RobustECD: Enhancement of Network Structure for Robust Community Detection (Appendix)

Jiajun Zhou, Zhi Chen, Min Du, Lihong Chen, Shanqing Yu,
Guanrong Chen, *Fellow, IEEE*, and Qi Xuan, *Member, IEEE*

APPENDIX A
DISCUSSION ON EDGE DELETION IN ROBUSTECD-SE

Similar to *RobustECD-GA*, these two parameters (β_a, β_d) are also available in *RobustECD-SE*. We extensively study the effect of budget parameters in *RobustECD-SE* to support our decision mentioned in Sec. 3.3.1, i.e., we only consider edge addition in *RobustECD-SE* and neglect edge deletion.

Fig. 1 show the relative improvement rate of time in *RobustECD-SE* involving edge deletion. From the comparison results, one can observe that edge deletion brings up extra time consumption in most cases.

Moreover, the impact of edge addition/deletion is shown in Fig. 2, from which one can observe that *RobustECD-SE* with extra edge deletion has the same or worse effect as that with only edge addition in most cases.

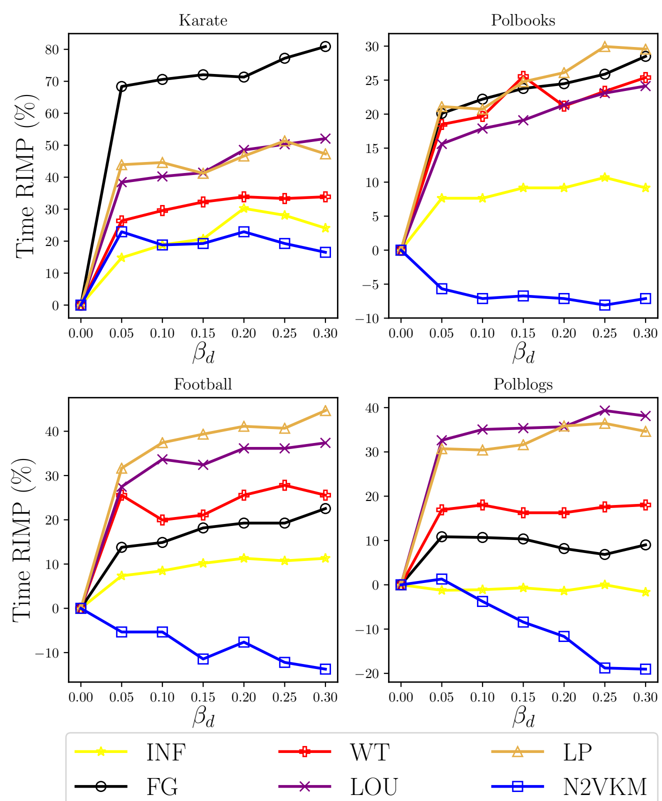

Fig. 1. Extra time consumption of edge deletion in *RobustECD-SE*.

APPENDIX B
ADVERSARIAL NETWORKS DETAILS

Adversarial perturbation is treated as malicious noise, which can be generated via adversarial attack. We consider the following two community deception methods to deploy adversarial attack:

- $\mathcal{Q}$**-Attack** [1]. It is an evolutionary attack strategy based on genetic algorithms, in which the modularity is used to design the fitness function. This strategy deploys attack via negligible network rewiring, which doesn't change the degree of vertices, and achieves the state-of-the-art attack effect.
- $\mathcal{D}_m$**-Deception via Modularity** [2]. $\mathcal{D}_m$ is a community deception algorithm based on modularity, which can hide a target community via intra-community edge deletion and inter-community edge addition.

We generate adversarial networks for all small benchmark networks. Details of attack setup are shown in TABLE 1.

For two large-scale networks, we generate networks with missing data as follows:

1) Select a certain number (x) of vertices as community seeds via weighted random sampling, in which the sampling weights are associate with vertex degree;
2) Extract h-hop ego-networks of these seed vertices, and remove those vertices that have multiple community labels;
3) Aggregate these ego-networks to form a connected subgraphs.

Details of parameter setup are reported in TABLE 2.

APPENDIX C
PARAMETER SENSITIVITY OF ROBUSTECD-GA

In this subsection, we discuss the impact of key parameters on the performance of *RobustECD-GA*. Fig. 3 shows the parameter sensitivity of *RobustECD-GA*, from which one can see that *RobustECD-GA* performs well in small-scale

