Cycle-Aware Hypothesis Testing for Community Structure via Renewal Non-Backtracking Random Walks

Behnaz Moradi-Jamei [1] **Heman Shakeri** [2] **Michael Higgins** [3]

Abstract

We study goodness-of-fit testing for community structure in networks. Classical spectral tests for the Erdős–Rényi null rely on adjacency information alone and can lose power in sparse or weak-signal regimes. We propose RNBRW-GOF, a cycle-aware spectral test based on renewal non-backtracking random walk retracing probabilities. The method replaces the adjacency matrix by a centered and normalized matrix of RNBRW edge weights and rejects for large values of its largest eigenvalue. On the theory side, we prove exact null centering under edge exchangeability, crude second-moment control, and finite-walk concentration for empirical RNBRW weights. We also establish sparse-regime local dependence decay for bounded truncated scores and derive truncation and normalization results that support the use of RNBRW-based spectral calibration. These results do not yield a full universality theorem for the final matrix statistic; accordingly, inference is theory-motivated and simulation-calibrated via a Bartlett-type correction together with a Tracy–Widom type 1 reference law. Across simulated Erdős–Rényi nulls, the corrected statistic is reasonably calibrated in moderate and large graphs. Under assortative stochastic block alternatives, RNBRW-GOF improves over adjacency-based spectral testing in several sparse and weak-signal regimes and remains competitive with a non-backtracking spectral baseline, which is itself highly competitive in intermediate regimes. Overall, the results position RNBRW-GOF as a practically calibrated, cycle-aware goodness-of-fit test for assortative community structure.

[1]James Madison University, Harrisonburg, Virginia, USA [2]University of Virginia, Charlottesville, Virginia, USA [3]Department of Statistics, Kansas State University, Manhattan, Kansas, USA. Correspondence to: Behnaz Moradi-Jamei <moradibx@jmu.edu>.

Proceedings of the 43^{rd} International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

1. Introduction

Detecting and validating community structure is a central problem in network analysis. Across social, biological, and information networks, nodes that belong to the same latent group tend to interact more frequently with one another than with nodes outside the group. This motivates statistical procedures for deciding whether an observed graph is consistent with a homogeneous null model or instead exhibits evidence of community structure. In this paper, we study this question as a hypothesis testing problem rather than a clustering problem: given an observed graph $G = (V, E)$, we test $H_0 : G \sim G(n, p)$ versus $H_1 :$ G exhibits assortative community structure. A substantial literature studies goodness-of-fit testing for network structure under this formulation. In particular, spectral tests based on the adjacency matrix or its normalization have provided a tractable route to testing the Erdős–Rényi null against structured alternatives. For example, Lei et al. (2016) proposed a goodness-of-fit test based on the largest eigenvalue of a normalized adjacency matrix, while Bickel & Sarkar (2013) developed a test based on the largest eigenvalue of a centered and rescaled adjacency matrix and derived its asymptotic null distribution under the Erdős–Rényi model. Other approaches include exact Monte Carlo goodness-of-fit procedures for block models (Karwa et al., 2016), sequential testing frameworks for community extraction (Zhao et al., 2011), and eigenvalue-based criteria for multi-type random networks (Golub & Jackson, 2012). These methods are statistically principled, but they rely primarily on edge-density information and therefore may lose sensitivity when community structure is weak or the graph is sparse. A complementary line of work exploits *cyclic structure* in networks. If nodes within the same community are more densely and more coherently connected, then short cycles should occur more frequently within communities than across them. This idea has been used algorithmically to improve community detection through cycle-aware weighting and loop-based heuristics (De Meo et al., 2013; Lai et al., 2010; Klymko et al., 2014; Radicchi et al., 2004; Newman & Girvan, 2004; Zhang et al., 2008). Recent work by Moradi et al. (2018) introduced the *renewal non-backtracking random walk* (RNBRW), a scalable procedure that assigns larger weights to edges that are more likely to complete cycles. Empirically,

RNBRW weighting improves community detection performance, especially in sparse networks, by incorporating information beyond raw edge counts.

Our goal is to bring this cycle-sensitive information into the hypothesis testing framework. We develop a new goodness-of-fit test that replaces adjacency-only spectral information with a matrix built from RNBRW retracing probabilities. RNBRW is a non-backtracking random walk that terminates and restarts whenever it completes a cycle; the terminal edge of each run is recorded as a *retracing edge*. Aggregating these retracing events yields edge weights that quantify local cyclic structure. Under the null hypothesis of no community structure, these weights are approximately homogeneous, whereas under block-structured alternatives they become systematically distorted by within-community cycle enrichment. This makes RNBRW a natural basis for testing departures from the Erdős–Rényi model, though rejection should be interpreted only as cycle-sensitive evidence against a homogeneous null rather than as a uniform certificate of assortative community structure; degree heterogeneity may also induce rejection. We return to the graph-versus-Monte-Carlo distinction in Section 3.

Contributions. Our contributions are threefold. First, we introduce RNBRW-GOF, a cycle-aware largest-eigenvalue goodness-of-fit test for

$$H_0 : G \sim G(n, p),$$
$$H_1 : G \text{ exhibits assortative community structure.}$$

We also give a simple balanced-SBM calculation showing that, conditional on an edge being present, within-community and between-community edges can have different expected short-cycle completion counts even though their binary adjacency value is identical. Second, we establish partial null-side results that motivate the statistic: exact centering, crude second-moment control, finite-walk concentration, and sparse-regime local dependence and truncation results that clarify the roles of graph randomness and RNBRW Monte Carlo error. These results do not yield a Tracy–Widom universality theorem for the final matrix statistic, so the final calibration remains simulation-based. Third, we provide finite-sample evidence that Bartlett-corrected RNBRW-GOF attains usable Type I error control in moderate and large graphs, improves over adjacency-based spectral testing in several sparse and weak-signal assortative SBM regimes, and remains competitive with a non-backtracking comparator.

Taken together, these results suggest that cyclic structure provides a useful and statistically meaningful signal for network inference. By integrating RNBRW with a spectral goodness-of-fit framework, we obtain a test that is computationally feasible, sensitive to higher-order structure, and particularly effective when traditional edge-based methods are least informative. The remainder of the paper is organized as follows. Section 2 introduces the testing setup and baseline models. Section 3 reviews RNBRW and develops the null-side properties of the RNBRW edge weights. Section 4 defines the proposed RNBRW goodness-of-fit statistic. Section 5 presents finite-sample calibration, Type I error, and power results. Section 6 concludes with limitations and future directions.

Table 1. Status of the ingredients used by RNBRW-GOF.

Ingredient	Status here	Role in the procedure
Null mean of RNBRW weights	Proved	Centering of Y_{obs}
Second-moment control	Proved	Scaling and stability
Finite-walk concentration	Proved	Separates Monte Carlo error from graph randomness
Localized sparse dependence decay	Proved for truncated local scores	Supports the weak-dependence heuristic
Weak dependence of full oracle RNBRW matrix	Not proved	Motivates spectral calibration
Tracy–Widom law for final statistic	Simulation-calibrated	Final p-value computation

2. Problem setup

We study community structure assessment as a *hypothesis testing* problem rather than a graph partitioning problem. Given an observed undirected graph $G = (V, E)$ with $n = |V|$ nodes and adjacency matrix $A \in \{0, 1\}^{n\times n}$, our goal is to determine whether the network is adequately described by a homogeneous random graph model or instead exhibits evidence of nontrivial community structure. Formally, we consider the hypotheses $H_0 : G \sim G(n, p)$ versus $H_1 :$ G exhibits assortative community structure. Under this formulation, the question is not to estimate node memberships directly, but to test whether the observed graph departs significantly from a one-community null model.

2.1. Null model: Erdős–Rényi graph

Under the null hypothesis, the network is assumed to follow an Erdős–Rényi model $G(n, p)$, in which each possible edge appears independently with common probability p. Thus, for distinct nodes $i, j \in V$, $\mathbb{P}(A_{ij} = 1) = p$, and all edges are exchangeable under H_0. This corresponds to a homogeneous network with no latent community structure, so any systematic deviation from this edge-homogeneous behavior should provide evidence against the null.

A classical spectral benchmark for testing this null is the procedure of Bickel & Sarkar (2013). Let $A^\star_{ij} = \frac{A_{ij}-\hat{p}}{\sqrt{(n-1)\hat{p}(1-\hat{p})}}$, $i \neq j$, where $\hat{p}$ is the empirical edge density of the observed graph. Bickel and Sarkar showed

that, under the Erdős–Rényi null, the largest eigenvalue of $A^\star$ admits Tracy–Widom calibration asymptotically, leading to a goodness-of-fit test based on $\theta_{\mathrm{BS}} = n^{2/3}\big(\lambda_{\max}(A^\star) - 2\big)$. Our proposed method follows this spectral testing philosophy, but replaces adjacency-only information by cycle-sensitive RNBRW edge weights.

