Summary:

This paper proposes a generalisation of the family of linear networks (SGC, SSGC, APPNP, etc.) to heterophilous graphs. As several graph neural networks can be expressed as a linear combination of weights w_i with order-i diffusion (A^i) of the adjacency matrix A representing graph, selecting optimal weights w_i is an open topic in graph learning. In contrast to existing methods which work well with homophilous networks, this work studies an objective which can work well with heterophilous graphs. Authors demonstrate the least squares based solver based on Krylov subspaces which attempts to find a set of weights which produce a network which preserves node attributes. The authors provide a theoretical analysis based on an abstract two-class problem. They simulate distributions for the positive and negative class, for one-hop and two-hop convolution, and they show that if weights w_1 and w_2 have opposite signs, the probability of missclassification is decreased for heterophilous graphs. They show the same holds if the difference between even and odd hop numbers is taken. They show that obtaining weights by the least squares enjoys similar properties.

Positives:

- the problem of learning graph representations without class labels is interesting and timely
- the results on heterophilous graph benchmarks is SOTA
- the idea of reconstructing graph network propagated node attributes to be close to input node attributes seems intuitive

Negatives:

- writing can be improved in some parts of the paper
- results on homophilous benchmarks could be higher (but this is perhaps OK as authors propose a method that for heterophilous benchmarks)

Main comments:

1. In Eq. 4 and Eq. 5, do authors use class labels or attribute channels? The reviewer guesses it is attribute channels to ensure the model is an unsupervised model. This detail could be made clearer. Additionally, does Eq. 5 result in one set of weights or d set of weights assuming attributes have d channels? Could authors compare both cases?
2. Do authors ablate the impact of hyper-parameter r? Is there a sweet spot for r, or is it the best to select the largest r in each experiment?
3. Section 3.3 shows an extension of the proposed idea to optimisation over Chybyshev polynomials. Does this mean the authors could apply a simple network with weighted polynomials of degrees 1...r? If so, it would be interesting to see some result for such a problem.
4. Can authors comment on the complexity of the approach? Are Krylov subspaces used because they can find weights for consecutive diffusions $A, A^2, A^3, \ldots$ of the linear network design or they are somehow more central to heterophily?

5. Section 3.4 provides a theoretical analysis. Are μ representing attribute centers? If p and q are inter- and intra-class probabilities of edge connectivity, does this mean p represents heterophily and q represents homophily?
6. According to Eq. 14 and Theorem 3.2, authors notice that $w_1 < 0$ and $w_2 > 0$ leads to a lower probability of missclassification on heterophilous graphs. In Figure 3 authors show the same holds true if the sign of weights for even and odd hops differs. Does the least squares of Eq. 5 produce such a case? It would be interesting to see it further analysed. Does theory of Section 3.4 directly extend to multi-hop setting for example by considering decomposing multi-hop setting into several stacked layers of one- and two-hop convolutions, or is the multi-hop case hard to analyse?
7. Table 1 and Table 2 generally support the idea that Eq. 4 and Eq. 5 can recover filter weights useful for heterophylous benchmarks. In Table 2 the proposed method performs a bit worse than APPNP on CORNELL. In Table 1 the proposed method is better than APPNP on PHOTO. But PHOTO is homophilious while CORNELL is heterophilious. Are there some other factors at play here beyond homophily and heterophily? Perhaps the low number of edges and nodes of CORNELL also plays the role? Could it be that heterophily should be somehow connected with the notion of local heterophily (in local graph neighbourhoods) and global heterophily (no local edge groups)? Or are there any other criteria to consider in such a case?
8. Is there any simple strategy that could combine homophilous and heterophilious graph classification strategies?

In conclusion, authors study an interesting and timely problem. For rebuttal, could authors provide required explanations and experiments supporting them?

Minor comments:

1. Dimensions of w appear to differ in Eq. 4 and Eq. 5. Do authors mean r-1 or r for the dimension of w? The same goes for Eq. 8.
2. Page 7 appears to have issues with floats (big gap between two figures).
3. Can authors provide the definition of Runge phenomenon or at least a reference?
4. While p and q are described in Section 3.4, figure 1 should be described better or at least cross-referenced with Section 3.4 and discussed briefly in captions.