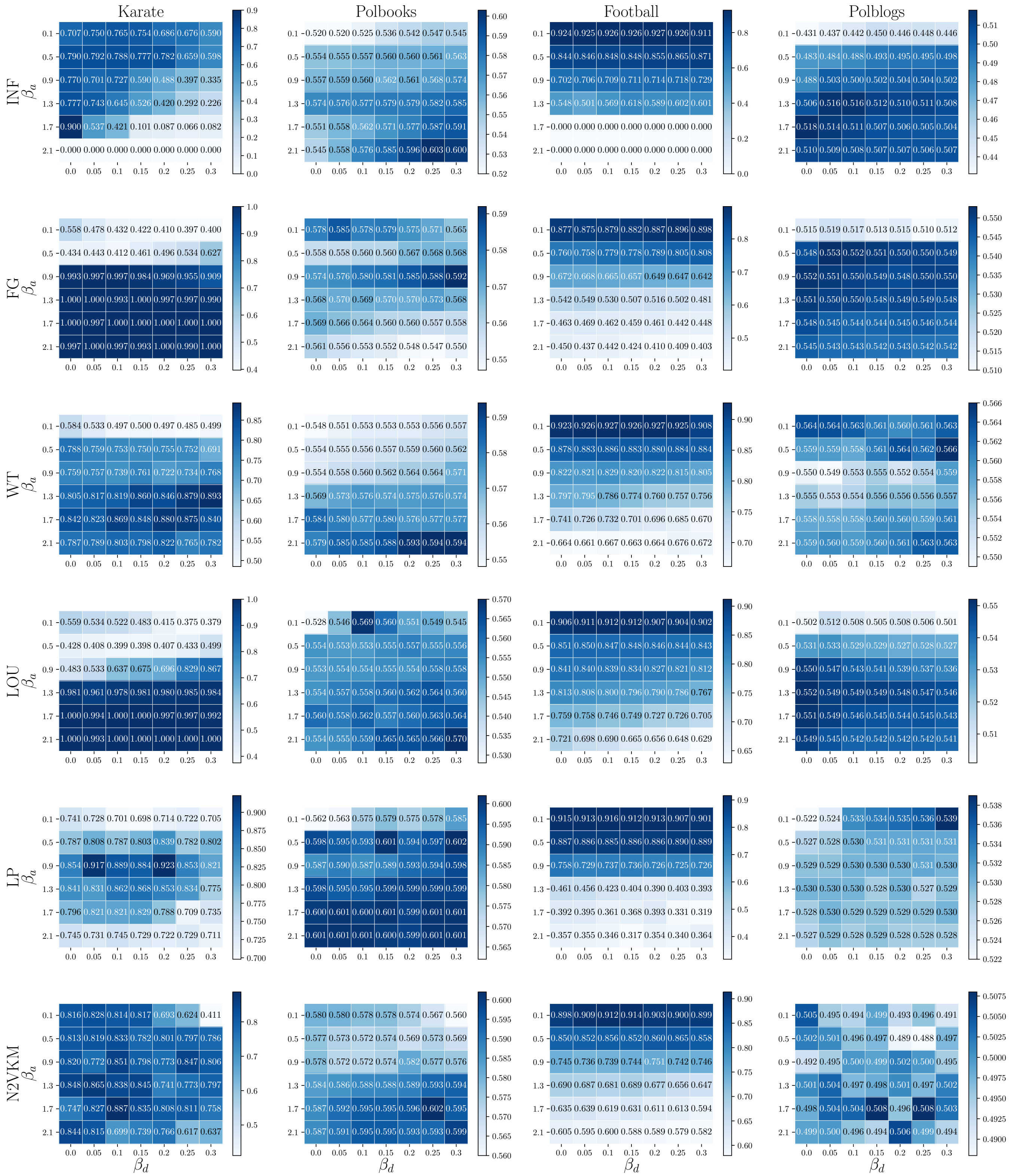

Fig. 2. The impact of both β_a and β_d in *RobustECD-SE* in term of NMI.

networks like Karate, Polbooks and Football. In general, *RobustECD-GA* is not strictly sensitive to different parameter settings in most cases, since that the sensitivity curves are smooth over the interval $[0.04, 0.28]$. Moreover, with the increase of network scale, like Polblogs, *RobustECD-GA* performs poorly. As a reasonable explanation, *RobustECD-GA* is based on evolutionary computation, which is capable of finding the optimal modification scheme in a small range of solution space. And with the increase of budget, the solution space becomes larger and *RobustECD-GA* is prone to fall into

