

Counting motifs in the human interactome

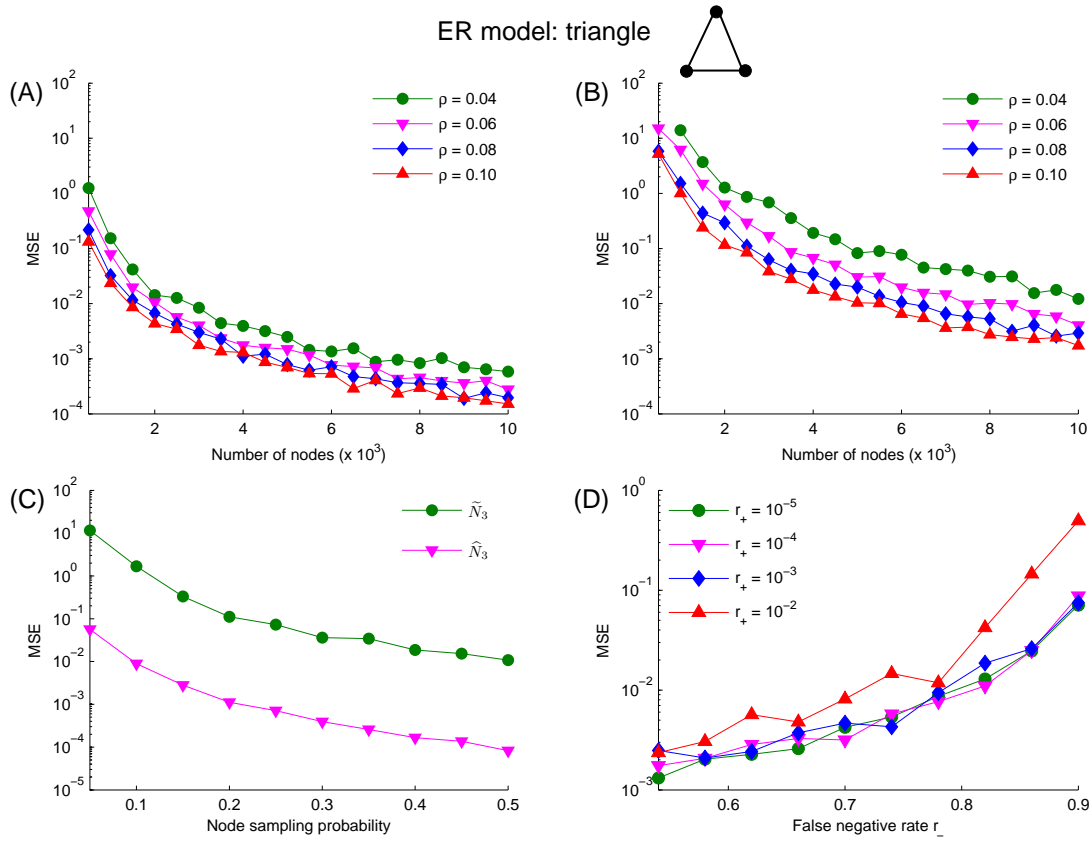
Ngoc Hieu Tran, Kwok Pui Choi, and Louxin Zhang

Supplementary Information

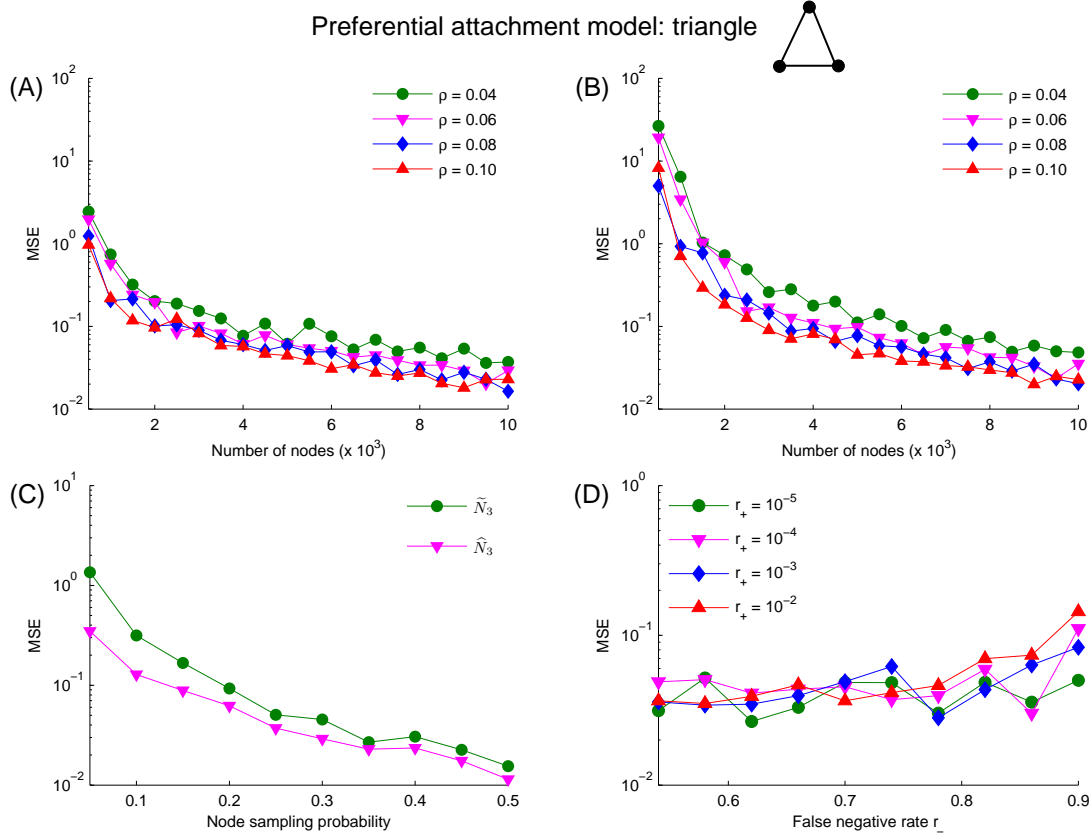
Contents

Supplementary Figures S1-S12	2
Supplementary Tables S1-S2	14
Supplementary Notes	16
Supplementary Note 1. Random network models	16
Supplementary Note 2. Comparison of \tilde{N}_1 and CCSB estimator	17
Supplementary Note 3. The computational efficiency of the sampling approach	19
Supplementary Note 4. Effect of non-uniform sampling on motif estimation	20
Supplementary Methods	22
Proof of Theorem 2	22
Derivation of bias-corrected estimator $\tilde{N}_{\mathcal{M}}$	27