2.2. Alternative model: stochastic block structure

Under the alternative, the graph is assumed to have latent communities. A canonical model for this setting is the stochastic block model (SBM), in which each node i is assigned a latent block label $g_i \in \{1, \dots, k\}$ and edge probabilities depend on the block memberships of the endpoints: $\mathbb{P}(A_{ij} = 1 \mid g_i, g_j) = \omega_{g_i g_j}$. The matrix $\Omega = (\omega_{ab})_{a,b=1}^{k}$ governs the connectivity pattern between and within communities. In the assortative case relevant for community detection, within-community probabilities exceed between-community probabilities, so that $\omega_{aa} > \omega_{ab}, \qquad a \neq b.$ Such alternatives induce not only edge-density heterogeneity but also cycle-level heterogeneity. To see the local mechanism, consider a balanced two-block SBM with within- and between-community probabilities $p_{\mathrm{in}} > p_{\mathrm{out}}$, and let $C_{ij} = \sum_{k \neq i,j} A_{ik}A_{jk}$ be the common-neighbor count. Conditional on $A_{ij} = 1$, a direct calculation gives

$$\Delta_{\triangle} := \mathbb{E}[C_{ij} \mid i, j \text{ same}, A_{ij} = 1] - \mathbb{E}[C_{ij} \mid i, j \text{ different}, A_{ij} = 1] = (p_{\mathrm{in}} - p_{\mathrm{out}})^2 = 2p_{\mathrm{in}}(p_{\mathrm{in}} - p_{\mathrm{out}}). \tag{1}$$

Under $p_{\mathrm{in}} = a/n$ and $p_{\mathrm{out}} = b/n$, this becomes $\Delta_{\triangle} = (a-b)^2/(2n) + O(n^{-2})$. Thus, among realized edges with the same adjacency value $A_{ij} = 1$, short-cycle statistics can still differ between within-community and between-community edges. RNBRW-GOF exploits this distinction through retracing probabilities, and our experiments focus on assortative SBM-type alternatives.

3. RNBRW and null-side properties

We now summarize the renewal non-backtracking random walk (RNBRW) construction and the null-side properties that motivate our test statistic. RNBRW is a cycle-sensitive variant of a non-backtracking random walk: it evolves as a non-backtracking walk until it first returns to a previously visited node, at which point the walk terminates, records the final edge as a *retracing edge*, and restarts from a new randomly selected directed edge. The resulting retracing frequencies quantify local cyclic structure in the graph.

More precisely, each RNBRW run proceeds as follows:

1. Choose a directed edge $\overrightarrow{v_0 v_1}$ uniformly at random from the directed edge set $\vec{E}$.
2. Initialize the walk as $w = [v_0, v_1]$.
3. For $k = 1, 2, \dots,$ move from v_k to a neighboring node $v_{k+1} \neq v_{k-1}$.
4. If v_{k+1} is already present in w, terminate the walk and record $\overrightarrow{v_k v_{k+1}}$ as the retracing edge; otherwise append v_{k+1} to w and continue.

Dead-end handling. If the walker is at node v_k and there is no admissible non-backtracking move, i.e., every neighbor of v_k is equal to v_{k-1} or no such neighbor exists, then the current RNBRW run is terminated without recording a retracing edge. A new run is then restarted from a uniformly selected directed edge. This convention is particularly relevant in sparse graphs and ensures that every RNBRW run is well defined. Intuitively, edges that are more likely to terminate an RNBRW are more strongly associated with short cycles. Since within-community edges are typically more enriched in cycles than between-community edges, RNBRW retracing probabilities provide a natural cycle-aware summary for testing community structure.

3.1. RNBRW edge weights

Let R_{ij} denote the number of successful RNBRW runs whose retracing edge is e_{ij}, and let $N_{\mathrm{ret}} = \sum_{e_{uv} \in E} R_{uv}$ be the total number of successful retracing events. We define $Y_{ij} = \frac{R_{ij}}{N_{\mathrm{ret}}}$, so that the RNBRW edge weights form a probability distribution over realized edges. If $N_{\mathrm{ret}} = 0$, then no RNBRW retracing event is observed and the normalized edge-weight matrix is undefined; in that case, we treat the graph as a conservative non-rejection and do not compute the RNBRW-GOF statistic.

Number of RNBRW runs. Throughout the experiments, the RNBRW weights are estimated from N independent RNBRW runs. In all simulations reported here, we set $N = c_{\mathrm{walk}}\, m, \qquad c_{\mathrm{walk}} = 20,$ where $m = |E|$. This choice scales the RNBRW sampling effort with graph size and keeps the estimator variance reasonably stable across (n, p) settings. In each graph replicate, RNBRW edge weights were normalized by the number of successful retracing events, consistent with the original RNBRW construction. From an implementation standpoint, this choice makes the RNBRW sampling budget scale linearly with the number of observed edges through $N = c_{\mathrm{walk}} m$, so the additional computational cost of the method is controlled by the walk multiplier c_{walk}.

3.2. Finite-walk estimation error

For a fixed graph G, let $\pi_{ij}(G) = \mathbb{P}(\text{a successful RNBRW run on } G \text{ retraces } e_{ij} \mid G)$ denote the graph-dependent conditional retracing probability of edge e_{ij}. Given N_{ret} successful retracing events, the retracing counts are multinomial with cell probabilities $\{\pi_{ij}(G)\}$, so $Y_{ij} = \pi_{ij}(G) + O_{\mathbb{P}}(N_{\mathrm{ret}}^{-1/2}), \qquad \|Y - \Pi(G)\|_{\max} = O_{\mathbb{P}}\Big(\sqrt{\frac{\log n}{N_{\mathrm{ret}}}}\Big)$. A more explicit concentration statement is given in Appendix A.2.

3.3. Mean and variance under the null

The first basic property is that, under the Erdős–Rényi null, every edge has the same expected retracing probability.

Lemma 3.1. *Under the Erdős–Rényi model $G(n,p)$, $\mathbb{E}[Y_{ij}] = \frac{1}{\binom{n}{2}} = \frac{2}{n(n-1)}$.*

Proof. For any graph realization, the RNBRW retracing proportions satisfy $\sum_{e_{ij}\in E} Y_{ij} = 1$. Taking expectations and using exchangeability under $G(n,p)$ gives $\sum_{e_{ij}\in E} \mathbb{E}[Y_{ij}] = 1$. Since all edges have the same expectation under H_0, each must equal $1/\binom{n}{2} = 2/(n(n-1))$. □

We also require crude but useful variance control. The manuscript provides upper and lower bounds that are sufficient for motivating the normalization of the test statistic.

Lemma 3.2. *Under the Erdős–Rényi model $G(n,p)$, the RNBRW retracing proportion satisfies* $\mathrm{Var}(Y_{ij}) \le \mathbb{E}[Y_{ij}] = \frac{2}{n(n-1)}$.

Proof. Since $0 \le Y_{ij} \le 1$, we have $Y_{ij}^2 \le Y_{ij}$. Therefore, $\mathrm{Var}(Y_{ij}) = \mathbb{E}[Y_{ij}^2] - \mathbb{E}[Y_{ij}]^2 \le \mathbb{E}[Y_{ij}^2] \le \mathbb{E}[Y_{ij}]$. Applying Lemma 3.1 gives $\mathrm{Var}(Y_{ij}) \le \mathbb{E}[Y_{ij}] = \frac{2}{n(n-1)}$. □

By Cauchy–Schwarz and Lemma 3.2, for distinct edges $e \ne f$ we have $\left|\mathrm{Cov}(Y_e, Y_f)\right| \le \sqrt{\mathrm{Var}(Y_e)\,\mathrm{Var}(Y_f)} \le 2/(n(n-1)) = O(n^{-2})$; after the normalization in (3), this yields pairwise standardized-entry covariances of at most order $O(n^{-1})$, but does not by itself imply correlation decay.

Lemma 3.1 and Lemma 3.2 justify centering and scaling the RNBRW matrix under H_0. The variance bound is intentionally crude: it records that the RNBRW edge weights have controlled second moments under H_0, and no matching lower bound is needed in the sequel. The remaining issue is dependence, since RNBRW retracing events are not exactly independent across edges. The next assumption records the working correlation-decay input used to motivate spectral calibration.

3.4. A conditional weak-dependence assumption under the null

Assumption 3.3 (Disjoint-edge weak dependence)**.** Let $G(n,p)$ be an Erdős–Rényi graph, and let Y_e denote the normalized RNBRW retracing weight associated with edge e. Suppose that, for disjoint edge pairs,

$$\max_{e\cap f=\varnothing} |\mathrm{Corr}(Y_e, Y_f)| \le r_n(p), \qquad r_n(p) \to 0. \tag{2}$$

Remark 3.4 (Scope of Assumption 3.3). Assumption 3.3 is a working weak-dependence condition for the full normalized RNBRW matrix, not a theorem proved here. The appendix proves a fixed-depth local analogue in the sparse regime and separates graph-level dependence from empirical-normalization dependence. These results support the weak-dependence heuristic but do not establish a tailored Tracy–Widom theorem, so the calibration below remains theory-motivated and simulation-calibrated.

