

Graph comparison via the nonbacktracking spectrum

Andrew Mellor^{*} and Angelica Grusovin

Mathematical Institute, University of Oxford, Woodstock Road, Oxford OX2 6GG, United Kingdom



(Received 12 December 2018; published 23 May 2019)

The comparison of graphs is a vitally important, yet difficult task which arises across a number of diverse research areas including biological and social networks. There have been a number of approaches to define graph distance, however, often these are not metrics (rendering standard data-mining techniques infeasible) or are computationally infeasible for large graphs. In this work we define a new pseudometric based on the spectrum of the nonbacktracking graph operator and show that it cannot only be used to compare graphs generated through different mechanisms but can reliably compare graphs of varying size. We observe that the family of Watts-Strogatz graphs lie on a manifold in the nonbacktracking spectral embedding and show how this metric can be used in a standard classification problem of empirical graphs.

DOI: [10.1103/PhysRevE.99.052309](https://doi.org/10.1103/PhysRevE.99.052309)

I. INTRODUCTION

Comparing graph structures is a fundamental task in graph theory. In particular, the need to identify similar graph structure arises to determine common function, behavior, or generative process. This need is universal across disciplines and applications range from image processing [1], chemistry [2,3], and social network analysis [4,5].

For small graphs, identifying two identical graph structures (up to an isomorphism, or relabelling of vertices) is a trivial and computationally feasible task due to the limited number of possible configurations. For connected graphs with four vertices there are six possible graph configurations. However, for graphs with ten vertices there are 11 716 571 possible configurations. Although there is not a closed formula for the number of graph configurations with n vertices, it is clear that it soon becomes intractable to enumerate or compare large graphs through brute-force.

The problem of determining whether two graphs are isomorphic is equivalent to finding a permutation matrix P such that $AP = PB$, where A and B are the adjacency matrices of the two graphs. Despite the difficulty in the task [6] there have been a number of attempts to define graph distance by minimising $\|AP - PB\|$ for a suitably defined matrix norm $\|\cdot\|$. The matrices A and B can also take different forms to shift from a local perspective (adjacency matrix) to consider global properties, where A and B capture the number of paths between vertices. Examples of such distances include the chemical distance [3], CKS distance [7], and edit distance [8,9]. Recent work has focused on reducing the space over which to find a permutation matrix [10], which makes the problem tractable, although still computationally restrictive for large graphs.

Beyond scalability issues, one of the drawbacks of these methods is that they require graphs to be of the same order (i.e., the same number of vertices). For many applications this may be desirable, especially where the addition of a single vertex or edge can have drastic consequences on graph function (in graphs of chemical compounds for example). However for other applications we may be interested in a more relaxed equivalence of graph structure. For example, a ring graph of 100 vertices is structurally similar to a ring graph of 200 vertices, however the two cannot be compared using traditional isomorphism arguments (except to trivially show they are not isomorphic).

Contrasting approaches in this direction use a number of graph properties, or *features*, to characterize graph structure [11–15]. The most promising of these make use of the graph spectrum, either of the adjacency matrix directly, or of a version of the graph Laplacian [16–20]. The spectra of graphs or graph operators are useful since the eigenvalues and eigenvectors characterize the topological structure of a graph in a way which can be interpreted physically. Eigenvalues of the graph Laplacian describe how a quantity (information, heat, people, etc.) localized at a vertex can spread across the graph. These eigenvalues also dictate the stability of dynamics acting across the graph [21]. Beyond the Laplacian, recent work has investigated the spectral properties of higher-order operators such as the nonbacktracking graph operator (described in the next section) [22]. For a more detailed (but not complete) examination of graph distances we refer the reader to a recent survey [23].

In this article we present a new method to compare graph structure using the distribution of eigenvalues of the nonbacktracking operator in the complex plane. In Sec. II we describe the nonbacktracking operator and investigate some of its spectral properties before introducing the distributional nonbacktracking spectral distance (d-NBD) in Sec. III. We show empirical results for both synthetic and real graphs in Sec. IV before discussing the significance of the results and future research in Sec. V.

^{*}mellor@maths.ox.ac.uk

II. THE NONBACKTRACKING OPERATOR AND SPECTRUM

Consider a simple undirected binary graph $G = (V, E)$ where $V \subset \mathbb{N}$ is a set of vertices and $E \subset V^2$ is a set of edges. Here we use the word *simple* to mean that the graph contains no self-loops or multiple edges between vertices. Let $n = |V|$ be the number of vertices and $m = |E|$ be the number of edges.