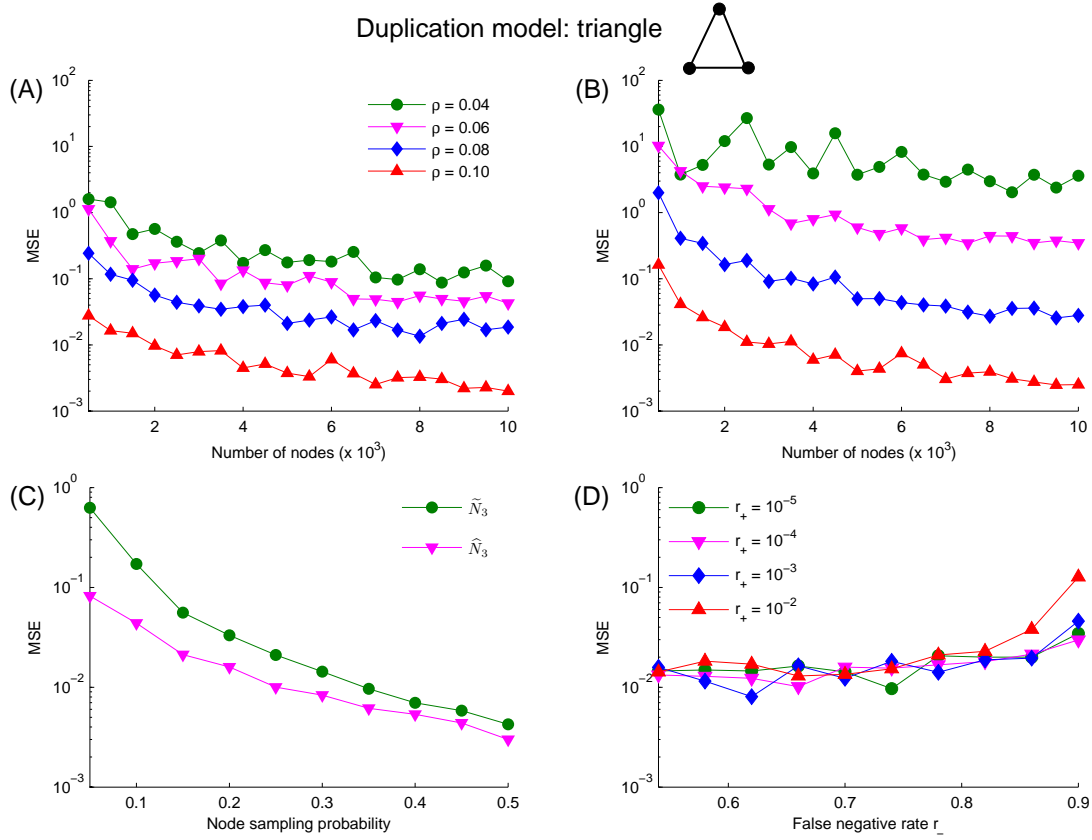
Supplementary Figures S1-S12



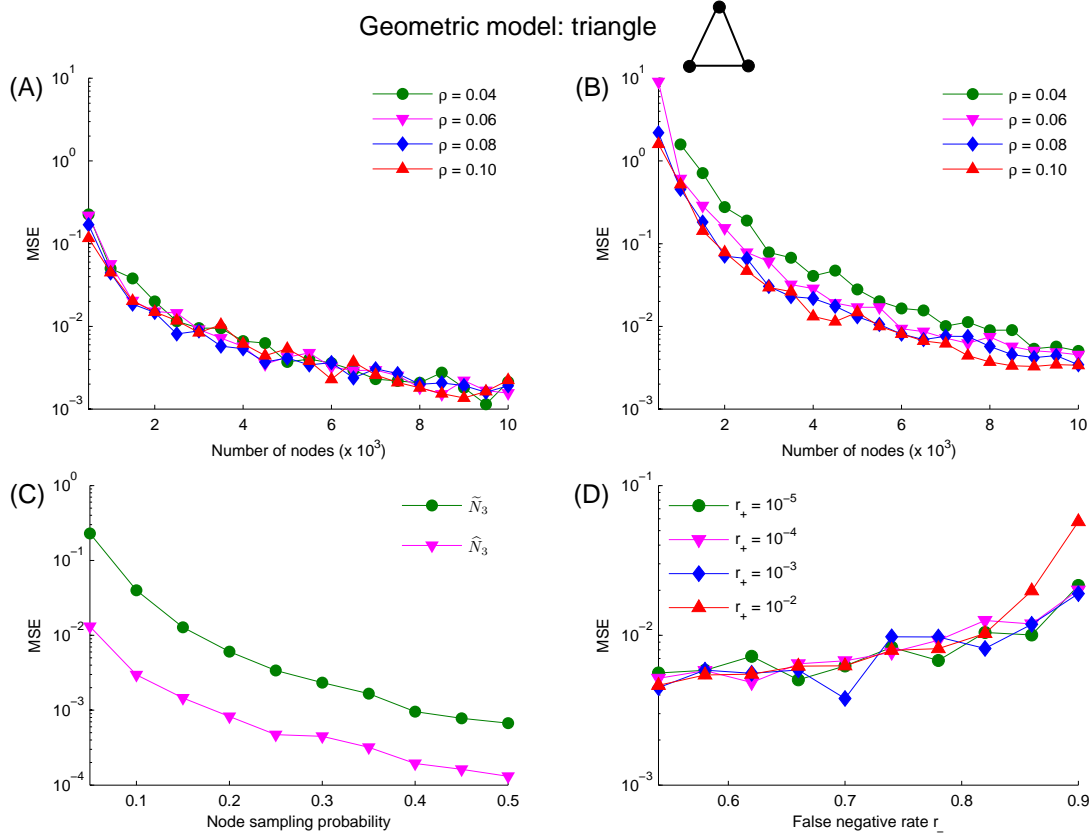
Supplementary Figure S1. Plots of $\text{MSE}(\hat{N}_3)$ and $\text{MSE}(\tilde{N}_3)$ for counting triangles in the ER model. Both $\text{MSE}(\hat{N}_3)$ and $\text{MSE}(\tilde{N}_3)$ depend on node number n , edge density ρ , and node sampling probability p . $\text{MSE}(\tilde{N}_3)$ also depends on the link error rates r_- and r_+ . (A) $\text{MSE}(\tilde{N}_3)$ changes with n and ρ when p is fixed at 0.1. (B) $\text{MSE}(\tilde{N}_3)$ changes with n and ρ when p, r_- and r_+ are fixed at 0.1, 0.85, and 0.00001, respectively. (C) $\text{MSE}(\hat{N}_3)$ and $\text{MSE}(\tilde{N}_3)$ change with p when n, ρ, r_- , and r_+ are fixed at 5,000, 0.1, 0.85, and 0.00001, respectively. (D) $\text{MSE}(\tilde{N}_3)$ changes with r_+ and r_- when n, ρ , and p are fixed at 5,000, 0.1, 0.1, respectively.



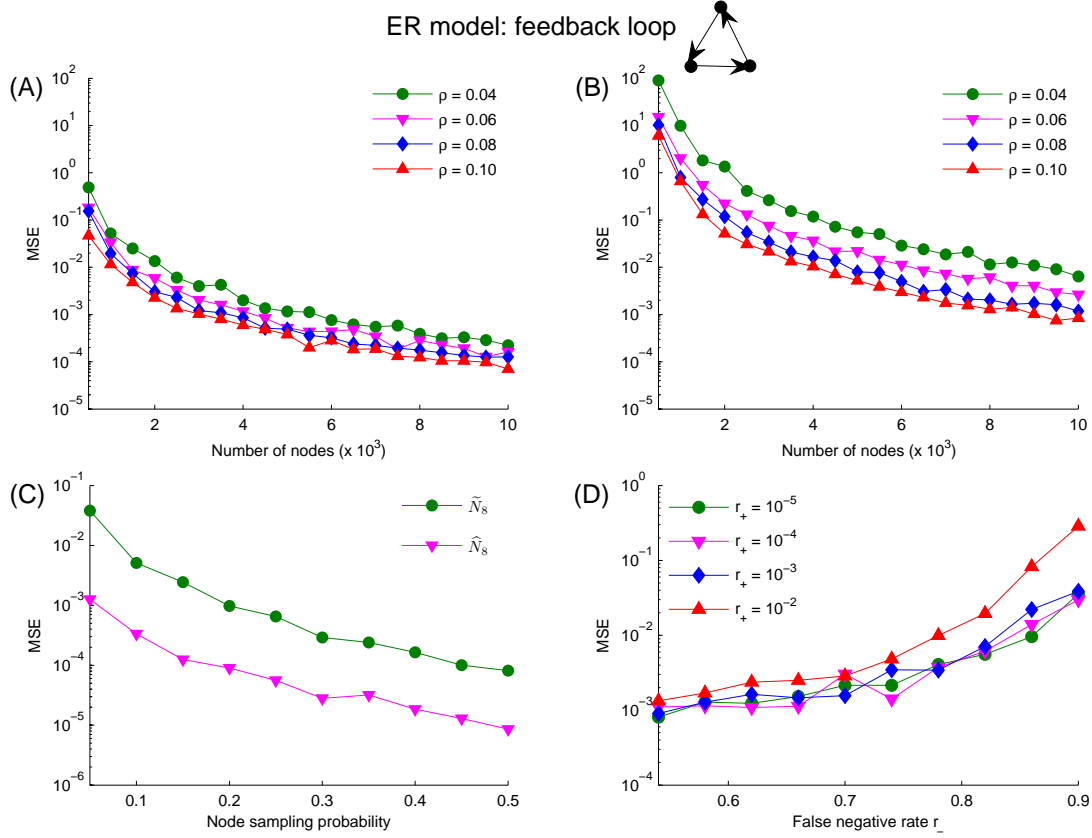
Supplementary Figure S2. Plots of $\text{MSE}(\hat{N}_3)$ and $\text{MSE}(\tilde{N}_3)$ for counting triangles in the preferential attachment model. (A) $\text{MSE}(\hat{N}_3)$ changes with n and ρ when p is fixed at 0.1. (B) $\text{MSE}(\tilde{N}_3)$ changes with n and ρ when p , r_- and r_+ are fixed at 0.1, 0.85, and 0.00001, respectively. (C) $\text{MSE}(\hat{N}_3)$ and $\text{MSE}(\tilde{N}_3)$ change with p when n , ρ , r_- , and r_+ are fixed at 5,000, 0.1, 0.85, and 0.00001, respectively. (D) $\text{MSE}(\tilde{N}_3)$ changes with r_+ and r_- when n , ρ , and p are fixed at 5,000, 0.1, 0.1, respectively.



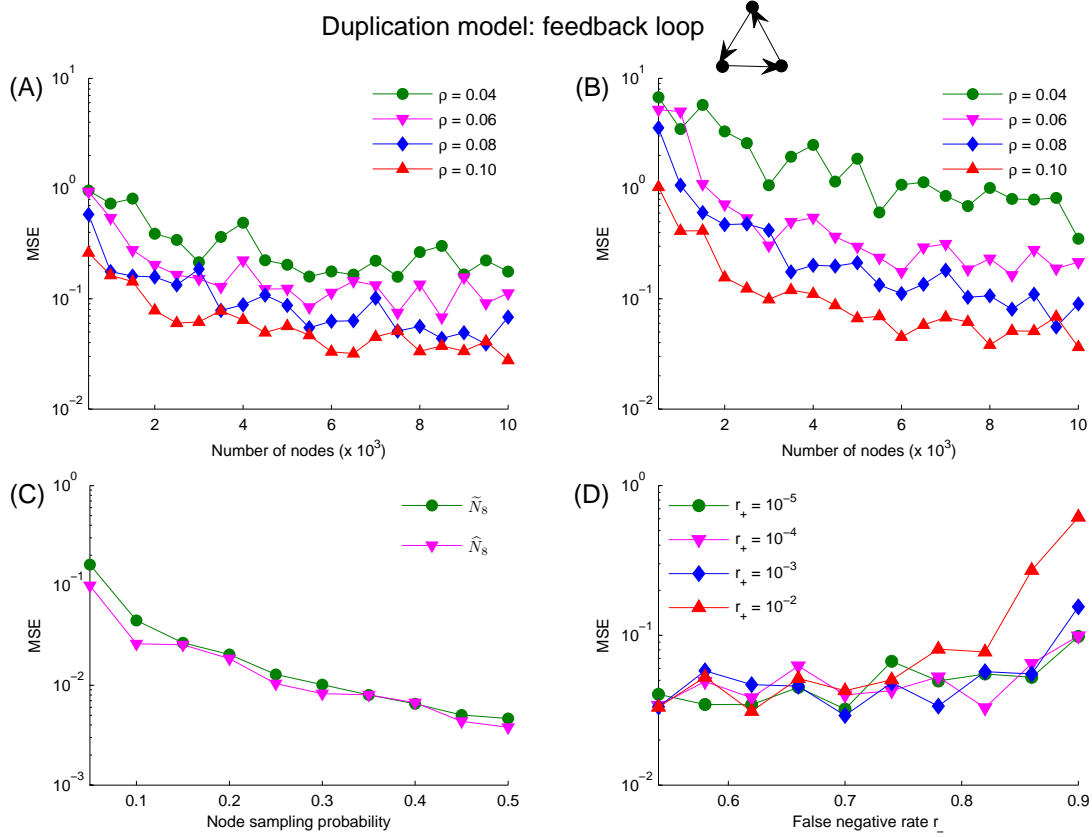
Supplementary Figure S3. Plots of $\text{MSE}(\hat{N}_3)$ and $\text{MSE}(\tilde{N}_3)$ for counting triangles in the duplication model. (A) $\text{MSE}(\hat{N}_3)$ changes with n and ρ when p is fixed at 0.1. (B) $\text{MSE}(\tilde{N}_3)$ changes with n and ρ when p , r_- and r_+ are fixed at 0.1, 0.85, and 0.00001, respectively. (C) $\text{MSE}(\hat{N}_3)$ and $\text{MSE}(\tilde{N}_3)$ change with p when n , ρ , r_- , and r_+ are fixed at 5,000, 0.1, 0.85, and 0.00001, respectively. (D) $\text{MSE}(\tilde{N}_3)$ changes with r_+ and r_- when n , ρ , and p are fixed at 5,000, 0.1, 0.1, respectively.