Interpretation of the null-side theory. Assumption 3.3 concerns disjoint-edge correlations in the full normalized RNBRW matrix and is not proved in this paper. Appendix B proves localized sparse-regime analogues for bounded truncated scores, records truncation and normalization bounds, and isolates the additional dependence induced by empirical normalization. These results support the plausibility of weak pairwise dependence for localized RNBRW-type functionals, but they do not establish Assumption 3.3 for the full oracle retracing probabilities.

4. RNBRW goodness-of-fit statistic

We now define the proposed goodness-of-fit test based on RNBRW edge weights. The key idea is that, under the null hypothesis of no community structure, RNBRW retracing proportions should behave approximately homogeneously across edges after centering and normalization, whereas under structured alternatives they should exhibit a spectral departure driven by cycle enrichment within communities.

4.1. RNBRW-GOF statistic

Let Y_{obs} denote the $n \times n$ symmetric matrix of RNBRW retracing proportions, where $(Y_{obs})_{uv}$ = proportion of successful RNBRW retracing events for which edge (u,v) is the retracing edge. Under the null hypothesis $H_0 : G \sim G(n,p)$, Lemma 3.1 implies that the expected retracing proportion of each edge is $\mu_{\mathrm{null}} = \frac{1}{\binom{n}{2}}$. Define the null mean matrix $M_{\mathrm{mean,null}}$ by $(M_{\mathrm{mean,null}})_{uv} = \begin{cases} \mu_{\mathrm{null}}, & u \ne v, \\ 0, & u = v. \end{cases}$

To form the test statistic, we first center and normalize the observed RNBRW matrix. Let $\widehat{\mathrm{Var}}_{\mathrm{null}}(\hat{p})$ be an estimate of the variance of the RNBRW edge weights under H_0, obtained from Erdős–Rényi simulations at the plug-in edge density $\hat{p} = m/\binom{n}{2}$.

We then define

$$Y_{std} = \frac{Y_{obs} - M_{\mathrm{mean,null}}}{\sqrt{(n-1)\,\widehat{\mathrm{Var}}_{\mathrm{null}}(\hat{p})}}. \tag{3}$$

The normalization in (3) follows the same scaling used in our simulation study and places the RNBRW matrix on a scale comparable to classical spectral statistics from random matrix theory. In finite samples, however, residual weak dependence among RNBRW edge weights can still influence the extreme eigenvalues. To improve null calibration in practice, we therefore add a small symmetric perturbation whose magnitude is tied to the observed graph density.

Let $c_{obs} = n\hat{p}$. Construct a symmetric matrix Δ_{obs} with diagonal entries $(\Delta_{obs})_{uu} = 0$ and off-diagonal entries $(\Delta_{obs})_{uv} = -\frac{4c_{obs}^2}{n^4} Z_{uv}, \quad u < v$, where $Z_{uv} \overset{i.i.d.}{\sim} N(0,1)$ and $(\Delta_{obs})_{vu} = (\Delta_{obs})_{uv}$. We then define the perturbed standardized matrix

$$\widehat{Y} = Y_{std} + \Delta_{obs}. \tag{4}$$

Our proposed RNBRW goodness-of-fit statistic is the largest-eigenvalue statistic

$$\theta_{raw} = n^{2/3}\big(\lambda_{\max}(\widehat{Y}) - 2\big). \tag{5}$$

Large values of θ_{raw} indicate a cycle-sensitive departure from the Erdős–Rényi null; under the assortative SBM alternatives studied here, this departure is interpreted as evidence of community structure. Under H_0, the null-side results of Section 3 justify centering, scaling, and finite-walk concentration of the RNBRW matrix, but they do not imply a tailored Tracy–Widom universality theorem for $\widehat{Y}$. Accordingly, the practical calibration below should be interpreted as theory-motivated and simulation-calibrated. Following the same philosophy as Bickel & Sarkar (2013), we therefore optionally apply a Bartlett-type correction.

Let $\mu_{\theta,\mathrm{null}}(\hat{p})$ and $\sigma^2_{\theta,\mathrm{null}}(\hat{p})$ denote the mean and variance of θ_{raw} estimated under the null by simulation at the plug-in edge density $\hat{p}$, and let μ_{TW_1} and $\sigma^2_{\mathrm{TW}_1}$ denote the theoretical mean and variance of the Tracy–Widom type 1 distribution. The corrected statistic is

$$\theta = \frac{\theta_{raw} - \mu_{\theta,\mathrm{null}}(\hat{p})}{\sqrt{\sigma^2_{\theta,\mathrm{null}}(\hat{p})}} \sqrt{\sigma^2_{\mathrm{TW}_1}} + \mu_{\mathrm{TW}_1}. \tag{6}$$

In particular, the practical calibration is conditional on the plug-in density $\hat{p}$, rather than unconditional over graph-density randomness under $G(n,p)$. Under H_0, the calibration is supported by the null-side results of Section 3, which establish centered first-order structure and controlled second moments for RNBRW edge weights, together with the working disjoint-edge correlation-decay formulation in Assumption 3.3, while the Tracy–Widom step is handled through finite-sample simulation calibration. We then compute the p-value by comparing θ (or θ_{raw} if no correction is used) to the Tracy–Widom type 1 reference distribution. Thus, the practical test is theory-motivated but simulation-calibrated.

4.2. Implementation

Algorithm 1 summarizes the testing pipeline; implementation details and sensitivity settings are deferred to Appendix C.

Algorithm 1 RNBRW-GOF

Require: Observed graph $G = (V, E)$ with $n = |V|$, $m = |E|$; walk multiplier c_{walk}; level α
$\hat{p} \leftarrow m/\binom{n}{2}, \quad N \leftarrow c_{\mathrm{walk}}\, m$
Run N RNBRW trials and form Y_{obs}
if $N_{\mathrm{ret}} = 0$ **then**
  return conservative non-rejection
end if
Form Y_{std} via (3)
Optionally set $\widehat{Y} \leftarrow Y_{std} + \Delta_{obs}$ via (4)
Compute $\theta_{raw} = n^{2/3}\big(\lambda_{\max}(\widehat{Y}) - 2\big)$
Bartlett-correct θ_{raw} via (6)
Reject H_0 if $\theta > q_{1-\alpha}(\mathrm{TW}_1)$

5. Null calibration and empirical performance

We now evaluate the finite-sample behavior of the proposed RNBRW-GOF statistic under both the null and structured alternatives. Our simulation study has two goals: (i) to assess how well the null distribution of the statistic is approximated by the Tracy–Widom type 1 law after finite-sample calibration, and (ii) to quantify the resulting Type I error control and power of the test across a range of graph sizes and sparsity regimes. This section is adapted from the simulation methodology and empirical results in the manuscript, with emphasis on hypothesis testing performance rather than recursive clustering.

5.1. Simulation methodology under the null

To evaluate the RNBRW-GOF test under H_0, we generated Erdős–Rényi graphs $G_k \sim G(n,p), \quad k = 1, \ldots, N_{\mathrm{iters}}$, for a range of graph sizes n and edge probabilities p. In the main null-study grid reported here, we considered graph sizes $n \in \{20, 50, 100\}$, and null edge probabilities $p \in \{0.05, 0.2, 0.35, 0.5\}$. In addition, for the Tracy–Widom alignment diagnostic in Figure 1, we included an additional larger-graph setting with $n = 500$ and $p = 0.5$ to illustrate the improvement in finite-sample calibration as the graph size increases. The experiments reported here span both sparse and relatively dense Erdős–Rényi regimes. The theoretical results in this paper, however, are limited to exact null centering under exchangeability, crude second-moment control, and finite-walk concentration, together with a weak-overlap assumption used to motivate the calibration. Accordingly, the proposed procedure should be interpreted as theory-motivated and simulation-calibrated across the full experimental grid. For each (n,p) configuration, we estimated the null variance using $N_{\mathrm{boot},1} = 200$ Erdős–Rényi graphs, collected $N_{\mathrm{iters}} = 1000$ raw test statistics for empirical null evaluation, and estimated the Bartlett-correction moments from an additional $N_{\mathrm{boot},2} = 200$ null graphs. For each simulated graph, we computed the RNBRW retracing-proportion matrix, centered and standardized it using the null mean and an estimated null variance, applied the perturbation term from Section 4, and then formed the raw largest-eigenvalue statistic $\theta_{raw,k} = n^{2/3}\big(\lambda_{\max}(\widehat{Y}_k) - 2\big)$. For each

graph replicate, the RNBRW retracing-proportion matrix was computed from $N = c_{\text{walk}}m$ independent RNBRW runs with $c_{\text{walk}} = 20$. Thus, the practical computational overhead relative to adjacency-only baselines is governed primarily by the walk multiplier c_{walk}, which we treat as a user-controlled accuracy–cost parameter. The same null simulations were also used to estimate the quantities required for finite-sample calibration: $\widehat{\text{Var}}_{\text{null}}(\hat{p})$, $\mu_{\theta,\text{null}}(\hat{p})$, and $\sigma^2_{\theta,\text{null}}(\hat{p})$. In particular, the Bartlett-type correction was implemented by matching the empirical mean and variance of the raw statistic under H_0 to the theoretical mean and variance of the Tracy–Widom type 1 distribution:

$$\theta_{\text{corrected},k} = \frac{\theta_{\text{raw},k} - \mu_{\theta,\text{null}}(\hat{p})}{\sqrt{\sigma^2_{\theta,\text{null}}(\hat{p})}}\sqrt{\sigma^2_{\text{TW}_1}} + \mu_{\text{TW}_1}. \quad (7)$$

For a nominal significance level α_{test}, the empirical Type I error was then computed as the proportion of corrected statistics exceeding the corresponding $(1 - \alpha_{\text{test}})$ Tracy–Widom quantile. All simulations involving the perturbation matrix Δ_{obs} were run with fixed random seeds for reproducibility. In particular, the sensitivity analysis in Appendix D shows that the perturbation scale has essentially no effect on the reported Type I error and power in the tested regime, whereas the RNBRW walk budget induces a visible calibration–power tradeoff. This indicates a practical calibration–power tradeoff in the RNBRW sampling budget, but in the weakest regime considered the absolute power remains low because the signal is near or below the finite-sample detectability boundary.

