General response:

First of all, thank you for the thoughtful comments. Overall, while most of the issues the reviewers highlighted can be addressed without much difficulty, e.g. clarifying some parts of the manuscript and using a different set of controls for the MC-MLE estimation process, undoubtedly, the most substantive issue shared by the reviewers was the fact that we did not included an application using real data.

Following both reviewers recommendation, we have now added a new section with an empirical application of the little ERGMs in which we analyze a novel data set obtained from a lab experiment. In this fully featured example, we fit various ERGMitos and highlight key features that are either elusive to implement with other methods (like arbitrarily constraining sample spaces or applying transformations to ERGM terms), or overall unthinkable (e.g. calculating standard errors by means of bootstrapping of ERGMs [1000 replicates in less than 2 minutes]).

We have made a couple of other major revisions:

-An important improvement to this version of the paper is the inclusion of the Robbins-Monroe stochastic approximation algorithm (RM). Since the ergm R package already features the RM algorithm, we included this in a re-ran our simulation study.

-We have removed thetree figure that summarized the software-error rates of ergm vs ERGMito, because: (1) the statnet team fixed an important bug that was causing the majority of the software failures, and (2) we redefined the set of analyzed samples. Previously we compared all the situations in which BOTH MLE and MC-MLE did not failed to converge; now, following Handcoock (2003), a single drawn is included only if: (a) at least one graph is not fully connected, and (b) there is at least one triangle. This effectively discards cases that are right on the boundary of the support of the sufficient statistic, which consequently implies that we focused on those cases in which a MLEs have a higher chance to exists.

As a general response to Reviewer #1, we significantly changed the language used to describe ‘degeneracy issues’ and now refer to the convex-hull problem, based on their recommendation.

Details on the responses to each reviewer follow.

Reviewer #1

Large networks have dominated research in the social network field for the last 15 years or so, but in fact we know that such networks are not representative of the settings in which people live and work. Most social interaction occurs in (often very) small groups, whose structure is thus of central importance. That small networks have been given short shrift in recent network research is, thus, concerning. This paper aims to improve that situation by observing that a well-known approach to statistical network modeling (ERGMs) can be effectively applied to small networks by pooling, capitalizing on the fact that the ERGM likelihood can be computed exactly in the small-network case. Since many of the challenges associated with using ERGMs are related to the need to approximate the likelihood, this setting makes many typically difficult aspects of ERGM work quite facile. Moreover, many inferential questions associated with ERGM inference are trivially resolved by the use of network samples (i.e., samples of multiple networks, rather than samples from single networks), making classical statistical ideas immediately and trivially applicable. Thus, there are many reasons to push forward on this previously overlooked part of the network landscape.

I like this paper a lot, and strongly support its publication. I have some quibbles and suggestions, as noted below, but nothing terribly extensive. My main disappointment is the lack of an illustrative empirical example, which would drive home the potential utility of the approach to those who might otherwise be skeptical. It might also have been a good idea to note that the other major alternatives right now for statistical analysis of small networks are conditional uniform graph tests and related techniques, which are useful but are very limited in their ability to test for more than one type of structural bias at a time. Employing ERGMs means being able to control for multiple factors at once, and this is likely to be a killer app for small-graph ERGMs in practice.

Response: Thank you for the detailed review and helpful suggestions, and overall support for this paper. We have added an empirical example, and highlight the benefits of this approach that the reviewer mentions above, in a new section of the paper.

Minor comments:

(1) p4. "Degenerate models occur when the observed graph statistics lie in a region on or near the boundary of the support, and can be stressed when estimation depends on Monte Carlo Integration." This is not actually degeneracy, in the original sense of Strauss (1986) who to my knowledge first used the term in this context. Degeneracy, in Strauss's sense, involves an asymptotic phenomenon in which probability mass becomes concentrated on a vanishingly small number of graphs as N increases. (In fact, his development is a bit more specific than that, and is close to the idea of unbounded changescore growth used by Butts (2011). But this was his main point.) Degeneracy is a property of the model, and has nothing to do with data whatsoever. Unfortunately, some later work (including by researchers who should have known better) conflated degeneracy in the original sense with the "convex hull problem," i.e., the phenomenon in which the MLE fails to exist (in practice, diverges) when the observed data lie on the convex hull of the potential values that can be taken by the sufficient statistics. This issue is not specific to ERGMs, or even to dependence models, or even really to models at all: it's a property of the MLE, and does not occur with estimators such as e.g. Bayes estimators with informative priors.