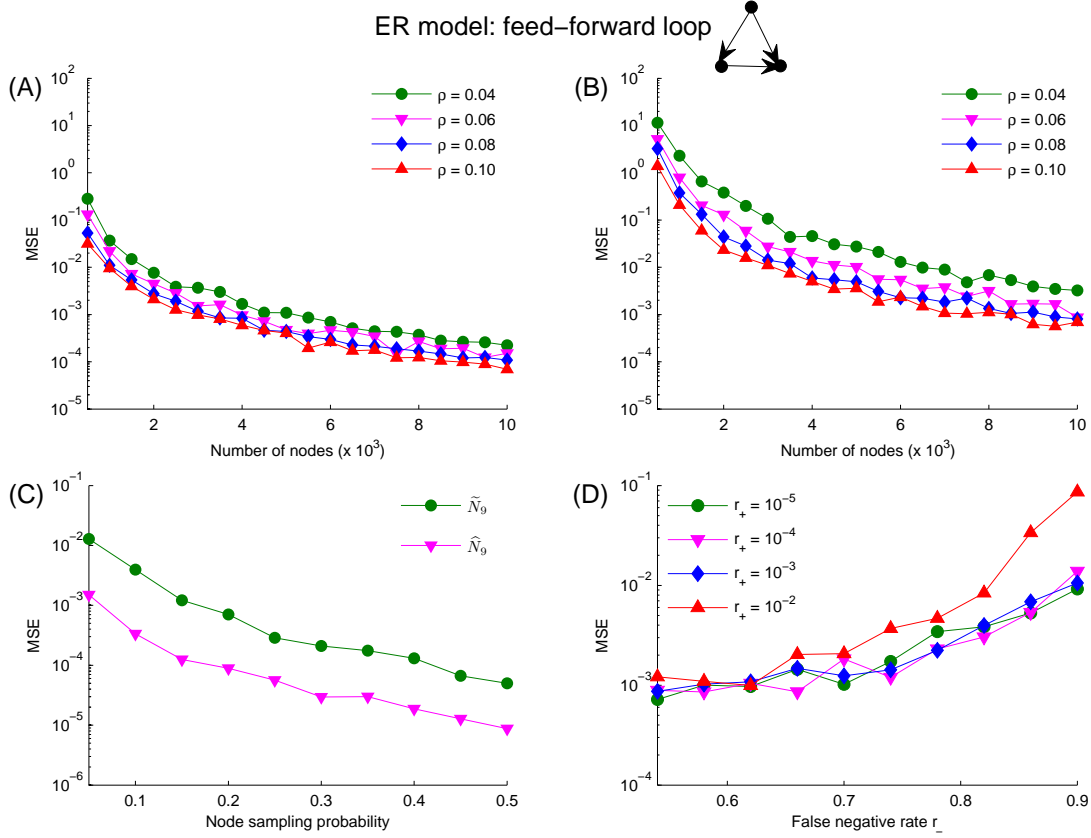
Supplementary Figure S4. Plots of $\text{MSE}(\hat{N}_3)$ and $\text{MSE}(\tilde{N}_3)$ for counting triangles in the geometric model. (A) $\text{MSE}(\hat{N}_3)$ changes with n and ρ when p is fixed at 0.1. (B) $\text{MSE}(\tilde{N}_3)$ changes with n and ρ when p , r_- and r_+ are fixed at 0.1, 0.85, and 0.00001, respectively. (C) $\text{MSE}(\hat{N}_3)$ and $\text{MSE}(\tilde{N}_3)$ change with p when n , ρ , r_- , and r_+ are fixed at 5,000, 0.1, 0.85, and 0.00001, respectively. (D) $\text{MSE}(\tilde{N}_3)$ changes with r_+ and r_- when n , ρ , and p are fixed at 5,000, 0.1, 0.1, respectively.



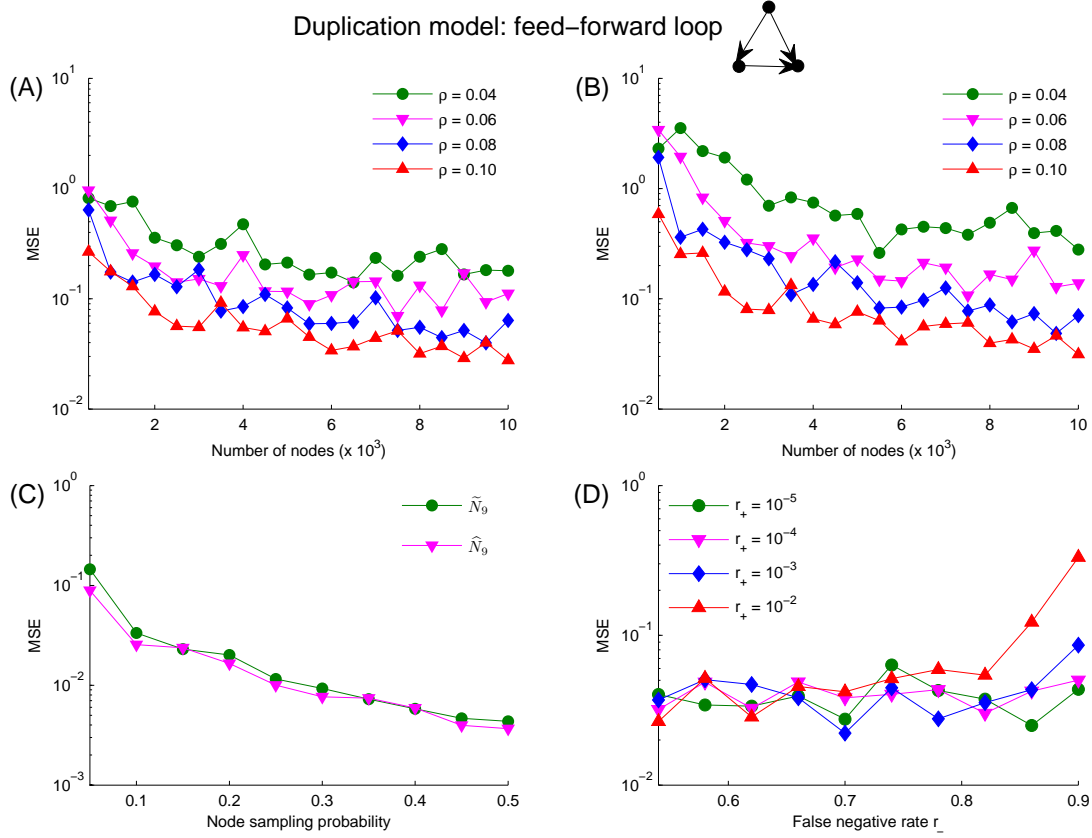
Supplementary Figure S5. Plots of $\text{MSE}(\hat{N}_8)$ and $\text{MSE}(\tilde{N}_8)$ for counting the occurrences of the feed-back loop in the ER model. (A) $\text{MSE}(\hat{N}_8)$ changes with n and ρ when p is fixed at 0.1. (B) $\text{MSE}(\tilde{N}_8)$ changes with n and ρ when p , r_- and r_+ are fixed at 0.1, 0.85, and 0.00001, respectively. (C) $\text{MSE}(\hat{N}_8)$ and $\text{MSE}(\tilde{N}_8)$ change with p when n , ρ , r_- , and r_+ are fixed at 5,000, 0.1, 0.85, and 0.00001, respectively. (D) $\text{MSE}(\tilde{N}_8)$ changes with r_+ and r_- when n , ρ , and p are fixed at 5,000, 0.1, 0.1, respectively.



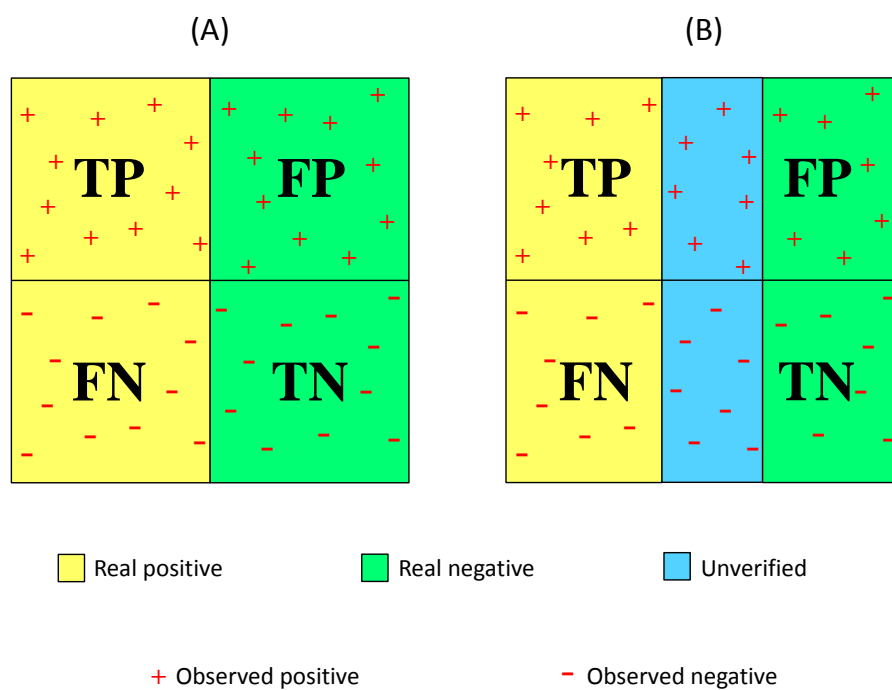
Supplementary Figure S6. Plots of $\text{MSE}(\hat{N}_8)$ and $\text{MSE}(\tilde{N}_8)$ for counting the occurrences of the feed-back loop in the duplication model. (A) $\text{MSE}(\hat{N}_8)$ changes with n and ρ when p is fixed at 0.1. (B) $\text{MSE}(\tilde{N}_8)$ changes with n and ρ when p , r_- and r_+ are fixed at 0.1, 0.85, and 0.00001, respectively. (C) $\text{MSE}(\hat{N}_8)$ and $\text{MSE}(\tilde{N}_8)$ change with p when n , ρ , r_- , and r_+ are fixed at 5,000, 0.1, 0.85, and 0.00001, respectively. (D) $\text{MSE}(\tilde{N}_8)$ changes with r_+ and r_- when n , ρ , and p are fixed at 5,000, 0.1, 0.1, respectively.



