Improving Community Detection via Community Association Strength Sores

WAW 2025

Jordan Barrett*, **Ryan DeWolfe***, Bogumił Kamiński†, Paweł Prałat*, Aaron Smith‡, and François Théberge¶.

*Toronto Metropolitan University
†SGH Warsaw School of Economics
‡University of Ottawa
¶Tutte Institute for Mathematics and Computing

July 2025

Motivation

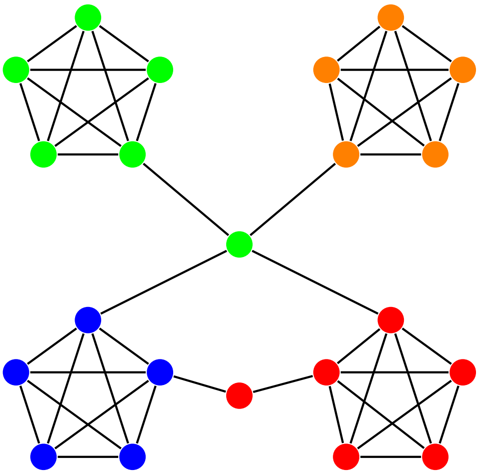


Figure: An example of a partition.

Agenda

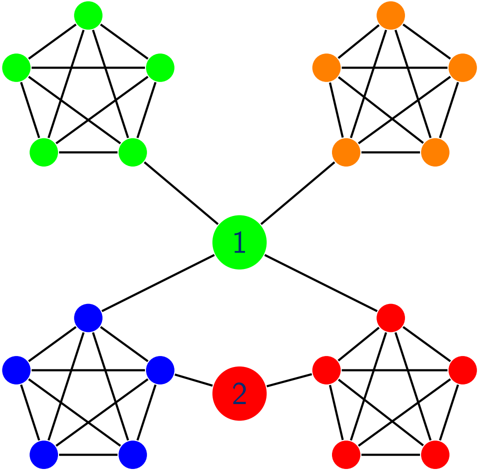
1. CAS Scores
2. Properties of CAS Scores
3. Applications:
   3.1 Improving Partitons
   3.2 Detecting Outliers
   3.3 Overlapping Communities

The Scores

Proposed Scores

Internal Edge Fraction:

$$\mathrm{IEF}(v, C) := \frac{\deg_C(v)}{\deg(v)}.$$



$$IEF(1, \text{"}GREEN\text{"}) = \frac{1}{4}$$

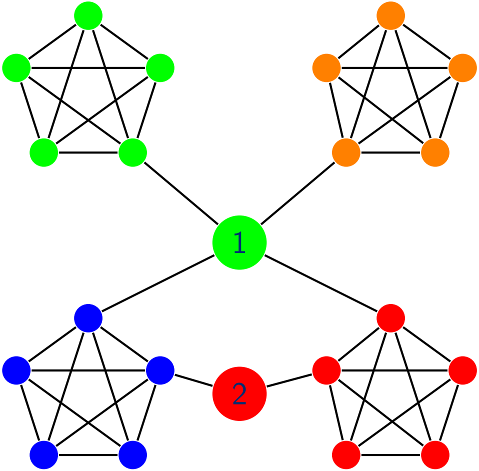
$$IEF(2, \text{"}RED\text{"}) = \frac{1}{2}$$

$$IEF(2, \text{"}BLUE\text{"}) = \frac{1}{2}$$

Proposed Scores

Normalized Internal Edge Fraction:

$$\mathrm{NIEF}(v, C) := \max \left\{ \mathrm{IEF}(v, C) - \frac{vol(C)}{volV}, 0 \right\}.$$



$NIEF(1, \text{"}GREEN\text{"}) = 0$

$NIEF(2, \text{"}RED\text{"}) = 0.24$

$NIEF(2, \text{"}BLUE\text{"}) = 0.26$

Proposed scores

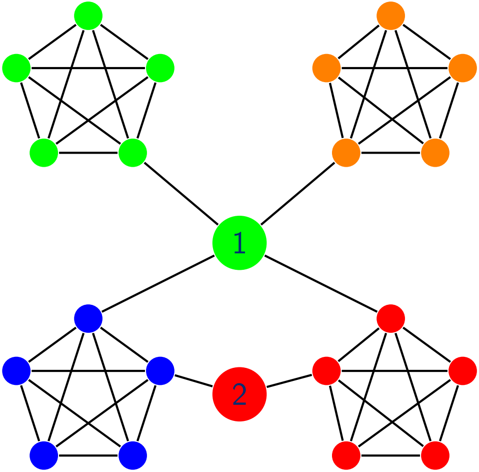
The motivation for P.