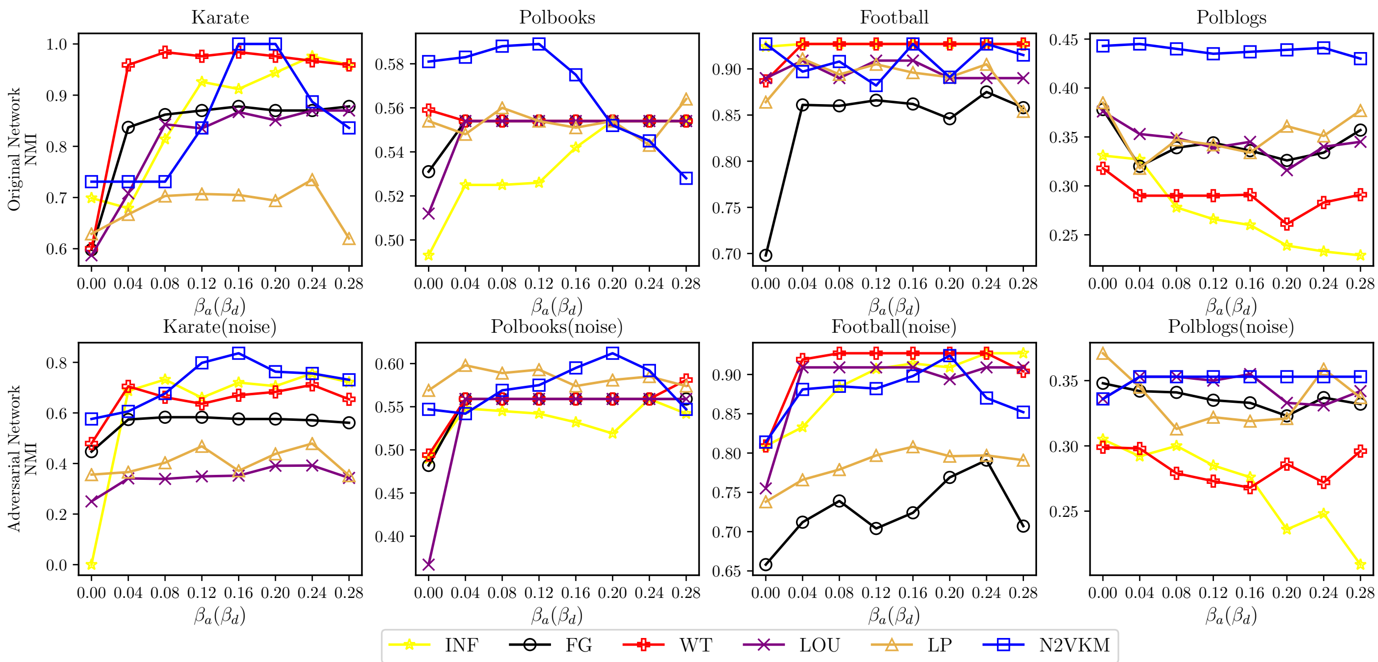

Fig. 3. The impact of modification budgets β_a and β_d on the performance of *RobustECD-GA*.

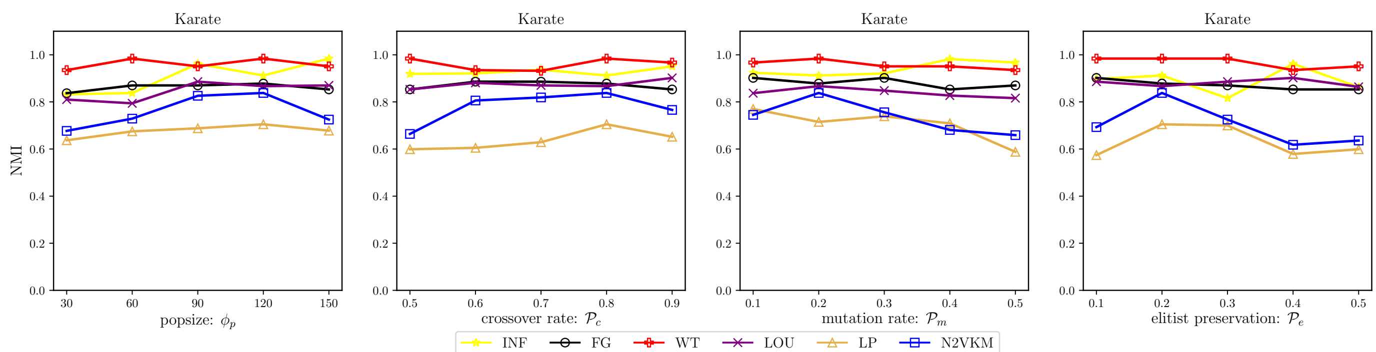

Fig. 4. The impact of GA paratemeters on the performance of *RobustECD-GA*.

TABLE 1
Details of attack setup.

Adversarial network	Attack method	Parameter	
		$\mathcal{S}$	β
Karate(noise)	$\mathcal{Q}$-Attack	LOU	5
Polbooks(noise)	$\mathcal{Q}$-Attack	LOU	20
Football(noise)	$\mathcal{D}_m$	WT	100
Polblogs(noise)	$\mathcal{Q}$-Attack	LOU	100

TABLE 2
Details of parameter setup in subgraph extraction.

Subgraph of real network	missing data	Parameter	
		x	h
Amazon-sub	$\surd$	15	4
DBLP-sub	$\surd$	10	3

local optimum, leading to a poor performance.

In addition, we also conduct experiments to show the sensitivity with respect to the parameters in genetic algorithm, as shown in Fig. 4. As we can see, *RobustECD-GA* is roughly not sensitive to different parameter settings in most cases. Note that the curves of LP and N2VKM have relatively obvious fluctuation, since that both of the two community detection algorithms have greater randomness.

Appendix D
Vertex similarity details

In this subsection, we briefly summarize the definition of similarity indices used in this paper.

- Common Neighbors (CN). It is defined as:

$$\mathcal{H}_{ij} = |\Gamma_i \cap \Gamma_j|, \tag{1}$$

where Γ_i denotes the set of 1-hop neighbors of vertex v_i and $|\cdot|$ is the cardinality of the set. It's easy to calculate CN by the adjacency matrix $\mathcal{A}$, i.e., $\mathcal{H}_{ij} = (\mathcal{A}^2)_{ij}$.