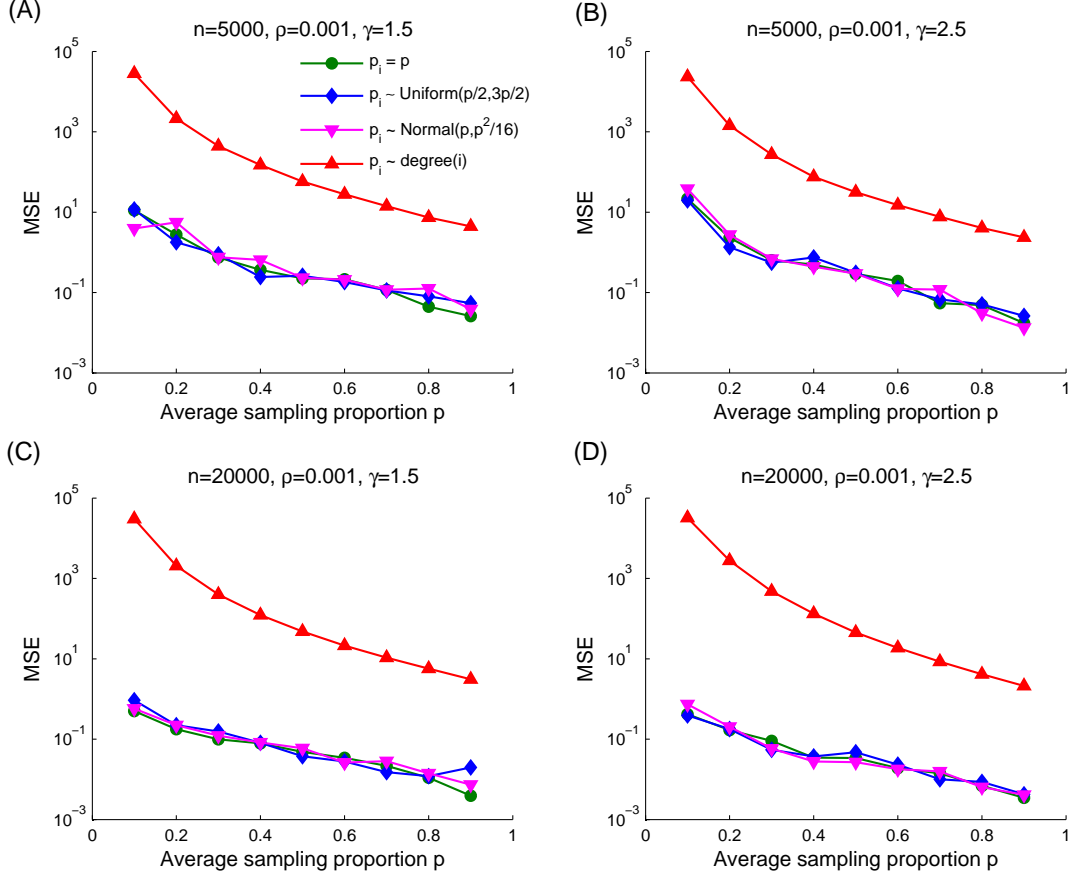
Supplementary Figure S7. Plots of $\text{MSE}(\hat{N}_9)$ and $\text{MSE}(\tilde{N}_9)$ for counting the occurrences of the feed-forward loop in the ER model. (A) $\text{MSE}(\hat{N}_9)$ changes with n and ρ when p is fixed at 0.1. (B) $\text{MSE}(\tilde{N}_9)$ changes with n and ρ when p , r_- and r_+ are fixed at 0.1, 0.85, and 0.00001, respectively. (C) $\text{MSE}(\hat{N}_9)$ and $\text{MSE}(\tilde{N}_9)$ change with p when n , ρ , r_- , and r_+ are fixed at 5,000, 0.1, 0.85, and 0.00001, respectively. (D) $\text{MSE}(\tilde{N}_9)$ changes with r_+ and r_- when n , ρ , and p are fixed at 5,000, 0.1, 0.1, respectively.



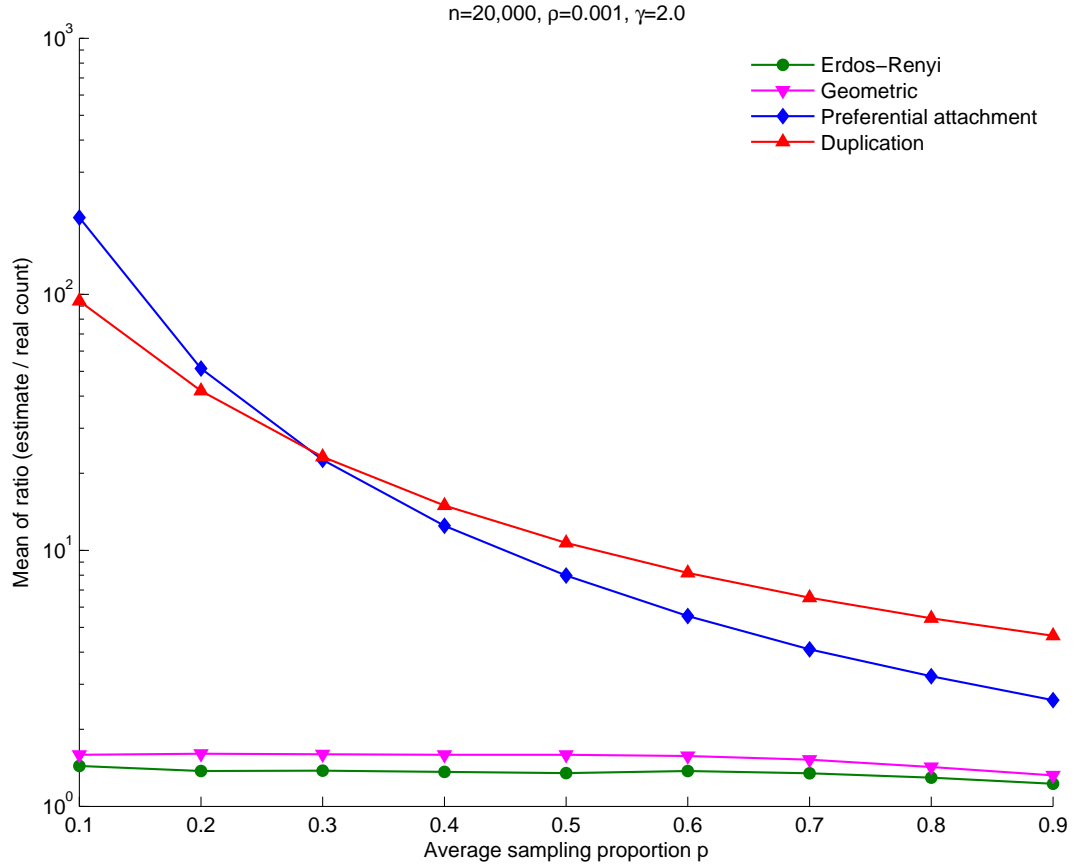
Supplementary Figure S8. Plots of $\text{MSE}(\hat{N}_9)$ and $\text{MSE}(\tilde{N}_9)$ for counting the occurrences of the feed-forward loop in the duplication model. (A) $\text{MSE}(\hat{N}_9)$ changes with n and ρ when p is fixed at 0.1. (B) $\text{MSE}(\tilde{N}_9)$ changes with n and ρ when p , r_- and r_+ are fixed at 0.1, 0.85, and 0.00001, respectively. (C) $\text{MSE}(\hat{N}_9)$ and $\text{MSE}(\tilde{N}_9)$ change with p when n , ρ , r_- , and r_+ are fixed at 5,000, 0.1, 0.85, and 0.00001, respectively. (D) $\text{MSE}(\tilde{N}_9)$ changes with r_+ and r_- when n , ρ , and p are fixed at 5,000, 0.1, 0.1, respectively.



