Fast Computing von Neumann Entropy for Large-scale Graphs via Quadratic Approximations

Hayoung Choi,1,* Jinglian He,1,† Hang Hu,1,‡ and Yuanming Shi1,§

1School of Information Science and Technology, ShanghaiTech University

(Dated: November 28, 2018)

The von Neumann graph entropy (VNGE) can be used as a measure of graph complexity, which can be the measure of information divergence and distance between graphs. Since computing VNGE is required to find all eigenvalues, it is computationally demanding for a large-scale graph. We propose novel quadratic approximations for computing the von Neumann graph entropy. Modified Taylor and Radial projection approximations are proposed. Our methods reduce the cubic complexity of VNGE to linear complexity. Computational simulations on random graph models and various real network datasets demonstrate the superior performance.

PACS numbers: 89.70.CF, 89.75.Fb, 03.67.-a

I. INTRODUCTION

Immense data are involved in human life today. As these data grow increasingly intricate, a decent representation of its anomalous and complicated structure is needed. Graphs have the sophisticated capability of representing and summarizing irregular structural features. Hence, graph-based learning become an emerging and promising method with numerous applications in social network, biology, physics, and chemistry [1–4].

With the boosting quantity of graph data, it is crucial to find a way to manage them effectively. Therefore, graph similarity as a typical way of presenting the relationship between graphs have been vastly applied [5, 6]. For example, Sadreazami et al. proposed an intrusion detection methodology based on learning graph similarity with a graph Laplacian matrix [7]; Lin and Kung presented MatchEdge function to score similarity between two directed acyclic graphs [8]; also, Yanardag and Vishwanathan gave a general framework to smooth graph kernels based on graph similarity [9]. However, all the above approaches relied on presumed models, and thus limited their ability of being applied on comprehending the general concept of divergences and distances between graphs.

Meanwhile, graph entropy [10] which is a model-free approach has been actively used as a way to quantify the structural complexity of a graph [11]. By regarding the eigenvalues of the normalized combinatorial Laplacian of a graph as a probability distribution, we can generate its Shannon entropy.

By giving the density matrix of a graph, which is the representation of a graph in a quantum mechanical state, we can calculate its von Neumann entropy [12]. Bai et al. proposed algorithm to solve depth-based complexity characterization of graph using von Neumann entropy [13]; Liu et al. gave method of detecting bifurcation network event based on von Neumann entropy [14]. Other applications including graph clustering [15], network analysis [16], and structural reduction of multiplex networks [17]. But the time complexity of calculating exact value is high.

The von Neumann entropy, which was introduced by John von Neumann, is the extension of classical entropy concepts to the field of quantum mechanics [18]. He introduced the notion of the density matrix, which facilitated the extension of the tools of classical statistical mechanics to the quantum domain in order to develop a theory of quantum measurements.

Denote the trace of a square matrix $A$ as $\text{tr}(A)$. A density matrix is a Hermitian positive semidefinite in $\mathbb{R}^{n \times n}$ with unite trace. The density matrix $\rho$ is a matrix that describes the statistical state of a system in quantum mechanics. The density matrix is especially useful for dealing with mixed states, which consist of a statistical ensemble of several different quantum systems.

The von Neumann entropy of a density matrix $\rho$, denoted by $H(\rho)$, is defined as

$$\text{H}(\rho) = -\text{tr}(\rho \ln \rho) = -\sum_{i=1}^{n} \lambda_i \ln \lambda_i,$$

where $\lambda_1, \ldots, \lambda_n$ are eigenvalues of $\rho$. It is conventional to define $0 \ln 0 = 0$. This definition is a proper extension of both the Gibbs entropy and the Shannon entropy to the quantum case.

A. Von Neumann Graph Entropy

In this article we consider only undirected simple graphs with non-negative edge weights. Let $G = (V, E, W)$ denote a graph with the set of vertices $V$ and the set of edges $E$, and the weight matrix $W$. The combinatorial graph Laplacian matrix of $G$ is defined as $L(G) = S - W$, where $S$ is a diagonal matrix and its diagonal entry $s_i = \sum_{j=1}^{n} W_{ij}$ [19]. The density matrix
of a graph $G$ is defined as
\[
\rho_G = \frac{1}{\text{tr}(L)} L,
\]
where $\frac{1}{\text{tr}(L)}$ is a trace normalization factor. Note that $\rho_G$ is a positive semidefinite matrix with unit trace. The von Neumann entropy for a graph $G$, denoted by $H(G)$, is defined as
\[
H(G) := H(\rho_G),
\]
where $\rho_G$ is the density matrix of $G$ [20, 21].

It can be proven that for any graph $G$ with $|V| = n$, it holds that $H(G) \leq \ln(n-1)$, and the equality holds when $G$ is a complete graph [22]. Obviously, the definition of von Neumann graph entropy is based on the combinatorial graph Laplacian matrix. There are variants of von Neumann graph entropy proposed in the prior literature, including the normalized graph Laplacian matrix [23, 24] and the generalized graph Laplacian matrix of directed graphs [25, 26].

However, these alternatives fail to consider approximation justification and are also shown to be suboptimal. Although various definitions have been proposed, computing the exact value based on different definition is still expensive, especially for a large-scale graph. Computing von Neumann graph entropy requires the entire eigenspectrum $\{\lambda_i\}_{i=1}^n$ of $\rho_G$. This calculation can be done with time complexity $O(n^3)$ [27, 28], making it computationally impractical for large-scale graphs. It is required to find an efficient method to compute von Neumann entropy for large-scale graphs. Although the von Neumann graph entropy have been proved to be an feasible approach in the computation of Jensen-Shannon distance between any two graphs from a graph sequence [17]. However, in the process of machine learning and data mining tasks, a sequence of large graphs will be involved. Therefore, it is of great significance to improve numerical algorithms that approximate the von Neumann entropy of large-scale graphs faster than the previous $O(n^3)$ approach. More details about the application of Jensen-Shannon distance will be shown in Section IV. To tackle this challenge about computational inefficiency, Chen et al. [29] proposed a fast algorithm for computing von Neumann graph entropy, which uses the quadratic polynomial to approximate the term $-\lambda_i \ln \lambda_i$ rather than extracting the eigenspectrum $\{\lambda_i\}_{i=1}^n$ of $\rho_G$. It was shown that the proposed approximation is more efficient than the exact algorithm based on the singular value decomposition. Our work was inspired by their approach. However, our proposed method have superior performance in random graphs as well as real datasets with time complexity $O(n + m)$.