Sensitivity to walk budget and perturbation. We use $N = c_{\text{walk}}m$ RNBRW runs with $c_{\text{walk}} = 20$. Appendix D shows that increasing c_{walk} primarily reduces Monte Carlo variability and can improve power in weak-signal regimes, but may also make finite-sample null calibration somewhat more anti-conservative. By contrast, varying the perturbation scale has negligible effect in the tested regime. In the weak-signal SBM setting $n = 100$ with $(p, q) = (0.017, 0.015)$ at $\alpha = 0.01$, increasing c_{walk} improves stability and power; additional sensitivity results are reported in Appendix D.

Table 2. Calibration ablation under the Erdős–Rényi null for the representative setting $n = 100$, $p = 0.5$. Entries are empirical rejection probabilities at nominal levels $\alpha \in \{0.10, 0.05, 0.01\}$ for four variants: the raw unperturbed RNBRW statistic, the Bartlett-only variant, the perturbation-only variant, and the full RNBRW-GOF procedure.

Method	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$
Raw RNBRW ($\theta^{(0)}_{\text{raw}}$)	0.242	0.150	0.064
Bartlett only	0.112	0.067	0.021
Perturbation only	0.262	0.151	0.043
Full RNBRW-GOF	0.110	0.049	0.011

5.2. Empirical null calibration and Tracy–Widom alignment

We first examined the empirical distribution of the RNBRW-GOF statistic under the null hypothesis. As expected for an extreme-eigenvalue statistic, the uncorrected finite-sample distribution can deviate from its asymptotic Tracy–Widom limit, especially for small n. We therefore apply a Bartlett-type correction, following Bickel & Sarkar (2013), to improve the finite-sample approximation. After correction, the empirical null distribution of the RNBRW-GOF statistic aligns substantially better with the Tracy–Widom type 1 reference law. Figure 1 shows this alignment for $n = 100$ and $n = 500$ under the Erdős–Rényi null with $p = 0.5$. For $n = 100$, the corrected empirical distribution already tracks the Tracy–Widom law well: the empirical 0.99 quantile of the Bartlett-corrected statistic is approximately 2.01, which is very close to the Tracy–Widom type 1 reference value of approximately 2.02. For $n = 500$, the agreement is even stronger, especially in the upper tail, consistent with improved asymptotic behavior as the graph size increases and supporting the use of Tracy–Widom calibration in moderate-to-large graphs. Overall, these diagnostics indicate that the Bartlett-corrected RNBRW-GOF statistic is well approximated by the Tracy–Widom type 1 law for the graph sizes displayed in Figure 1, with the agreement improving further at the additional large-graph setting $n = 500$.

Bootstrap robustness in a representative null setting. For the representative null setting $n = 100$, $p = 0.5$, we also compared the main Tracy–Widom/Bartlett calibration with direct parametric bootstrap under fitted Erdős–Rényi and degree-corrected Chung–Lu nulls using the same RNBRW pipeline. This robustness check supports the main calibration story: the fitted Erdős–Rényi bootstrap gives similar finite-sample calibration, while the degree-corrected Chung–Lu bootstrap yields substantially more conservative calibration in this homogeneous setting. For reference, the fitted Erdős–Rényi bootstrap used $B = 930$ resamples and the fitted Chung–Lu bootstrap used $B = 1000$ resamples; full empirical rejection rates and bootstrap critical values are reported in Appendix D.3.

5.3. Type-I error behavior

Using the calibrated null distribution described above, we evaluated the empirical Type I error of the RNBRW-GOF test across multiple values of n, p, and α_{test}. The main pattern is that the test is conservative in the most challenging null regimes, namely very small and very sparse Erdős–Rényi graphs. In particular, the manuscript observes near-zero rejection rates in extremely sparse small-sample settings such as $n = 20$ and $p = 0.05$. This conservatism is statistically preferable to anti-conservative behavior, though it also indicates that asymptotic calibration is less reliable

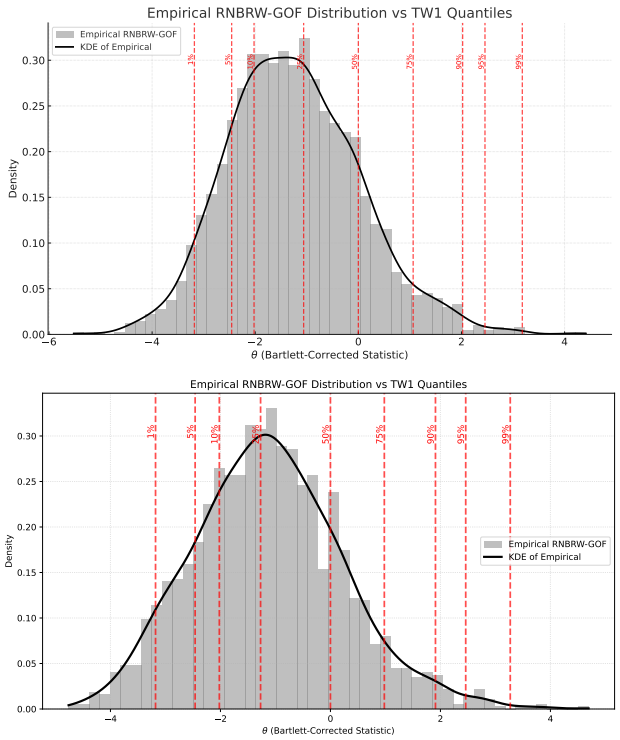

Figure 1. Empirical distributions of the Bartlett-corrected RNBRW-GOF statistic θ under the Erdős–Rényi null model with $p = 0.5$, shown for two graph sizes: $n = 100$ (top) and $n = 500$ (bottom). In each panel, the histogram and kernel density estimate are overlaid with selected Tracy–Widom type 1 (TW_1) quantiles, shown as dashed vertical lines. For $n = 100$, the empirical upper tail already agrees closely with the TW_1 reference, with empirical and theoretical 0.99 quantiles nearly identical. For $n = 500$, the agreement is even sharper, especially in the upper tail, illustrating improved finite-sample calibration as the graph size increases.

in that regime and may come at the expense of power. As the graph size increases, the empirical Type I error becomes better aligned with nominal significance levels, even when the graph remains relatively sparse. The manuscript specifically notes improved calibration for $p = 0.05$ once $n \geq 50$, with the best agreement observed at moderate significance levels such as $\alpha = 0.1$. In denser graphs, the test generally maintains reasonable null control, although the quality of calibration varies somewhat with (n, p) and the chosen significance level. Taken together, these results indicate that the RNBRW-GOF test is usable as a calibrated hypothesis test in moderate-to-large networks, while the smallest sparse regimes remain finite-sample limited. Accordingly, the present implementation should not be viewed as a well-calibrated small-sample sparse-graph test; rather, its most reliable regime of applicability is moderate-to-large graphs, including sparse graphs once n is sufficiently large for the Tracy–Widom calibration to stabilize.

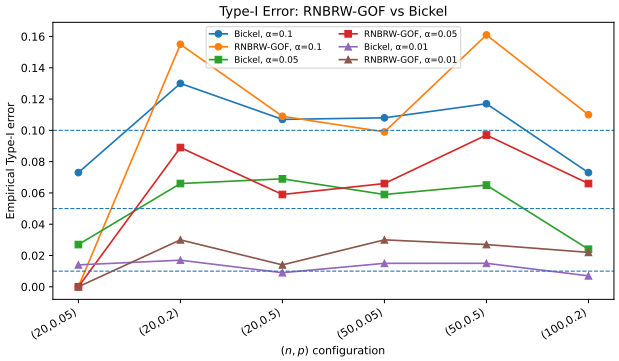

Figure 2. Empirical Type I error under the Erdős–Rényi null for RNBRW-GOF and the adjacency-based spectral baseline of Bickel & Sarkar (2013). Dashed horizontal lines indicate nominal significance levels. RNBRW-GOF is conservative in very small sparse graphs, while calibration improves for moderate and large networks.