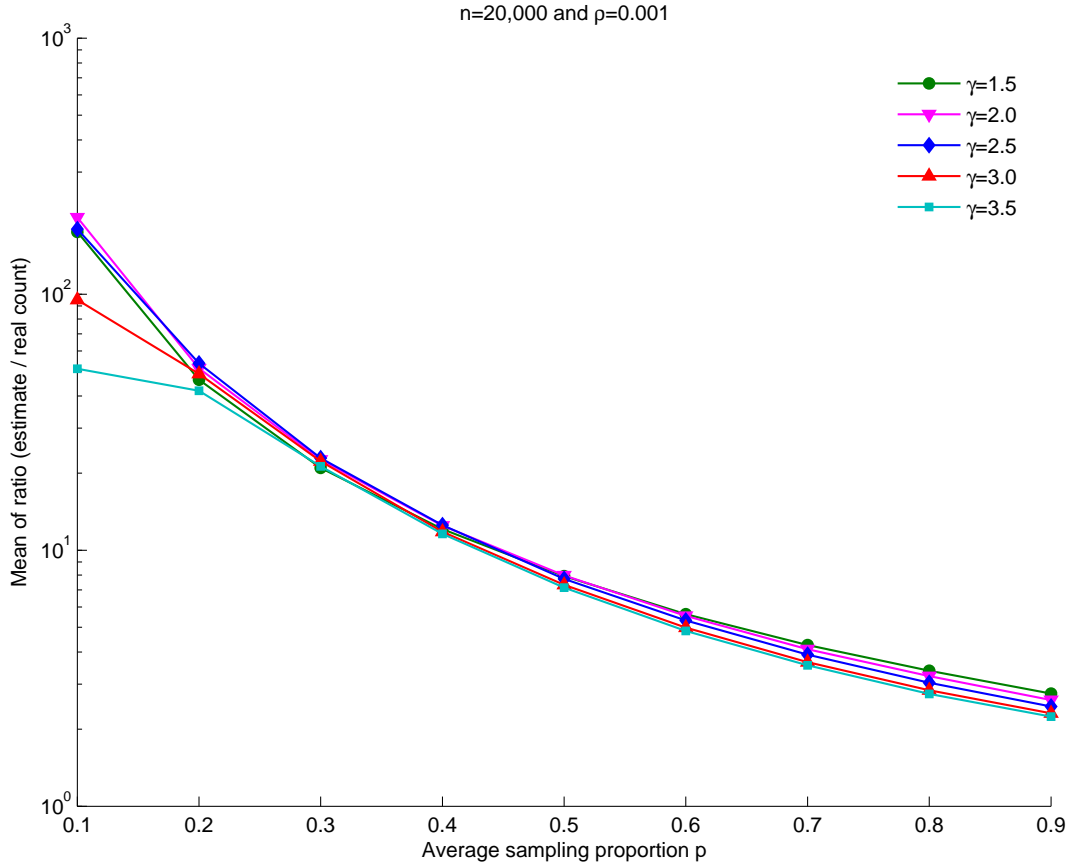
Supplementary Figure S9. Limitations of gold-standard reference sets in biological networks data. (A) In an ideal classification problem, gold-standard positive and negative sets are fully known. (B) Gold-standard sets for biological networks data are limited.



Supplementary Figure S10. Plots of $\text{MSE}(\hat{N}_3)$ for triangles with respect to four different sampling schemes and the average sampling proportion p . Random networks were generated from the preferential attachment model with different parameters n, ρ, γ . Subnetworks were drawn using four sampling schemes: $p_i = p$ (i.e., uniform node sampling), $p_i \sim \text{Uniform}(\frac{p}{2}, \frac{3p}{2})$, $p_i \sim \text{Normal}(\mu = p, \sigma^2 = \frac{p^2}{16})$, $p_i = np \frac{d_i}{\sum_{j=1}^n d_j}$, $i = 1, 2, \dots, n$.

















Supplementary Figure S11. Plots of the mean of the ratio $\frac{\hat{N}_3}{N_3}$ for triangles with respect to the four network models and the average sampling proportion p . Random networks were generated from the ER, geometric, preferential attachment, and duplication models, with $n = 20,000$ nodes, link density $\rho = 0.001$, power-law exponent $\gamma = 2.0$ (for preferential attachment and duplication models). Subnetworks were drawn using the degree-bias sampling scheme with $p_i = np \frac{\text{degree}(i)}{\sum_{j=1}^n \text{degree}(j)}$, $i = 1, 2, \dots, n$.



Supplementary Figure S12. Plots of the mean of the ratio $\frac{\hat{N}_3}{N_3}$ for triangles with respect to the power-law exponent γ and the average sampling proportion p . Random networks were generated from the preferential attachment model with 20,000 nodes and link density $\rho = 0.001$. Subnetworks were drawn using the degree-bias sampling scheme with $p_i = np \frac{d_i}{\sum_{j=1}^n d_j}$, $i = 1, 2, \dots, n$.















Supplementary Tables S1-S2

Supplementary Table S1. The corresponding functions $f_{\mathcal{M}}()$ for the motifs under study

Motif	$f_{\mathcal{M}}()$
1 	$f_1(\mathbf{A}[i_1, i_2]) = a_{i_1 i_2}$
2 	$f_2(\mathbf{A}[i_1, i_2, i_3]) = a_{i_1 i_2} a_{i_2 i_3} + a_{i_2 i_3} a_{i_3 i_1} + a_{i_3 i_1} a_{i_1 i_2}$
3 	$f_3(\mathbf{A}[i_1, i_2, i_3]) = a_{i_1 i_2} a_{i_2 i_3} a_{i_3 i_1}$
4 	$f_4(\mathbf{A}[i_1, i_2]) = a_{i_1 i_2} + a_{i_2 i_1}$
5 	$f_5(\mathbf{A}[i_1, i_2, i_3]) = (a_{i_2 i_1} a_{i_1 i_3} + a_{i_3 i_1} a_{i_1 i_2}) + (a_{i_1 i_2} a_{i_2 i_3} + a_{i_3 i_2} a_{i_2 i_1}) + (a_{i_1 i_3} a_{i_3 i_2} + a_{i_2 i_3} a_{i_3 i_1})$
6 	$f_6(\mathbf{A}[i_1, i_2, i_3]) = a_{i_1 i_2} a_{i_1 i_3} + a_{i_2 i_1} a_{i_2 i_3} + a_{i_3 i_1} a_{i_3 i_2}$
7 	$f_7(\mathbf{A}[i_1, i_2, i_3]) = a_{i_2 i_1} a_{i_3 i_1} + a_{i_1 i_2} a_{i_3 i_2} + a_{i_1 i_3} a_{i_2 i_3}$
8 	$f_8(\mathbf{A}[i_1, i_2, i_3]) = a_{i_1 i_2} a_{i_2 i_3} a_{i_3 i_1} + a_{i_3 i_2} a_{i_2 i_1} a_{i_1 i_3}$
9 	$f_9(\mathbf{A}[i_1, i_2, i_3]) = a_{i_1 i_2} a_{i_1 i_3} (a_{i_2 i_3} + a_{i_3 i_2}) + a_{i_2 i_1} a_{i_2 i_3} (a_{i_1 i_3} + a_{i_3 i_1}) + a_{i_3 i_1} a_{i_3 i_2} (a_{i_1 i_2} + a_{i_2 i_1})$
10* 	$f_{10}(\mathbf{A}[i_1, i_2, i_3, i_4]) = \sum a_{ik} a_{il} (a_{jk} + a_{jl}) + a_{jk} a_{jl} (a_{ik} + a_{il})$
11* 	$f_{11}(\mathbf{A}[i_1, i_2, i_3, i_4]) = \sum a_{ik} a_{il} a_{jk} a_{jl}$
12* 	$f_{12}(\mathbf{A}[i_1, i_2, i_3, i_4]) = \sum (a_{ik} a_{jl} + a_{il} a_{jk}) (a_{kl} + a_{lk})$
13* 	$f_{13}(\mathbf{A}[i_1, i_2, i_3, i_4]) = \sum (a_{ik} a_{jl} + a_{il} a_{jk}) (a_{ij} + a_{ji})$
14* 	$f_{14}(\mathbf{A}[i_1, i_2, i_3, i_4]) = \sum a_{ik} a_{il} a_{kj} a_{lj} + a_{jk} a_{jl} a_{ki} a_{li}$

* The sum is taken over all possible combinations $(i < j)$ chosen from $\{i_1, i_2, i_3, i_4\}$, and $(k < l)$ being the two remaining nodes.

Supplementary Table S2. The functions $W_{\mathcal{M}}()$ for the motifs under study

Motif	$W_{\mathcal{M}}$
1 	$\binom{n}{2}r_+$
2 	$2(n-2)r_+rN_1 + 3\binom{n}{3}r_+^2$
3 	$r_+r^2N_2 + (n-2)r_+^2rN_1 + \binom{n}{3}r_+^3$
4 	$2\binom{n}{2}r_+$
5 	$2(n-2)r_+rN_4 + 6\binom{n}{3}r_+^2$
6 	$(n-2)r_+rN_4 + 3\binom{n}{3}r_+^2$
7 	$(n-2)r_+rN_4 + 3\binom{n}{3}r_+^2$
8 	$r_+r^2N_5 + (n-2)r_+^2rN_4 + 2\binom{n}{3}r_+^3$
9 	$r_+r^2(N_5 + 2N_6 + 2N_7) + 3(n-2)r_+^2rN_4 + 6\binom{n}{3}r_+^3$
10 	$2r_+r^2\left(\binom{N_4}{2} + (N_6 + N_7)(n-3)\right) + 6\binom{n-2}{2}r_+^2rN_4 + 24\binom{n}{4}r_+^3$
11 	$r_+r^3N_{10} + r_+^2r^2\left(\binom{N_4}{2} + (N_6 + N_7)(n-3)\right) + 2\binom{n-2}{2}r_+^3rN_4 + 6\binom{n}{4}r_+^4$
12 	$r_+r^2\left(2\binom{N_4}{2} + (N_5 + 2N_7)(n-3)\right) + 6\binom{n-2}{2}r_+^2rN_4 + 24\binom{n}{4}r_+^3$
13 	$r_+r^2\left(2\binom{N_4}{2} + (N_5 + 2N_6)(n-3)\right) + 6\binom{n-2}{2}r_+^2rN_4 + 24\binom{n}{4}r_+^3$
14 	$r_+r^3(N_{12} + N_{13}) + r_+^2r^2\left(2\binom{N_4}{2} + (N_5 + N_6 + N_7)(n-3)\right) + 4\binom{n-2}{2}r_+^3rN_4 + 12\binom{n}{4}r_+^4$

$$r = 1 - r_- - r_+.$$

Supplementary Notes

Supplementary Note 1. Random network models

Four widely used network models are considered.

Erdős-Renyi (ER) model

This model has edge density as its parameter. In the ER model with parameter ρ , a random network \mathcal{G} with n nodes is generated by adding a link between each pair of nodes independently and uniformly with probability ρ .

The node degrees of a random network \mathcal{G} generated from the ER model follow a Poisson distribution in which all nodes tend to have similar degrees.

Preferential attachment model

This model has two parameters: an initial seed network and the number of links incident to the newly added node. In the preferential attachment model with parameters \mathcal{G}_0 and l , a random network \mathcal{G} with n nodes is generated from \mathcal{G}_0 by adding one node at a time until the resulting network has n nodes. At each iteration step, a new node with l incident links is added to the current network. The neighbors of the newly added node are chosen with probabilities proportional to their degrees in the current network.