- The probability of an edge from v into C in a resampling of G is $w(C)$.
- There are $deg(v)$ edges from v.
- $1 - P(v, C)$ is the probability that at least $deg_C(v)$ edges are into C after a resampling.
- Let $F(\cdot; n, p)$ be the CDF of the binomial distribution with parameters n and p.
- $P(v, c) = F\left(\deg_C(v) - 1; \deg(v), \frac{vol(C)}{vol(V)}\right)$

Proposed Scores

P:

$$\mathrm{P}(v, C) := F\left(\deg_C(v) - 1; \deg(v), \frac{vol(C)}{vol(V)}\right).$$



$P(1, \text{"}GREEN\text{"}) = 0.28$

$P(2, \text{"}RED\text{"}) = 0.55$

$P(2, \text{"}BLUE\text{"}) = 0.58$

Properties

Properties of CAS

1. All scores are 0 if there are no edges into a community.
2. For a fixed $vol(C)$, all scores are monotone increasing with deg_C.

Both of these properties are intuitive. A further research direction is finding a larger set of intuitive properties that could narrow the set of acceptable CAS scores.

Properties of CAS

Consider the following graph transformations:

1. Add a new community this is disjoint to the original graph.
2. Create a copy of each edge. (Since the CAS scores only depend on $deg_C(v)$ and $vol(C)$, this can be replaced with doubling each of these values by creating new edges.)

Transformation	IEF	NIEF	P
1	Unchanged	Increases	Increases
2	Unchanged	Unchanged	"More Extreme"

Applications

Data

We use **ABCD** (Kamiński et al., 2021), **ABCD** + **o** (Kamiński et al., 2023), and **ABCD** + **o**$^{\mathbf{2}}$ (Barrett et al., 2025) synthetic graphs for evaluation.

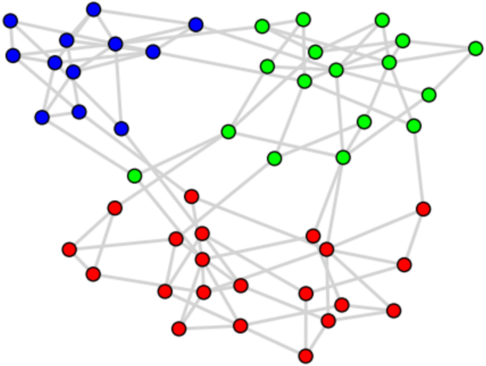


Figure: Example ABCD graph with $n = 50, \xi = 0.2$.

Data

We also consider the Football graph (Girvan and Newman, 2002) as an example of real data with known outliers.

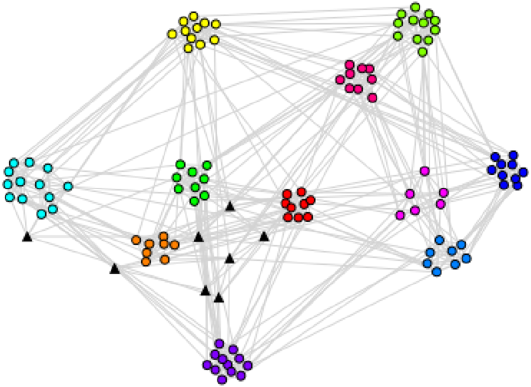


Figure: The football graph colored by conference.

Improving Partitions

Improved ECG

The very successful ECG (Poulin and Théberge, 2019) community detection algorithm works as follows:

1. Perform 1-iteration of the Louvain algorithm k times to get k partitions.
2. Weight each edge uv as the average of the indicator $\chi(u$ and v are in the same community) from step 1.
3. Run Louvain or Leiden on this weighted graph.

Improved ECG

We rewrite step 2 using CAS scores.

1. Perform 1-iteration of the Louvain algorithm k times to get k partitions.
2. Weight each edge uv as the average of $CAS(uv)$ from step 1.
3. Run Louvain or Leiden on this weighted graph.

