pattern detection for the analysis of motifs in networks. *Lecture Notes in Computer Science* **3737**, 89–104 (2005).
- Pržulj, N. Biological network comparison using graphlet degree distribution. *Bioinformatics* **23**, e177–e183 (2007).
- Milo, R. *et al.* Network motifs: simple building blocks of complex networks. *Science* **298**, 824–827 (2002).
- Picard, F., Daudin, J.-J., Koskas, M., Schbath, S. & Robin, S. Assessing the exceptionality of network motifs. *Journal of Computational Biology* **15**, 1–20 (2008).
- Milenkovič, T. & Pržulj, N. Uncovering biological network function via graphlet degree signatures. *Cancer informatics* **6**, 257 (2008).
- Benson, A. R., Gleich, D. F. & Leskovec, J. Higher-order organization of complex networks. *Science* **353**, 163–166 (2016).
- Barabási, A.-L. & Oltvai, Z. N. Network biology: understanding the cell’s functional organization. *Nature reviews genetics* **5**, 101–113 (2004).
- Tran, N. H., Choi, K. P. & Zhang, L. Counting motifs in the human interactome. *Nature communications* **4** (2013).
- Wuchty, S., Oltvai, Z. N. & Barabási, A.-L. Evolutionary conservation of motif constituents in the yeast protein interaction network. *Nature genetics* **35**, 176–179 (2003).
- Hayes, W., Sun, K. & Pržulj, N. Graphlet-based measures are suitable for biological network comparison. *Bioinformatics* **29**, 483–491 (2013).

24. Zhang, Y. & Xuan, J. de los Reyes, B. G., Clarke, R. & Resson, H. W. Network motif-based identification of transcription factor-target gene relationships by integrating multi-source biological data. *BMC bioinformatics* **9**, 1 (2008).
25. Alon, U. Network motifs: theory and experimental approaches. *Nature Reviews Genetics* **8**, 450–461 (2007).
26. Newman, M. E. Clustering and preferential attachment in growing networks. *Physical review E* **64**, 025102 (2001).
27. Muscoloni, A. & Cannistraci, C. V. A nonuniform popularity-similarity optimization (nPSO) model to efficiently generate realistic complex networks with communities. *arXiv preprint arXiv 1707.07325* (2017).
28. Zhou, Y., Cheng, H. & Yu, J. X. Graph clustering based on structural/attribute similarities. *Proceedings of the VLDB Endowment* **2**, 718–729 (2009).
29. Pan, Y., Li, D.-H., Liu, J.-G. & Liang, J.-Z. Detecting community structure in complex networks via node similarity. *Physica A: Statistical Mechanics and its Applications* **389**, 2849–2857 (2010).
30. Cannistraci, C. V., Alanis-Lobato, G. & Ravasi, T. From link-prediction in brain connectomes and protein interactomes to the local-community-paradigm in complex networks. *Scientific reports* **3**, 1613 (2013).
31. Daminelli, S., Thomas, J. M., Durán, C. & Cannistraci, C. V. Common neighbours and the local-community-paradigm for topological link prediction in bipartite networks. *New Journal of Physics* **17**, 113037 (2015).
32. Boutin, F. & Hascoët, M. Cluster validity indices for graph partitioning, In *Information Visualisation, 2004. IV2004. Proceedings. Eighth International Conference on*. 376–381 (IEEE).
33. Jaccard, P. Distribution de la flore alpine dans le bassin des Dranses et dans quelques régions voisines. *Bull Soc Vaudoise Sci Nat* **37**, 241–272 (1901).
34. Cannistraci, C. V., Alanis-Lobato, G. & Ravasi, T. Erratum: From link-prediction in brain connectomes and protein interactomes to the local-community-paradigm in complex networks. *Scientific Reports* **5** (2015).
35. Wasserman, S. & Faust, K. Social network analysis: Methods and applications. Vol. 8 (Cambridge university press, 1994).
36. Rossi, R. A. & Ahmed, N. K. An Interactive Data Repository with Visual Analytics. *ACM SIGKDD Explorations Newsletter* **17**, 37–41 (2016).
37. Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N. & Barabási, A.-L. The large-scale organization of metabolic networks. *Nature* **407**, 651–654 (2000).
38. Clauset, A., Shalizi, C. R. & Newman, M. E. Power-law distributions in empirical data. *SIAM review* **51**, 661–703 (2009).
39. Newman, M. E. Assortative mixing in networks. *Physical review letters* **89**, 208701 (2002).
40. Xulvi-Brunet, R. & Sokolov, I. M. Changing correlations in networks: assortativity and dissortativity. *Acta Physica Polonica B* **36**, 1431 (2005).
41. Vinh, N. X., Epps, J. & Bailey, J. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research* **11**, 2837–2854 (2010).
42. Kötter, R. Online retrieval, processing, and visualization of primate connectivity data from the CoCoMac database. *Neuroinformatics* **2**, 127–144 (2004).
43. Isella, L. *et al.* What's in a crowd? Analysis of face-to-face behavioral networks. *Journal of theoretical biology* **271**, 166–180 (2011).
44. Csardi, G. & Nepusz, T. The igraph software package for complex network research. *InterJournal, Complex Systems* **1695**, 1–9 (2006).
45. Chen, J., Hsu, W., Lee, M. L. & Ng, S.-K. NeMoFinder: Dissecting genome-wide protein-protein interactions with meso-scale network motifs, In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. 106–115 (ACM).
46. Smoot, M. E., Ono, K., Ruscheinski, J., Wang, P.-L. & Ideker, T. Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics* **27**, 431–432 (2011).
47. Newman, M. E. Finding community structure in networks using the eigenvectors of matrices. *Physical review E* **74**, 036104 (2006).
48. Colizza, V., Pastor-Satorras, R. & Vespignani, A. Reaction-diffusion processes and metapopulation models in heterogeneous networks. *Nature Physics* **3**, 276 (2007).
49. Bader, D. A., Meyerhenke, H., Sanders, P. & Wagner, D. Graph partitioning and graph clustering, In *10th DIMACS Implementation Challenge Workshop*.

Acknowledgements

The authors thank to Hawoong Jeong (KAIST) for theoretical discussions. This work was supported by the Bio-Synergy Research Project (NRF-2012M3A9C4048758) of the Ministry of Science, ICT and Future Planning through the National Research Foundation.

Author Contributions

J.C. and D.L. conceived and designed the study. J.C. collected network data and wrote code for the experiments. J.C. and D.L. wrote the manuscript.

Additional Information

Supplementary information accompanies this paper at <https://doi.org/10.1038/s41598-018-30845-4>.

Competing Interests: The authors declare no competing interests.

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



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2018