ABSTRACT

Title of Dissertation:	MEASURING NETWORK DEPENDENCIES FROM NODE ACTIVATIONS Rachael T.B. Sexton Doctor of Philosophy,
Dissertation Directed by:	Professor Mark D. Fuge Department of Mechanical Engineering

My abstract for this dissertation.

MEASURING NETWORK DEPENDENCIES FROM NODE ACTIVATIONS

by

Rachael T.B. Sexton

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy

Advisory Committee:
Professor Mark D. Fuge, Chair/Advisor
Professor Jordan L. Boyd-Graber
Professor Maria K. Cameron
Professor Michelle Girvan
Professor Vincent P. Lyzinski

Acknowledgements

I would like to acknowledge. . .

Dedication

to my friends and family

Foreward

Fwd content

Preface

Preface content

Table of Contents

List of Tables

List of Figures

Chapter 1:

Part I

Background: Graph and Hypergraph Structures as Inner Product Spaces

For each, describe the meaning of

1. Observation model
2. Vector space representation
3. Inner product (linear kernel)
4. Induced Norms

A wide variety of fields show consistent interest in inferring latent network structure from observed interactions, from human cognition and social infection networks, to marketing, traffic, finance, and many others. [6, 7].

However, an increasing number of authors are noting a lack of agreement in how to approach the metrology of this problem. This includes rampant disconnects between the theoretical and methodological network analysis sub-communities[1], treatment of error as purely aleatory, rather than epistemic [8], or simply ignoring measurement error in network reconstruction, entirely [6].

Intro part 1

A salient point is brought up by Torres et al. [3], in that many times our observations are more appropriately thought of as *different classes of incidence structures.*[3] The dependencies in the data generation and observation processes, as well as are assumptions are difficult to preserve when we model data as a simple graph that is better represented as, say, a hypergraph.

A common type of data used in network recovery is often called *co-occurrence* data, where nodes are observed as being in an "on" or "off" state in any given data point.

From these, we might wish to recover the latent relationships or dependencies between them. An argument could be made in the style of Torres et al. [3], that this data is fundamentally bipartite (and, thus, hypergraphical) in nature. However, in a large number of cases, we observe such co-occurrences as the result of underlying dynamical processes, like diffusion, on a carrier graph. *Further*, an increasing amount of literature is dedicated to capturing statistical properties of very large graphs, e.g. through variations on random-walk sampling. If our only snapshot of a large graph originates from samples generated on it by random walk (and other diffusive dynamic sampling strategies), then we must be able to perform *and normalize the practice of* epistemic uncertainty of edge existence from co-occurrence data.

How do we provide the needed metrological foundation to direct future research in this area? I hope that by unifying the interpretations of several classes of network recovery techniques into a single, non-parametric framework, we can re-unify methodology and theory. Much like the ubiquity of kernel density estimates for exploratory data analysis, with the right tools analysts can better reason about what they don't know, while researchers can use that reported, mutually understood experience to work on extending the things we *can know*: i.e. what they *actually want to measure*.

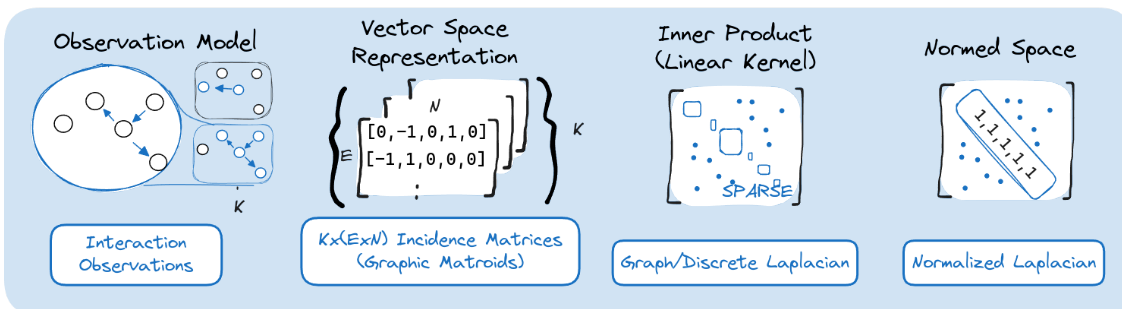

Figure 2.1: Edge Relation Observational Model

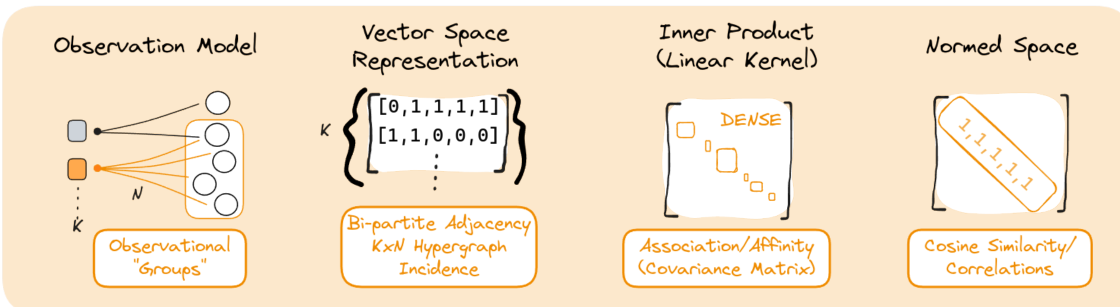

Figure 3.1: Hyperedge Relation Observational Model

Chapter 4: Review: Projections and Inverse Problems

I.e. trying to "recover" one from the other. In a way, proceduralizing the precautions described in Torres et al. [3].