II. QUADRATIC APPROXIMATIONS

For a Hermitian matrix $A \in \mathbb{C}^{n \times n}$ it is true that $\text{tr} f(A) = \sum_i f(\lambda_i)$ where $\lambda_i$ are all eigenvalues of $A$.

Since $H(\rho) = -\text{tr}(\rho \ln \rho)$, one natural approach to approximate the von Neumann entropy of a density matrix is to use a Taylor series expansion to approximate the logarithm of a matrix. It is required to calculate $\text{tr}(\rho^j)$, $j \leq N \in \mathbb{N}$ for some $N$. Indeed, $\ln(I_n - A) = -\sum_{j=1}^\infty A^j / j$ for a symmetric matrix $A$ whose eigenvalues are all in the interval $(-1, 1)$. Assuming that all eigenvalues of a density matrix $\rho$ are nonzeros, we have
\[
H(\rho) = -\ln \lambda_{\text{max}} + \sum_{j=1}^\infty \frac{1}{j} \text{tr}(\rho(I_n - (\lambda_{\text{max}})^{-1}\rho)^j),
\]
where $\lambda_{\text{max}}$ is the maximum eigenvalue of $\rho$. We refer to [30] for more details. However, the computational complexity is $O(n^3)$, so it is impractical as $n$ grows.

In this article, we propose quadratic approximations to approximate the von Neumann entropy for large-scale graphs. It is noted that only $\text{tr}(\rho^2)$ is needed to compute them. We consider various quadratic polynomials $f_{\text{app}}(x) = c_2 x^2 + c_1 x + c_0$ to approximate $f(x) = -x \ln x$ on $(0, 1]$ ($f(0) = 0$). Then $H(\rho) = \text{tr}(f(\rho)) \approx \text{tr}(f_{\text{app}}(\rho)) = \sum_{i=1}^n f_{\text{app}}(\lambda_i) = c_2 \text{tr}(\rho^2) + c_1 \rho_0 + c_0$. Such approximations are required to be considered on the only values in $[0, 1]$ such that their sum is 1, which are eigenvalues of a given density matrix.

Recall that $\text{tr}(\rho^2)$ is called purity of $\rho$ in quantum information theory. The purity gives information on how much a state is mixed. We denote the purity of $\rho$ as $\gamma(\rho^2)$ or simply $\gamma$. For a given graph $G$, the purity of $\rho_G$ can be computed efficiently due to the sparsity of the $\rho_G$ as follows.

Lemma 1. For a graph $G = (V, E, W) \in \mathcal{G}$,
\[
\text{tr}(\rho_G^2) = \frac{1}{(\text{tr}(L))^2} \left( \sum_{i \in V} S_i^2 + 2 \sum_{(i,j) \in E} L_{ij}^2 \right).
\]

Proof. Since $L, \rho_G$ are symmetric, it follows that
\[
\text{tr}(\rho_G^2) = \|\rho_G\|^2_F = \frac{1}{(\text{tr}(L))^2} \|L\|^2_F
= \frac{1}{(\text{tr}(L))^2} \sum_{1 \leq i,j \leq n} L_{ij}^2
= \frac{1}{(\text{tr}(L))^2} \left( \sum_{1 \leq i \leq n} L_i^2 + 2 \sum_{1 \leq i < j \leq n} L_{ij}^2 \right).
\]

It is trivial that $\text{tr}(\rho_G^2)$ only depends on the edge weights in $G = (V, E, W)$, resulting in linear computation complexity $O(n + m)$, where $|V| = n$ and $|E| = m$.

We denote the maximum eigenvalue of a given density matrix as $\lambda_{\text{max}}$. When the eigenvalues of $\rho$ are given, we denote $\hat{H}(\rho)$ and $\gamma(\rho)$ as $\hat{H}(\lambda_1, \ldots, \lambda_n)$ and $\gamma(\lambda_1, \ldots, \lambda_n)$, respectively. We will use them interchangeably.
**Theorem 1.** The following are true.

1. \(2\lambda_{\text{max}}(1 - \lambda_{\text{max}}) \leq H(\lambda_{\text{max}}, 1 - \lambda_{\text{max}}) \leq H(\rho).\)
2. \(1/n \leq \gamma(\rho) \leq \gamma(\lambda_{\text{max}}, 1 - \lambda_{\text{max}}).\)
3. \(\text{tr}(-\rho^2 \ln \rho) \leq -\gamma \ln \gamma.\)

**Proof.** Let \(\lambda_n \leq \ldots \leq \lambda_1\) be eigenvalues of a given density matrix \(\rho\).

1. Although this is one of known properties, we provide a proof. Let \(f(x) = -x \ln x\). It is true that \(f(p_1 + p_2) \leq f(p_1) + f(p_2)\) for all \(p_1, p_2 \geq 0\), since
   \[
   f(p_1) + f(p_2) - f(p_1 + p_2) = -p_1 \ln p_1 - p_2 \ln p_2 + (p_1 + p_2) \ln (p_1 + p_2)
   = p_1 \ln (p_1 + p_2) - \ln p_1 + p_2 \ln (p_1 + p_2) - \ln p_2 \geq 0.
   \]
   Thus, by induction we have
   \[
   f(\lambda_2 + \cdots + \lambda_n) \leq f(\lambda_2) + \cdots + f(\lambda_n).
   \]
   Therefore, \(f(\lambda_1) + f(\lambda_2 + \cdots + \lambda_n) \leq f(\lambda_1) + \cdots + f(\lambda_n) = H(\rho)\).
   The first inequality holds from the fact that \(\lambda_{\text{max}} - 1 \geq \ln \lambda_{\text{max}}\). Note that \(x - 1 \geq \ln x \) for all \(x > 0\). Thus,
   \[
   -\lambda_{\text{max}} \ln \lambda_{\text{max}} - (1 - \lambda_{\text{max}}) \ln (1 - \lambda_{\text{max}})
   \geq -\lambda_{\text{max}}(\lambda_{\text{max}} - 1) - (1 - \lambda_{\text{max}})(-\lambda_{\text{max}})
   = 2\lambda_{\text{max}}(1 - \lambda_{\text{max}}).
   \]
   (2) Since \(\text{tr}(\rho^2) \leq \lambda_1^2 + (\lambda_2 + \cdots + \lambda_n)^2\), clearly \(\text{tr}(\rho^2) \leq \lambda_1^2 + (1 - \lambda_1)^2\).
   (3) It is known that the purity satisfies \(1/n \leq \gamma \leq 1\). Since \(f(x) = -x \ln x\) is concave on \([0, 1]\), by Jensen’s inequality, it follows that
   \[
   f(\gamma) = f \left( \sum_{i=1}^{n} \lambda_i^2 \right) \geq \sum_{i=1}^{n} \lambda_if(\lambda_i) = \sum_{i=1}^{n} -\lambda_i^2 \ln \lambda_i,
   \]
   where \(\lambda_i\) are the spectrum of \(\rho\). Note that the last term can be expressed as \(\text{tr}(-\rho^2 \ln \rho)\).

