Review of the preprint, “Linking parasitism to network centrality and the impact of sampling bias in its interpretation.”

In this preprint manuscript, the authors assess geohelminth parasitism in a population of free-ranging Japanese macaques. The authors first test for a relationship between network measures of macaque centrality and the intensity of helminth parasitism (quantified by eggs per gram of fecal sediment; EPG), finding that, when considering the whole population of macaques, strength and eigenvector centrality are positively associated with parasitism, and macaque age is negatively associated. The authors then mimic sampling conditions in which only adult females, only juvenile macaques, or a random subset of the population are considered in analysis, finding that the relationship between centrality and parasitism is obscured or lost under these conditions.

The effect of inadequate and/or biased subsampling of a population on subsequent network structure is a significant concern in network studies, with a number of publications devoted to understanding this relationship. This study is an important extension of such previous work, focusing on how subsampling a population affects network *inference*, a largely understudied issue in disease ecology. That being said, I have a few major comments about this preprint, in addition to more minor comments.

Major comments:

1. My biggest comment here is concerning the appropriateness of the bounds used in the network construction.
    a. Firstly, the authors use proximity to construct their macaque social network in a study of parasites which are environmentally transmitted. Individuals within one meter of each other at one time are considered connected in the network, but there is not assessment of potential network connectivity between individuals in the same location with a time lag, despite the authors’ acknowledgement that “there is a time-lag between infection of one host and transmission to the next host through the environment” (pg 25). This calls into question the appropriateness of proximity as the edge definition without considering the role of specific locations/environmental exposures. Presumably, many individuals should be connected in the network through time-lagged proximity events, which would likely greatly affect results. The authors further assume that close proximity events are associated with resources, yet there may be a mismatch between locations in which individuals aggregate/are in close proximity and locations in which they perform behaviors most conducive to environmental parasite transmission. The authors should provide a justification for their choice of edge definition for this study of environmentally transmitted parasites and discuss any limitations to this choice.
    b. Secondly, the network is constructed over a period of 3 months of behavioral observations, but there is no justification provided for this choice. Indeed, the authors later discuss potential usefulness of a much longer period of observations (i.e., 16 months). However, in the context of parasite transmission,

networks should generally be constructed over a time period which aligns with transmission of the parasite(s) in question. How long do the parasites under study survive and transmit in the environment? If only for a few days, for example, then edges should be aggregated over a similarly short time period to represent the interactions occurring during the "transmission period" of the parasite (e.g., see White et al 2017, *Biological Reviews*). The authors should provide a justification for their choice of aggregating interactions over a 3 month period and use caution in suggesting that aggregating over longer time periods would be more appropriate without further epidemiological justification.

2. The modeling approach needs additional explanation and clarification in the methods.
    a. I was never certain exactly which models were run with which variables. I would recommend at least a table outlining the models which were run. Alternatively, the authors might consider a methods flow chart-style figure to show the different modeling approaches in the context of the different networks constructed.
    b. I think there is perhaps an error in the description of the randomization procedure (pg 16): as written, it sounds as though there were many observations compared to many randomizations. Shouldn't there be one observation compared to the randomization-based null distribution? Or did the authors fit 1000 "observed" models?
3. I have several questions about EPG as the outcome variable.
    a. The authors state that EPG is "not always indicative of true worm infection intensity," but do not later discuss how this limitation might affect their results.
    b. Furthermore, the methods and first entries in the dataset (which are presented in the supplementary materials) suggest that EPG is extremely variable within the same individual over time. How might this variability affect results?
    c. It appears that EPG was pooled across parasite species for modeling. The authors provide a statistical justification, but not a biological/epidemiological one. Do all parasites examined have identical epidemiological characteristics? If not, wouldn't they be expected to have different relationships between social structure and transmission?
4. I'm a little confused about the interpretation of the results when focusing on just adult females or juveniles. The authors state that age was significantly associated with parasitism (juveniles having higher EPG than older individuals); does it not then follow that we would struggle to find a relationship between centrality and parasitism when focusing on only juveniles? I imagine there is less variability in the EPG measures for juveniles, if they tend to have higher EPG, which would reduce the power for detecting an effect in just this subset of the population. In contrast, because sex was not significantly associated with parasitism, we might expect to have higher power for detecting a relationship between centrality and parasitism with this subset. Indeed, the results seem to support this to some extent (e.g., strength was still significantly associated with EPG for females only). I would suggest the authors could discuss how changes in statistical power with subsampling may affect inference.
5. Overall, I think the manuscript—and especially the introduction—could be streamlined. The real target of the manuscript is only arrived at by the end of the 7th paragraph and I