5.4. Power under community-structured alternatives

We next evaluated the power of the RNBRW-GOF test under assortative stochastic block model alternatives. The central empirical conclusion is that the RNBRW-based test has strong ability to detect departures from the Erdős–Rényi null, and that its advantage is most pronounced when the alternative is difficult for purely edge-based tests—namely, when graphs are sparse or the community signal is weak. This is consistent with the mechanism in (1): conditional on $A_{ij} = 1$, adjacency-only information is fixed, while short-cycle participation can still differ between within-community and between-community edges. We also compare RNBRW-GOF to a non-backtracking spectral baseline based on the Ihara–Hashimoto matrix on directed edges, using the second-largest real-part eigenvalue as the informative statistic and calibrating it under a matched Erdős–Rényi null by bootstrap quantiles. This provides a natural sparse-graph comparator and helps isolate where RNBRW-based retracing information provides value beyond the standard non-backtracking spectrum. Accordingly, the present empirical evidence should be interpreted for assortative SBM-type departures from the Erdős–Rényi null rather than for arbitrary forms of community structure. Relative to the non-backtracking baseline, RNBRW-GOF is competitive overall, improves clearly over the adjacency-only baseline in several sparse and weak-signal settings, and remains comparable to the non-backtracking method in the weakest settings shown in Figure 3. At the same time, the non-backtracking baseline remains highly competitive in several intermediate regimes, so the empirical evidence supports complementarity rather than uniform dominance. For the sparse two-block settings with $n = 100$, it is useful to report the detectability-scale index

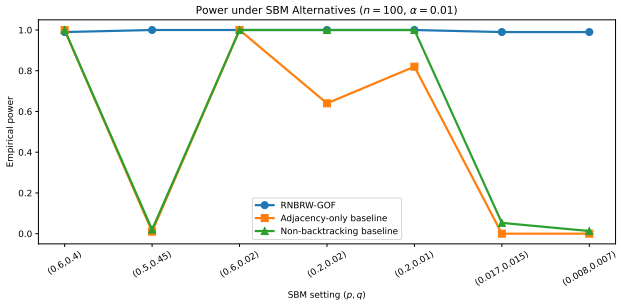

Figure 3. Empirical power under stochastic block model alternatives for RNBRW-GOF, the adjacency-based spectral baseline of Bickel & Sarkar (2013), and a non-backtracking spectral baseline based on the Ihara–Hashimoto matrix on directed edges. The non-backtracking statistic is the second-largest eigenvalue by real part and is calibrated under a matched Erdős–Rényi null by bootstrap quantiles. This comparison helps distinguish gains due to cycle-sensitive RNBRW retracing from gains due merely to moving beyond the raw adjacency operator.

$\rho_{\mathrm{det}} = \frac{(a-b)^2}{2(a+b)}, \qquad a = np, \;\; b = nq.$(8) For the two weakest settings in Figure 3, namely $(p, q) = (0.017, 0.015)$ and $(0.008, 0.007)$, this gives $\rho_{\mathrm{det}} \approx 6.25 \times 10^{-3}$ and $\rho_{\mathrm{det}} \approx 3.33 \times 10^{-3}$, respectively. These values lie far below the classical sparse two-block benchmark $\rho_{\mathrm{det}} = 1$, so the very low empirical power in those cases should be interpreted as a below-boundary stress test rather than as evidence of a limitation specific to RNBRW-GOF.

5.5. Summary of empirical findings

The simulation study supports four conclusions. First, RNBRW-GOF can be calibrated under the Erdős–Rényi null using simulation-based moment estimation and a Bartlett-type correction. Second, after correction, the empirical null distribution aligns reasonably well with the Tracy–Widom type 1 law for moderate and large graphs. Third, the test is conservative in extremely small sparse networks, where asymptotic approximations are least accurate. Fourth, under assortative SBM alternatives, RNBRW-GOF often improves on the adjacency-only baseline in sparse or weak-signal regimes while remaining competitive with a non-backtracking comparator that is itself highly competitive in several intermediate settings.

6. Discussion

We introduced RNBRW-GOF, a cycle-aware goodness-of-fit test for community structure in networks. The method combines RNBRW retracing probabilities with a largest-eigenvalue spectral statistic, extending adjacency-based testing procedures to incorporate higher-order graph structure. The main motivation is that cyclic organization carries information about community structure that is not fully captured by edge density alone, especially in sparse or weak-signal regimes. On the theory side, we prove exact null centering, crude second-moment control, finite-walk concentration, and sparse-regime local dependence results supporting RNBRW-based spectral calibration, but these do not yield a weak-dependence theorem for the full oracle RNBRW matrix or a tailored Tracy–Widom universality theorem for the final statistic, so calibration remains theory-motivated and simulation-calibrated. Empirically, this is sufficient to obtain reasonable null behavior in moderate and large graphs and improved power over the adjacency-only baseline in several sparse and weak-signal assortative SBM regimes, while a non-backtracking comparator remains highly competitive in several intermediate settings. At the same time, the method can be conservative in extremely small and sparse networks, where asymptotic approximations are least reliable. Our empirical study focuses on assortative SBM-type alternatives, so the present claims under H_1 should be interpreted for that class rather than as a uniform guarantee over arbitrary forms of community structure. Under degree heterogeneity, RNBRW weights may also respond to hub-driven cycle concentration, while under disassortative or nearly bipartite structure, the present one-sided largest-eigenvalue statistic need not be the most sensitive direction of departure. Within that scope, the results suggest that cycle information can materially improve goodness-of-fit testing for assortative community structure. Two natural next diagnostics are an empirical characterization of the overlap remainder $r_n(p)$ and a robustness study of the plug-in null-variance estimator $\widehat{\mathrm{Var}}_{\mathrm{null}}(\hat{p})$ in sparse regimes. On the theory side, a full Tracy–Widom universality theorem for the final RNBRW statistic remains open. A key obstacle is that the full oracle matrix $\Pi(G)$ is not yet controlled at the operator-norm level by the present local-dependence and truncation arguments, which leaves open the passage from localized weak-overlap estimates to edge fluctuations of the largest eigenvalue.

Impact Statement

This paper presents work whose goal is to advance the methodology of hypothesis testing for network data. The proposed method may be useful in scientific settings where detecting higher-order graph structure is important. We do not foresee immediate harmful uses beyond those already associated with graph-based inference methods.

References

Bickel, P. J. and Sarkar, P. Hypothesis testing for automated community detection in networks. *arXiv preprint arXiv:1311.2694*, 2013.

De Meo, P., Ferrara, E., Fiumara, G., and Provetti, A. Enhancing community detection using a network weighting strategy. *Information Sciences*, 222:648–668, 2013.

Golub, B. and Jackson, M. O. Does homophily predict

consensus times? testing a model of network structure via a dynamic process. *Review of Network Economics*, 11 (3), 2012.

Karwa, V., Pati, D., Petrović, S., Solus, L., Alexeev, N., Raič, M., Wilburne, D., Williams, R., and Yan, B. Exact tests for stochastic block models. *arXiv preprint arXiv:1612.06040*, 2016.

Klymko, C., Gleich, D., and Kolda, T. G. Using triangles to improve community detection in directed networks. *arXiv preprint arXiv:1404.5874*, 2014.

Lai, D., Lu, H., and Nardini, C. Enhanced modularity-based community detection by random walk network preprocessing. *Physical Review E*, 81(6):066118, 2010.

Lei, J. et al. A goodness-of-fit test for stochastic block models. *The Annals of Statistics*, 44(1):401–424, 2016.

Moradi, B., Shakeri, H., Poggi-Corradini, P., and Higgins, M. New methods for incorporating network cyclic structures to improve community detection. *arXiv preprint arXiv:1805.07484*, 2018.

Newman, M. E. and Girvan, M. Finding and evaluating community structure in networks. *Physical review E*, 69 (2):026113, 2004.

Radicchi, F., Castellano, C., Cecconi, F., Loreto, V., and Parisi, D. Defining and identifying communities in networks. *Proceedings of the National Academy of Sciences of the United States of America*, 101:2658–2663, 2004. doi: 10.1073/pnas.0400054101.

Zhang, P., Wang, J., Li, X., Li, M., Di, Z., and Fan, Y. Clustering coefficient and community structure of bipartite networks. *Physica A: Statistical Mechanics and its Applications*, 387(27):6869–6875, 2008. doi: 10.1016/j.physa.2008.09.006.

Zhao, Y., Levina, E., and Zhu, J. Community extraction for social networks. *Proceedings of the National Academy of Sciences*, 108(18):7321–7326, 2011.

A. Finite-walk concentration details

Definition A.1 (Graph-dependent retracing probabilities)**.** For a fixed graph G, let $\pi_{ij}(G) = \mathbb{P}(\text{a successful RNBRW run on } G \text{ retraces } e_{ij} \mid G)$, and let $\Pi(G) = (\pi_{ij}(G))$ denote the corresponding conditional retracing-probability matrix over successful retracing events.