With the option of any CAS score and two options to symmetrise:

$$CAS_{or}(uv) = CAS(u, C_v) + CAS(v, C_u) - CAS(u, C_v) \cdot CAS(v, C_u)$$

$$CAS_{and}(uv) = CAS(u, C_v) \cdot CAS(v, C_u)$$

Improved ECG

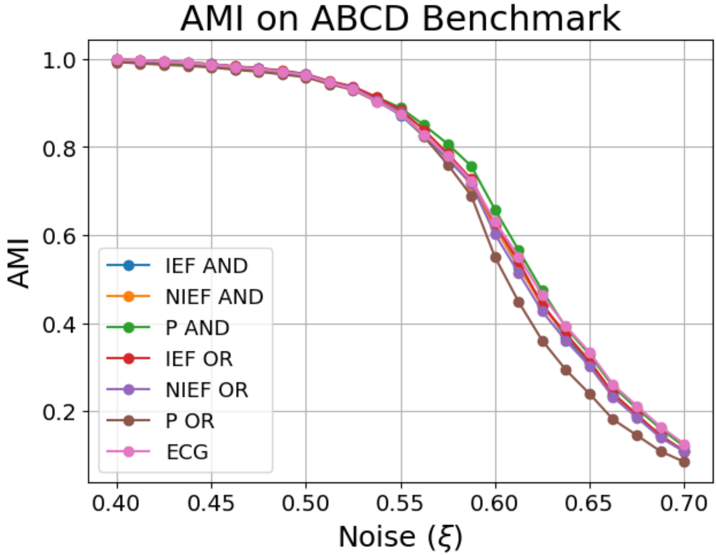


Figure: AMI of the proposed CAS-ECG methods.

Improved ECG

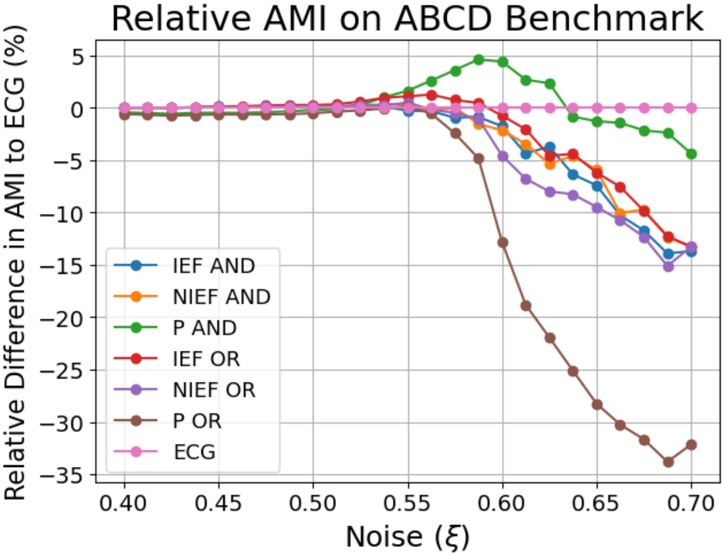


Figure: AMI of the proposed CAS-ECG methods compared to ECG.

Detecting Outliers

Detecting Outliers

- Suppose some nodes are not strongly associated to any community (outliers).
- We test if the maximum CAS to any community can predict if a node is an outlier.

$$outlier(v) = \max_{C \in \mathcal{C}} CAS(v, C)$$

Detecting Outliers

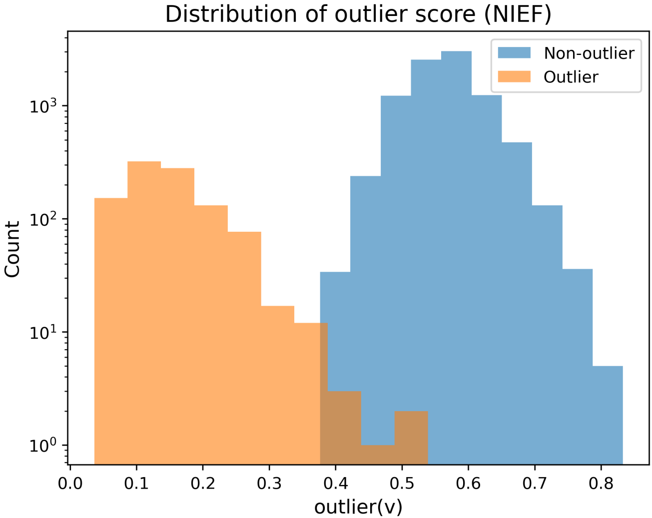


Figure: Histogram of *outlier* scores on an ABCD+o graph with $xi = 0.4$.