\(\square\)

In this article we consider the following quadratic approximations for von Neumann entropy: (A) FINGER-\(\hat{H}\); (B) Taylor-\(\hat{T}\); (C) Modified Taylor-\(\tilde{T}\); (D) Radial Projection-\(\hat{R}\). Note that they all can be computed by the purity of density matrix of a graph. Additionally (A) and (C) need the maximum eigenvalue as well.

![FIG. 1: FINGER-\(\hat{H}\): graphs of the quadratic function in (1) with different \(\lambda_{\text{max}}\) are shown.](image)

**A. FINGER-\(\hat{H}\)**

Chen et al.\[29\] proposed a fast incremental von Neumann graph entropy (FINGER) to reduce the cubic complexity of von Neumann graph entropy to linear complexity. They considered the following quadratic polynomial to approximate \(f(x) = -x \ln x\) (see Fig. 1).

\[
q(x) = -(\ln \lambda_{\text{max}})(1-x) \quad \text{on } [0, 1].
\]

Then FINGER, denoted by \(\hat{H}\), is defined as

\[
\hat{H}(G) := \sum_{i=1}^{n} q(\lambda_i) = -(\ln \lambda_{\text{max}})(1-\text{tr}(\rho_G^2)),
\]

where \(\lambda_1, \ldots, \lambda_n\) are the eigenvalues of \(\rho_G\).

Note that since all eigenvalues are smaller than or equal to the maximum eigenvalue, we mainly consider the interval \([0, \lambda_{\text{max}}]\) instead of \([0, 1]\). Now let us show that the approximation FINGER is always smaller than the exact von Neumann entropy.

**Lemma 2.** The following are true.

1. \(f - q\) is concave on \([0, \lambda_{\text{max}}]\).
2. \(f(x) - q(x) \geq \lambda_{\text{max}}f(x)\) on \([0, \lambda_{\text{max}}]\).

**Proof.** It is trivial for \(\lambda_{\text{max}} = 1\). Suppose that \(\lambda_{\text{max}} \neq 1\).

1. Note that \(f(x) < 1/e < 1/2\) on all \(0 \leq x \leq 1\). So, 
   \(\lambda_{\text{max}} \ln \lambda_{\text{max}} < 1/2\). Then \(\lambda_{\text{max}} < \frac{1}{2 \ln \lambda_{\text{max}}}\), implying, 
   \(1 - 2x\ln \lambda_{\text{max}} \leq 0\) for all \(0 \leq x \leq \lambda_{\text{max}}\). Thus,
   \[
   (f - q)(x) = \frac{-1 - 2x\ln \lambda_{\text{max}}}{x} < 0.
   \]
(2) When \( x = 0 \) or \( x = \lambda_{\text{max}} \), it is trivial. Since \((1 - \lambda_{\text{max}}) f(x) - q(x) = x(\lambda_{\text{max}} - 1) \ln x + (\ln \lambda_{\text{max}})(1 - x)\), it suffices to show \( \phi(x) = (\lambda_{\text{max}} - 1) \ln x + \ln \lambda_{\text{max}}(1 - x) > 0 \) for all \( 0 < x < \lambda_{\text{max}} \). By observing the tangent line at \( x = 1 \) for \( y = x \ln x \), it is easy to check that \( \lambda_{\text{max}} - 1 < \lambda_{\text{max}} \ln \lambda_{\text{max}} \). Then
\[
\frac{\lambda_{\text{max}} - 1}{\ln \lambda_{\text{max}}} > \lambda_{\text{max}}.
\]
Thus, \( \phi'(x) = \frac{\lambda_{\text{max}} - 1}{\ln \lambda_{\text{max}}} - \ln \lambda_{\text{max}} < 0 \) for all \( 0 < x < \lambda_{\text{max}} \). Since \( \phi(\lambda_{\text{max}}) = 0 \) and \( \phi'(x) < 0 \) for all \( 0 < x < \lambda_{\text{max}} \), it follows that \( \phi(x) > 0 \) for all \( 0 < x < \lambda_{\text{max}} \).

\[\square\]

**Theorem 2.** For any density matrix \( \rho \), it holds that

1. \( \hat{H}(\lambda_1, \ldots, \lambda_n) \geq \hat{H}(\lambda_{\text{max}}, 1 - \lambda_{\text{max}}) \).
2. \( H(\rho) \geq \hat{H}(\rho) \) The equality holds if and only if \( \lambda_{\text{max}} = 1 \).
3. \( H(\rho) - \hat{H}(\rho) \geq \lambda_{\text{max}}H(\rho) \).

**Proof.** (1) Let \( \phi(x) = x(1 - x) \). Then it is easy to check that \( \phi(t_1) + \phi(t_2) - \phi(t_1 + t_2) \geq 0 \) for all \( t_1, t_2 \geq 0 \). Thus, \( q(t_1 + t_2) \leq q(t_1) + q(t_2) \). By induction, it holds that
\[
q(t_1 + \cdots + t_n) \leq q(t_1) + \cdots + q(t_n)
\]
for all \( t_i \geq 0, i = 1, \ldots, n \). Then it follows that
\[
\hat{H}(\lambda_1, \ldots, \lambda_n) = \sum_{i=1}^{n} q(\lambda_i) = q(\lambda_1) + \sum_{i=2}^{n} q(\lambda_i)
\leq q(\lambda_1) + q\left(\sum_{i=2}^{n} \lambda_i\right)
= q(\lambda_{\text{max}}) + q(1 - \lambda_{\text{max}}),
\]
where \( \lambda_1 \geq \cdots \geq \lambda_n \) are spectrum of \( \rho \).