Importantly, trivial models that cannot be properly degenerate (e.g., the homogeneous Bernoulli graphs) will still have non-existent MLEs for some observations (e.g., empty or complete graphs), and by turns well-defined MLEs exist for degenerate models (e.g., the edge-triangle model with a positive triangle parameter). These concepts are connected only incidentally, and as best I can surmise they became conflated because in both cases you can in practice wind up with very bad (and highly concentrated) graph distributions. However, the unfortunate mixing of these ideas is an error that needs to be undone, and one should not therefore refer to the convex hull problem as "degeneracy." "The convex hull problem," "non-existence of the MLE," or any other sensible descriptor is fine with me. But let us save degeneracy for the distinct class of phenomena to which it applies. (Phenomena that are especially inapplicable in this setting, where by definition we are concerned with graphs of very small size!)

Response: Thank you for your helpful suggestions. We have reviewed the literature that you point to above, and have revised the manuscript accordingly, in lines 69-76.

(2) p5. Again, do not refer to the convex hull problem as "inference degeneracy." Yes, I know that some very smart and famous people have made the mistake, but let's not perpetuate it.

Response: We have revised the section as suggested (lines 104-107).

(3) p6, 121-125. Again, please distinguish (even if the people you are citing did not) between models that lead to degenerate graph distributions and failure of the MCMC-MLE to exist due to the observation being on or outside the convex hull of the simulated statistics. The latter has nothing to do with degeneracy in its original sense, and is a property of the estimation procedure rather than the model.

Response: We have revised the section as suggested (lines 130-141).

(4) p7. Projectivity is a red herring, as was explained by e.g. Schweinberger et al. (2017), and has no bearing on the suitability of a network model for practical use (or anything else). With respect to generalization across network sizes, I think your take is far too cautionary. Krivitsky et al. (2011) (whom you cite) provide a very effective approach to correcting for size scaling that has been found to work extremely well (including, in my experience, in networks having very strong dyadic dependence). See also Krivitsky and Kolaczyc (2015) for simple corrections that preserve reciprocity and transitivity in directed graphs, and Butts and Almquist (2015) for corrections to mean degree with arbitrary power law scaling in N. So long as one chooses a reference measure with the appropriate scaling properties for one's problem, one can generally extrapolate across (or pool over) different values of N without major issue (assuming that there are not other major sources of uncontrolled heterogeneity, but that is a different matter). (All that said, I agree that these issues are probably not a concern in this setting.)

Response: Excellent point, many thanks for all the suggested references, we have now included them in the manuscript. Furthermore, we moved this point to the discussion section and highlighted the fact that it may only be relevant in the case of using ERGMitos to fit larger networks by taking samples (an application that we do not tackle in this paper) (lines 627-636).

(5) p8. The text seems to imply that the authors' package is the only tool that allows estimation via exact calculation of the ERGM likelihood. This option is available in the ergm package (and has been for many years). To be fair, it isn't set up to facilitate pooling, but it seems reasonable to acknowledge this (since you are using the ergm package later in the paper).

Response: You are right. The ergm package IS the backbone of this paper's R package, and we should be more explicit about it, including the fact that MLEs can be fitted in the ergm package without pooling the data. We have added a comment on this and stress some of the new features that the ergmito package brings to the table (lines 207-215).

(6) p13. The thinning interval used for the ergm MCMC-MLE routine is probably far too low here. It is often wise to scale the number of iterations in terms of the square of the number of nodes (which would be much larger than the setting chosen for these aggregate networks). It is not always necessary to use values this large for good results, but often it is. So this may have something to do with the high MCMC-MLE failure rate. I would also wonder if an equivalent failure rate is seen when using the Robins-Monroe algorithm (which is an option in the ergm package), as this is often more stable.

Response: During the development of the revised version of the paper, we noticed that errors that we previously observed in the ergm package were solved (almost all of them), after a bug-fix was introduced by the statnet team. Now, when we re-run our analyses, the effective number of software errors with cryptic user messages is zero. In the handful of cases in which ergm returns an error (only ~ 100/20,000), the error message to the user is clear. Following your recommendation, we increased the thinning, sample and burn-in to 10x the default values in cases where ergm reported an error, yet, after the bug fix, we rarely need to re-estimate the model. As suggested, we explored estimations based on the Robbins-Monroe algorithm [RM] and have added this to the paper. The RM does have a high error rate, but this seems to be related to a bug in the ergm package (we note in the paper that this algorithm is not as extensively developed in ergm), but again, most of the errors are associated with cases in which either of the target statistics was on the boundary.