Lemma A.2. *Fix a graph G and an edge e_{ij}. Let N_{ret} denote the number of successful RNBRW runs, and let $Y_{ij} = \frac{R_{ij}}{N_{\mathrm{ret}}}$, where R_{ij} is the number of successful runs whose retracing edge is e_{ij}. Then $\mathbb{E}[Y_{ij} \mid G, N_{\mathrm{ret}}] = \pi_{ij}(G)$, $\mathrm{Var}(Y_{ij} \mid G, N_{\mathrm{ret}}) = \frac{\pi_{ij}(G)(1-\pi_{ij}(G))}{N_{\mathrm{ret}}}$, and, for every $\varepsilon > 0$, $\mathbb{P}(|Y_{ij} - \pi_{ij}(G)| > \varepsilon \,|\, G, N_{\mathrm{ret}}) \le 2\exp(-2N_{\mathrm{ret}}\varepsilon^2)$. Consequently, $Y_{ij} = \pi_{ij}(G) + O_{\mathbb{P}}(N_{\mathrm{ret}}^{-1/2})$ conditionally on G.*

Proof. Conditionally on G and N_{ret}, the retracing counts over successful RNBRW runs are multinomial with cell probabilities $\{\pi_{ij}(G)\}$. Equivalently, $R_{ij} \mid G, N_{\mathrm{ret}} \sim \mathrm{Binomial}(N_{\mathrm{ret}}, \pi_{ij}(G))$, so $\mathbb{E}[Y_{ij} \mid G, N_{\mathrm{ret}}] = \pi_{ij}(G)$ and $\mathrm{Var}(Y_{ij} \mid G, N_{\mathrm{ret}}) = \frac{\pi_{ij}(G)(1-\pi_{ij}(G))}{N_{\mathrm{ret}}}$. The Hoeffding bound then applies to the empirical proportion based on the N_{ret} successful retracing events, yielding the stated inequality. □

Corollary A.3. *For any $\varepsilon > 0$, $\mathbb{P}(\|Y - \Pi(G)\|_{\max} > \varepsilon \,|\, G, N_{\mathrm{ret}}) \le 2\binom{n}{2}\exp(-2N_{\mathrm{ret}}\varepsilon^2)$.*

In particular, $\|Y - \Pi(G)\|_{\max} = O_{\mathbb{P}}\Big(\sqrt{\frac{\log n}{N_{\mathrm{ret}}}}\Big)$ conditionally on G.

Proof. Apply the previous lemma to each off-diagonal entry and use a union bound over the $\binom{n}{2}$ undirected edges. □

B. Localized dependence decay and empirical normalization

Lemma B.1 (Dependence decay for bounded truncated local scores in the sparse regime)**.** *Fix $c > 0$ and an integer $L \ge 1$, and let $G_n \sim G(n, c/n)$. For each undirected edge e, let $T_e^{(L)}(G_n) = \Phi_L\big(\mathcal{N}_L(e; G_n)\big) \in [0,1]$ be a bounded L-truncated local score, where $\mathcal{N}_L(e; G_n)$ denotes the radius-L neighborhood of e and Φ_L is a measurable function of that rooted neighborhood. Then there exists a constant $C_{L,c} < \infty$ such that for any two fixed disjoint edges $e \ne f$, $\big|\mathrm{Cov}(T_e^{(L)}(G_n), T_f^{(L)}(G_n))\big| \le C_{L,c}\, n^{-1}$.*

Proof sketch. Let $\mu_n^{(L)}$ denote the joint law of $\big(\mathcal{N}_L(e; G_n), \mathcal{N}_L(f; G_n)\big)$ and let $\nu_n^{(L)}$ denote the product of the corresponding marginals. Couple the simultaneous breadth-first explorations of the two radius-L neighborhoods with two independent copies. If the explored neighborhoods do not overlap, then the two local scores are driven by disjoint families of Bernoulli edge indicators, so under the coupling the joint exploration agrees with the product exploration. Therefore, $\|\mu_n^{(L)} - \nu_n^{(L)}\|_{\mathrm{TV}} \le \mathbb{P}(\mathcal{N}_L(e; G_n) \cap \mathcal{N}_L(f; G_n) \ne \varnothing)$.

Since $p = c/n$, a radius-L neighborhood is stochastically dominated by a depth-L Galton–Watson exploration with $\mathrm{Poisson}(c)$ offspring. Hence, for fixed L, these neighborhoods have uniformly bounded second moments, and for

fixed disjoint edges the overlap probability is $O_{L,c}(n^{-1})$. Because $0 \le T_e^{(L)}, T_f^{(L)} \le 1$, the covariance is controlled by total variation: $\big|\mathrm{Cov}(T_e^{(L)}(G_n), T_f^{(L)}(G_n))\big| \le \|\mu_n^{(L)} - \nu_n^{(L)}\|_{\mathrm{TV}} \le C_{L,c}\, n^{-1}$. □

Lemma B.2 (Adjacent realized edge pairs in the sparse regime)**.** *Fix $c > 0$ and let $G_n \sim G(n, c/n)$. Let $A_n := \#\big\{\{e, f\} \subset E(G_n) : e \sim f\big\}$, where $e \sim f$ means that e and f share exactly one vertex. Then $A_n = O_{\mathbb{P}}(n)$.*

Proof. Each adjacent unordered realized edge pair is counted exactly once by its shared vertex, so $A_n = \sum_{v=1}^{n} \binom{d_{G_n}(v)}{2}$. Since $d_{G_n}(v) \sim \mathrm{Bin}(n-1, c/n)$, $\mathbb{E}\Big[\binom{d_{G_n}(v)}{2}\Big] = \binom{n-1}{2}\left(\frac{c}{n}\right)^2 = O(1)$. Therefore, $\mathbb{E}[A_n] = \sum_{v=1}^{n} \mathbb{E}\Big[\binom{d_{G_n}(v)}{2}\Big] = O(n)$, and Markov's inequality gives $A_n = O_{\mathbb{P}}(n)$. □

Truncation identity for oracle retracing masses.

Definition B.3 (Cycle-completion time and truncated oracle retracing masses)**.** Let G be a realized graph, and consider one RNBRW run started from a uniformly chosen directed edge. Let τ_{cyc} denote the first time the walk completes a cycle, with the convention $\tau_{\mathrm{cyc}} = \infty$ if the run terminates at a dead end before any cycle completion. For an undirected edge $e \in E(G)$, define the full oracle retracing mass $q_e(G) := \mathbb{P}(\text{the RNBRW run retraces } e \,|\, G)$, and the depth-$L$ truncated oracle retracing mass $q_e^{(\le L)}(G) := \mathbb{P}(\tau_{\mathrm{cyc}} \le L \text{ and the RNBRW run retraces } e \,|\, G)$. Also define the long-excursion tail mass $R_L(G) := \mathbb{P}(L < \tau_{\mathrm{cyc}} < \infty \,|\, G)$.

Lemma B.4 (Truncation error equals tail mass)**.** *For every realized graph G and every $L \ge 1$, $0 \le q_e(G) - q_e^{(\le L)}(G) \le R_L(G)$ for every $e \in E(G)$, and $\sum_{e \in E(G)} \big(q_e(G) - q_e^{(\le L)}(G)\big) = R_L(G)$. Consequently, $\sum_{e \in E(G)} \big|q_e(G) - q_e^{(\le L)}(G)\big| = R_L(G)$.*

Proof. Fix a realized graph G and an edge $e \in E(G)$. The event $\{\text{the RNBRW run retraces } e\}$ splits disjointly into $\{\tau_{\mathrm{cyc}} \le L \text{ and the run retraces } e\} \cup \{L < \tau_{\mathrm{cyc}} < \infty \text{ and the run retraces } e\}$. Therefore $q_e(G) = q_e^{(\le L)}(G) + \mathbb{P}(L < \tau_{\mathrm{cyc}} < \infty \text{ and the run retraces } e \,|\, G)$, which implies $0 \le q_e(G) - q_e^{(\le L)}(G) \le R_L(G)$.