(2) Let \( \lambda \) be an eigenvalue of \( \rho \). Since \( \lambda \leq \lambda_{\text{max}} \), it follows that
\[
f(\lambda) - q(\lambda) = -\lambda \ln \lambda + (\ln \lambda_{\text{max}})\lambda(1 - \lambda)
\geq -\lambda \ln \lambda + (\ln \lambda)\lambda(1 - \lambda)
= -\lambda^2 \ln \lambda \geq 0.
\]
Note that \( 0 = \lambda_n = \cdots = \lambda_2 \) and \( \lambda_1 = 1 \) if and only if \( \lambda_i^2 \ln \lambda_i = 0 \) for all \( 1 \leq i \leq n \).

(3) By Lemma 2 (2) we can find the following error bound for FINGER.

**Remark.** By Theorem 2 (1) and Theorem 1 (1) it holds that \( H(\rho) \geq H(\lambda_{\text{max}}, 1 - \lambda_{\text{max}}) \) and \( \hat{H}(\rho) \geq \hat{H}(\lambda_{\text{max}}, 1 - \lambda_{\text{max}}) \). However, it is questionable if it holds that \( H(\rho) - \hat{H}(\rho) \geq H(\lambda_{\text{max}}, 1 - \lambda_{\text{max}}) - \hat{H}(\lambda_{\text{max}}, 1 - \lambda_{\text{max}}) \). If this is true, then it is possible to show more interesting result as follows:

\[
H(\rho) - \hat{H}(\rho) = \sum_{i=1}^{n} f(\lambda_i) - q(\lambda_i)
\geq f(\lambda_{\text{max}}) - q(\lambda_{\text{max}}) + f(1 - \lambda_{\text{max}}) - q(1 - \lambda_{\text{max}})
= -\lambda_{\text{max}} \ln \lambda_{\text{max}} + \lambda_{\text{max}}(1 - \lambda_{\text{max}}) \ln \lambda_{\text{max}}
- (1 - \lambda_{\text{max}}) \ln (1 - \lambda_{\text{max}}) + \lambda_{\text{max}}(1 - \lambda_{\text{max}}) \ln (1 - \lambda_{\text{max}})
= -(1 - \lambda_{\text{max}})^2 \ln (1 - \lambda_{\text{max}}) - (\lambda_{\text{max}})^2 \ln \lambda_{\text{max}}
\geq \frac{\ln 2}{2}.
\]
That is, \( \hat{H} \) always has error greater than \( \ln \sqrt{2} \). For the dataset in Fig. 5 it shows that the smallest error for the real dataset is 0.6397 > \( \ln \sqrt{2} \).

The following results were established in [29].

**Theorem 3.** For any \( G \in \mathcal{G} \), let \( \lambda_{\text{max}} \) and \( \lambda_{\text{min}} \) be the largest and smallest positive eigenvalue of \( \rho_G \), respectively. If \( \lambda_{\text{max}} < 1 \), then
\[
-(1 - \text{tr} (\rho_G^2)) \frac{\ln \lambda_{\text{max}}}{1 - \lambda_{\text{min}}} \leq H(G) \leq -(1 - \text{tr} (\rho_G^2)) \frac{\ln \lambda_{\text{min}}}{1 - \lambda_{\text{max}}}.
\]
The bounds become exact and \( H = \ln(n - 1) \) when \( G \) is a complete graph with identical edge weight.

**Corollary 1.** For any \( G \in \mathcal{G} \), let \( n_+ \) denote the number of positive eigenvalues of \( \rho_G \). If \( n_+ \sim n \) and \( \lambda_{\text{max}} \sim \lambda_{\text{min}} \), then
\[
\frac{H}{\ln n} - (1 - \text{tr} (\rho_G^2)) \to 0 \text{ as } n \to \infty.
\]

**B. Taylor-T**

Since the sum of all eigenvalues of density matrix is 1, the average of them is 1/n. As \( n \) gets bigger, the average gets closer to 0. Thus, for a large-scale \( n \times n \) density matrix, many of eigenvalues must be on \([0, 1/n]\). So, it is reasonable to use Taylor series for \( f \) at \( x = 1/n \) instead of \( x = 0 \). In fact, since \( f'(0) \) does not exist, there does not exist Taylor series for \( f \) at \( x = 0 \).

**Lemma 3.** Let \( \lambda_{\text{max}} \) be the maximum eigenvalue of a density matrix \( \rho \). Then, \( \frac{1}{n} \leq \lambda_{\text{max}} \leq 1 \). Especially, \( \lambda_{\text{max}} = 1 \) if and only if \( H(\rho) = 0 \). Also it is ture that \( \lambda_{\text{max}} = \frac{1}{n} \) if and only if \( H(\rho) = \ln(n) \).

**Proof.** Since \( \text{tr} \rho = 1 \), if \( \lambda_{\text{max}} = 1 \) then all eigenvalues are 0 except \( \lambda_{\text{max}} \). It is known that \( \hat{H}(\rho) \leq \ln(n) \) and the equality holds when \( \lambda_i = 1/n \) for all \( i \). \[\square\]
We can propose the quadratic Taylor approximation for $f$ at $x = 1/n$ as follows.

$$q(x) = f\left(\frac{1}{n}\right) + f'(\frac{1}{n}) \left(x - \frac{1}{n}\right) + \frac{f''(\frac{1}{n})}{2!} \left(x - \frac{1}{n}\right)^2$$

$$= -\frac{n}{2} x^2 + (\ln n) x + \frac{1}{2n}.$$ 

Using such approximation, Taylor, denoted by $T$, is defined as

$$T(G) = -\frac{n}{2} \text{tr}(\rho^2_G) + \ln n - \frac{1}{2}.$$ 

As Fig. 2 shows, The function $q$ is very similar to the function $f$ near $x = 1/n$. However, as the maximum eigenvalue gets closer to 1, the error becomes very large. Note that $f(\lambda_{\text{max}}) - q(\lambda_{\text{max}}) = -\lambda_{\text{max}} \ln \lambda_{\text{max}} + \frac{n}{2} \lambda_{\text{max}}^2 - (\ln n) \lambda_{\text{max}} - \frac{1}{2n} \to \infty$ as $n \to \infty$ for $\lambda_{\text{max}} \approx 1$. Alternatively, this approximation needs to be modified. We use the information about $\lambda_{\text{max}}$ in order to reduce the error.

