Review of "Heat diffusion distance processes : a statistically founded method to analyze graph data sets"

Jiaxin Hu

March 7, 2022

This work proposes two notions of distances, heat kernel distance (HKD) and heat persistence distance (HPD), to compare the pair of graphs. Confidence band for the distance estimate and the two sample test to detect the significant difference between graphs are constructed based on HKD and HPD.

Though the target problem is very interesting, this paper lacks sufficient descriptions for the main method, constructing confidence band and two sample test using HKD and HPD, as well as formal theorems for the consistency and supportive numerical experiments. In contrast, this paper takes a large space for the extended theoretical analysis on general empirical processes with proofs. The main method is also introduced from the general process rather than the proposed HKD and HPD processes. Therefore, I believe reconstructing the paper exposure in a methodological flavor with more detailed method description, theorems, and informative numerical analyses may better tackle the graph comparison problem.

Specific comments are below:

1. The method to construct the confidence band and two sample test statistics with HKD and HPD is unclear. The author describes the constructions in terms of the general empirical process rather than the proposed HKD and HPD processes. More instructions on how to construct the confidence band and test statistics using HKD and HPD would be helpful. In addition, independent theorems for the consistency of confidence band and two sample test with HKD and HPD should be provided.

2. Some simulation results are confusing, and more explanations for the simulation setup and results would be helpful. As far as I am concerned, the simulations for two sample tests consider the hypotheses

$$H_0 : ER - ER \;\leftrightarrow\; H_1 : ER - SBM, \quad \text{and} \quad H_0 : Disk - Disk \;\leftrightarrow\; H_1 : Disk - Annulus,$$

where the test statistics are functions of HKD and HPD, respectively. Intuitively, a larger difference between the HKD or HPD under the null H_0 and alternative H_1 leads to a easier task. As Figure 1 shows, the distributions of HKD under $ER - ER$ and $ER - SBM$ are very close with each other over time t while Figure 3 shows the distributions of HPD under $Disk - Disk$ and $Disk - Annulus$ are much more separable. Then, the $Disk - Annulus$ testing would be easier. However, Figure 4(b) indicates the $ER - SBM$ testing with HKD

achieves power 1 with sample size 100, and Figure 5(b) indicates the $Disk - Annulus$ testing with HPD achieves power 1 with sample size more than 120. Hence, the results in Figures 4, 5 seem counter-intuitive. Therefore, more explanations are necessary to make the simulation results more convincible.

3. The experiments lack the comparisons with other methods. As mentioned in introduction, there are several methods tackle the graph comparison problem with and without node correspondence. However, the current simulation focuses on the verification of the proposed method, except the comparison with intractable Neyman-Pearson test. More comparisons with current methods should be added to emphasize the advantage of the purposed method. In addition, real data analysis should be added to illustrate the practical usage of method.

4. Assumptions for the proposed method may be restrictive. First, the paper assumes the graphs have bounded non-negative weights, which excludes the graphs with Gaussian and Poisson weights. Second, the paper assumes the graphs satisfy the heat equation (first equation in Section 2.2). However, the applicability of such heat equation assumption is not discussed. Therefore, more discussions about the assumptions should be provided to strengthen the paper.

Therefore, I would be hesitant to decide that this paper is suitable for the venue.