Summing over all realized edges $e \in E(G)$ and using that exactly one retracing edge is recorded on the event $\{\tau_{\mathrm{cyc}} < \infty\}$, we obtain $\sum_{e \in E(G)} \big(q_e(G) - q_e^{(\le L)}(G)\big) = \mathbb{P}(L < \tau_{\mathrm{cyc}} < \infty \,|\, G) = R_L(G)$. Since every summand is nonnegative, $\sum_{e \in E(G)} \big|q_e(G) - q_e^{(\le L)}(G)\big| = R_L(G)$. □

Lemma B.5 (Normalized truncation error bound)**.** *Let $s(G) = \mathbb{P}(\tau_{\mathrm{cyc}} < \infty \mid G) = \sum_{e \in E(G)} q_e(G)$, and define the normalized full and truncated oracle retracing probabilities by $\pi_e(G) := \frac{q_e(G)}{s(G)}$, $\pi_e^{(\le L)}(G) := \frac{q_e^{(\le L)}(G)}{s(G) - R_L(G)}$, whenever $s(G) > 0$ and $s(G) - R_L(G) > 0$. Then $\sum_{e \in E(G)} \big|\pi_e(G) - \pi_e^{(\le L)}(G)\big| \le \frac{2R_L(G)}{s(G)}$. In particular, on the event $\{s(G) \ge s_0\}$ with $s_0 > 0$, $\sum_{e \in E(G)} \big|\pi_e(G) - \pi_e^{(\le L)}(G)\big| \le \frac{2}{s_0} R_L(G)$.*

Proof. Write $a_e = q_e(G)$, $b_e = q_e^{(\le L)}(G)$, $s = \sum_e a_e$, $t = \sum_e b_e = s - R_L(G)$. Then $\sum_e \big|\frac{a_e}{s} - \frac{b_e}{t}\big| \le \sum_e \big|\frac{a_e - b_e}{s}\big| + \sum_e b_e \big|\frac{1}{s} - \frac{1}{t}\big|$. By Lemma B.4, $\sum_e |a_e - b_e| = R_L(G)$, $\sum_e b_e = t = s - R_L(G)$. Hence $\sum_e \big|\frac{a_e}{s} - \frac{b_e}{t}\big| \le \frac{R_L(G)}{s} + t \cdot \frac{|s-t|}{st} = \frac{R_L(G)}{s} + \frac{R_L(G)}{s} = \frac{2R_L(G)}{s(G)}$. The final claim follows immediately on $\{s(G) \ge s_0\}$. □

Definition B.6 (Success probability)**.** For a fixed graph G, let $s(G) = \mathbb{P}(\text{a single RNBRW run on } G \text{ is successful} \mid G)$, that is, the conditional probability that the run ends with a retracing edge rather than a dead end.

Lemma B.7 (Covariance decomposition for empirical RNBRW proportions)**.** *Fix a graph G, and perform N independent RNBRW runs on G. Let $\pi_e(G)$ denote the conditional probability that a successful run retraces edge e, let R_e be the number of successful runs retracing edge e, let $N_{\mathrm{ret}} = \sum_e R_e$, and define $Y_e = \frac{R_e}{N_{\mathrm{ret}}}$ on the event $\{N_{\mathrm{ret}} > 0\}$. Then, conditional on G and N_{ret}, $(R_e)_{e \in E} \sim \mathrm{Multinomial}\big(N_{\mathrm{ret}}, (\pi_e(G))_{e \in E}\big)$, and therefore $\mathbb{E}[Y_e \mid G, N_{\mathrm{ret}}] = \pi_e(G)$, and $\mathrm{Cov}(Y_e, Y_f \mid G, N_{\mathrm{ret}}) = \frac{\pi_e(G)\mathbf{1}\{e=f\} - \pi_e(G)\pi_f(G)}{N_{\mathrm{ret}}}$. Consequently, on $\{N_{\mathrm{ret}} > 0\}$, $\mathrm{Cov}(Y_e, Y_f \mid G, N_{\mathrm{ret}} > 0) = \big(\pi_e(G)\mathbf{1}\{e = f\} - \pi_e(G)\pi_f(G)\big)\,\mathbb{E}\Big[\frac{1}{N_{\mathrm{ret}}} \,\Big|\, G, N_{\mathrm{ret}} > 0\Big]$.*

In particular, for distinct edges $e \ne f$, $\mathrm{Cov}(Y_e, Y_f \mid G, N_{\mathrm{ret}} > 0) = -\pi_e(G)\pi_f(G)\,\mathbb{E}\Big[\frac{1}{N_{\mathrm{ret}}} \,\Big|\, G, N_{\mathrm{ret}} > 0\Big]$.

Moreover, if along the graph sequence under consideration one has $Ns(G) \to \infty$, then $\mathbb{E}\Big[\frac{1}{N_{\mathrm{ret}}} \,\Big|\, G, N_{\mathrm{ret}} > 0\Big] = \frac{1}{Ns(G)} + O\big((Ns(G))^{-2}\big)$, and hence for $e \ne f$, $\mathrm{Cov}(Y_e, Y_f \mid G, N_{\mathrm{ret}} > 0) = -\frac{\pi_e(G)\pi_f(G)}{Ns(G)} + O\big((Ns(G))^{-2}\big)$.

Proof. Conditional on G, each RNBRW run has one of the following outcomes: failure, or success with retracing edge e. Hence the full collection of counts is multinomial with cell probabilities $\big(1 - s(G), (s(G)\pi_e(G))_{e \in E}\big)$. Conditioning further on the total number N_{ret} of successful runs yields $(R_e)_{e \in E} \mid G, N_{\mathrm{ret}} \sim \mathrm{Multinomial}\big(N_{\mathrm{ret}}, (\pi_e(G))_{e \in E}\big)$. Therefore, $\mathbb{E}[R_e \mid G, N_{\mathrm{ret}}] = N_{\mathrm{ret}}\pi_e(G)$, and $\mathrm{Cov}(R_e, R_f \mid$

$G, N_{\mathrm{ret}}) = N_{\mathrm{ret}}\big(\pi_e(G)\mathbf{1}\{e = f\} - \pi_e(G)\pi_f(G)\big)$. Dividing by N_{ret}^2 gives $\mathbb{E}[Y_e \mid G, N_{\mathrm{ret}}] = \pi_e(G)$ and $\mathrm{Cov}(Y_e, Y_f \mid G, N_{\mathrm{ret}}) = \frac{\pi_e(G)\mathbf{1}\{e=f\} - \pi_e(G)\pi_f(G)}{N_{\mathrm{ret}}}$.

Now condition on the event $\{N_{\mathrm{ret}} > 0\}$. Since $\mathbb{E}[Y_e \mid G, N_{\mathrm{ret}}] = \pi_e(G)$ does not depend on N_{ret}, the conditional law of total covariance gives $\mathrm{Cov}(Y_e, Y_f \mid G, N_{\mathrm{ret}} > 0) = \mathbb{E}[\mathrm{Cov}(Y_e, Y_f \mid G, N_{\mathrm{ret}}) \,|\, G, N_{\mathrm{ret}} > 0]$.

Substituting the previous display yields $\mathrm{Cov}(Y_e, Y_f \mid G, N_{\mathrm{ret}} > 0) = \big(\pi_e(G)\mathbf{1}\{e = f\} - \pi_e(G)\pi_f(G)\big)\,\mathbb{E}\Big[\frac{1}{N_{\mathrm{ret}}} \,\Big|\, G, N_{\mathrm{ret}} > 0\Big]$. For $e \neq f$, this reduces to the stated negative formula.

Finally, conditional on G, $N_{\mathrm{ret}} \sim \mathrm{Binomial}(N, s(G))$. Write $\mu = Ns(G)$. On the event $\{N_{\mathrm{ret}} > 0\}$, a second-order Taylor expansion of $x \mapsto 1/x$ around μ gives

$$\frac{1}{N_{\mathrm{ret}}} = \frac{1}{\mu} - \frac{N_{\mathrm{ret}} - \mu}{\mu^2} + \frac{(N_{\mathrm{ret}} - \mu)^2}{\xi^3},$$

for some random ξ between N_{ret} and μ. Since $\mathrm{Var}(N_{\mathrm{ret}} \mid G) = Ns(G)(1 - s(G)) = O(\mu)$, and $\mathbb{P}(N_{\mathrm{ret}} = 0 \mid G) = (1 - s(G))^N \leq e^{-\mu}$, the conditioning on $\{N_{\mathrm{ret}} > 0\}$ changes the first reciprocal moment only by an exponentially small amount. Hence, when $\mu = Ns(G) \to \infty$,

$$\mathbb{E}\left[\frac{1}{N_{\mathrm{ret}}} \,\middle|\, G, N_{\mathrm{ret}} > 0\right] = \frac{1}{Ns(G)} + O\big((Ns(G))^{-2}\big).$$

Substituting this into the covariance identity proves the final display. □

Scope. Lemma B.1 controls dependence decay for bounded fixed-depth local scores associated with disjoint edges in the sparse regime. Lemma B.4 shows that passing from truncated oracle retracing masses to the full oracle quantities reduces to controlling the long-excursion tail mass, and Lemma B.5 gives the corresponding normalized ℓ_1 bound for the oracle retracing probabilities. Lemma B.2 shows that adjacent realized edge pairs are only $O_{\mathbb{P}}(n)$ in number under $G(n, c/n)$, so they can be handled separately by crude bookkeeping bounds when summing pairwise contributions on the natural order-n scale. Lemma B.7 isolates the additional dependence induced purely by the empirical normalization $Y_e = R_e/N_{\mathrm{ret}}$ and shows that, for distinct edges, this contribution is conditionally negative and lower order whenever $Ns(G) \to \infty$. With the simulation choice $N = c_{\mathrm{walk}}m$, this condition becomes $c_{\mathrm{walk}}\, m\, s(G) \to \infty$, which is a regime assumption rather than an automatic consequence of sparsity alone. Extending these arguments to the fully normalized oracle RNBRW retracing probabilities would still require controlling long non-backtracking excursions, handling the global normalization, and treating adjacent-edge interactions at the oracle level.