Detecting Outliers

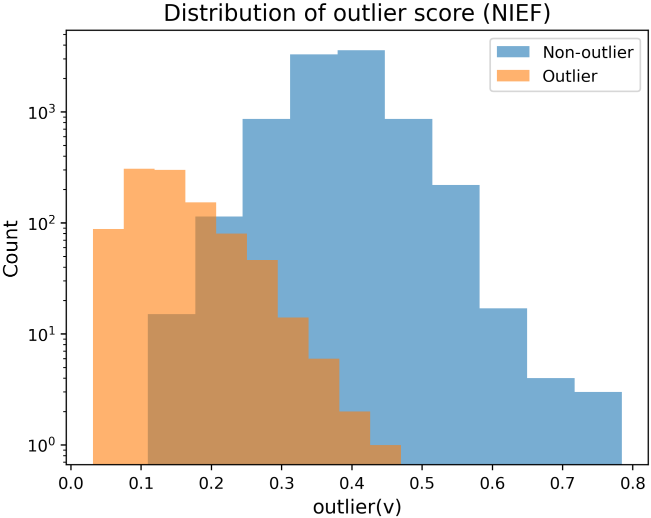


Figure: Histogram of *outlier* scores on an ABCD+o graph with $xi = 0.6$.

Detecting Outliers

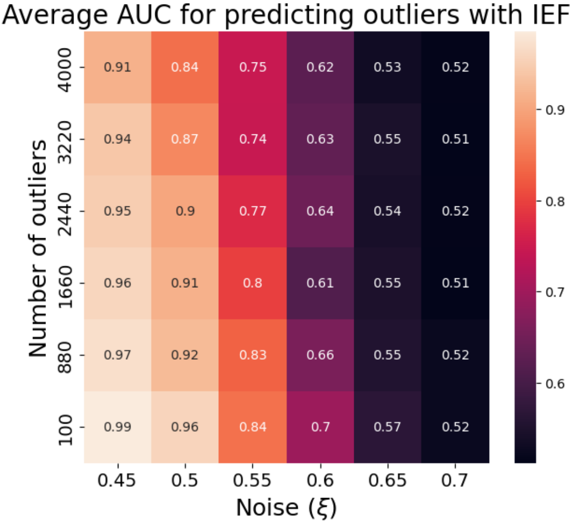
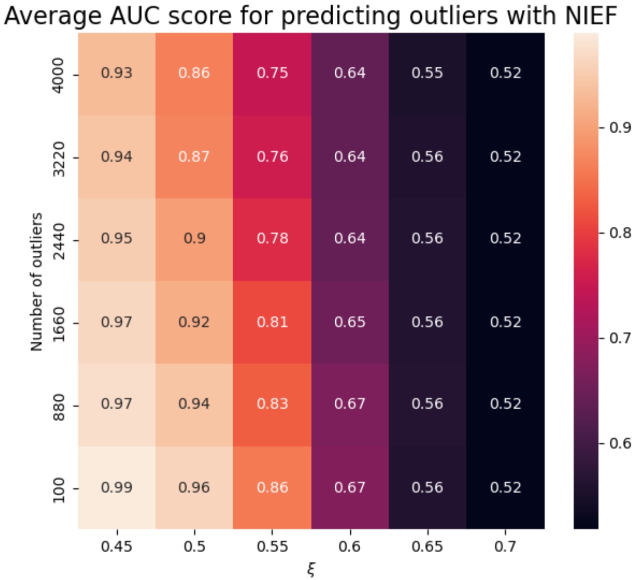
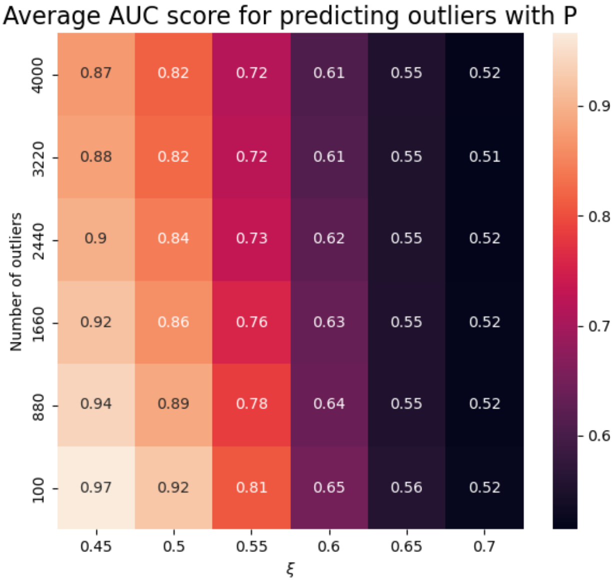


Figure: Classifing ABCD+o outliers with CAS.