Networks generated from the preferential attachment model are scale-free, having power-law degree distributions.

Duplication model

This model has two parameters: an initial seed network and the probability that a link is added between the duplicate node and a neighbor of the copied node. In the duplication model with parameters \mathcal{G}_0 and p_{dup} , a random network \mathcal{G} is generated from \mathcal{G}_0 by adding a node at a time until the resulting network reaches n nodes. At each iteration step, a node u is chosen uniformly at random to duplicate. The duplicate node u' is connected to each neighbor of u with probability p_{dup} in addition to being connected to the node u .

The duplication model is used to model gene duplication and mutation events in biological evolution. Networks generated from the duplication model also have the scale-free property.

Geometric model

This network model has a distance threshold δ as its parameter. In the geometric network model with parameter δ , a random undirected network \mathcal{G} of n nodes is generated by (i) uniformly placing n points in a unit cube and (ii) adding a link between two nodes if the distance between the corresponding points is less than δ .

The random networks generated in this model do not necessarily have the power-law degree distribution.

Supplementary Note 2. Comparison of \tilde{N}_1 and CCSB estimator

In this section we compare our estimator \tilde{N}_1 and the CCSB estimator for estimating the number of links in PPI networks. In each of these two approaches, the number of links in the observed subnetwork \mathcal{G}^{obs} is first scaled up by the factor $\binom{n}{2}/\binom{n^{\text{obs}}}{2}$ to estimate the number of links in the entire network. The bias caused by the error rates, however, is handled differently. Let \tilde{N}_{CCSB} denote the CCSB estimator. We have:

$$\tilde{N}_1 = \frac{\hat{N}_1 - \binom{n}{2}r_+}{1 - r_+ - r_-},$$

$$\tilde{N}_{\text{CCSB}} = \frac{\hat{N}_1 \times \text{precision}}{\text{sensitivity}}.$$

The quality parameters are defined as follows:

- TP: true positives;
- FP: false positives;
- FN: false negatives;
- TN: true negatives;
- precision = $\frac{\text{TP}}{\text{TP}+\text{FP}}$;
- false discovery rate $r_d = \frac{\text{FP}}{\text{TP}+\text{FP}}$;
- sensitivity = $\frac{\text{TP}}{\text{TP}+\text{FN}}$;
- false negative rate $r_- = \frac{\text{FN}}{\text{TP}+\text{FN}} = 1 - \text{sensitivity}$;
- false positive rate $r_+ = \frac{\text{FP}}{\text{FP}+\text{TN}}$.

In an ideal classification problem (Supplementary Fig. S9A), that is, when the gold standard positive and negative sets are fully known, we have:

$$\hat{N}_1 = \text{TP}+\text{FP}, \quad N_1 = \text{TP}+\text{FN}, \quad \binom{n}{2} - N_1 = \text{FP}+\text{TN}. \quad (\star)$$

Since

$$N_1 = \text{TP}+\text{FN} = \hat{N}_1 \times \text{precision} + N_1 \times (1 - \text{sensitivity}),$$

the CCSB estimator $\tilde{N}_{\text{CCSB}} = \frac{\hat{N}_1 \times \text{precision}}{\text{sensitivity}}$ is derived.

On the other hand, we have

$$\hat{N}_1 = \text{TP}+\text{FP} = N_1(1 - r_-) + \left(\binom{n}{2} - N_1 \right) r_+,$$

giving our estimator $\tilde{N}_1 = \frac{\hat{N}_1 - \binom{n}{2} r_+}{1 - r_+ - r_-}$. Thus, \tilde{N}_1 and \tilde{N}_{CCSB} are mathematically equivalent.

However, for biological networks data, gold-standard positive and negative sets are limited and biased (Supplementary Fig. S9B). Thus, using gold-standard sets to infer the quality parameters is not a reliable approach as the equations in (\star) no longer holds. The empirical framework using multiple experimental assays proposed by CCSB offers more accurate estimates of the quality parameters. In such cases, it can be shown that \tilde{N}_1 and \tilde{N}_{CCSB} are still almost the same. In particular, using the error rates r_- , r_+ , and r_d , the two estimators can be rewritten as

$$\begin{aligned}\tilde{N}_1 &= \frac{1}{1 - r_+ - r_-} \left(\frac{\binom{n}{2}}{\binom{n^{\text{obs}}}{2}} N_1^{\text{obs}} - \binom{n}{2} r_+ \right), \\ \tilde{N}_{\text{CCSB}} &\simeq \frac{1}{1 - r_-} \frac{\binom{n}{2}}{\binom{n^{\text{obs}}}{2}} N_1^{\text{obs}} (1 - r_d).\end{aligned}$$

Since r_+ is much smaller than r_- , we have $1 - r_+ - r_- \simeq 1 - r_-$. Then

$$\tilde{N}_1 - \tilde{N}_{\text{CCSB}} \simeq \frac{\binom{n}{2}}{1 - r_-} \left(\frac{N_1^{\text{obs}} r_d}{\binom{n^{\text{obs}}}{2}} - r_+ \right) = \frac{\binom{n}{2}}{1 - r_-} \left(\frac{\text{FP}^{\text{obs}}}{\binom{n^{\text{obs}}}{2}} - \frac{\text{FP}^{\text{obs}}}{\binom{n^{\text{obs}}}{2} - \text{P}^{\text{obs}}} \right),$$

where FP^{obs} and P^{obs} respectively denote the number of false positive links and real positive links in the observed subnetwork \mathcal{G}^{obs} . It should be noted that biological networks are quite sparse, that is, $\text{P}^{\text{obs}} \ll \binom{n^{\text{obs}}}{2}$. Hence, \tilde{N}_1 and \tilde{N}_{CCSB} are almost the same.

However, using multiple experimental assays to accurately estimate the quality of detected interactions is not always possible due to time and cost. As a result, one must use gold-standard sets to make inference although they are currently limited and biased (Supplementary Fig. S9). In such cases, \tilde{N}_1 is better than \tilde{N}_{CCSB} because the estimated false positive rate r_+ is more reliable than the estimated false discovery rate r_d .

Supplementary Note 3. The computational efficiency of the sampling approach

One may count the occurrences of a motif in a fully known network by exhaustive enumeration. However, this direct approach is often time-consuming and even intractable for widely used huge networks such as world-wide-web and social networks. Our estimator $\hat{N}_{\mathcal{M}}$ provides a fast sampling-estimating method for counting motif occurrences. Take the triangle motif for instance. A naive method for counting the number of triangles in a network of n nodes may have time complexity $O(n^3)$, whereas the sampling-estimating approach has time complexity $O((pn)^3 \times (\text{no. of sampling times}))$, where p is the node sampling probability. Fig. 4 shows how the computational efficiency and the mean square error (MSE) of $\hat{N}_{\mathcal{M}}$ depend on the node sampling probability p and the number of sampling times, indicating that it can achieve within $\sim 1\%$ deviation from the true count by using no more than 50% of the computing time as compared with the naive triangle counting method (Fig. 4A).

Supplementary Note 4. Effect of non-uniform sampling on motif estimation

Our proposed estimators are derived based on that the observed subnetwork \mathcal{G}^{obs} is the outcome of a uniform node sampling process. More specifically, independent, identically distributed (iid) Bernoulli random variables X_i with parameter $0 < p < 1$ are used to denote the event whether the node i is sampled ($X_i = 1$) or not ($X_i = 0$), $i = 1, 2, \dots, n$. In practice, the observed subnetwork may be sampled by consideration of protein properties and hence it does not follow the uniform sampling scheme. Here, we investigate how other sampling schemes affect the accuracy of our proposed estimation.

One may likely select important proteins for Y2H experiment to study PPIs. First, we assume that proteins are selected independently, but with different probability p_i . In other words, $X_i \sim \text{Bernoulli}(p_i)$, $0 < p_i < 1$, $i = 1, 2, \dots, n$. To keep the average proportion of sampled nodes to be p ($0 < p < 1$), we consider the following two schemes:

1. Each p_i is randomly drawn from a uniform distribution such that

$$p_i \sim \text{Uniform}\left(\frac{p}{2}, \frac{3p}{2}\right).$$

2. Each p_i is randomly drawn from a normal distribution such that

$$p_i \sim \text{Normal}\left(\mu = p, \sigma^2 = \frac{p^2}{16}\right).$$

Here, the variance is chosen so that randomly generated values of p_i are not likely to be negative and the mean of p_i is p in each sampling scheme. When $p_i \geq 1$, we take it to be 1.

Biological networks are often modeled as scale-free, that is, most of the nodes have low degrees, whereas a small number of nodes have significantly high degrees. Highly connected proteins may likely play many important functions through their vast repertoire of interactions with other proteins. Thus, one may select proteins with probability proportional to their degree in the corresponding PPI network. To take such bias selection into account, we also consider the following sampling scheme:

3. The network node i ($1 \leq i \leq n$) is sampled with probability

$$p_i = \min\left\{np \frac{d_i}{\sum_{j=1}^n d_j}, 1\right\},$$

where d_i is the degree of the node i in the corresponding network \mathcal{G} . Such a sampling scheme also has the average sampling proportion p .

We examined the effect of the above three non-uniform schemes on the proposed motif estimation by simulation. We generated random scale-free networks of different order (n) and power-law exponent (γ) from the preferential attachment model, where $\gamma = 1.5, 2, 2.5, 3, 3.5$ and $n = 5,000, 6,000, \dots, 30,000$. The link density ρ was fixed at 0.001 so that the resulting networks are similar to real PPI networks studied in this work. We also considered different

values of the average sampling proportion $p = 0.1, 0.2, \dots, 0.9$. For each p and each non-uniform sampling scheme, we first generated a sampling vector (p_1, p_2, \dots, p_n) . For each sampling vector, we sampled 50 subnetworks and estimated the number of links and triangles using the estimator $\hat{N}_{\mathcal{M}}$.