Spectral analysis of graphs is typically conducted using the Laplacian matrix $L = A - D$ where A is the adjacency matrix which encodes G , and D is a diagonal matrix of vertex degrees with $D_{ii} = \sum_j A_{ij}$ and zeros elsewhere. In this work we chose instead to use a different linear graph operator, namely, the nonbacktracking operator. The nonbacktracking (or Hashimoto) matrix B is a $2m \times 2m$ matrix defined on the set of directed edges of G . Here each undirected edge $u \leftrightarrow v$ is replaced by two directed edges $u \rightarrow v$ and $v \rightarrow u$. The nonbacktracking matrix is then given by

$$B_{(u \rightarrow v), (x \rightarrow y)} = \begin{cases} 1 & \text{if } v = x \text{ and } y \neq u \\ 0 & \text{otherwise.} \end{cases}$$

The nonbacktracking matrix is asymmetric and the entries of B^k describe the number of nonbacktracking walks of length $k + 1$ across G . A nonbacktracking walk is a walk across the graph which may visit a vertex multiple times but at no point makes a traversal $i \rightarrow j \rightarrow i$. Of special note is $(B^k)_{ii}$, which counts the nonbacktracking cycles of length $k + 1$ which start and end at edge i [24]. The total number of nonbacktracking cycles of length $k + 1$ is therefore captured in the trace of B^k . Note that in constructing the nonbacktracking matrix we have constructed an equivalent directed graph and so this method can also be applied to directed graphs directly.

Spectral properties

The spectral properties of the nonbacktracking matrix are well understood, especially for regular graphs [25–27]. As B is asymmetric the eigenvalues can also take complex values. In Fig. 1 we show the spectrum in the complex plane of three real-world graphs and three synthetic graphs of varying size.

The first observation is that the bulk of the eigenvalues λ_k lie within the disk $|\lambda_k| < \sqrt{\rho(B)}$ where $\sqrt{\rho(B)} = \max_k |\lambda_k|$ is the spectral radius of B . This is commonly used as a heuristic for the bulk of the spectrum, however in particular cases more is known. For random regular graphs of degree d there are n pairs of conjugate eigenvalues which lie exactly on a circle of radius $\sqrt{d - 1}$, and for a stochastic block model the expected magnitude of the eigenvalues can be shown to be less than or equal to $\sqrt{\rho(B)}$ in the limit $n \rightarrow \infty$ [31].

Another property of the spectrum of B is that any tree structure (whether connected to the graph, or disconnected) contributes zero eigenvalues to the spectrum. This is a consequence of any nonbacktracking walk becoming “stuck” at the leaves of the tree. Furthermore, any unicyclic component gives rise to eigenvalues in $\{0\} \cup \{\lambda : |\lambda| = 1\}$. These two properties are easily explained as we make the connection between the number of nonbacktracking cycles in the graph and the eigenvalues of B .

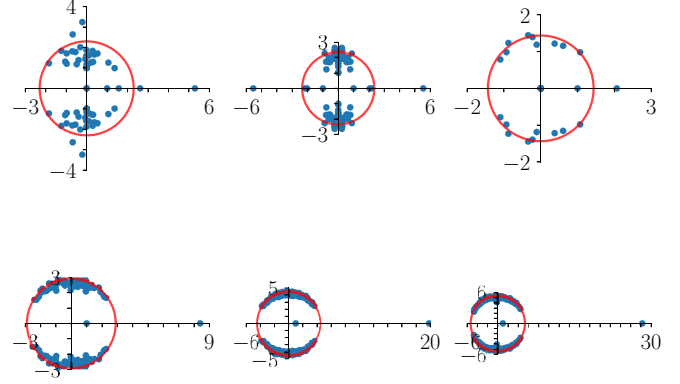


FIG. 1. The nonbacktracking spectrum of real and synthetic networks in the complex plane. Each eigenvalue $\lambda_k = \alpha_k + \beta_k i$ is represented by the point (α_k, β_k) . (Top, left to right) The Karate Club graph [28], Davis Southern Women graph [29], and Florentine Families graph [30]. (Bottom, left to right) An Erdős-Rényi graph with edge connection probability $p = 0.2$ for 50, 100, and 150 vertices. The red circle marks the curve $\alpha^2 + \beta^2 = \sqrt{\rho(B)}$ where $\rho(B)$ is the spectral radius of B .

For any matrix A the trace of A^k can be calculated as

$$\text{tr } A^k = \sum_{i=1}^n \lambda_i^k,$$

where $(\lambda_i)_{i=1}^n$ are the eigenvalues of A . This result is trivial for symmetric matrices which can be diagonalised, however the result also holds for asymmetric matrices. We can therefore write

$$\text{tr } B^k = \sum_{i=1}^{2m} \lambda_i^k, \quad (1)$$

noting that $\text{tr } B^k$ captures a count of all nonbacktracking cycles of length k . This means that the eigenvalues of B encode all information regarding the number of nonbacktracking cycles in G . Using this relation it is then clear that tree structures contribute zero eigenvalues since they cannot be part of a nonbacktracking cycle.