Nevertheless, we revised the set of realizations of the data generating process included in the analysis, in particular, we now look at cases in which: (a) at least one of the networks included is not fully connected, and (b) at least one network has one transitive triad. In practice this means that we have discarded cases in which all of either edgecounts of transitive triads in the sample lie on the bound of the support. An immediate effect of this is that the RM dramatically reduced the software-type errors from ~ 5,000 to 3.

These changes are integrated in a substantial rewriting of section 5 (Simulation study).

(7) p21. Again, careful on the use of "degeneracy."

Response: We have revised the section as suggested. (line 589-591)

(8) p22. One has to be careful about the idea that one can fit an ERGM to a large graph by taking random subgraphs, fitting equivalent ERGMs to those subgraphs, and then using the resulting parameters as an estimate of the same model on the whole graph. This only works for projective models (as Shalizi and Rinaldo observe), which includes approximately no models of substantive importance. It is possible that there is a way do what you describe with an appropriate correction procedure, but to my knowledge there is no simple way of doing so at this time. (There is a way to legitimately do something somewhat like that by conditioning on the other parts of the graph when estimating locally, but that's another matter.…)

Response: Thank you for pointing that out. We haven’t explored how to fit large graphs by partitioning/sampling the network yet. We have amended the manuscript pointing out these difficulties and will keep this in mind for future research (lines 627-636).

Reviewer #2:

The paper proposes the use of exhaustive enumerations as the method to obtain maximum likelihood estimates (MLE) for exponential random graph models (ERGM) on small networks. Through simulation studies, the author(s) compared the simulation based parameter estimation approaches with the proposed method with the support of the implementation of "ergmito" package in R. These simulation studies demonstrate that ergmito is an efficient and effective estimation method for ERGM for small networks with 4 or 5 nodes. The paper discussed the possible use of ergmito in pooling estimates on sampled (and potentially larger networks). I can see this as a nice simulation study of ergmito, and hoping it can be demonstrated with an empirical example to show its potential contribution for analysing small networks.

Response: Thank you for the feedback. We have incorporated an empirical example, as suggested.

I do have a few comments in the way how the simulation and comparisons are conducted that I hope the author(s) can clarify or elaborate.

(1) The use of simulation based methods for ERGM parameter estimation is due to the un-tractable normalising constant, or in other words, if it is tractable (as demonstrated by the paper for small networks), there is no theoretical reason why we should use simulation. So it is not a surprising finding that ergmito can provide more accurate MLE, as it should.

Response: That is true, it is exactly what we were expecting, however, our simulation study shows that the approximations are actually very good overall, which be believe is worthwhile noticing. Yet, the simulation study also allowed us to stress the fact that using exact likelihoods provides a great benefit regarding speed and, at least at some level, provided some idea regarding power analysis for small ERGMs.

(2) Pooling ERGM estimates on network of different sizes. As the author(s) noted that ERGMs are defined and constrained on the number of nodes which defines the sample space. Even for small networks, the difference in sample space is huge, a factor of 256 in ratio between networks with 4 nodes and 5 nodes. The same parameter would produce very different network structures for networks of different sizes, and I suspect this is more relevant to small networks and the Markov model (or edge + triangle) model. For the purpose of simulations, I do not see the benefit of pooling networks of the two sizes. But this might be misleading in practice if readers have not considered the consequences of pooling appropriately. For example, pooling a random subnetworks of sizes of 4 or 5 from a larger network without considering the sampling and tie dependence structure can produce misleading pooling results. The snowball sampling approach may produce more reliable pooling estimates, but the snowballs may be too big for ergmito. The more traditional hierarchical linear models are much more appropriate, but the assumptions are the small networks are independent from one another.

Response: An important advantage of pooling data for modeling is that it reduces the chances of observing models with observed sufficient statistics near to boundary of its support. In a lot of cases, fitting small networks independently may lead to non-existence of MLE. It is because of this, and (as you pointed out as well) the assumption that the small networks are independent draws from a population, that pooling becomes a good idea (and we would argue, the right approach). We have addressed the concerns that are raised about pooling networks of different sizes in our response to Reviewer 1’s point #4.