think much of the introductory material, while interesting, is perhaps beyond the scope of the objectives of the study. To tighten up the manuscript and improve readability, I'd recommend streamlining and reducing the length of the introduction.

Minor comments:

Throughout: Just a quick note to perhaps add line numbers to future submissions. I've had some difficulty in describing the precise locations of minor comments without line numbers, so I apologize for any confusion or lack of clarity in the following comments. I have chosen to give the page and paragraph number for the following comments, with "paragraph 1" (abbreviated "P") always the first text on the page, regardless of if this is a new paragraph on that page.

Pg 3, P 2: The flow of logic in the paragraph was difficult to follow and citations were missing. Is the objective of this paragraph to introduce the idea that social behavior affects transmission? This is a paragraph that could be drastically altered to streamline the introduction (see major comments).

Pg 4, P 2: I did not follow the mention of young males dispersing. Recommend taking this point further and explaining the relevance if the authors choose to keep this sentence/paragraph.

Pg 5, P 1: Use caution with vague/non-specific language such as "other factors" here. It decreases the readability of the manuscript. Other examples would be on page 26, paragraph 2: "For various reasons, numerous studies…"; page 29, conclusions paragraph: repeated use of "different" is vague and nonspecific to the direction of changes.

Pg 6, P 2: Recommend rephrasing the sentence with the very long parenthetical statement for readability.

Pg 6, P 2: The authors repeatedly cite the same two papers here and in the discussion (Smith et al 2017 and Silk et al 2015). These are excellent choices, but additional relevant papers include Davis et al 2018, *Animal Behaviour* and Gilbertson et al 2021, *Methods in Ecology and Evolution*. In the interest of full disclosure: this reviewer is an author on one of these papers, but I do feel they are both particularly relevant to this manuscript (perhaps especially Davis et al 2018).

Pg 9, P 2: Recommend using "macaques" throughout for consistent terminology; I think there were two instances where the authors chose to use "monkey" instead.

Pg 9, P 2: Does the focus group of macaques ever use spaces previously occupied by the other group on the island? I'm wondering if there could be any environmentally-mediated transmission relationships between the two groups.

Pg 10, P 1: Should the Kappa index results perhaps be given in the results rather than the methods?

Pg 12, P 1 and 2: Paragraph one states that samples were sent to Kyoto University; paragraph 2 then states that "we" processed the samples. Did the authors perform parasitological testing themselves, or was this work performed separately by the Primate Research Institute? I see that two of the authors are affiliated with the Primate Research Institute, but it was unclear if testing was performed by the authors or not, as phrased.

Pg 13, P 3: The authors do not need to provide information about how the data was laid out. However, please define "SP" in "(SP 1)."

Pg 14, P 1: The Git repository does not need to be given in the main text as it is listed in a data availability statement later. However, the authors should provide information about how/where this repository will be versioned and archived in the data availability statement. For example, *BES* journals require Github code to be archived and versioned with Zenodo upon acceptance for publication.

Pg 14, P 2: The different models for the different centrality measures and model averaging descriptions were very unclear. Recommend additional explanation and a table or figure to show the different models (see major comments). In addition, in GLMMs, were individuals and dates included as random intercepts or also as random slopes?