The last property of the spectrum we detail is derived from the Ihara determinant formula [32–34]. For any finite and undirected graph,

$$\det(B - \lambda I_n) = (\lambda^2 - 1)^{|E| - |V|} \det(Q_\lambda),$$

where $Q_\lambda = (D - I_n) - \lambda A + \lambda^2 I_n$ and D is the matrix with vertex degrees on the diagonal with zeros elsewhere. Using a linearization of the quadratic polynomial Q_λ [35] one can show that the $2n \times 2n$ matrix

$$B' = \begin{pmatrix} A & I - D \\ I & 0 \end{pmatrix} \quad (2)$$

shares the same eigenvalues as Q_λ , i.e.,

$$\det(Q_\lambda) = \det(B' - \lambda I_n).$$

Therefore, all eigenvalues of B , save for $|E| - |V|$ eigenvalues of both ± 1 , are eigenvalues of the smaller matrix B' . This result has very practical implications for calculating the spectrum of B given that n can often be substantially smaller than

m . The matrix B' provides a means to calculate a significant proportion of the spectrum of B in a fashion that scales linearly with the total number of vertices.

III. DISTRIBUTIONAL NONBACKTRACKING SPECTRAL DISTANCE

Much like the spectrum of the graph Laplacian has found use in community detection [36], dynamics on graphs [21], and graph clustering [37] our aim is to exploit the nonbacktracking spectrum to differentiate between graph structures. Recent work [22,38] has shown that the two-cores extracted from two graphs are isomorphic when the set of all nonbacktracking cycles in each graph are equal (this is referred to as the length spectrum). This means that should we be able to enumerate all possible nonbacktracking cycles we can effectively compare two graph structures. In practice enumerating all possible cycles of all possible lengths is infeasible, and even so it remains unclear how two such sets should be compared.

Torres *et al.* [22] address this issue by considering a “relaxed” length spectrum, calculating the *number* of cycles of length k as opposed to the set of cycles themselves. As shown in Eq. (1) the number of these cycles is captured completely by the spectrum of B . Based on this argument the distance between the number of cycles (and therefore eigenvalues) should provide a reasonable approximation to the graph isomorphism.

In Ref. [37] the number of nonbacktracking cycles of length k , c_k are used directly as an embedding of the graph. Each graph is embedded as a vector $\vec{v}_1 = [c_3, c_4, c_5, c_6, c_7, \ln(c_{2m})]$ with the graph distance subsequently defined as the Euclidean distance between vectors. These distances correlate well with the graph edit distance and perform marginally better than a truncated Laplacian spectrum embedding when applied to a computer vision task.

In contrast, the approach in Ref. [22] uses the largest k eigenvalues of B (in order of magnitude) to construct a feature vector. In particular for an ordered sequence of eigenvalues $(\lambda_i)_{i=1}^{2m}$ the graph is embedded as

$$\vec{v} = [\alpha_1, \dots, \alpha_k, \beta_1, \dots, \beta_k], \quad (3)$$

where $\lambda_s = \alpha_s + \beta_s i$ and again defining graph distance using the Euclidean distance between embeddings. Taking the largest eigenvalues is an intuitive step as these contribute the most to the cycle counts. There are, however, many instances where the magnitude of the eigenvalues are approximately equal (seen in Fig. 1) and so the method is heavily sensitive to the choice of k and the method of eigenvalue calculation. This is especially problematic for a d -regular random graph where all nonreal eigenvalues of B are conjugate pairs on the circle of radius $\sqrt{d-1}$.

The two previous approaches to nonbacktracking distance are limited in their ability compare graphs of different size. Consider two graphs of sizes n_1 and n_2 . Naturally, as we increase the number of vertices and edges the number of possible cycles will increase and therefore any cycle counting distance will diverge as $|n_1 - n_2| \rightarrow \infty$. Furthermore, to compare the eigenvalues of the nonbacktracking matrices directly we are restricted to choosing the largest $\min\{2n_1, 2n_2\}$

eigenvalues. We are therefore neglecting a potentially significant part of the spectrum for the larger graph, i.e., we are using a truncated spectrum. The issues surrounding the truncated spectrum become clear for d -regular random graphs. Since the bulk of all eigenvalues lie on the circle $|\lambda| = \sqrt{d-1}$ we cannot order them by magnitude. We therefore need a secondary ordering, say in descending order of $\Re(\lambda)$. The full spectrum will cover all eigenvalues on the circle, however, the truncated spectrum will lie on an arc of $|\lambda| = \sqrt{d-1}$. An effective comparison is therefore not possible.

This motivates us to consider a property of the spectrum that can be compared across graphs of different size, namely the spectral distribution. In this sense we hope to cluster graphs which share the same properties of generating mechanism, regardless of size. For computational efficiency we consider the eigenvalues of

$$B' = \begin{pmatrix} A & I - D \\ I & 0 \end{pmatrix},$$

where now A is the two-core of the original graph (vertices of degree one are iteratively removed), and D is the corresponding degree matrix. This means that we are operating on a reduced spectrum with no zero eigenvalues and $|E| - |V|$ eigenvalues of 1 or -1 omitted. The spectral radius of B' scales with the largest vertex degree, however the corresponding eigenvalues are not localized around high-degree vertices to the same extent as they are with the graph Laplacian. This is due to walks on B' not being permitted to return to high-degree vertices immediately after leaving them. This suggests a rescaling of the eigenvalues

$$|\lambda_i| \rightarrow \log_{\rho(B)} |\lambda_i|.$$

The rescaled eigenvalues $(\hat{\lambda})_{i=1}^{2n}$ then lie exclusively in the disk $|\hat{\lambda}| \leq 1$ with the bulk of the eigenvalues distributed in $|\hat{\lambda}| \leq \frac{1}{2}$.