The challenge of fitting larger networks by splitting them into small-graphs was also raised by Reviewer 1, and we have revised the discussion to be more explicit about the context in which ergmitos are better suited (lines 627-636).

3. Two of the five simulated networks (Fig 2) are troublesome. In the first one, ALL ties are among females. The same apply to the 4th one where all ties are defined on males. The counts of Edgecounts and Homopholy are identical (100% collinear), I'm not sure how an Edgecounts+Homopholy model is obtained.

Response: This highlights the value of pooling data: the MLE will be less likely to be non-existent as long as one of the networks in the sample of pooled data is not (a) fully connected/empty or (b) has all vertices of the same “color” (i.e. gender, age, etc.). Thus, ERGMitos that used this pooled approach are most useful in contexts where where we have independent samples of small networks from the same population, that can be pooled in one model. To better illustrate this, we have added new tables (tables 1, 5, and 7 in pages 9, 26, and 28 respectively) illustrating the target statistics used for these pooled samples, that we hope gives the reader a clearer idea of what pooling the data means in practice.

4. The Edgecounts+Homopholy models are dyadic independent models, and one can derive the MLE analytically with or without simulation based MLE or ergmito, so I don't see the exact advantage of using either. The simulation study using edges+ttriads model is more of an interest here.

Response: Good point. We have sought to clarify in the revised version of the paper that the fivenets example is illustrative, and the addition of a new empirical application of the method (section 6) compliments this with more complex models.

5. The comparison using sample sizes less than 30 seems to be too small to make valid statistical comparisons. The same applies to the Type 1 error rates comparison (and again I don't know what the reason behind mixing 4 nodes and 5 nodes networks).

Response: As far as we understand, there are no other studies in which a form of power analysis has been conducted to identify what is the right number of networks to collect. We believe that including samples with less than 30 networks in these analyses is relevant because this may be common in practice (e.g., studies may have data on 10 or 20 team or other small group networks they wish to analyze).

We have addressed the rationale for combining networks of size 4 and 5 in our pooled models in Point 2 above, as well as in response to Reviewer #1 point 4. In our new applied example (section 6), we demonstrate the use of various types of terms that help control for network size when pooling data.

6. The accuracy of the simulation based MLE is quite sensitive to the estimation/simulation algorithm settings (e.g. number of iterations, the thinning applied, etc.) For example, the scale of the parameter updates may depend on a factor for "step size", and tiny step size may make estimation more accurate while sacrificing computation time. Please provide the exact settings you used here.

Response: We have included more details regarding the settings used for MC-MLE. (lines 304-310)

7. The computation time comparisons in Fig 10 shows ergmito is faster for networks of sizes 4 or 5. As we know when the network size increases, the exact enumeration will be intractable, it will be great to know what is the maximum size ergmito can handle (for a given hardware specification), and what sized network ergmito and simulation based MLE have similar computational performance based on today's hardware. Is network sizes up to 5 the limit ergmito can handle?

Response: Thank you for this suggestion. We have included more details regarding the hardware specification. Overall, for undirected graphs, which can also be fitted with ergmito, a regular computer can handle networks with up to 8 vertices. For directed graphs, networks up to 5 vertices can be safely (without freezing the computer) fitted. Larger directed graphs (say, 6 vertices) can be fitted as long as the model is simple, e.g. edgecounts and triangles only. We summarize these limitations in lines 154-155 and 224-225. We have also added a new section in the appendix providing some details on the hardware settings used for the paper (page 42).

8. Large part of the paper is simulation study, and the discussion described the possible future use of ergmito in practice. I do see the value of small network research, but a bit puzzled on what exact research question ergmito is designed to answer? (I know it is ERGM for small networks, but I hope it can do more than that, while not quite convinced that the pooling or the potential sampling in the most appropriate method). It will be nice if the paper can demonstrate an practical example of using ergmito, even on some existing data (which can be better as comparisons can be made to existing results whether it is method or empirical findings).

Response: This is a very good point, and as mentioned above, we have tackled this by including a fully featured analysis with experimental data of small teams (Section 6). Three key benefits that we highlight of using this framework are: (1) creating arbitrary interaction effects, (2) applying arbitrary transformations to the sufficient statistics (like one would do in a GLM setting), and (3) calculating standard errors using bootstrap in a finite time-frame.

A minor point, that Fig 1 Covariate Effect for Incoming Ties should be a tie, instead of a star of size 2.

Response: That is a good point. We have changed the figure to follow your advice.