What is missing for a full null limit theorem

The results in this paper are intentionally partial. They establish exact null centering, crude second-moment control, finite-walk concentration, and localized sparse-regime dependence bounds for truncated local scores. These ingredients justify the proposed centering and normalization and support the simulation-calibrated use of the RNBRW statistic, but they do not yield a full Tracy–Widom-type limit theorem for the final matrix statistic.

A full null limit theorem would require several additional ingredients.

First, one would want an operator-norm concentration bound for the Monte Carlo error in the empirical RNBRW matrix, strengthening the current entrywise concentration result to a matrix concentration statement conditional on the realized graph. Such a result would separate finite-walk sampling error from graph-level spectral fluctuations on the natural operator scale.

Second, one would need an operator-level approximation theorem for the oracle RNBRW retracing-probability matrix itself. At a technical level, this would require controlling long non-backtracking excursions, the global normalization inherent in retracing probabilities, and adjacent-edge interactions in operator norm rather than only through localized pairwise estimates.

Third, after identifying the correct reference operator for the oracle RNBRW matrix under the null, one would need the appropriate edge-fluctuation theorem for that reference object in the relevant sparsity regime. Only at that stage would a limiting extreme-eigenvalue law become available for the reference model.

Finally, one would need a perturbative transfer argument showing that the oracle RNBRW matrix differs from the reference operator by a remainder that is negligible on the edge-fluctuation scale. Once such a representation is established, the passage from the reference edge law to the RNBRW statistic would be a standard comparison or delta-method step.

We do not attempt to prove these ingredients here. Accordingly, the RNBRW calibration used in the paper should be interpreted as theory-motivated and simulation-calibrated rather than as a fully proved Tracy–Widom-calibrated limit theorem.

C. Algorithmic summary

Algorithm 2 RNBRW-GOF test

Require: Observed graph $G = (V, E)$ with $n = |V|$, $m = |E|$; null variance estimate $\widehat{\mathrm{Var}}_{\mathrm{null}}$; optional null calibration parameters $\mu_{\theta,\mathrm{null}}, \sigma^2_{\theta,\mathrm{null}}$

Ensure: Test statistic θ and p-value

Compute the observed edge density $\hat{p} \leftarrow m/\binom{n}{2}$
Run RNBRW on G to obtain the retracing-proportion matrix Y_{obs}
if $N_{\mathrm{ret}} = 0$ **then**
  return non-rejection
end if
Set $\mu_{\mathrm{null}} \leftarrow 1/\binom{n}{2}$ and form $M_{\mathrm{mean,null}}$
Compute $Y_{std} = \frac{Y_{obs} - M_{\mathrm{mean,null}}}{\sqrt{(n-1)\,\widehat{\mathrm{Var}}_{\mathrm{null}}}}$
Set $c_{obs} \leftarrow n\hat{p}$ and generate the symmetric perturbation matrix Δ_{obs}
Form $\widehat{Y} \leftarrow Y_{std} + \Delta_{obs}$
Compute $\theta_{raw} = n^{2/3}\big(\lambda_{\max}(\widehat{Y}) - 2\big)$
if Bartlett correction is available **then**
  Compute $\theta = \frac{\theta_{raw} - \mu_{\theta,\mathrm{null}}}{\sqrt{\sigma^2_{\theta,\mathrm{null}}}} \sqrt{\sigma^2_{\mathrm{TW}_1}} + \mu_{\mathrm{TW}_1}$
else
  $\theta \leftarrow \theta_{raw}$
end if
Compute the p-value from the Tracy–Widom type 1 upper tail

Reproducibility and randomness. The perturbation matrix Δ_{obs} is generated from i.i.d. Gaussian draws and therefore introduces an explicit randomized component into the finite-sample statistic. In all experiments, we fix the random seed used for perturbation generation, as well as the seeds used for graph simulation and RNBRW sampling, so that every reported result is exactly reproducible. The role of Δ_{obs} is purely calibration-oriented: it is a small density-dependent adjustment intended to improve finite-sample spectral behavior when weak residual correlations remain in the standardized RNBRW matrix. Thus, the practical procedure should be interpreted as simulation-calibrated rather than purely theory-calibrated.

D. Ablation tables and sensitivity analyses

Additional implementation sensitivity results are collected here. The calibration ablation under the representative null setting is reported in the main text (Table 2); this appendix focuses on walk-budget and perturbation sensitivity.

Table 3. Sensitivity of RNBRW-GOF to the RNBRW walk budget and perturbation scale.

Setting	Empirical Type I error	Empirical power
$c_{\mathrm{walk}} = 5$	0.011	0.009
$c_{\mathrm{walk}} = 10$	0.032	0.019
$c_{\mathrm{walk}} = 20$	0.028	0.023
$c_{\mathrm{walk}} = 50$	0.028	0.026
$\gamma = 0.0$	0.012	0.023
$\gamma = 0.5$	0.012	0.023
$\gamma = 1.0$	0.012	0.023
$\gamma = 2.0$	0.012	0.023

D.1. Sensitivity to the walk budget and perturbation magnitude

We next studied the robustness of RNBRW-GOF to two implementation choices: the RNBRW walk budget and the perturbation magnitude. For the walk budget, we varied $N = c_{\mathrm{walk}}\, m, \quad c_{\mathrm{walk}} \in \{5, 10, 20, 50\}$. For the perturbation study, we replaced Δ_{obs} by a scaled version $\Delta_{obs}(\gamma) = \gamma\, \Delta_{obs}, \quad \gamma \in \{0, 0.5, 1, 2\}$.

Table 3 summarizes the resulting empirical Type I error and power for the weak-signal SBM setting $n = 100$ with within- and between-community probabilities $(p, q) = (0.017, 0.015)$ at nominal level $\alpha = 0.01$. This regime is intended to probe near-boundary detection performance. For the perturbation-sensitivity rows, the null Type I error is evaluated at the denser Erdős–Rényi setting $p = 0.5$, matching the implementation used to generate the appendix table. The results show that the perturbation scale has negligible empirical effect in the tested regime, supporting the interpretation of the perturbation step as a finite-sample calibration device rather than a source of signal. In contrast, the RNBRW walk budget affects both null calibration and power: increasing c_{walk} improves power under the weak-signal SBM alternative, but also yields more anti-conservative null rejection in finite samples. This indicates a practical calibration–power tradeoff in the choice of RNBRW sampling intensity.

D.2. Diagnostic variants used in the experiments

To separate the contribution of cycle-aware weighting from that of finite-sample calibration, we also consider three diagnostic variants of the proposed statistic. First, the unperturbed raw statistic is $\theta^{(0)}_{\mathrm{raw}} = n^{2/3}\big(\lambda_{\max}(Y_{std}) - 2\big)$. Second, the Bartlett-only variant applies the moment correction in (6) to $\theta^{(0)}_{\mathrm{raw}}$. Third, the perturbation-only variant uses $\theta^{(\Delta)}_{\mathrm{raw}} = n^{2/3}\big(\lambda_{\max}(Y_{std} + \Delta_{obs}) - 2\big)$ without Bartlett correction. The full RNBRW-GOF procedure corresponds to the corrected statistic θ in (6). Comparing these variants helps determine whether the observed empirical gains

Table 4. Empirical rejection probabilities under the Erdős–Rényi null for the representative setting $n = 100$, $p = 0.5$, based on 1000 null replicates. The table compares the main Tracy–Widom/Bartlett calibration to direct parametric bootstrap under fitted Erdős–Rényi and degree-corrected Chung–Lu nulls.

Method	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$
TW/Bartlett	0.110	0.049	0.011
Bootstrap-ER	0.087	0.036	0.007
Bootstrap-Chung–Lu	0.008	0.001	0.000

arise from RNBRW cycle information itself or from the calibration devices used to stabilize the finite-sample null distribution.

D.3. Direct parametric-bootstrap robustness under the null

To further assess finite-sample calibration, we supplemented the Tracy–Widom/Bartlett calibration with a direct parametric-bootstrap robustness check under fitted null models. We focus on the representative null setting

$$n = 100, \qquad p = 0.5,$$

which is also used in the main calibration ablation. For each observed Erdős–Rényi replicate, we computed the RNBRW-GOF statistic using the same RNBRW weighting, centering, normalization, perturbation, and walk budget as in the main pipeline, and then calibrated the resulting statistic in three different ways:

1. Tracy–Widom calibration after Bartlett correction;
2. direct parametric bootstrap under a fitted Erdős–Rényi null;
3. direct parametric bootstrap under a fitted degree-corrected Chung–Lu null.

We then estimated empirical rejection probabilities at nominal levels

$$\alpha \in \{0.10, 0.05, 0.01\}$$

over 1000 null replicates.

Table 4 reports the resulting empirical Type I error. The Tracy–Widom/Bartlett calibration remains close to nominal, while the direct bootstrap under the fitted Erdős–Rényi null is similarly calibrated and the fitted Chung–Lu bootstrap is substantially more conservative in this homogeneous setting. This supports the use of Tracy–Widom/Bartlett calibration as the main homogeneous-null calibration scheme in the experiments, with direct bootstrap serving as a useful robustness diagnostic rather than a replacement.