Figure 2 shows the rescaled spectrum for Erdős-Rényi graphs of varying size with edge connection probabilities $p = 0.2$ (top) and $p = 0.8$ (bottom). Here we can clearly distinguish between the two parameter regimes and we see that these distributions are consistent across graphs irrespective of size.

We capture the distribution of the eigenvalues through the empirical cumulative spectral density given by

$$F(r, \theta) = \frac{1}{2n} \sum_{i=1}^{2n} \mathbb{1}_{\{|\hat{\lambda}_i| \leq r\}} \mathbb{1}_{\{0 \leq \arg(\hat{\lambda}_i) \leq \theta\}}. \quad (4)$$

This is defined on $r \in [0, 1]$, $\theta \in [0, \pi]$ and for $n > 0$. A special case to mention is when the two-core of a graph is empty. In this case we set $F(r, \theta) = 0 \forall r, \theta$ by convention [39]. Since the complex eigenvalues of B (and therefore B') are conjugate pairs we gain no further information by considering the lower half of the complex plane and is hence omitted. This naturally leads us to define the distributional nonbacktracking spectral distance (d-NBD) as the distance between empirical spectral densities. For two graphs G_1 and G_2 with rescaled spectral densities F_1 and F_2 respectively, the d-NBD

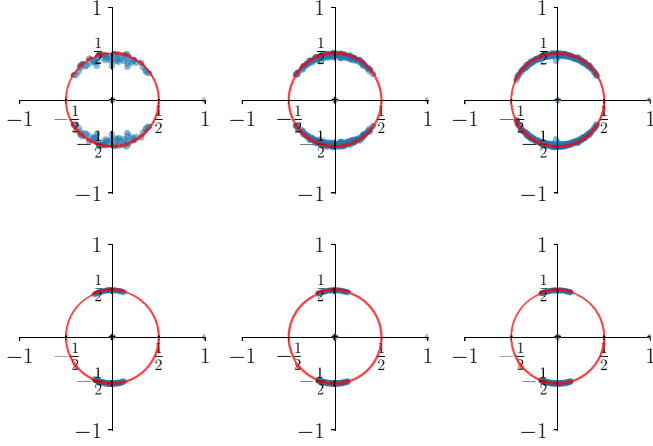


FIG. 2. The rescaled nonbacktracking spectrum of B' for Erdős-Rényi graphs of size 50, 100, 150 (left to right) and edge connection probability 0.2 (top) and 0.8 (bottom). The rescaled spectrum is persistent as the number of vertices in the graph increases, however for small graphs there are finite size effects. In this example, the two parameter regimes are distinguished by the range over which the argument of the eigenvalues take.

is given by

$$d(G_1, G_2) = \frac{1}{\pi} \left(\int_0^\pi \int_0^1 |F_1(r, \theta) - F_2(r, \theta)|^2 dr d\theta \right)^{\frac{1}{2}}, \quad (5)$$

i.e., $d(G_1, G_2)$ is proportional to the p -norm $\|F_1 - F_2\|_p$ with $p = 2$.

The d-NBD satisfies many desirable properties. It is non-negative, symmetric, and satisfies the triangle inequality (via the Minkowski inequality). While it is evident that $d(G, G) = 0$ it is not the case that $d(G_1, G_2) = 0$ implies that $G_1 = G_2$, only that they share the same two-core. The d-NBD is therefore a *pseudometric*.

Finally, we define a k^2 -dimensional embedding of a graph through a discretization of the empirical spectral density. Here a graph is represented by

$$\begin{aligned} \vec{v} = & [F(r_1, \theta_1), \dots, F(r_k, \theta_1), \\ & F(r_1, \theta_2), \dots, F(r_k, \theta_2), \\ & \dots \\ & F(r_k, \theta_1), \dots, F(r_k, \theta_k)], \end{aligned} \quad (6)$$

where $(r_1, \theta_1) = (0, 0)$ and $(r_k, \theta_k) = (1, \pi)$. This embedding is useful in the next section to visualise the d-NBD by plotting the graph embeddings in low-dimensional space.

IV. RESULTS

In this section we give a number of results from both synthetic and real-world graph examples.