Detecting Outliers



Detecting Outliers



Detecting Outliers

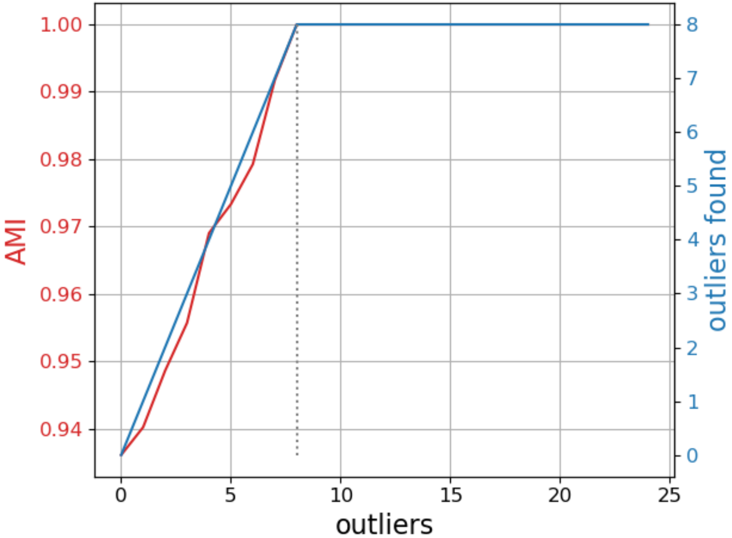


Figure: Classifying Football outliers with P.

Overlapping Communities

Overlapping Communities

1. We start with some set of (possibly overlapping communities) $\mathcal{C} = \{C_1, C_2, \dots C_k\}$.
2. Construct a new collection of communities $\mathcal{C}'$ where $C_i' = \{v : CAS(v, C_i \geq \tau)\}$.

We use ego-split (Epasto et al., 2017) to find the initial communities, and we find $\tau \in [0.075, 0.25]$ improves the communities when compared to the true labels with oNMI (McDaid et al., 2013).

Overlapping Communities

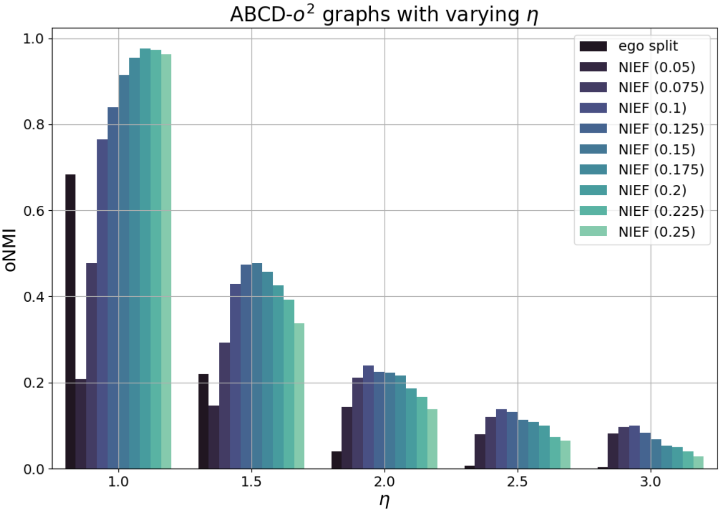


Figure: Using NIEF to post-process Ego-split. ($\xi = 0.35$)