C. Modified Taylor-$\hat{T}$

Consider the quadratic approximation, $q$, to approximate $f(x) = -x \ln x$ such that it holds

$$q\left(\frac{1}{n}\right) = f\left(\frac{1}{n}\right), \quad q'\left(\frac{1}{n}\right) = f'\left(\frac{1}{n}\right), \quad q(\lambda_{\text{max}}) = f(\lambda_{\text{max}}),$$

assuming that $\lambda_{\text{max}} \neq \frac{1}{n}$, we have

$$q(x) = \sigma x^2 + \left(\ln n - 1 - \frac{2\sigma}{n}\right)x + \frac{\sigma + n}{n^2}, \quad (2)$$

where

$$\sigma = -n\lambda_{\text{max}} \ln \left(n\lambda_{\text{max}} + n\lambda_{\text{max}} - 1 \right) \frac{1}{n(\lambda_{\text{max}} \frac{1}{n})^2}.$$ 

Using such approximation, the Modified Taylor, denoted by $\hat{T}$, is defined as

$$\hat{T}(G) = \sigma \left(\text{tr}(\rho^2_G) - \frac{1}{n}\right) + \ln n.$$ 

Lemma 4. The following are true.

(1) $q$ is concave on $[0, \lambda_{\text{max}}]$.

(2) Let $t = n\lambda_{\text{max}}$. By Lemma 3, $t > 1$. Since

$$\frac{1}{1 - 2\sigma} = \frac{1}{n} - \frac{(t - 1)^2}{2(t \ln t - t + 1) - 1},$$ 

it is enough to show that $(t - 1)^2 > 2(t \ln t - t + 1)$. Recall that $t \ln t - t + 1 > 0$ for all $t > 1$.

Let $\phi(t) = t^2 - 2t \ln t$. Since $t > 1$, $\phi(t) = 2t - 2 \ln t - t > 2(t - 1) - \ln t > 0$ for all $t > 1$. From the fact $\phi(1) = 0$, clearly, $\phi(t) > 0$ for all $t > 1$. So, $\frac{1}{n} \leq - \frac{1}{2\sigma}$. Taking $t = n\lambda_{\text{max}}$, we have

$$\lambda_{\text{max}} - \frac{1}{1 - 2\sigma} = \frac{t}{n} - \frac{(t - 1)^2}{2(t \ln t - t + 1) - 1} \geq \frac{2(1 - t)}{2n(t \ln t - t + 1) - n(t - 1)^2}.$$ 

Theorem 4. For any density matrix $\rho$, it holds that

$$\hat{T}(\rho) \geq H(\rho).$$ 

Proof. Let $h(x) = q(x) - f(x)$. Clearly, $h(1/n) = h(\lambda_{\text{max}}) = 0$ and $h'(1/n) = 0$. We show that $h(x) \geq 0$ for three different intervals: (i) $[0, \frac{1}{2\sigma}]$, (ii) $[\frac{1}{2\sigma}, \frac{1}{4\sigma}]$, and (iii) $[\frac{1}{4\sigma}, \lambda_{\text{max}}]$. (i) Since $h'' = 2\sigma + \frac{1}{4\sigma}$. Lemma 4 (2) implies that $h$ is convex on $[0, \frac{1}{2\sigma}]$. Since $h(1/n) = h'(1/n) = 0$, $h(x) \geq 0$ on $[0, \frac{1}{2\sigma}]$. (ii) In the similar way, it holds that $h(x) \geq 0$ on $[\frac{1}{2\sigma}, \frac{1}{4\sigma}]$. (iii) Since $h$ is concave on $[\frac{1}{4\sigma}, \lambda_{\text{max}}]$, by the definition of concavity, it follows that

$$h\left(-\frac{1}{2\sigma} t + \lambda_{\text{max}}(1-t)\right) \geq h\left(-\frac{1}{2\sigma} \right) t + h(\lambda_{\text{max}})(1-t) \geq 0$$ 

for all $0 \leq t \leq 1$. The last inequality holds from the fact that $h(-\frac{1}{2\sigma}) \geq 0$ and $h(\lambda_{\text{max}}) = 0$. Thus, $h(x) \geq 0$ on $[\frac{1}{4\sigma}, \lambda_{\text{max}}]$. Therefore, by (i), (ii), (iii) it holds that $h(x) \geq 0$ on $[0, \lambda_{\text{max}}]$.
D. Radial projection-R

We denote the simplex of positive probability as \( \Delta_n \), i.e., \( \Delta_n := \{ (\lambda_1, \lambda_2, \ldots, \lambda_n) \mid \sum_{i=1}^{n} \lambda_i = 1, \lambda_i \geq 0 \} \). Also we denote \( \mathcal{E} := (1/n, \ldots, 1/n) \). Clearly, \( \mathcal{E} \in \Delta_n \). The Shannon entropy of \( \lambda \in \Delta_n \) is defined as \( S(\Lambda) = -\sum_{i=1}^{n} \lambda_i \ln \lambda_i \). It is well-known that it holds \( S(\Lambda) \leq \ln n \), with equality if \( \Lambda = \mathcal{E} \). That is, \( \mathcal{E} \) is the only point where the entropy is maximum.

As Fig. 3 is shown, the simplex of positive probability \( \Delta_3 \) can be geometrically a part of plane in \( \mathbb{R}^3 \). The color stands for the value of Shannon entropy at each point. One can see that as \( \Lambda \in \Delta_3 \) gets closer to \( \mathcal{E} \), the entropy gets bigger and bigger. Our main observation is that if two points on \( \Delta_3 \) have same (Euclidean) distances from \( \mathcal{E} \), then their purity are same. In general it holds for any \( n \).