A. Synthetic graphs

To investigate the properties of the d-NBD we consider three random graph models; the Erdős-Rényi graph (ER), the Watts-Strogatz graph (WS), and the d -regular random graph

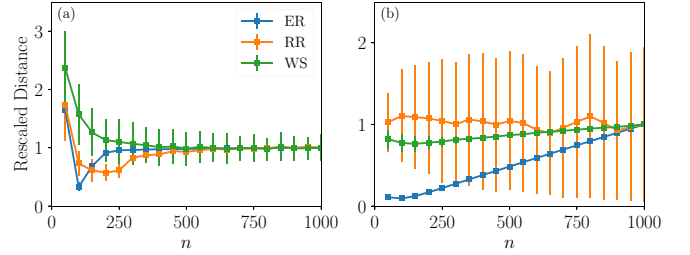


FIG. 3. Graph distance sensitivity for graphs generated from the same model with varying number of vertices. Here the models used are ER with edge connection probability $p = 0.25$, WS with $k = 4$ neighbor connections and rewiring probability $p = 0.1$, and RR graphs with degree $d = 5$. Distances are rescaled by the average distance $\langle d(G_{100}^*, G_{1000}^{(i)}) \rangle_i$ so that models can be overlaid and trends are preserved. (a) The average rescaled d-NBD $\langle d(G_{100}^*, G_n^{(i)}) \rangle_i$ between a graph with 100 vertices and graphs with identical model parameters. The distance is both stable and nonincreasing for all graph types considered. (b) Graph distance using only the largest k eigenvalues. Since the smallest graph is of size 50, we must choose $k < 100$ to be able to compare all graphs. Here the graph distance increases as n increases, except for the random regular graph.

(RR). One of the main advantages of the d-NBD over other graph distances is that ability to compare graphs of varying size. Figure 3 illustrates the stability of the d-NBD for graphs of size 50 to 1000. Considering a benchmark graph with 100 vertices we subsequently calculate the distances to an ensemble of graphs all generated from the same model. That is, we generate a reference graph G_{100}^* from the model and for each $n \in [50, 100, \dots, 1000]$ we sample 200 graphs $(G_n^{(i)})_{i=1}^{200}$ from that model and of that size before computing $d(G_{100}^*, G_n^{(i)})$ for each graph. Due to randomness in each model the distances are small but nonzero [Fig. 3(a)], however, this distance remains constant as n grows. The distances do however become noisier for small n as the graph properties for smaller graphs are more susceptible to stochastic fluctuations.

In contrast, taking the largest k eigenvalues of B' does not provide a stable distance as n varies [Fig. 3(b)]. For both ER and WS graphs the distance from the reference graph is monotonically increasing with n . The distance for the regular graph is however stable. This can be attributed to the fact that the largest degree, which correlates with the largest eigenvalue, is fixed. The d-NBD therefore is able to compare graphs generated from the same model, even if the degree distribution changes.

A further test we can consider on synthetic graphs is whether the d-NBD has a continuous dependence on the model parameters. For this purpose we consider the Watts-Strogatz model since we can consider dependency in both the nearest neighbor connection parameter k and the edge rewiring parameter p . We can interpolate between regularized ring structure ($p = 0, k \geq 2$) to completely randomized connections. Similarly we can interpolate between sparsely connected graphs ($k = 2$) to the complete graph ($k = n - 1$).

By considering a 10 000-dimensional embedding of these graphs, via Eq. (6), we can see that the space of Watts-Strogatz networks lies on a manifold (Fig. 4). The embedding has been reduced to three dimensions using principle component analysis and each point is taken as the average position of

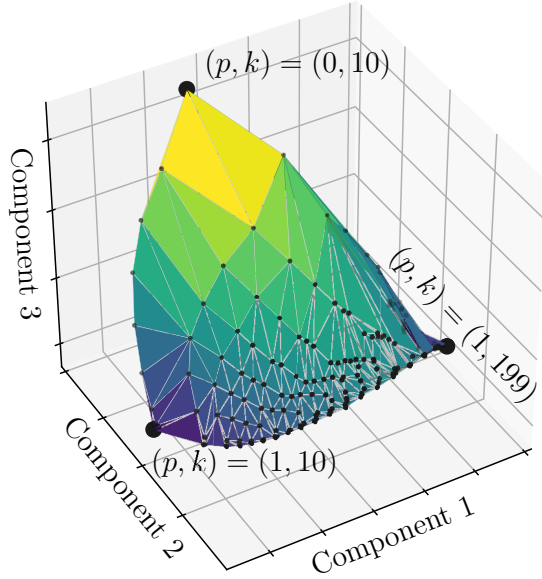


FIG. 4. An embedding of Watts-Strogatz graphs with 200 vertices and with rewiring parameter $p \in [0, 1]$ and vertex neighbors $k \in [2, 199]$. Each point is the average embedding of 100 samples from a particular model. The three dimensions displayed at the three principle components extracted from a 10 000-dimensional embedding. Note that when $k = 199$ we have a complete graph and so $(p, k) = (0, 199)$ is identical to $(p, k) = (1, 199)$.

100 graphs with 200 vertices. This observation gives further credit to the distance as changes in the underlying graph model give predictable changes in the underlying embedding. It must be stressed that the changes in the embedding are nontrivial functions of the parameter changes. This is clear by the uneven spacing of points in Fig. 4 which were generated by an even spacing of parameter space.

These results on synthetic graphs have shown that the d-NBD captures, but does not necessarily inform about, the generating process for these graphs and is able to detect subtle changes in the underlying model parameters. By their very nature, these synthetic graphs are generated from simple mechanisms. Is the d-NBD able to capture more complex generative mechanisms?

