Supplementary File

I. Parallelization of LCAAG

Due to the fact that $n_V \ll n_E$, LCAAG takes much more time than the algorithms proposed to detect communities of nodes according to the above complexity analysis. For the purpose of accelerating LCAAG, we implement it in a parallelization manner such that its efficiency can be boosted by the increase in the number of threads. In particular, as indicated by Eq. (19) in the main document, solving the derivative of $\hat{\boldsymbol{\alpha}}_{ij}$ is independent for different links in E. In other words, the update of $\hat{\boldsymbol{\alpha}}$ allows for parallelization. Regarding $\hat{\boldsymbol{\sigma}}$, $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\lambda}}$, it is also possible to update them simultaneously with different threads, as their updates are only related to $\hat{\boldsymbol{\alpha}}$. In doing so, the parallelization of LCAAG can be achieved.

II. Evaluation Metrics

NMI is an information-theoretic measure to indicate the matching degree between identified clusters and ground truth. Assuming that $\boldsymbol{Z} = \{z_k\}$ $(1 \leq k \leq n_C)$ is a known set of ground truth clusters, n_{c_k} is the number of links grouped in the k-th community as indicated by by $\hat{\boldsymbol{\alpha}}$, n_{z_k} is the number of links in z_k, and n_{c_k,z_l} is the number of common links in c_k and z_l. NMI is defined as:

$$NMI = \frac{\sum_{k=1}^{n_C}\sum_{l=1}^{n_C} n_{c_k,z_l}\ln(\frac{n_E n_{c_k,z_l}}{n_{c_k} n_{z_l}})}{\sqrt{(\sum_{k=1}^{n_C} n_{c_k}\ln\frac{n_{c_k}}{n_E})(\sum_{l=1}^{n_C} n_{z_l}\ln\frac{n_{z_l}}{n_E})}} \tag{1}$$

As for Accuracy, assuming that a mapping function $f : c_k \mapsto z_l$ is to indicate that z_l is the corresponding ground truth of c_k. Then, the Accuracy metric is defined as:

$$Accracy = \frac{\sum_{k=1}^{n_C} n_{c_k,f(c_k)}}{\sum_{k=1}^{n_C} n_{c_k}} \tag{2}$$

Since even for the most popular clustering similarity measures, their vulnerability to critical biases are still existed. Hence, we additionally adopt JI to indicate the detection accuracy more objectively. JI is an evaluation metric by comparing the similarity and difference between the communities identified and corresponding ground truth. Its definition is given as follows.

$$JI(\mathbf{C},\mathbf{Z}) = \frac{\sum_{k=1}^{n_C}\sum_{l=1}^{n_C} n_{c_k,z_l}}{\sum_{k=1}^{n_C}\sum_{l=1}^{n_C}(n_{c_k} + n_{z_l} - n_{c_k,z_l})} \tag{3}$$

According to the definitions of NMI, Accuracy and JI, we note that the better detected communities match with the ground truth, the larger the scores of NMI, Accuracy and JI are. If all detected communities match perfectly with those of ground truth, the NMI, Accuracy and JI scores will be the value of 1, which is the maximum value they can take.

TABLE I
Parameter Settings of All Comparing Algorithms

Algorithm	Parameter Setting
LCAAG	ϵ=0.0001, l_{max}=100
FSPGA	bias=0, λ=1
MISAGA	bias=0.5, λ=1
DHCD	α=1, β=0.1, maximum iterations=1000
SSB	α=260, β=1, convergence=0.001
niMM	maximum iterations=100
GBAGC	maximum iterations=100
CESNA	sa=0.05, sb=0.3
MARINE	Q=5, maximum iterations=100, d=128, alpha=1
FCM	maximum iterations=1000, convergence=0.005, n components=0.7

III. Parameter Setting

To conduct a fair comparison, we have tuned the performances of all comparing algorithm so as to ensure that they are compared at their best performances. More specifically, we explicitly adopt the parameter settings recommended by their original work in the experiments. But for algorithms without such a recommendation, their performances are further tuned by varying parameter values to respond best for the tasks. The parameter setting of each algorithm is given in TABLE I. Regarding LCAAG, its parameter setting is only involved in generating the skeleton of G, as we need to sample the prior distributions of C, $\mathbf{D}$ and $\mathbf{A}$ from their Dirichlet distributions, which are determined by the hyperparameters $\boldsymbol{\sigma}$, $\boldsymbol{\mu}$ and $\boldsymbol{\lambda}$. In the experiments, we simply set the values of all these hyperparameters to 1 by considering them as non-informative priors [1].

IV. Significance analysis

To further indicate the statistical significance of experimental results in performance comparison, we have conducted the Friedman test [2]. As an alternative for repeated measures analysis of variances, the Friedman test is a non-parametric test for testing the difference between several related samples. Assuming that the null hypothesis is that there is no difference among the performances of all comparing algorithms, such a null hypothesis is rejected if the test statistic, denoted as χ_F^2, is larger than a critical value χ_α^2 with the critical level α, and we could conclude that the null hypothesis is rejected with the confidence $(1 - \alpha)$. The definition of χ_F^2 is given as:

$$\chi_F^2 = \frac{12}{nm(m+1)}\sum_{j=1}^{m} R_j^2 - 3n(m+1) \tag{4}$$

where n is the number of comparing algorithms, m is the number of evaluation groups, and R_j is the Friedman rank of the j-th comparing algorithm. In our work, we have $n = 9$ and $m = 15$ according to Table III in the main document. By substituting the ranks into (4), we have $\chi_F^2 = 69.76$. Since

TABLE II
EXPERIMENTAL RESULTS OF WILCOXON SIGNED-RANK TEST

Pair of Comparing Algorithms	R_{pos}	R_{neg}	p**-value**
LCAAG vs. FSPGA	91	0	0.00074
LCAAG vs. MISAGA	107	13	0.00379
LCAAG vs. DHCD	116	4	0.00074
LCAAG vs. SSB	115.5	4.5	0.00082
LCAAG vs. niMM	120	0	0.00032
LCAAG vs. GBAGC	120	0	0.00032
LCAAG vs. CESNA	105	0	0.00048
LCAAG vs. MARINE	120	0	0.00032
LCAAG vs. FCM	120	0	0.00032

our degree of freedom is $n - 1$, the corresponding critical value in the chi-square table is 15.507 with the confidence level at 95%. Given the fact that $\chi_F^2 \gg 15.507$, the null hypothesis is rejected with the confidence level at 95%. Hence, the performances of all comparing algorithms are significantly different.

In addition to the Friedman test, we have also performed the one-tailed Wilcoxon signed-rank test [2] to confirm whether the performance of LCAAG is significantly better than the other comparing algorithms. There are three indicators, i.e., R_{pos}, R_{neg} and p-value, calculated by the Wilcoxon signed-rank test. In particular, R_{pos} and R_{neg} are the sums of positive and negative ranks respectively, and they would be similar if the null hypothesis is true. The results of the Wilcoxon signed-rank test are presented in Table II with a confidence level at 95%. We note that the p-values are always less than 0.05 when we compare LCAAG with the other algorithms. For the AG clustering task, we believe that LCAAG achieves significantly better accuracy than state-of-the-art comparing algorithms with a confidence level at 95%.

REFERENCES

[1] A. E. Gelfand and D. K. Dey, "Bayesian model choice: asymptotics and exact calculations," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 56, no. 3, pp. 501–514, 1994.

[2] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *The Journal of Machine learning research*, vol. 7, pp. 1–30, 2006.