**Lemma 5.** Let \( \Lambda = (\lambda_1, \lambda_2, \ldots, \lambda_n) \in \Delta_n \). Then the following are true.

1. \( ||E - \Lambda||^2 = \sum_{i=1}^{n} \lambda_i^2 - \frac{1}{n} \).
2. \( \langle E, \mathcal{E} - \Lambda \rangle = 0 \). (\( \langle \cdot, \cdot \rangle \) is the usual inner product.)

**Proof.** It is easy to check that

\[
||E - \Lambda||^2 = \sum_{i=1}^{n} \left( \lambda_i - \frac{1}{n} \right)^2 = \sum_{i=1}^{n} \lambda_i^2 - \frac{1}{n}.
\]

Lemma 5 states that two points on \( \Delta_n \) have the same distance from \( \mathcal{E} \) if and only if the distances from the origin are identical. Then we can find \( \tilde{\Lambda} \in \Delta_n \) whose entropy can be computed much easily such that

\[
||E - \Lambda||_2 = ||E - \tilde{\Lambda}||_2.
\]

There are infinitely many directions from \( \mathcal{E} \) to find \( \tilde{\Lambda} \in \Delta_n \). Among them we pick \( \mathbf{w} = (c, \frac{1-c}{n-1}, \ldots, \frac{1-c}{n-1}) \in \Delta_n \) with \( c = \lambda_{\text{max}} \).

We consider the line segment \( \ell(t) = (\mathbf{w} - \mathcal{E})t + \mathcal{E} \), \( 0 \leq t \leq 1 \), i.e.,

\[
\ell(t) = \left( \frac{1-t}{n} + tc, \frac{1-t}{n} + \frac{t(1-c)}{n-1}, \ldots, \frac{1-t}{n} + \frac{t(1-c)}{n-1} \right).
\]

Since \( \Delta_n \) is convex, \( \ell(t) \in \Delta_n \) for all \( 0 \leq t \leq 1 \). For each \( 0 \leq t \leq 1 \), we have

\[
S(\ell(t)) = -\left( \frac{1-t}{n} + tc \right) \ln \left( \frac{1-t}{n} + tc \right) - (n-1) \left( \frac{1-t}{n} + \frac{t(1-c)}{n-1} \right) \ln \left( \frac{1-t}{n} + \frac{t(1-c)}{n-1} \right).
\]

Lemma 5 implies that

\[
||E - \ell(t)||^2 = t^2 ||\mathbf{w} - \mathcal{E}||^2 = \frac{(cn-1)^2}{n(n-1)} t^2.
\]

We solve the following equation for \( 0 \leq t \leq 1 \):

\[
||E - \Lambda||_2 = ||E - \ell(t)||_2.
\]

Then the solution, say \( t_0 \), is

\[
t_0 = \frac{\sqrt{n(n-1)}}{cn-1} \sqrt{\sum_{i=1}^{n} \lambda_i^2 - \frac{1}{n}}.
\]

Since \( ||E - \Lambda||_2 = ||E - \ell(t_0)||_2 \), we have that

\[
S(\Lambda) \approx S(\ell(t_0)) = -\left( \frac{1-t_0}{n} + t_0c \right) \ln \left( \frac{1-t_0}{n} + t_0c \right) - (n-1) \left( \frac{1-t_0}{n} + \frac{t_0(1-c)}{n-1} \right) \ln \left( \frac{1-t_0}{n} + \frac{t_0(1-c)}{n-1} \right).
\]

In fact, putting \( t_0 \) into the right side the constant \( c \) can be cancelled. Thus, this approximation does not need the maximum eigenvalue.

In the similar way we propose the quadratic approximation for the von Neumann graph entropy, called Radial Projection. Radial projection, denoted by \( R \), is defined as

\[
R(G) = -\left( \sqrt{\frac{n-1}{n} \kappa_G + \frac{1}{n}} \right) \times \ln \left( -\sqrt{\frac{n-1}{n} \kappa_G + \frac{1}{n}} \right) \left( -\frac{1}{\sqrt{n(n-1) \kappa_G + \frac{1}{n}}} \right),
\]

where \( \kappa_G = [\gamma(\rho_G) - \frac{1}{n}]^{1/2} \).
E. Weighted mean

Denote the weighted mean of \( a, b \in \mathbb{R} \) as \( a \# b \). That is, \( a \# b = ta + (1 - t)b \). In the similar way we consider \( H(\rho) \# T(\rho) \) and \( H(\rho) \# R(\rho) \). By Theorem 2 (2), it is shown that FINGER-\( T \) is always smaller than the exact von Neumann entropy. On the other hand, by Theorem 4 Modified Taylor-\( T \) is always greater than the exact von Neumann entropy. Even though it is not proved mathematically, Fig. 3, 5, 6 show that Radial projection-\( R \) are greater than the exact von Neumann entropy. The weighted mean of them can be computed to improve the approximations.

We solve a optimization problem to find optimal \( t^* \). For example, consider \( H \# T \). Given large quantity of real data sets, the approximation of von Neumann entropy using \( H \) and \( T \) were calculated as the input values \( x_T(i) \) and \( y_T(i) \) while the actual von Neumann entropy value were also calculated as output values \( y(i) \) for \( i = 1, \ldots, N \) (\( N \) is the number of data sets). Then the optimization problem is given as follows:

\[
  t^* = \arg \min_{0 \leq t \leq 1} J(t),
\]

where the cost function is

\[
  J(t) = \frac{1}{N} \sum_{i=1}^{N} (t x_T(i) + (1 - t) y_T(i) - y(i))^2.
\]

We use the gradient descent method to find optimal \( t^* \). Initially, \( t_0 = \frac{1}{2} \) is given. In each step of gradient descent, \( t_j \) is updated with the function:

\[
  t_j = t_{j-1} - \alpha J'(t_{j-1}),
\]

where \( j \) denotes iteration times, and \( \alpha \) is the step size, which is set to be \( 10^{-6} \).

Using the optimal values \( t^* \) solved by the gradient descent method, we call \( H \#_{0.3824} T \) and \( H \#_{0.2794} R \) as Improved Modified Taylor and Improved Radial Projection, respectively.

III. EXPERIMENTS

In this section results from various experiments with data sets are provided. All experiments were conducted by MATLAB R2016 on a 16-core machine with 128GB RAM.