B. Empirical graphs

Graph distances are useful in practice to be able to determine either the function of a graph or the (closely related) mechanisms for graph formation. To illustrate how the d-NBD can distinguish between these mechanisms we consider a number of graph collections (detailed in Table I) drawn from different fields.

For example, we consider a sample of 26 graphs from the Facebook100 dataset which describes the social connections within universities in the U.S. These networks are human-made (although technologically assisted) and capture a number of human behaviours such as assortativity and clustering [40]. In contrast the metabolic networks dataset captures the reactions between metabolites within a number of different organisms. These graphs are generated based on chemical and biological interaction, although arguably there

TABLE I. Empirical datasets consider in this section, in addition to three synthetic graph models for comparison. Each dataset contains a collection of graphs of varying size, however within each collection all graphs are assumed to be generated from the same mechanism. For synthetic graphs we consider graphs with mean size 200 and variance 40. ER graphs have rewiring parameter $p = 0.1$, the WS graphs have $k = 20$ and $p = 0.1$, and the BA graphs have incoming vertices connect to $m = 2$ vertices with the system initialised with a single vertex. All graphs are taken to be simple and undirected.

Dataset		Count	max(n)	min(n)
Facebook100 (FB)	[40]	26	4563	769
World Subways (SW)	[41]	15	433	82
Autonomous Systems (AS)	[42]	30	3144	2948
Metabolic Networks (MN)	[43]	10	1593	505
Erdős-Rényi (ER)	[21]	30	244	151
Watts-Strogatz (WS)	[21]	30	244	151
Barabási-Albert (BA)	[21]	30	244	151

has been human interference in the design and curation of such graphs. A useful distance measure should therefore be able to distinguish between these different graph types by having small intra-type distances and relatively large inter-type distances.

Figure 5 shows a two-dimensional projection of the spectral embedding Eq. (6) for each graph, each of which is coloured by the graph collection it came from. Visually it is evident that the embedding performs well as each data collection forms its own isolated cluster. This is even the case for network collections, such as the Facebook100 graphs where the largest graph is approximately six times as large as the smallest graph in the collection.

To assess the equality of the clustering more formally we conduct a k -nearest neighbor classification [45] with the number of nearest neighbors $k = 10$. We compare the clustering of the d-NBD against two benchmarks. First, the non-backtracking spectral distance (NBD) [22] which is simply the Euclidean distance between the largest eigenvalues. Since the smallest graph we consider has 82 vertices we can only

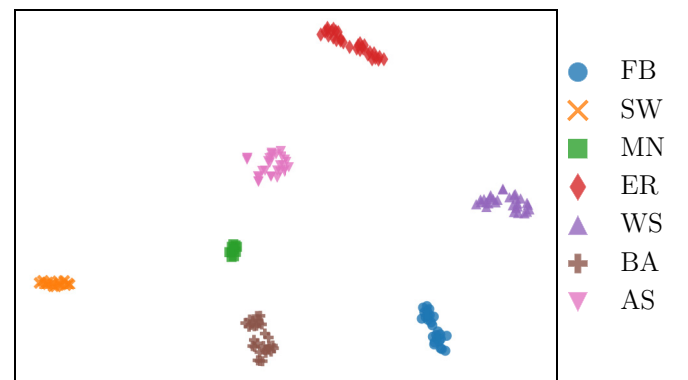


FIG. 5. A two-dimensional projection of the spectral embedding (6) of the graphs listed in Table I. The projection was created using the t-SNE algorithm [44]. The samples from each graph collection are well clustered with zero overlap between them.

TABLE II. The clustering performance of the distributional non-backtracking spectral distance (d-NBD), the nonbacktracking spectral distance (NBD), and the Laplacian (LAP) distance. Results are averaged over a 10-fold cross-validation. For the graph collections considered, the d-NBD performs without fault achieving perfect accuracy, precision, and recall. Both the NBD and LAP clusterings also show relatively high accuracy but a number of graphs are consistently labeled incorrectly.

	Train			Test		
	Rec.	Prec.	Acc.	Rec.	Prec.	Acc.
d-NBD	1.00	1.00	1.00	1.00	1.00	1.00
NBD	0.96	0.92	0.94	0.97	0.95	0.95
LAP	0.92	0.80	0.84	0.92	0.79	0.84

consider up to 164 eigenvalues. We also compare against the spectrum of the graph Laplacian (LAP), again taking up to 82 of the largest eigenvalues. Similar to the nonbacktracking matrix, the graph Laplacian (in particular the spectrum) has been previously used to characterize graph structure [37].

We assess the quality of the clustering using both the training and testing accuracy, precision, and recall. The results are presented in Table II.

To prevent overfitting we average the results over a 10-fold cross validation. The d-NBD achieves perfect performance for both the training and test data. Arguably this may not be a difficult classification task and so these results should be interpreted as a flawless measure. However, even in this “simple” task both the NBD and Laplacian methods struggle to classify a number of graphs correctly. This is likely due to the methods not being able to utilize the full spectrum when comparing graphs of dissimilar size. We can alternatively pose this problem as a supervised learning task, however strong performance in the unsupervised setting gives little motivation for us to do so.

