NetSciX2026: International School and Conference on Network Science
February 17^{th}-20^{th}, 2026 - Auckland, New Zealand

Beyond NMI: Reliable Community Detection Evaluation Using Kappa Index and F-Score

Keywords: Normalized mutual information; Kappa index; F-score; Evaluation; Community structures detection

Extended Abstract

Problem Statement. Community structures are critical towards understanding not only the network topology but also how the network functions[1]. However, how to evaluate the quality of detected community structures is still challenging and remains unsolved. The most widely used metric, normalized mutual information (NMI,[2]), exhibits finite size effect—producing non-zero values for random independent partitions. Its variants rNMI[3] and cNMI[4] introduce reverse finite size effects and violate proportionality assumptions. In addition, NMI-type metrics have the problem of ignoring importance of small communities. These systematic biases lead to contradictory method rankings and compromise algorithm development. Additionally, NMI-type metrics cannot evaluate individual communities of interest, limiting applications requiring targeted analysis (e.g., identifying terrorist networks or biological pathways).

Methodology. We transform unsupervised evaluation into supervised classification through optimal label mapping using integer linear programming. Decision variables $x_{ij} \in \{0,1\}$ indicate mapping of computed community i to ground-truth j, with cost function $\ell_{ij} = |B_i \cup A_j| - |B_i \cap A_j|$. The assignment problem is solved via Hungarian algorithm. After optimal mapping, we apply kappa index for overall performance ($\kappa = \frac{\text{accuracy} - \text{expected accuracy}}{1 - \text{expected accuracy}}$) and F-score for individual community assessment.

Experimental Results. (1) Metric reliability validated: Kappa exhibits correct theoretical behavior while NMI shows systematic deviations, and F-score reveals method-specific strengths for different community types invisible to aggregate metrics. (2) InfoMap, Louvain, and ModBP are evaluated on LFR benchmarks and real networks, demonstrating systematic ranking reversals: InfoMap ranked 1st under NMI but 3rd under kappa; (2) Bias mechanism revealed: InfoMap generates 3-4 times more communities than ground truth, artificially inflating NMI through finite size effects; (3) These results demonstrate that NMI-based method rankings may be fundamentally incorrect, with serious implications for community detection research.

Conclusions and Implications. We identified fundamental flaws in NMI-type metrics and demonstrated that kappa index and F-score enable reliable, bias-free evaluation of community detection performance. Method rankings significantly differ under reliable evaluation, suggesting previous comparative studies may have reached incorrect conclusions about algorithm effectiveness. Our approach provides both reliable overall assessment and individual community evaluation capability. The computational efficiency makes it practical for large-scale networks. These findings have major implications for community detection research, potentially affecting algorithm selection and development directions. We recommend adopting kappa index and F-score to ensure robust scientific progress in the field.

References

[1] Santo Fortunato. "Community detection in graphs". In: *Physics reports* 486.3-5 (2010), pp. 75–174.

[2] Alexander Strehl and Joydeep Ghosh. "Cluster ensembles—a knowledge reuse framework for combining multiple partitions". In: *Journal of machine learning research* 3.Dec (2002), pp. 583–617.

[3] Pan Zhang. "Evaluating accuracy of community detection using the relative normalized mutual information". In: *Journal of Statistical Mechanics: Theory and Experiment* 2015.11 (2015), P11006.

[4] Darong Lai and Christine Nardini. "A corrected normalized mutual information for performance evaluation of community detection". In: *Journal of Statistical Mechanics: Theory and Experiment* 2016.9 (2016), p. 093403.

Table 1: **Ranking Reversals.** Method rankings under NMI vs. Kappa evaluation on real networks.

Network	Method	NMI	Kappa
Email-Eu-core	InfoMap	1st	2rd
	Louvain	2nd	1st
	ModBP	3rd	3nd
CiteSeer	InfoMap	1st	3rd
	Louvain	2nd	1st
	ModBP	3rd	2nd

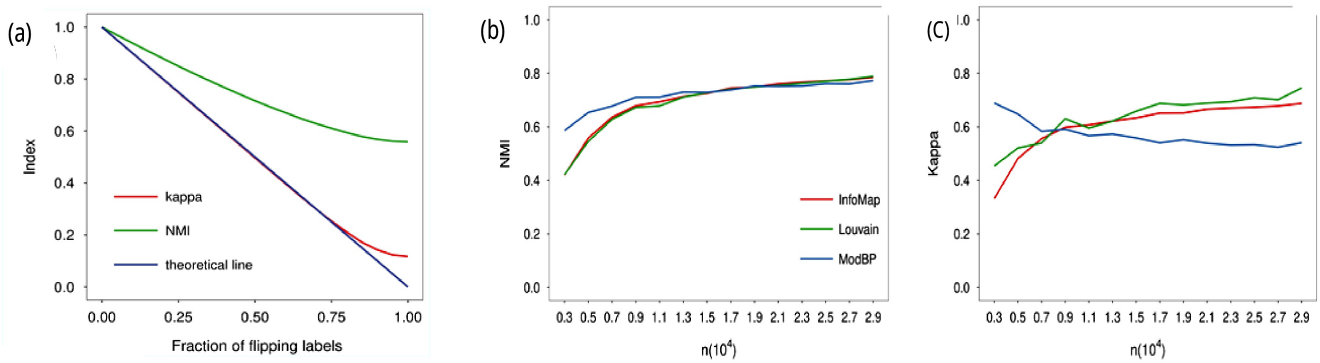

Figure 1: **Evaluation Metric Bias and Its Impact.** (a) Kappa index vs. NMI behavior in label flipping experiments, showing kappa's theoretically correct linear decline compared to NMI's systematic deviation. (b-c) Method ranking reversals on LFR networks: InfoMap drops from 1st (NMI) to 2nd (Kappa) while Louvain rises to 1st, demonstrating how metric choice fundamentally affects algorithm assessment.