A. Random graphs

Three random graph models are considered: (i) the Erdős-Rényi (ER) model \([31, 32]\) - the ER model represents two closely related models that were introduced independently and simultaneously. Here we use the ER model which was proposed by Gilbert \([32]\). \( G(n, p) \) is denoted as a model with \( n \) nodes, and each pair of nodes were linked independently with probability \( 0 \leq p \leq 1 \); (ii) the Barabási-Albert (BA) model \([33]\) - the BA model is a special case of the Price's model. It can generate scale-free graphs in which the degree distribution of the graph follows the power law distribution; and (iii) the Watts-Strogatz (WS) model \([34]\) - the WS model generates graphs with small world properties, given a network with \( N \) nodes and the mean degree \( K \), initially nodes are linked as a regular ring where each node is connected to \( K/2 \) nodes in each side, then rewire the edges with probability \( 0 \leq p \leq 1 \). The approximation error is defined as \([\text{Exact} - \text{Approximation}]\). The results are averaged over 10 random trials.

B. Real-world datasets

The real-world datasets in various fields are considered \([35-37]\). We use 137 different number of unweighted
networks and 48 different number of weighted networks in different fields for simulations. The detailed information about datasets on Fig. 5 and Fig. 6 can be found at [https://github.com/Hang14/RDJ](https://github.com/Hang14/RDJ). Fig. 5 and Fig. 6 show the scatter points of the von Neumann entropy (y-axis) versus the quadratic approximations (x-axis) for both the unweighted and weighted real-world datasets. It demonstrates that Modified Taylor and Radial Projection have better performances than FINGER. Improved Modified Taylor and Improved Radial Projection show the strongest performance.

C. Time comparison

Recall that computing von Neumann graph entropy requires \( O(n^3) \) computational complexity. In order to accelerate its computation, we use the quadratic approximations for the function \( f(x) = -x \ln x \). Then each approximation can be computed by the purity of the density matrix for a given graph. Lemma 1 shows that computing the purity requires \( O(n + m) \) computational complexity, where \( |V| = n \) and \( |E| = m \). However, FINGER and Modified Taylor additionally need to compute the maximum eigenvalue whose time complexity is \( O(n^2) \) [27, 28]. On the other hand, Radial projection shows the best performance with no maximum eigenvalue. When the original graph is a complete graph then \( |E| = n^2 \) which is also the upper bound of \( |E| \). However, in real life scenario, complete graph is rare, instead, sparse graph are more commonly seen, therefore, the time complexity can remain in a rather linear form.

IV. APPLICATIONS

One major application of von Neumann graph entropy (VNGE) is the computation of Jensen-Shannon distance (\( \text{JSdist} \)) between any two graphs from a graph sequence [17]. Given a graph sequence \( G \), the Jensen-Shannon distance of any two graphs \( G = (V, E, W) \in \mathcal{G} \) and \( G' = (V, E', W') \in \mathcal{G} \) is defined as

\[
\text{JSdist}(G, G') = \sqrt{\frac{1}{2}[H(G) + H(G')] - \frac{1}{4}H(\overline{G})},
\]

where \( \overline{G} = \frac{1}{2}G + G' \) is the averaged graph of \( G \) and \( G' \) such that \( \overline{W} = \frac{W + W'}{2} \). The Jensen-Shannon distance has been proved to be a valid distance metric in [38, 39].

The Jensen-Shannon distance has been applied into many fields including network analysis [17] and machine learning [40]. Especially, it can be used in anomaly detection and bifurcation detection [14, 29]. Demonstrated the validation of using FINGER-\( \mathbf{H} \) for computing VNGE. Comparing to the state-of-art graph similarity methods, FINGER-\( \mathbf{H} \) yields superior and robust performance for anomaly detection in evolving Wikipedia networks and router communication networks, as well as bifurcation analysis in dynamic genomic networks. Note that the simulations show that our proposed methods show better performance than FINGER-\( \mathbf{\bar{H}} \).

V. CONCLUSION

We proposed quadratic approximations for efficiently estimating the von Neumann entropy of large-scale graphs. It reduces the computation of VNGE from cubic complexity to linear complexity for a given graph.
FINGER, Taylor, Modified Taylor, and Radial Projection and considered. We found several theoretical results for each approximation including inequalities between approximations and the exact von Neumann entropy. Moreover, the novel quadratic approximation - Radial Projection was proposed. Even the Radial Projection does not need to compute the maximum eigenvalue. Moreover, using the weighted mean with optimal value $t^*$, we improved the approximations as well. Computational simulations demonstrated that the proposed method outperforms the state-of-art method for random graphs as well as real datasets.