Supplementary Fig. S10 demonstrates how $\text{MSE}(\hat{N}_3)$ for triangle count depends on sampling scheme and the average sampling proportion p . First, for each of the four sampling schemes considered, $\text{MSE}(\hat{N}_3)$ decreases when p increases. This confirms that $\text{MSE}(\hat{N}_3)$ is asymptotically unbiased for all the three sampling schemes. Second, the first two non-uniform sampling schemes are not significantly different from the uniform-node sampling scheme ($p_i = p$). When the third non-uniform sampling scheme is used, $\text{MSE}(\hat{N}_3)$ is significantly higher than that of the other schemes. This is not surprising because bias selection towards highly connected nodes should lead to over-estimation.

Supplementary Fig. S11 further confirms the over-estimation bias since the mean of the ratio of estimate to real count (i.e., \hat{N}_3/N_3) is larger than one. Moreover, it can be seen clearly from the figure that the over-estimation bias is significantly higher for networks generated from the preferential attachment and duplication models because of their scale-free degree distribution.

Supplementary Fig. S12 shows how the over-estimation bias depends on the scale-free parameter γ of networks when setting $n = 20,000$ and $\rho = 0.001$ for the third non-uniform sampling scheme. The mean of the ratio \hat{N}_3/N_3 slightly decreases as the power-law exponent γ increases. Interestingly, the ratio mean does not change much when we change γ . This shows that estimation is robust against different choices of exponents in the power law. Supplementary Fig. S12 also demonstrates how the mean of the ratio \hat{N}_3/N_3 depends on the average sampling proportion p . In particular, when more than 60% of nodes in \mathcal{G} are sampled, the estimate is less than five times the real count.

Supplementary Methods

Proof of Theorem 2

Theorem 2 *Let $q = 1 - p$. We have*

$$\text{Var} \left(\frac{\hat{N}_1}{N_1} \right) = \frac{2q}{p} \frac{N_2}{N_1^2} [1 + O(n^{-1})] + \frac{(1+p)q}{N_1 p^2} [1 + O(n^{-1})] + O(n^{-1}).$$

The convergence rate of $\frac{N_2}{N_1^2}$ as $n \rightarrow \infty$ depends on the structure of the underlying network \mathcal{G} . More specifically, we have

- *ER model: $\frac{N_2}{N_1^2} = O(n^{-1})$*
- *Preferential attachment model: $\frac{N_2}{N_1^2} = O(\log n/n)$*
- *Geometric model: $\frac{N_2}{N_1^2} = O(n^{-1})$*
- *Partial duplication model: let β be the approximated exponent of the power-law degree distribution of \mathcal{G} ,*

$$\frac{N_2}{N_1^2} = \begin{cases} O((\log n)^{-2}) & \text{if } \beta = 2, \\ O(n^{2-\beta}) & \text{if } 2 < \beta < 3, \\ O(\log n/n) & \text{if } \beta = 3, \\ O(n^{-1}) & \text{if } \beta > 3. \end{cases}$$

Hence, $\text{Var} \left(\frac{\hat{N}_1}{N_1} \right) \rightarrow 0$ as $n \rightarrow \infty$ and the estimator \hat{N}_1 is consistent.

Proof. Let β_0 be the number of pairs of edges in \mathcal{G} which have no common nodes. Recall that N_2 is the number of pairs of edges in the original network that have exactly one common neighbors. We have

$$\beta_0 + N_2 = \binom{N_1}{2}. \quad (\text{S1})$$

Define

$$\begin{aligned} \Gamma_0 &= \{(i, j, k, l) : 1 \leq i < j \leq n; 1 \leq k < l \leq n; \{i, j\} \cap \{k, l\} = \emptyset\}; \\ \Gamma_1 &= \{(i, j, k, l) : 1 \leq i < j \leq n; 1 \leq k < l \leq n; |\{i, j\} \cap \{k, l\}| = 1\}. \end{aligned}$$

Observe that

$$2\beta_0 = \sum_{(i,j,k,l) \in \Gamma_0} a_{ij} a_{kl}, \quad (\text{S2})$$

$$2N_2 = \sum_{(i,j,k,l) \in \Gamma_1} a_{ij} a_{kl}. \quad (\text{S3})$$

Let

$$\mu = E \frac{X_1 X_2}{n^{\text{obs}}(n^{\text{obs}} - 1)} = \frac{1 - q^n - npq^{n-1}}{n(n-1)}, \quad (\text{S4})$$

$$\alpha_0 - \mu^2 = \text{Cov} \left(\frac{X_1 X_2}{n^{\text{obs}}(n^{\text{obs}} - 1)}, \frac{X_3 X_4}{n^{\text{obs}}(n^{\text{obs}} - 1)} \right), \quad (\text{S5})$$

$$\alpha_1 - \mu^2 = \text{Cov} \left(\frac{X_1 X_2}{n^{\text{obs}}(n^{\text{obs}} - 1)}, \frac{X_2 X_3}{n^{\text{obs}}(n^{\text{obs}} - 1)} \right), \quad (\text{S6})$$

$$\alpha_2 - \mu^2 = \text{Var} \left(\frac{X_1 X_2}{n^{\text{obs}}(n^{\text{obs}} - 1)} \right). \quad (\text{S7})$$

Let Z and ξ be independent with $Z \sim \text{Binomial}(n-4, p)$ and $\xi \sim \text{Bernoulli}(p)$. For $m = 1, 2, 3$, define $\gamma_m = E[(Z+m)^{-2}(Z+m+1)^{-2}]$. It follows that

$$\begin{aligned} \alpha_1 &= E \left(\frac{X_1 X_2 X_3}{(n^{\text{obs}})^2 (n^{\text{obs}} - 1)^2} \right) \\ &= p^3 E [(Z + \xi + 3)^{-2} (Z + \xi + 2)^{-2}] \\ &= p^4 E [(Z + 4)^{-2} (Z + 3)^{-2}] + p^3 q E [(Z + 3)^{-2} (Z + 2)^{-2}] \\ &= p^4 \gamma_3 + p^3 q \gamma_2. \end{aligned} \quad (\text{S8})$$

Similarly,

$$\alpha_2 = p^4 \gamma_3 + 2p^3 q \gamma_2 + p^2 q^2 \gamma_1 \quad (\text{S9})$$

$$\alpha_0 = p^4 \gamma_3. \quad (\text{S10})$$

We compute the variance of \widehat{N}_1 as follows:

$$\begin{aligned} & \frac{\text{Var}(\widehat{N}_1)}{n^2(n-1)^2} \\ &= \sum_{1 \leq i < j \leq n} a_{i,j} \text{Var} \left(\frac{X_i X_j}{n^{\text{obs}}(n^{\text{obs}} - 1)} \right) + \sum_{(i,j,k,l) \in \Gamma_0} a_{i,j} a_{k,l} \text{Cov} \left(\frac{X_i X_j}{n^{\text{obs}}(n^{\text{obs}} - 1)}, \frac{X_k X_l}{n^{\text{obs}}(n^{\text{obs}} - 1)} \right) \\ & \quad + \sum_{(i,j,k,l) \in \Gamma_1} a_{i,j} a_{k,l} \text{Cov} \left(\frac{X_i X_j}{n^{\text{obs}}(n^{\text{obs}} - 1)}, \frac{X_k X_l}{n^{\text{obs}}(n^{\text{obs}} - 1)} \right) \\ &= N_1(\alpha_2 - \mu^2) + \sum_{(i,j,k,l) \in \Gamma_0} a_{i,j} a_{k,l}(\alpha_0 - \mu^2) + \sum_{(i,j,k,l) \in \Gamma_1} a_{i,j} a_{k,l}(\alpha_1 - \mu^2) \\ &= N_1(\alpha_2 - \mu^2) + 2\beta_0(\alpha_0 - \mu^2) + 2N_2(\alpha_1 - \mu^2) \\ &= 2p^3 q \gamma_2 N_2 + p^2 q [2p \gamma_2 + q \gamma_1] N_1 + [p^4 \gamma_3 - \mu^2] N_1^2. \end{aligned}$$

In the second equality, we used symmetry consideration, (S5), (S6) and (S7). In the third equality, we used (S2) and (S3). In the fourth equality, we use (S1), (S8), (S9), and (S10).

Summarizing, we have

$$\frac{\text{Var}(\widehat{N}_1)}{N_1^2} = n^4 [1 + O(n^{-2})] \left\{ \frac{2p^3 q \gamma_2 N_2}{N_1^2} + \frac{p^2 q [2p \gamma_2 + q \gamma_1]}{N_1} + [p^4 \gamma_3 - \mu^2] \right\}. \quad (\text{S11})$$

Combining the following lemma 1 and (S11), we have

$$\text{Var} \left(\frac{\hat{N}_1}{N_1} \right) = \frac{2q}{p} \frac{N_2}{N_1^2} [1 + O(n^{-1})] + \frac{(1+p)q}{N_1 p^2} [1 + O(n^{-1})] + O(n^{-1}).$$

Lemma 1 Recall the notations Z and the γ_1 's. We have

- (i) $n^4 \gamma_2 = p^{-4} [1 + O(n^{-1})]$,
- (ii) $n^4 [2p\gamma_2 + q\gamma_1] = (1+p)p^{-4} [1 + O(n^{-1})]$,
- (iii) $n^4 [p^4 \gamma_3 - \mu^2] = O(n^{-1})$.

Proof. Since $\varphi(x) = (x+2)^{-2}(x+3)^{-2}$ is convex, we apply Jensen's inequality to obtain a lower bound on

$$\begin{aligned} \gamma_2 = E [(Z+2)^{-2}(Z+3)^{-2}] &\geq [(n-4)p+2]^{-2}[(n-4)p+3]^{-2} \\ &= (np)^{-4} - 2(5-8p)(np)^{-5} + O(n^{-6}). \end{aligned}$$

For upper bound, we proceed as

$$\begin{aligned} \gamma_2 &\leq E [(Z+1)(Z+2)(Z+3)(Z+4)]^{-1} \\ &= E \left\{ \frac{1}{6(Z+1)} - \frac{1}{2(Z+2)} + \frac{1}{2(Z+3)} - \frac{1}{6(Z+4)} \right\} \\ &= 1/6 E \int_0^1 (1-t)^3 t^Z dt \\ &= 1/6 \int_0^1 (1-t)^3 (q+pt)^{n-4} dt \\ &= p^{-4} \int_q^1 (1-u)^3 u^{n-4} du \\ &= (np)^{-4} + 6(np)^{-5} + O(n^{-6}). \end{aligned}$$

Hence part (i) follows from these lower and upper bounds.