- Salton Index. It is defined as:

$$\mathcal{H}_{ij} = \frac{|\Gamma_i \cap \Gamma_j|}{\sqrt{k_i \times k_j}}, \tag{2}$$

where k_i denotes the degree of vertex v_i.
- Jaccard Index. It is defined as:

$$\mathcal{H}_{ij} = \frac{|\Gamma_i \cap \Gamma_j|}{|\Gamma_i \cup \Gamma_j|}. \tag{3}$$

- Hub Promoted Index (HPI). It is defined as:

$$\mathcal{H}_{ij} = \frac{|\Gamma_i \cap \Gamma_j|}{\min\{k_i, k_j\}}. \tag{4}$$

Under this measure, the links adjacent to hubs are likely to be assigned higher scores since the denominator is determined by the lower degree.
- Adamic-Adar Index (AA). It is defined as:

$$\mathcal{H}_{ij} = \sum_{z \in \Gamma_i \cap \Gamma_j} \frac{1}{\log k_z}, \tag{5}$$

where z is the common neighbor of v_i and v_j. The main assumption of this index is that the common neighbors of smaller degrees contribute more to the similarity.
- Resource Allocation Index (RA). It is defined as:

$$\mathcal{H}_{ij} = \sum_{z \in \Gamma_i \cap \Gamma_j} \frac{1}{k(z)}. \tag{6}$$

RA index is similar to AA, but punish more heavily on their common neighbors of high-degree.
- Local Path Index (LP). It is defined as:

$$\mathcal{H}_{ij} = (\mathcal{A}^2)_{ij} + \alpha(\mathcal{A}^3)_{ij}, \tag{7}$$

where α is a free parameter. LP considers the contribution of third-order neighbors on the basis of Common neighbors (CN), and degenerates to CN when $\alpha = 0$.
- RandomWalk with Restart (RWR). Consider a random walker starting from vertex v_i, who will iteratively moves to a random neighbor with probability c and return to vertex v_i with probability $1 - c$. Denote by q_{ij} the probability this random walker locates at vertex v_j in the steady state, we have

$$\boldsymbol{q}_i(t+1) = c\boldsymbol{P}^{\mathrm{T}}\boldsymbol{q}_i(t) + (1-c)\boldsymbol{e}_i, \tag{8}$$

where $\boldsymbol{P}$ is the transition matrix with $\boldsymbol{P}_{ij} = 1/k_i$ if v_i and v_j are connected, and $\boldsymbol{P}_{ij} = 0$ otherwise. The solution is straightforward, as

$$\boldsymbol{q}_i = (1-c)\left(I - c\boldsymbol{P}^{\mathrm{T}}\right)^{-1}\boldsymbol{e}_i, \tag{9}$$

The RWR index is thus defined as

$$\mathcal{H}_{ij}^{\mathrm{RWR}} = q_{ij} + q_{ji}, \tag{10}$$

where q_{ij} is the jth element of the vector $\boldsymbol{q}_i$.

Appendix E
Impact of single similarity index in RobustECD-SE

Fig. 5 and Fig. 6 show the results of all similarity indices (*RobustECD-SE(all)*) and single index (*RobustECD-SE(single)*) in real networks and adversarial networks, respectively. In Karate and Karate(noise) ($\beta_a \in (1.0, 2.0)$), *RobustECD-SE(single)*s with first-order similarity have relatively good performance while those with second-order and high-order similarity have relatively poor performance. Since that the scale of Karate is particularly small and first-order similarity are sufficient to capture structure features, so first-order similarity indices are complementary, second-order and high-order similarity are redundant or even negative in some cases. For Polbooks and Polbooks(noise), *RobustECD-SE(single)* with LP similarity receives the best performance in most cases. The specific definition of Local Path (LP) is $\mathcal{H} = \mathcal{A}^2 + \alpha\mathcal{A}^3$, where α is adjustable parameter. LP considers the contribution of third-order neighbors on the basis of Common neighbors (CN). So other similarity indices are redundant or even negative. For Football and Football(noise) ($\beta_a \in (0.0, 0.5)$), all *RobustECD-SE(single)*s have similarity results and achieve competitive performance against *RobustECD-SE(all)*. Since that Football network has a strong community structures, and single similarity index is sufficient to capture structure features. For polblogs and Polblogs(noise), *RobustECD-SE(single)*s achieve competitive performance against *RobustECD-SE(all)* except for those with Jaccard, Salton and high-order similarity, which turn out to be redundant.

Appendix F
Impact of combination of similarity indices in RobustECD-SE

Combined with the conclusions in Appendix E, we further explore how to design proper combinations of similarity indices. Table 3 and Table 4 report the results of several special combinations of similarity indices, from which we come to the following conclusions:

- *RobustECD-SE(combination:4)* obtains competitive results compared to *RobustECD-SE(all)*, i.e., after pruning half of the indices according to the category of similarity, *RobustECD-SE* still performs robustly.
- For very small networks like Karate, the combinations of first-order similarity indices (i.e., *RobustECD-SE(combination:1)*) are sufficient to achieve excellent performance while the combinations of higher-order similarity indices is counterproductive.
- For those networks with strong community structures like Football, different combinations obtain relatively consistent performances, and even a single similarity index is sufficient.