[1] A. Ortega, P. Frossard, J. Kovaevi, J. M. F. Moura, and P. Vanderheynst, “Graph signal processing: Overview, challenges, and applications,” Proc. IEEE, vol. 106, no. 5, pp. 808–828, May 2018.
[2] M. M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, and P. Vanderheynst, “Geometric deep learning: Going beyond euclidean data,” IEEE Signal Process. Mag., vol. 34, no. 4, pp. 18–42, July 2017.
[3] A. Sandryhaila and J. M. F. Moura, “Discrete signal processing on graphs: Graph fourier transform,” in Proc. IEEE Int. Conf. Acoustics Speech Signal Process. (ICASSP), May 2013, pp. 6167–6170.
[4] C. Donnat and S. Holmes, “Tracking network dynamics: a survey of distances and similarity metrics,” ArXiv e-prints, Jan. 2018.
[5] D. Berberidis and G. B. Giannakis, “Data-adaptive active sampling for efficient graph-cognizant classification,” IEEE Trans. Signal Process., vol. 66, no. 19, pp. 5167–5179, Oct 2018.
[6] A. Kapoor, H. Ahn, Y. Qi, and R. W. Picard, “Hyperparameter and kernel learning for graph based semi-supervised classification,” in Adv. Neural Inf. Process. Syst. (NIPS), pp. 627–634, 2006.
[7] H. Sadreazami, A. Mohammadi, A. Asif, and K. N. Plataniotis, “Distributed-graph-based statistical approach for intrusion detection in cyber-physical systems,” IEEE Transactions on Signal and Information Processing over Networks, vol. 4, no. 1, pp. 137–147, March 2018.
[8] I. J. Lin and S. Y. Kung, “Coding and comparison of dag’s as a novel neural structure with applications to on-line handwriting recognition,” IEEE Transactions on Signal Processing, vol. 45, no. 11, pp. 2701–2708, Nov 1997.
[9] P. Yanardag and S. V. N. Vishwanathan, “A structural smoothing framework for robust graph-comparison,” in Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2, Cambridge, MA, USA, 2015, NIPS’15, pp. 2134–2142, MIT Press.
[10] K. Anand and G. Bianconi, “Entropy measures for networks: Toward an information theory of complex topologies,” Phys. Rev. E, vol. 80, pp. 045102, Oct 2009.
[11] A. Li and Y. Pan, “Structural information and dynamical complexity of networks,” IEEE Trans. Inf. Theory, vol. 62, no. 6, pp. 3290–3339, June 2016.
[12] K. Anand, G. Bianconi, and S. Severini, “Shannon and von neumann entropy of random networks with heterogeneous expected degree,” Phys. Rev. E, vol. 83, pp. 036109, Mar 2011.
[13] L. Bai, E. R. Hancock, L. Han, and P. Ren, “Graph clustering using graph entropy complexity traces,” in Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012), Nov 2012, pp. 2881–2884.
[14] S. Liu, P. Chent, I. Rajapakse, and A. Hero, “First-order bifurcation detection for dynamic complex networks,” in Proc. IEEE Int. Conf. Acoustics Speech Signal Process. (ICASSP), April 2018, pp. 6912–6916.
[15] L. Bai, E. R. Hancock, and P. Ren, “Jensen-shannon graph kernel using information functionals,” in Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012), Nov 2012, pp. 2877–2880.
[16] C. Ye, R. C. Wilson, and E. R. Hancock, “Network analysis using entropy component analysis,” Journal of Complex Networks, vol. 6, no. 3, pp. 404–429, Sept 2017.
[17] Manlio De Domenico, Vincenzo Nicosia, Alexandre Arenas, and Vito Latora, “Structural reducibility of multilayer networks,” Nature Communications, vol. 6, no. 1, apr 2015.
[18] B. C. Hall, Quantum theory for Mathematicians, vol. 267, Springer, 2013.
[19] U. V. Luxburg, “A tutorial on spectral clustering,” Stat. Comput., vol. 17, no. 4, pp. 395–416, 2007.
[20] S. L. Braunstein, S. Ghosh, and S. Severini, “The laplacian of a graph as a density matrix: A basic combinatorial approach to separability of mixed states,” Ann. Comb., vol. 10, no. 3, pp. 291–317, Dec 2006.
[21] F. Passerini and S. Severini, “Quantifying complexity in networks: The von neumann entropy,” Int. J. Agent Technol. Syst. (IJATS), vol. 1, no. 4, pp. 58–67, 2009.
[22] Filippo Passerini and Simone Severini, “The von neumann entropy of networks,” SSRN Electronic Journal, 2008.
[23] J. B. Shi and J. Malik, “normalized cuts and image segmentation”, ieee trans., IEEE Trans.Pattern Anal.mach.intell, vol. 22, no. 8, pp. 888–905, 2000.
[24] Lin Han, Francisco Escolano, Edwin R. Hancock, and Richard C. Wilson, “Graph characterizations from von neumann entropy,” Pattern Recognition Letters, vol. 33, no. 15, pp. 1958–1967, nov 2012.
[25] C. Fan, “Laplacians and the cheeger inequality for directed graphs,” Annals of Combinatorics, vol. 9, no. 1, pp. 1–19, apr 2005.
[26] Cheng Ye, Richard C. Wilson, César H. Comin, Luciano da F. Costa, and Edwin R. Hancock, “Approximate von neumann entropy for directed graphs,” Physical Review E, vol. 89, no. 5, may 2014.
[27] Lloyd N. Trefethen and David Bau, Numerical Linear Algebra, SIAM, jan 1997.
[28] G. H. Golub and C. F. V. Loan, Matrix computations (3. ed.), Johns Hopkins University Press, 1996.
[29] P. Y. Chen, L. Wu, S. Liu, and I. Rajapakse, “Fast incremental von neumann graph entropy computation: Theory, algorithm, and applications,” ArXiv e-prints, May 2018.
[30] E.-M. Kontopoulou, G.-P. Dexter, W. Szpankowski, A. Grau, and P. Drineas, “Randomized Linear Alge-
bra Approaches to Estimate the Von Neumann Entropy of Density Matrices,” *ArXiv e-prints*, Jan. 2018.

[31] P. Erdős and A. Rényi, “On random graphs,” *Publications Mathematicae Debrecen*, vol. 6, pp. 290–297, 1959.

[32] E. N. Gilbert, “Random graphs,” *Ann. Math. Statist.*, vol. 30, no. 4, pp. 1141–1144, 12 1959.

[33] Réka Albert and Albert-László Barabási, “Statistical mechanics of complex networks,” *Rev. Mod. Phys.*, vol. 74, pp. 47–97, Jan 2002.

[34] Duncan J. Watts and Steven H. Strogatz, “Collective dynamics of small-world networks,” *Nature*, vol. 393, pp. 440–442, June 1998.

[35] R. A. Rossi and N. K. Ahmed, “The network data repository with interactive graph analytics and visualization,” in *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[36] Tore Opsahl and Pietro Panzarasa, “Clustering in weighted networks,” *Social Networks*, vol. 31, no. 2, pp. 155 – 163, 2009.

[37] Tore Opsahl, “Triadic closure in two-mode networks: Redefining the global and local clustering coefficients,” *Social Networks*, vol. 35, no. 2, pp. 159 – 167, 2013, Special Issue on Advances in Two-mode Social Networks.

[38] D. M. Endres and J. E. Schindelin, “A new metric for probability distributions,” *IEEE Trans. Information Theory*, vol. 49, no. 7, pp. 1858–1860, 2003.

[39] J. Briët and P. Harremoës, “Properties of classical and quantum jensen-shannon divergence,” *Phys. Rev. A*, vol. 79, pp. 052311, May 2009.

[40] P. K. Sharma, G. Holness, Y. Markushin, and N. Meikechi, “A family of chisini mean based jensen-shannon divergence kernels,” in *Int. Conf. Mach. Learn. Applications (ICMLA)*, Dec 2015, pp. 109–115.