Overlapping Communities

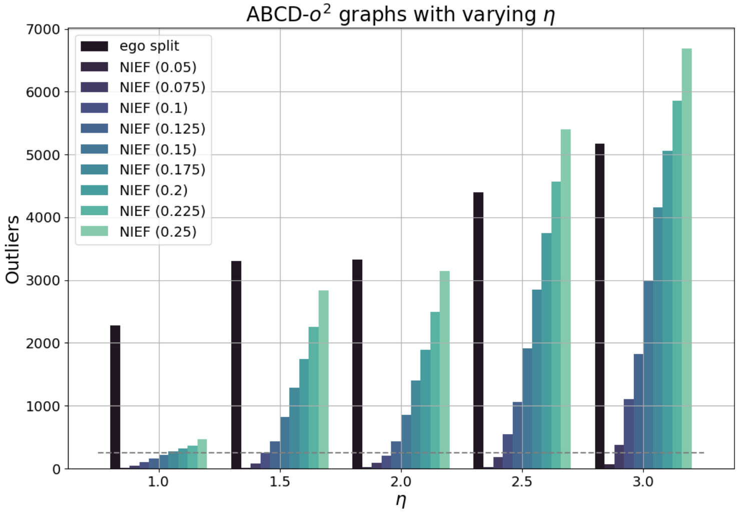


Figure: Using NIEF to post-process Ego-split. ($\xi = 0.35$)

Overlapping Communities

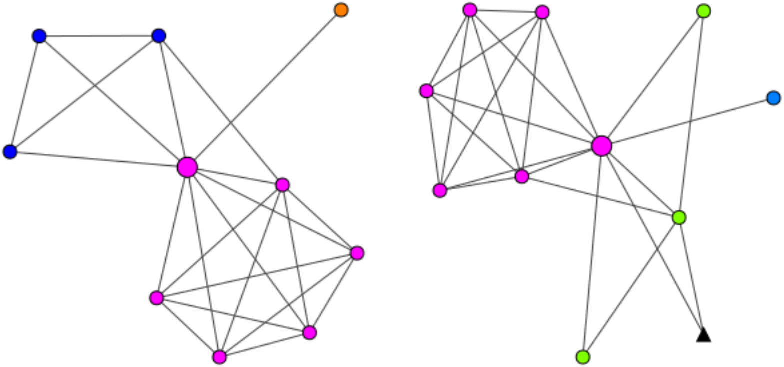


Figure: Ego-nets of nodes with potential overlap using CAS-ECG and P from the football graph.

Bonus Application: Layouts

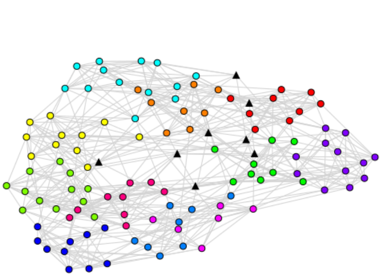
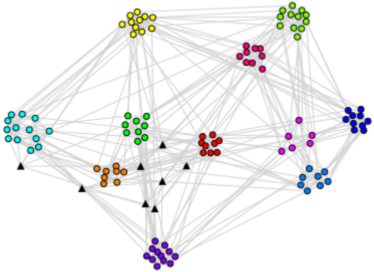



Figure: Force-direct layout of the football graph with (right) and without (left) edge weighting using CAS-ECG.

Summary

- There are several options for community association strength scores.
- They are useful for improving a variety of community detection tasks.
- Post-processing techniques appear to be a viable approach to outlier detection and overlapping communities.

References I

J. Barrett, R. DeWolfe, B. Kamiński, P. Prałat, A. Smith, and F. Théberge. The artificial benchmark for community detection with outliers and overlapping communities (**ABCD** + $\mathbf{o}^2$). In *Modelling and Mining Networks*, pages 125–140. Springer, Cham, 2025. doi: 10.1007/978-3-031-92898-7_9.

A. Epasto, S. Lattanzi, and R. Paes Leme. Ego-splitting framework: from non-overlapping to overlapping clusters. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, pages 145–154, New York, NY, USA, 2017. Association for Computing Machinery. doi: 10.1145/3097983.3098054.

M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826, 2002. doi: 10.1073/pnas.122653799.

References II

B. Kamiński, P. Prałat, and F. Théberge. Artificial benchmark for community detection (abcd)—fast random graph model with community structure. *Network Science*, 9(2):153–178, 2021. doi: 10.1017/nws.2020.45.

B. Kamiński, P. Prałat, and F. Théberge. Artificial benchmark for community detection with outliers (abcd+o). *Applied Network Science*, 8(1):25, 2023. doi: 10.1007/s41109-023-00552-9.

A. F. McDaid, D. Greene, and N. Hurley. Normalized mutual information to evaluate overlapping community finding algorithms, 2013. URL `https://arxiv.org/abs/1110.2515`.

V. Poulin and F. Théberge. Ensemble clustering for graphs. In *Complex Networks and Their Applications VII*, pages 231–243, Cham, 2019. Springer International Publishing.