References

[1] J. Chen, L. Chen, Y. Chen, M. Zhao, S. Yu, Q. Xuan, and X. Yang, "Ga-based q-attack on community detection," *IEEE Transactions on Computational Social Systems*, vol. 6, no. 3, pp. 491–503, 2019.
[2] V. Fionda and G. Pirro, "Community deception or: How to stop fearing community detection algorithms," *IEEE Transactions on Knowledge and Data Engineering*, vol. 30, no. 4, pp. 660–673, 2017.

TABLE 3
Community detection results under different combinations of similarity indices in the real networks.

Dataset	Method	Similarity indices								Community Detection						
		1				2		3		NMI						
		CN	Jaccard	Salton	HPI	AA	RA	LP	RWR	INF	FG	WT	LOU	LP	N2VKM	Avg RIMP
Karate	Original									0.699±0.000	0.598±0.000	0.600±0.000	0.587±0.000	0.689±0.283	0.705±0.175	—
	RobustECD-SE(all)	√	√	√	√	√	√	√	√	**0.825±0.220**	**1.000±0.000**	0.821±0.088	**1.000±0.000**	**0.847±0.136**	**0.834±0.114**	**38.95%**
	RobustECD-SE(combination:1)	√	√	√	√					**0.874±0.241**	**0.997±0.023**	**0.825±0.099**	0.933±0.138	**0.839±0.144**	**0.855±0.109**	**38.54%**
	RobustECD-SE(combination:2)					√	√			0.379±0.218	0.974±0.060	0.767±0.120	0.913±0.152	0.771±0.131	0.571±0.263	15.56%
	RobustECD-SE(combination:3)							√	√	0.341±0.191	0.942±0.099	0.784±0.144	0.899±0.119	0.729±0.125	0.112±0.190	1.97%
	RobustECD-SE(combination:4)			√	√		√		√	0.730±0.227	**0.997±0.023**	**0.855±0.119**	**0.975±0.080**	0.809±0.140	0.829±0.111	35.79%
	RobustECD-SE(single)			√						**0.825±0.255**	0.967±0.065	0.797±0.120	0.821±0.152	0.799±0.136	0.833±0.119	31.09%
Polbooks	Original									0.493±0.000	0.531±0.000	0.559±0.000	0.512±0.000	0.554±0.025	0.556±0.017	—
	RobustECD-SE(all)	√	√	√	√	√	√	√	√	**0.574±0.014**	**0.569±0.001**	**0.586±0.017**	0.560±0.011	**0.598±0.009**	**0.589±0.009**	**8.61%**
	RobustECD-SE(combination:1)	√	√	√	√					0.561±0.014	0.558±0.014	0.581±0.016	0.556±0.009	0.592±0.014	0.588±0.007	7.34%
	RobustECD-SE(combination:2)					√	√			0.562±0.018	0.568±0.007	0.584±0.021	0.561±0.018	0.594±0.013	0.582±0.021	7.82%
	RobustECD-SE(combination:3)							√	√	**0.577±0.016**	**0.576±0.014**	0.579±0.015	**0.571±0.013**	**0.604±0.017**	0.589±0.024	**9.26%**
	RobustECD-SE(combination:4)			√	√		√		√	0.572±0.011	0.566±0.010	**0.588±0.022**	**0.563±0.012**	0.591±0.015	**0.591±0.014**	8.46%
	RobustECD-SE(single)			√						0.562±0.016	0.551±0.025	0.575±0.018	0.560±0.019	0.586±0.020	0.589±0.013	6.95%
Football	Original									0.924±0.000	0.698±0.000	0.887±0.000	0.890±0.000	0.888±0.037	0.912±0.012	—
	RobustECD-SE(all)	√	√	√	√	√	√	√	√	**0.924±0.000**	**0.877±0.021**	**0.923±0.009**	**0.906±0.014**	**0.915±0.018**	0.898±0.021	**5.50%**
	RobustECD-SE(combination:1)	√	√	√	√					**0.925±0.001**	**0.869±0.023**	**0.920±0.009**	0.902±0.015	0.908±0.019	0.894±0.017	4.99%
	RobustECD-SE(combination:2)					√	√			0.924±0.001	0.860±0.026	0.918±0.012	0.898±0.019	0.913±0.016	0.890±0.017	4.67%
	RobustECD-SE(combination:3)							√	√	0.924±0.001	0.857±0.028	0.916±0.014	**0.908±0.019**	0.910±0.015	**0.899±0.017**	4.85%
	RobustECD-SE(combination:4)			√	√		√		√	0.924±0.001	0.867±0.021	0.918±0.009	0.903±0.017	**0.914±0.019**	0.892±0.021	4.98%
	RobustECD-SE(single)								√	0.924±0.001	0.864±0.030	0.912±0.015	0.903±0.019	0.911±0.014	**0.908±0.017**	**5.04%**
Polblogs	Original									0.330±0.001	0.378±0.000	0.318±0.000	0.376±0.000	0.375±0.053	0.458±0.067	—
	RobustECD-SE(all)	√	√	√	√	√	√	√	√	0.517±0.007	**0.551±0.006**	0.556±0.009	**0.551±0.005**	0.529±0.007	**0.499±0.006**	**45.64%**
	RobustECD-SE(combination:1)	√	√	√	√					0.501±0.006	0.544±0.005	0.540±0.007	0.518±0.020	0.522±0.004	0.496±0.003	41.80%
	RobustECD-SE(combination:2)					√	√			0.514±0.011	**0.555±0.009**	**0.559±0.009**	**0.555±0.008**	**0.531±0.007**	0.486±0.012	45.61%
	RobustECD-SE(combination:3)							√	√	**0.525±0.006**	0.548±0.007	**0.560±0.011**	0.549±0.006	**0.532±0.007**	**0.498±0.007**	**46.13%**
	RobustECD-SE(combination:4)			√	√		√		√	**0.523±0.008**	0.548±0.009	0.534±0.012	0.546±0.008	0.526±0.007	0.453±0.005	42.63%
	RobustECD-SE(single)						√			0.521±0.010	0.547±0.009	0.537±0.038	0.549±0.008	0.521±0.013	0.457±0.007	42.70%