Pg 15, P 1: AIC results should perhaps be in results, not methods?

Pg 15, P 2 heading: It is unclear, as written, how this heading and therefore section is different from the previous section. I think this is where a methods flow chart could be particularly helpful (see major comments).

Pg 15, P 2: I did not follow the "because we were confident in the observed edges" explanation. Could you explain further?

Pg 16, P 2: I would recommend stating the size of the subsampled networks, especially for the adult female and juvenile only networks.

Pg 16, P 2: Why Pearson's and not a non-parametric correlation measure (i.e., Spearman's)? Both Davis et al 2018 and Gilbertson et al 2021 (mentioned in minor comments for page 6) use ranked correlations rather than Pearson's.

Pg 17, Figure 1: I found figure 1 to be extremely "busy" and it was difficult to decipher what the authors intended the reader to understand from this figure. If limited for space, I would recommend swapping this figure for a methods flow-chart style figure.

Pg 19, P 1: “Full model” is never really defined. This is an example of how the methods are currently unclear, which has downstream impact on the interpretability of the results (see major comments).

Pg 20, Figure 2: The figure legend states that the authors constructed three different models for each network metric, but this is unclear in the main text. Further, the different colored lines are not readily visible in the figures, as currently presented.

Pg 20, Table 2: Symbols being defined (e.g., *p*) should generally be given in parentheses after the definition on first use. Additionally, the “name” for the summed output is excessively long and complicated. Please use a shorter abbreviation or name and define it. I should also add that the null model for LRT is also never defined, again adding to confusion about the methods. These comments for table 2 are relevant for other similar tables in the manuscript.

Pg 23, P 1: It would appear that the authors are using the partial networks to test for “false negative” results (Type II error), but I would think that false positives (Type I error) would also be interesting and informative. I would therefore think that including results for degree would be useful, rather than just strength and eigenvector centrality.

Pg 23, Table 5: “Equivalency” appears to be qualitatively defined (are the randomization results also showing a positive estimate?), without information about the effect size of the detected relationship. Why? Furthermore, I think giving the “equivalent” results as proportions/percents would be more understandable than the current numbers. **Importantly:** it appears that the centrality measures in table 2 are incorrect. The legend states these should be strength and eigenvector, but the column in the table shows degree and strength.

Pg 24, P 1: Recommend defining “social integration” as I found this language unclear.

Pg 25, P 2: Regarding individuals with “high degree but low frequencies of interaction” - how common are these individuals? You should be able to comment on this with the data you have.

Pg 25, P 2: Regarding the example of lice - the direct transmission mentioned here would seem to reinforce my concerns about the appropriateness of proximity as the edge definition in this study (see major comments).

Pg 26, P 3: Much of this paragraph is repetitive with the introduction. I would suggest cutting/trimming from one location to help streamline (see major comments).

Pg 27, P 2: Did the authors not normalize degree for network size? This is standard functionality in *igraph* and would be valuable for controlling for the effect of network size.

Pg 28, P 2: “Unlike with targeted removal” - this phrasing was confusing and seemed to contradict the previous paragraph stating that targeted removal was a problem for inference.

Pg 28, P 2: I am very surprised that random removals would have a stronger impact on inference than targeted/biased removals. I would suggest further interrogation of this result as it has major implications to sampling strategies. Furthermore, this finding contradicts the statement in the next paragraph that incomplete networks should have “little imbalance as to who gets excluded.”

Supplement, SP 2 and 5: I found these figures very difficult to interpret. I would recommend providing additional information for readers. For example, the legend states there should be vertical lines for the original result and the 95% confidence intervals, but there appear to be 4 vertical lines in each figure. Also, what does “statistically significant results” mean in the context of these figures? I needed a little more “hand-holding” here.

Throughout: I noticed a number of instances of grammatical errors or informal language throughout. I’d recommend perhaps having a native English speaker read and correct the manuscript. In addition, some in-text citations had doubled close parentheses; recommend fixing these.