V. CONCLUSIONS

In this article we have presented a method of graph comparison which can be used to compare graphs of different sizes. Through application to both synthetic and real graphs we have shown that the d-NBD can successfully identify the differences between graphs generated by different mechanisms. This distance performs better than distances defined on the truncated spectra of both the Laplacian and nonbacktracking matrix, however, this comes at a cost of having to compute the entire graph spectrum. This is not an issue for graphs of moderate size but can become computationally taxing for graphs formed of millions of vertices.

One possible remedy to this may be to consider the pseudospectra of such matrices. The pseudospectra gives some information on the location of eigenvalues and can be calculated

to a prescribed precision within the complex plane [46]. The pseudospectrum also has the added property that it gives a measure of the stability of the eigenvalues to small perturbations in the adjacency matrix. It should therefore give an indication to the effect of adding or removing an edge from the graph.

A limitation of the d-NBD is that it is ignorant to dangling nodes and tree structures due to its consideration of the two-core only. Pure tree structures (e.g. a spanning tree) are distinguished from other graphs, however the d-NBD is unable to distinguish between them and so alternate methods are needed. We can argue that many empirical graph structures of interest contain a two-core and that tree structure is perhaps better examined with more specialised (not walk-based) methods. Alternatively we can consider solutions which combine both the nonbacktracking and Laplacian spectrum in tandem, or to modify the nonbacktracking operator to allow backtracking only in the case of dangling nodes.

Another unexplored avenue is the use of graph distances to infer the graph generating mechanism rather than using the distance to distinguish between potentially different mechanisms. Approaches in this area are predominantly statistical, such as the work of Chen *et al.* [47] who use approximate Bayesian computation to select a likely graph generation mechanism. These approaches are heavily computational especially if the number of model choices is large. The d-NBD could be used as a precursor to such analysis by filtering possible models based on the graph distance to representative samples from these models.

The d-NBD has yet to be tested on a larger set of data which includes a wider range of graph categories (social, biological, etc.) but also with a diversity of graph collections within each category (Twitter, Facebook, and Instagram social graphs for instance). We anticipate that the d-NBD will continue to outperform other truncated spectral methods on the basis that more information is captured in the complete spectrum.

There also lies the more fundamental question of why spectral methods work in clustering graphs (both for the Laplacian and nonbacktracking operators). The nonbacktracking operator is closely related to the deformed Laplacian [48], also known as the Bethe Hessian [49]. This suggests there may be a physical interpretation to elucidate the underpinnings of such operators. What does a family of Laplacian-like operators tell us about the graph structure, and how many such operators are sufficient to characterize a graph? Connecting these two areas could lead to possible generalisations which could provide significant improvements in graph comparison.

ACKNOWLEDGMENTS

The authors thank Renaud Lambiotte for useful discussion and both Renaud and Ebrahim Patel for their comments on early versions of this article.

- [1] D. Conte, P. Foggia, C. Sansone, and M. Vento, *Int. J. Pattern Recogn. Artific. Intell.* **18**, 265 (2004).
- [2] F. H. Allen, *Acta Crystallogr. Sect. B: Struct. Sci.* **58**, 380 (2002).

- [3] V. Kvasnička, J. Pospíchal, and V. Baláz, *Theor. Chim. Acta* **79**, 65 (1991).
- [4] D. Koutra, J. T. Vogelstein, and C. Faloutsos, in *Proceedings of the SIAM International Conference on Data Mining* (SIAM, Philadelphia, PA, 2013), pp. 162–170.