TABLE 4
Community detection results under different combinations of similarity indices in the adversarial networks.

Dataset	Method	Similarity indices								Community Detection						
		1				2		3		NMI						
		CN	Jaccard	Salton	HPI	AA	RA	LP	RWR	INF	FG	WT	LOU	LP	N2VKM	Avg RIMP
Karate (noise)	Original									0.699±0.000	0.598±0.000	0.600±0.000	0.587±0.000	0.689±0.283	0.705±0.175	63.90%
	Attack									0.000±0.000	0.447±0.000	0.487±0.000	0.250±0.000	0.475±0.337	0.399±0.231	—
	RobustECD-SE(all)	√	√	√	√	√	√	√	√	0.425±0.336	**0.709±0.218**	0.576±0.075	0.484±0.173	**0.828±0.197**	**0.529±0.340**	53.31%
	RobustECD-SE(combination:1)	√	√	√	√					0.594±0.212	**0.844±0.179**	**0.702±0.151**	**0.598±0.121**	**0.828±0.106**	**0.744±0.165**	**82.06%**
	RobustECD-SE(combination:2)					√	√			0.052±0.083	0.343±0.142	0.542±0.128	0.324±0.134	0.573±0.160	0.063±0.123	-6.79%
	RobustECD-SE(combination:3)							√	√	0.000±0.000	0.532±0.173	0.600±0.140	0.465±0.144	0.621±0.137	0.039±0.112	11.46%
	RobustECD-SE(combination:4)			√	√		√		√	**0.608±0.266**	0.617±0.259	0.594±0.119	0.411±0.170	0.781±0.181	0.264±0.308	35.96%
	RobustECD-SE(single)			√						**0.628±0.319**	0.657±0.618	**0.707±0.173**	**0.579±0.155**	0.689±0.155	0.445±0.301	**57.19%**
Polbooks (noise)	Original									0.493±0.000	0.531±0.000	0.559±0.000	0.512±0.000	0.554±0.025	0.556±0.017	26.69%
	Attack									0.418±0.004	0.482±0.000	0.393±0.000	0.343±0.000	0.461±0.030	0.462±0.012	—
	RobustECD-SE(all)	√	√	√	√	√	√	√	√	**0.590±0.014**	**0.565±0.012**	**0.599±0.008**	0.564±0.010	**0.599±0.008**	**0.643±0.026**	**40.72%**
	RobustECD-SE(combination:1)	√	√	√	√					0.587±0.016	0.545±0.005	**0.599±0.012**	0.562±0.014	0.593±0.014	0.635±0.027	39.31%
	RobustECD-SE(combination:2)					√	√			**0.590±0.016**	0.560±0.015	0.593±0.008	0.576±0.016	0.593±0.016	0.595±0.062	38.93%
	RobustECD-SE(combination:3)							√	√	**0.594±0.013**	**0.586±0.018**	0.590±0.018	**0.578±0.018**	**0.599±0.010**	0.632±0.037	**41.51%**
	RobustECD-SE(combination:4)			√	√		√		√	0.587±0.015	0.554±0.013	0.591±0.008	0.565±0.016	0.594±0.015	**0.638±0.034**	39.57%
	RobustECD-SE(single)			√						0.580±0.022	0.548±0.009	0.592±0.016	**0.578±0.022**	0.586±0.019	0.615±0.035	38.64%
Football (noise)	Original									0.924±0.000	0.698±0.000	0.887±0.000	0.890±0.000	0.888±0.037	0.912±0.012	11.27%
	Attack									0.809±0.000	0.658±0.000	0.809±0.000	0.755±0.000	0.800±0.051	0.838±0.027	—
	RobustECD-SE(all)	√	√	√	√	√	√	√	√	0.809±0.000	0.762±0.024	**0.909±0.020**	**0.886±0.014**	**0.863±0.051**	**0.862±0.021**	**9.38%**
	RobustECD-SE(combination:1)	√	√	√	√					0.809±0.001	0.759±0.039	0.906±0.021	0.881±0.018	0.843±0.049	0.857±0.023	8.61%
	RobustECD-SE(combination:2)					√	√			0.809±0.004	0.750±0.036	0.898±0.027	0.873±0.031	0.841±0.049	0.839±0.030	7.64%
	RobustECD-SE(combination:3)							√	√	0.809±0.000	**0.768±0.035**	0.905±0.021	**0.887±0.019**	0.849±0.045	0.852±0.023	8.98%
	RobustECD-SE(combination:4)			√	√		√		√	0.809±0.001	**0.768±0.034**	**0.911±0.014**	0.883±0.020	0.857±0.049	**0.860±0.023**	**9.34%**
	RobustECD-SE(single)								√	0.809±0.000	0.767±0.048	0.906±0.021	0.877±0.027	**0.863±0.041**	**0.860±0.031**	9.20%
Polblogs (noise)	Original									0.330±0.001	0.378±0.000	0.318±0.000	0.376±0.000	0.375±0.053	0.458±0.067	8.60%
	Attack									0.303±0.001	0.348±0.000	0.299±0.000	0.336±0.000	0.340±0.061	0.434±0.034	—
	RobustECD-SE(all)	√	√	√	√	√	√	√	√	0.444±0.007	0.505±0.007	**0.558±0.005**	0.493±0.004	**0.500±0.005**	**0.469±0.023**	**46.69%**
	RobustECD-SE(combination:1)	√	√	√	√					0.444±0.010	**0.513±0.007**	0.539±0.009	**0.495±0.007**	0.499±0.005	**0.470±0.003**	46.10%
	RobustECD-SE(combination:2)					√	√			0.426±0.009	0.488±0.007	0.541±0.011	0.479±0.007	0.492±0.006	0.439±0.010	41.70%
	RobustECD-SE(combination:3)							√	√	**0.453±0.009**	0.499±0.007	**0.545±0.009**	0.487±0.007	0.496±0.006	0.450±0.009	44.95%
	RobustECD-SE(combination:4)			√	√		√		√	**0.458±0.009**	**0.511±0.007**	0.536±0.009	**0.497±0.007**	**0.501±0.006**	0.453±0.006	**46.15%**
	RobustECD-SE(single)						√			0.419±0.009	0.484±0.007	0.529±0.032	0.478±0.008	0.488±0.010	0.431±0.009	39.90%