In a similar way, we can prove that $\gamma_1 = (np)^{-4} [1 + O(n^{-1})]$ and $\gamma_3 = (np)^{-4} [1 + O(n^{-1})]$. Hence part (ii) follows.

Since $\mu^2 = n^{-4} [1 + O(n^{-1})]$, part (iii) also follows immediately from these bounds. □

We now derive the convergence rate of $\frac{N_2}{N_1^2}$ for ER model, the preferential attachment model, the partial duplication model, and the geometric model. For ER model we have

$$\begin{aligned} N_1 &\simeq \binom{n}{2} \rho, \text{ where } \rho \text{ is the density of } \mathcal{G}, \\ N_2 &\leq 3 \binom{n}{3}, \text{ (this is true for any network),} \\ \frac{N_2}{N_1^2} &\leq \frac{3 \binom{n}{3}}{[\binom{n}{2} \rho]^2} = \frac{n(n-1)(n-2)/2}{[n(n-1)\rho/2]^2} = \frac{2}{\rho^2} \frac{n-2}{n(n-1)} = O(n^{-1}). \end{aligned}$$

For the preferential attachment model, let n_k be the number of nodes of degree k , $p_k = n_k/n$, $1 \leq k \leq n$. Networks generated from the preferential attachment model described in Supplementary Note 2 have the power-law degree distribution with exponent 3, that is, $p_k \simeq Ck^{-3}$. We have

$$\begin{aligned}
N_1 &\simeq l \times n, \text{ where } l \text{ is the number of edges added at each iteration,} \\
N_2 &= \sum_{k=1}^n \binom{k}{2} n_k \\
&= n \sum_{k=1}^n \binom{k}{2} p_k \\
&\simeq n \sum_{k=1}^n \binom{k}{2} \frac{C}{k^3} \\
&\simeq \frac{C}{2} n (\log(n) - \pi^2/6), \\
\frac{N_2}{N_1^2} &\simeq \frac{\frac{C}{2} n (\log(n) - \pi^2/6)}{(nl)^2} = O\left(\frac{\log(n)}{n}\right).
\end{aligned}$$

For the geometric model, let k_i denote the degree of node i , $1 \leq i \leq n$. We have

$$\begin{aligned}
\frac{\sum_{i=1}^n k_i}{n} &= \bar{k} \simeq E(k_i) \simeq (n-1) \frac{4}{3} \pi \delta^3, \\
N_1 &= \frac{\sum_{i=1}^n k_i}{2} \simeq \frac{n(n-1)}{2} \frac{4}{3} \pi \delta^3, \\
N_2 &\leq 3 \binom{n}{3}, \\
\frac{N_2}{N_1^2} &\leq \frac{2}{(\frac{4}{3} \pi \delta^3)^2} \times \frac{n-2}{n(n-1)} = O(n^{-1}).
\end{aligned}$$

For the duplication model, it can be shown that

$$N_1 \simeq \begin{cases} \frac{n}{1-2p_{\text{dup}}} + C_0 n^{2p_{\text{dup}}} & \text{if } p_{\text{dup}} \neq \frac{1}{2}, \\ n \log(n) + C_0 n & \text{if } p_{\text{dup}} = \frac{1}{2}, \end{cases}$$

where constant C_0 is determined by the initial ER random network \mathcal{G}_0 . The degree distribution follows a power law, that is, $p_k \simeq Ck^{-\beta}$, where the exponent β satisfying the following equation

$$1 + p_{\text{dup}} = p_{\text{dup}} \beta + p_{\text{dup}}^{\beta-1}.$$

Let n_k be the number of nodes of degree k , $p_k = n_k/n$, $1 \leq k \leq n$, we have

$$N_2 = \sum_{k=1}^n \binom{k}{2} n_k \simeq \sum_{k=1}^n \binom{k}{2} n p_k \simeq n \sum_{k=1}^n \binom{k}{2} C k^{-\beta} = \frac{C}{2} n \sum_{k=1}^n \left(\frac{1}{k^{\beta-2}} - \frac{1}{k^{\beta-1}} \right).$$

For $p_{\text{dup}} = \frac{1}{2}$, we have $\beta = 2$ and

$$\begin{aligned} N_2 &\simeq \frac{C}{2}n(n - \sum_{k=1}^n \frac{1}{k}) < \frac{C}{2}n(n - \log(n+1)), \\ N_1 &\simeq n \log(n) + C_0 n, \\ \frac{N_2}{N_1^2} &< \frac{\frac{C}{2}n(n - \log(n+1))}{(n \log(n) + C_0 n)^2} = O(\frac{1}{(\log(n))^2}). \end{aligned}$$

For $p_{\text{dup}} < \frac{1}{2}$, we have $\beta > 2$. We consider the following cases.

Case 1. When $\beta > 3$, both $\sum_{k=1}^n \frac{1}{k^{\beta-2}}$ and $\sum_{k=1}^n \frac{1}{k^{\beta-1}}$ converge. This implies that

$$N_2 \simeq \frac{C}{2}n \sum_{k=1}^n (\frac{1}{k^{\beta-2}} - \frac{1}{k^{\beta-1}}) = O(n),$$

and

$$N_1 \simeq \frac{n}{1 - 2p_{\text{dup}}} + C_0 n^{2p_{\text{dup}}} = O(n).$$

Hence $\frac{N_2}{N_1^2} = O(n^{-1})$.

Case 2. When $\beta = 3$,

$$N_2 \simeq \frac{C}{2}n \sum_{k=1}^n (\frac{1}{k} - \frac{1}{k^2}) = O(n \log(n)),$$

and

$$N_1 \simeq \frac{n}{1 - 2p_{\text{dup}}} + C_0 n^{2p_{\text{dup}}} = O(n).$$

Hence, $\frac{N_2}{N_1^2} = O(\frac{\log(n)}{n})$.

Case 3. When $2 < \beta < 3$, $\sum_{k=1}^n \frac{1}{k^{\beta-1}}$ converges, and

$$\sum_{k=1}^n \frac{1}{k^{\beta-2}} < \int_0^n \frac{1}{x^{\beta-2}} dx = \frac{n^{3-\beta}}{3-\beta}.$$

Therefore,

$$N_2 < \frac{C}{2}n(\frac{n^{3-\beta}}{3-\beta} + \sum_{k=1}^n \frac{1}{k^{\beta-1}}) = O(n^{4-\beta}),$$

and

$$N_1 \simeq \frac{n}{1 - 2p_{\text{dup}}} + C_0 n^{2p_{\text{dup}}} = O(n).$$

This implies that $\frac{N_2}{N_1^2} = O(\frac{1}{n^{\beta-2}})$. □

Derivation of bias-corrected estimator $\tilde{N}_{\mathcal{M}}$

When the error rates r_+ and r_- are taken into account, the adjacency matrix \mathbf{A} in the expressions of $N_{\mathcal{M}}^{\text{obs}}$ and $\hat{N}_{\mathcal{M}}$ is replaced by $\tilde{\mathbf{A}}$. In particular $\hat{N}_{\mathcal{M}}$ can be written as

$$\hat{N}_{\mathcal{M}} = \frac{\binom{n}{m}}{\binom{n^{\text{obs}}}{m}} \sum_{1 \leq i_1 < i_2 < \dots < i_m \leq n} f_{\mathcal{M}}(\tilde{\mathbf{A}}[i_1, i_2, \dots, i_m]) X_{i_1} X_{i_2} \dots X_{i_m}.$$

Since the random variables X_i are independently and identically distributed, and they are also independent of $F_{i_1 i_2}^+$ and $F_{i_1 i_2}^-$, we have

$$E(\hat{N}_{\mathcal{M}}) = E\left(\frac{\binom{n}{m}}{\binom{n^{\text{obs}}}{m}} X_1 X_2 \dots X_m\right) E\left(\sum_{1 \leq i_1 < i_2 < \dots < i_m \leq n} f_{\mathcal{M}}(\tilde{\mathbf{A}}[i_1, i_2, \dots, i_m])\right).$$

As shown in Theorem 1, we have

$$\begin{aligned} E\left(\frac{\binom{n}{m}}{\binom{n^{\text{obs}}}{m}} X_1 X_2 \dots X_m\right) &= 1 - q^n - npq^{n-1} - \dots - \binom{n}{j} p^j q^{n-j} - \dots - \binom{n}{m-1} p^{m-1} q^{n-(m-1)} \\ &= 1 - \sum_{j=0}^{m-1} \binom{n}{j} p^j q^{n-j} \rightarrow 1 \text{ as } n \rightarrow \infty \end{aligned}$$

On the other hand, we can work out the second expectation using the fact that $F_{i_1 i_2}^+ \sim \text{Bernoulli}(r_+)$, $F_{i_1 i_2}^- \sim \text{Bernoulli}(r_-)$, and they are independent. In particular,

$$E\left(\sum_{1 \leq i_1 < i_2 < \dots < i_m \leq n} f_{\mathcal{M}}(\tilde{\mathbf{A}}[i_1, i_2, \dots, i_m])\right) = (1 - r_+ - r_-)^s N_{\mathcal{M}} + W_{\mathcal{M}},$$

where s is the number of links in \mathcal{M} and $W_{\mathcal{M}}$ is a function of n , the error rates r_- and r_+ , and $N_{\mathcal{M}'}$, for all sub-motifs \mathcal{M}' of \mathcal{M} . $W_{\mathcal{M}}$'s are given in Supplementary Table S2 for the motifs \mathcal{M} under study. Thus, to correct the bias, we replace $N_{\mathcal{M}'}$ by $\tilde{N}_{\mathcal{M}'}$, obtaining $\tilde{W}_{\mathcal{M}}$, and adjust $\hat{N}_{\mathcal{M}}$ to

$$\tilde{N}_{\mathcal{M}} = \frac{1}{(1 - r_+ - r_-)^s} (\hat{N}_{\mathcal{M}} - \tilde{W}_{\mathcal{M}}).$$