- [5] V. Lyzinski, D. E. Fishkind, M. Fiori, J. T. Vogelstein, C. E. Priebe, and G. Sapiro, *IEEE Trans. Pattern Anal. Mach. Intell.* **38**, 60 (2016).
- [6] It is currently unknown as to whether this problem lies in the class of P- or NP-hard problems [50].
- [7] G. Chartrand, G. Kubicki, and M. Schultz, *Aequat. Math.* **55**, 129 (1998).
- [8] M. R. Garey and D. S. Johnson, *Computers and Intractability*, Vol. 29 (W. H. Freeman, New York, 2002).
- [9] A. Sanfeliu and K.-S. Fu, *IEEE Trans. Syst. Man Cybernet.* **SMC-13**, 353 (1983).
- [10] J. Bento and S. Ioannidis, in *Proceedings of the SIAM International Conference on Data Mining* (SIAM, Philadelphia, PA, 2018), pp. 333–341.
- [11] A. E. Wegner, L. Ospina-Forero, R. E. Gaunt, C. M. Deane, and G. Reinert, *J. Complex Netw.* **6**, 887 (2018).
- [12] M. Ferrer, E. Valveny, F. Serratos, K. Riesen, and H. Bunke, *Pattern Recogn.* **43**, 1642 (2010).
- [13] G. W. Klau, *BMC Bioinform.* **10**, S59 (2009).
- [14] K. Riesen, M. Neuhaus, and H. Bunke, in *Proceedings of the International Workshop on Graph-Based Representations in Pattern Recognition* (Springer, Berlin, 2007), pp. 383–393.
- [15] J.-P. Onnela, D. J. Fenn, S. Reid, M. A. Porter, P. J. Mucha, M. D. Fricker, and N. S. Jones, *Phys. Rev. E* **86**, 036104 (2012).
- [16] M. De Domenico and J. Biamonte, *Phys. Rev. X* **6**, 041062 (2016).
- [17] R. C. Wilson and P. Zhu, *Pattern Recogn.* **41**, 2833 (2008).
- [18] P. Zhu and R. Wilson, in *Proceedings of the British Machine Conference* (Springer, Berlin, 2005), pp. 69–1.
- [19] H. Elghawalby and E. R. Hancock, in *International Conference Image Analysis and Recognition* (Springer, Berlin, 2008), Vol. SMC-13, pp. 517–526.
- [20] B. Luo, R. C. Wilson, and E. R. Hancock, *Pattern Recogn.* **36**, 2213 (2003).
- [21] M. Newman, *Networks: An Introduction* (Oxford University Press, Oxford, 2010).
- [22] L. Torres, P. Suarez-Serrato, and T. Eliassi-Rad, Graph distance from the topological view of non-backtracking cycles, *arXiv:1807.09592* (2018).
- [23] C. Donnat and S. Holmes, *Ann. Appl. Stat.* **12**, 971 (2018).
- [24] Nonbacktracking does not mean that a walk only visits distinct vertices, it only means that it does not immediately return to a vertex it has just arrived from.
- [25] B. D. McKay, *Linear Algebra Appl.* **40**, 203 (1981).
- [26] O. Angel, J. Friedman, and S. Hoory, *Trans. Amer. Math. Soc.* **367**, 4287 (2015).
- [27] A. Saade, F. Krzakala, and L. Zdeborová, *Europhys. Lett.* **107**, 50005 (2014).
- [28] W. W. Zachary, *J. Anthropol. Res.* **33**, 452 (1977).
- [29] A. Davis, B. B. Gardner, and M. R. Gardner, *Deep South: A Social Anthropological Study of Caste and Class* (University of South Carolina Press, Columbia, SC, 2009).
- [30] R. L. Breiger and P. E. Pattison, *Soc. Netw.* **8**, 215 (1986).
- [31] F. Krzakala, C. Moore, E. Mossel, J. Neeman, A. Sly, L. Zdeborová, and P. Zhang, *Proc. Natl. Acad. Sci. USA* **110**, 20935 (2013).
- [32] H. Bass, *Int. J. Math.* **3**, 717 (1992).
- [33] K.-I. Hashimoto, *Automorphic Forms and Geometry of Arithmetic Varieties* (Elsevier, Amsterdam, 1989), pp. 211–280.
- [34] M. D. Horton, H. Stark, and A. A. Terras, *Contemp. Math.* **415**, 173 (2006).
- [35] This is in the form of a quadratic eigenvalue problem (QEP).
- [36] L. Donetti and M. A. Muñoz, *J. Stat. Mech.* (2004) P10012.
- [37] P. Ren, R. C. Wilson, and E. R. Hancock, *IEEE Trans. Neural Netw.* **22**, 233 (2011).
- [38] D. Constantine and J.-F. Lafont, *J. Topol. Anal.* **1** (2018).
- [39] Alternatively we could set $F(r, \theta) = \infty$ so that all distances to the empty graph are infinite.
- [40] A. L. Traud, P. J. Mucha, and M. A. Porter, *Physica A* **391**, 4165 (2012).
- [41] C. Roth, S. M. Kang, M. Batty, and M. Barthélemy, *J. R. Soc. Interface* **9**, 2540 (2012).
- [42] J. Leskovec, J. Kleinberg, and C. Faloutsos, in *Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining* (ACM, New York, 2005), pp. 177–187.
- [43] M. Huss and P. Holme, *IET Syst. Biol.* **1**, 280 (2007).
- [44] L. v. d. Maaten and G. Hinton, *J. Mach. Learn. Res.* **9**, 2579 (2008).
- [45] N. S. Altman, *Amer. Stat.* **46**, 175 (1992).
- [46] L. N. Trefethen and M. Embree, *Spectra and Pseudospectra: The Behavior of Nonnormal Matrices and Operators* (Princeton University Press, Princeton, NJ, 2005).
- [47] S. Chen, A. Mira, and J.-P. Onnela, Flexible model selection for mechanistic network models via super learner, *arXiv:1804.00237* (2018).
- [48] P. Grindrod, D. J. Higham, and V. Noferini, *SIAM J. Matrix Anal. Appl.* **39**, 310 (2018).
- [49] A. Saade, F. Krzakala, and L. Zdeborová, *Advances in Neural Information Processing Systems* (MIT Press, Cambridge, MA, 2014), pp. 406–414.
- [50] L. Babai, in *Proceedings of the 48th annual ACM Symposium on Theory of Computing* (ACM, New York, 2016), pp. 684–697.