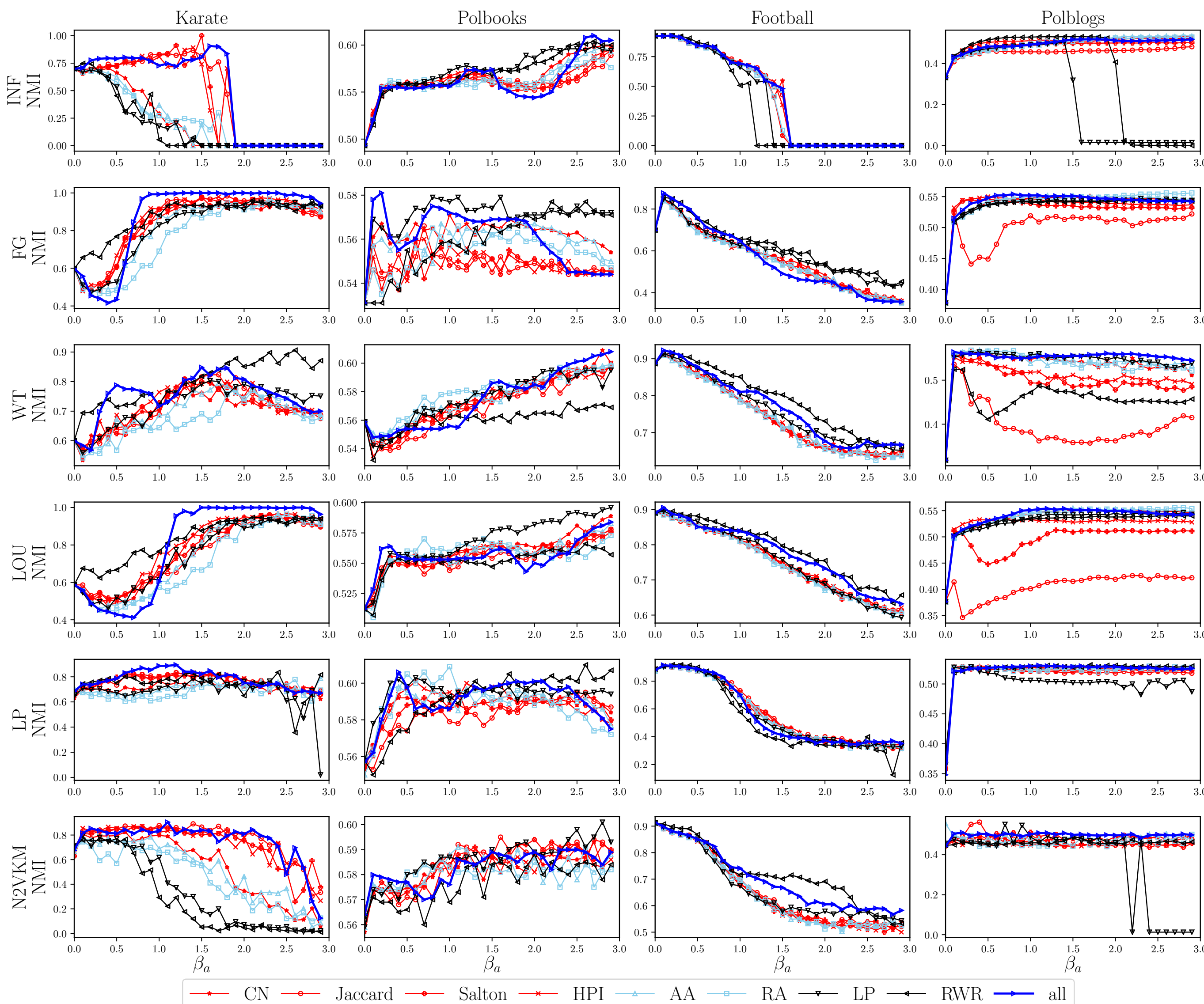

Fig. 5. The impact of single similarity in *RobustECD-SE* for real networks.

