Summary of Dissertation Revisions

Rachael T.B. Sexton
NIST
rachael.sexton@nist.gov

Following committee member feedback, several changes and improvements have been made to the dissertation document, primarily involving additions to the literature review, and a discussion on implications and constraints of the current experiment settings. Other changes have been added to improve readability of the text, and several new figures and tables have been added to summarize key concepts and results.

Literature Review Extensions

The committee requested an improved discussion of more recent progress in the network recovery space. A new section (4.2.2) has been added elucidating the distinction between network *model* frameworks and estimation *approaches*. This is relevant when discussing the most recent advances in network recovery that involve Greedy Coordinate Descent (GCD) [1] and its application to estimation using the principle of Minimum Description Length [2]. I classified GCD as an estimation approach, since it is compatible with any number of network model specifications, and the authors do not spend time on the network models themselves.

Consequently, I have added relevant context for our work, portraying Desire Path Densities as a modeling framework (much like stochastic Block Models or Markov Random Fields), while Forest Pursuit is an estimation approach (like GCD) that has good properties. Clarity on this point is added in Section 4.2.3.

Discussion Extensions

In terms of validating network recovery results, committee members requested elaboration on how various assumption made in the course of this work would impact the results. Section 10.1 has been added as a discussion of limitations and possible next steps for Forest Pursuit, in terms of relaxing the assumptions that bound our methods applicability.

Model Validation

First, the use future dyadic pairs (e.g. two-author papers) that occur after edge prediction would theoretically serve to validate real-world network recovery. I have added section 10.1.1, which discusses the difficulty with forecasting when false-negatives are increasingly prevalent with increased network size. However, I then build on this suggestion to recommend a future focus on assembling suites of *metamorphic tests* for these unsupervised algorithms [3].

Generative Model Comparisons

Next, the committee has brought attention to the design of *Forest Pursuit* as specifically addressing the case that domain knowledge indicates possible *spreading processes* (like random walks) are the cause of node activa-

tions. Since MENDR is designed to synthesize datasets using random walks, this could be seen as a source of bias, as soon as practitioners wish to use MENDR to infer performance of network recovery approaches in other generative settings.

I have added new background context to section 7.2 on the state of generative modeling for *correlated binary data*. These models fall into several categories and do not always include a mechanism for inference [4], [5], [6], [7]. I have clarified that the contribution of the RSFm generative model comes from *using* prior information on spreading processes, if that applies.

Furthermore, section 10.1.2 is now a broader discussion on how violations of that assumption might be investigated, going forward.

Other Updates

- New metrics for Optimal MCC and min-connected MCC have been added for best-case and maximally-sparse performance comparison.
- Various minor errors in index notation and ambiguous wording have been corrected.
- FP and FPi have been highlighted in Figures 6.2, 6.4, and 7.1 for clarity.
- Results tables 6.3 and 7.1 have been added using median(IQR) format, with maximum values highlighted for legibility and ease of comparison.
- Some figure scalings have been increased for legibility of node labels.

Bibliography

[1] T. P. Peixoto, "Scalable network reconstruction in subquadratic time," Jan. 2024, doi: 10.48550/ARXIV.2401.01404.

[2] T. P. Peixoto, "Network reconstruction via the minimum description length principle," *Phys. Rev. X 15, 011065 (2025)*, vol. 15, no. 1, p. 11065, May 2024, doi: 10.1103/physrevx.15.011065.

[3] X. Xie, Z. Zhang, T. Y. Chen, Y. Liu, P.-L. Poon, and B. Xu, "METTLE: A METamorphic Testing Approach to Assessing and Validating Unsupervised Machine Learning Systems," *IEEE Transactions on Reliability*, vol. 69, no. 4, pp. 1293–1322, Dec. 2020, doi: 10.1109/tr.2020.2972266.

[4] F. Leisch, A. Weingessel, and K. Hornik, "On the generation of correlated artificial binary data." Vienna University of Economics, Business, 1998. doi: 10.57938/6884F809-93BC-4497-AB2B-FC1611198F5B.

[5] T. L. Griffiths and Z. Ghahramani, "The Indian Buffet Process: An Introduction and Review," *J. Mach. Learn. Res.*, vol. 12, no. null, pp. 1185–1224, Jul. 2011.

[6] H. C. Nguyen, R. Zecchina, and J. Berg, "Inverse statistical problems: from the inverse Ising problem to data science," *Advances in Physics*, vol. 66, no. 3, pp. 197–261, Jun. 2017, doi: 10.1080/00018732.2017.1341604.

[7] Z. P. Neal, "Randomly sampling bipartite networks with fixed degree sequences," May 2023, doi: 10.48550/ARXIV.2305.04937.