Date: Mar 02, 2020
To: "George Vega Yon" vegayon@usc.edu
From: "Social Networks" son@elsevier.com
Subject: Decision on submission to Social Networks

Manuscript Number: SON_2019_554

Exponential Random Graph models for Little Networks

Dear Mr Vega Yon,

Thank you for submitting your manuscript to Social Networks.

I have completed my evaluation of your manuscript. The reviewers recommend reconsideration of your manuscript following major revision. I invite you to resubmit your manuscript after addressing the comments below. Please resubmit your revised manuscript by May 01, 2020.

When revising your manuscript, please consider all issues mentioned in the reviewers' comments carefully: please outline every change made in response to their comments and provide suitable rebuttals for any comments not addressed. Please note that your revised submission may need to be re-reviewed.

To submit your revised manuscript, please log in as an author at https://www.editorialmanager.com/SON/, and navigate to the "Submissions Needing Revision" folder.

Social Networks values your contribution and I look forward to receiving your revised manuscript.

Kind regards,
Martin Everett
Co-Editor

Social Networks

Editor and Reviewer comments:

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.

Minor comments:

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!)

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.

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.

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.)

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).

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.

p21. Again, careful on the use of "degeneracy."

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....)

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.

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.

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.

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.

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.

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).

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.

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?

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).

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

More information and support

FAQ: How do I revise my submission in Editorial Manager?
https://service.elsevier.com/app/answers/detail/a_id/28463/supporthub/publishing/

You will find information relevant for you as an author on Elsevier's Author Hub: https://www.elsevier.com/authors.

FAQ: How can I reset a forgotten password?
https://service.elsevier.com/app/answers/detail/a_id/28452/supporthub/publishing/kw/editorial+manager/

For further assistance, please visit our customer service site: https://service.elsevier.com/app/home/supporthub/publishing/. Here you can search for solutions on a range of topics, find answers to frequently asked questions, and learn more about Editorial Manager via interactive tutorials. You can also talk 24/7 to our customer support team by phone and 24/7 by live chat and email.