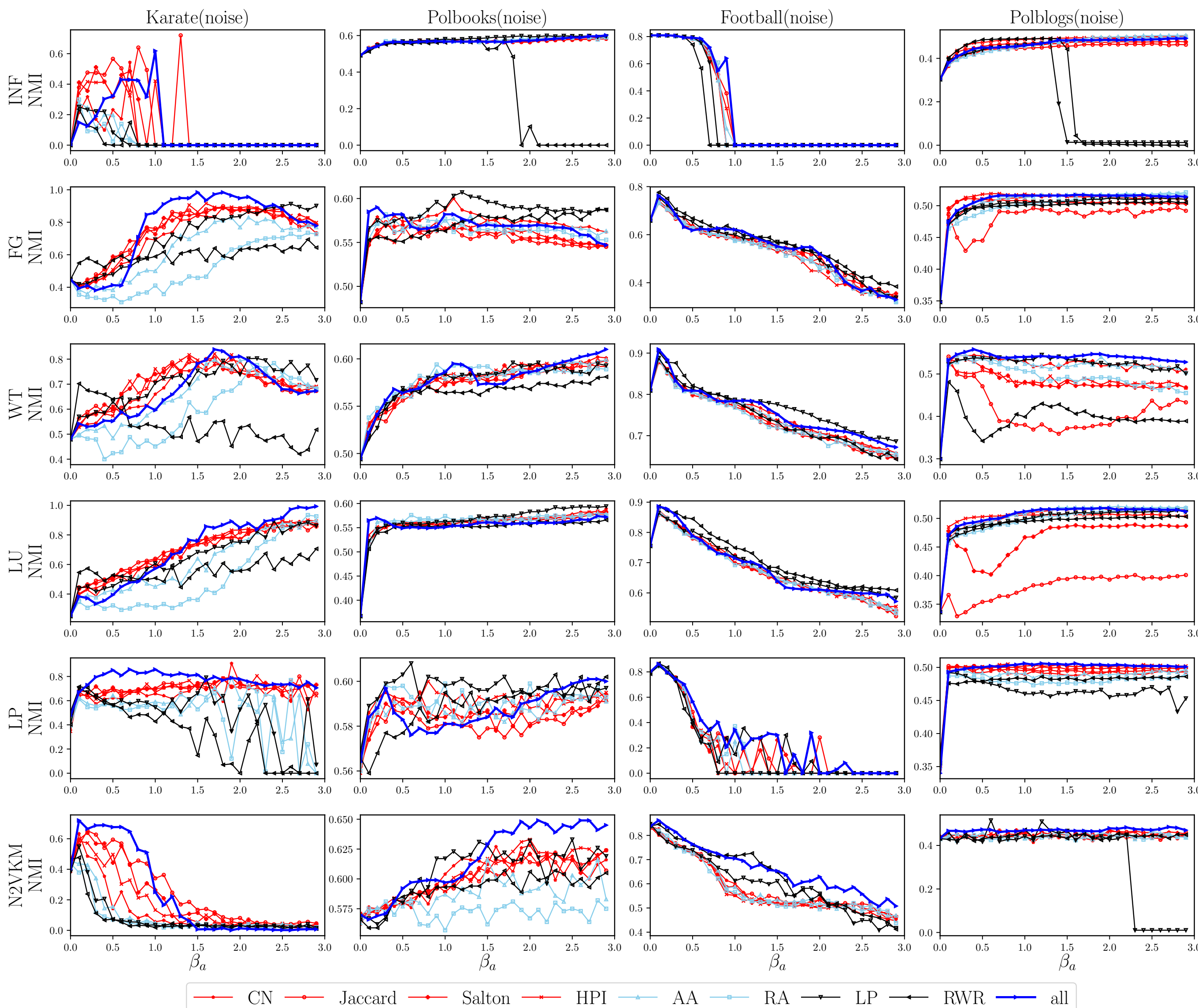

Fig. 6. The impact of single similarity in *RobustECD-SE* for adversarial networks.