> Takeaway: a way to organize existing algorithms, AND highlight unique set of problems we set out to solve

4.1 Connecting Graphs & Hypergraphs as Adjoint

i.e. Network Recovery as an Inverse Problem

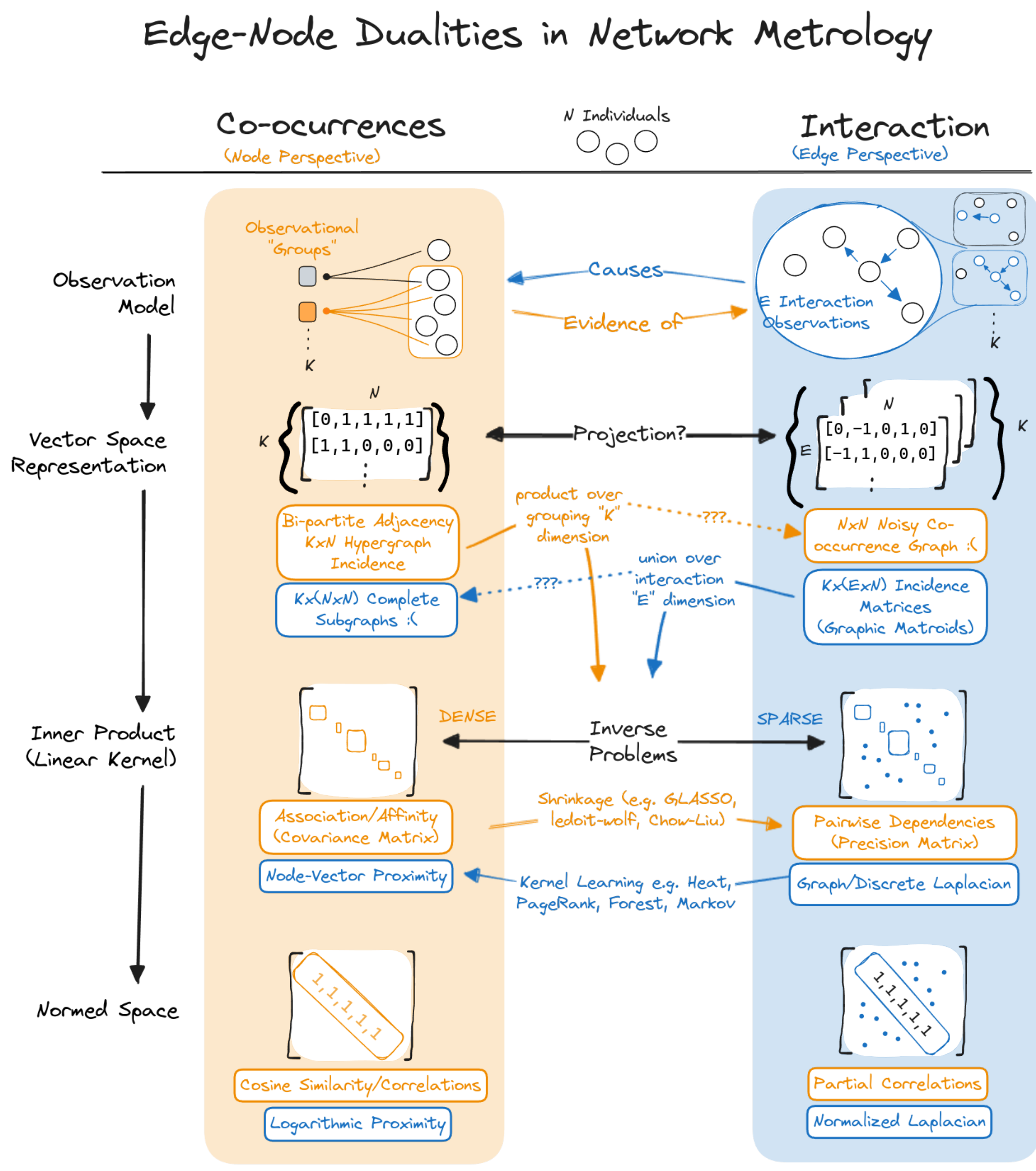

Figure 4.1: Relating Graphs and Hypergraph/bipartite structures as adjoint operators

4.2 Algorithmic Information Loss in Literature

- Observation-level loss (starting with the inner product or kernel)
- Non-generative model loss (no projection of data into model space)

- no uncertainty quantification

Sorting algorithms… *none address all three!*

i.e. MOTIVATES FOREST PURSUIT

Part II

Recovery from Bipartite Occurrence Records

Chapter 5: Nonparametric Network Recovery With Random Spanning Forests

filling the gap we saw in the literature

5.1 Generative Model Specification

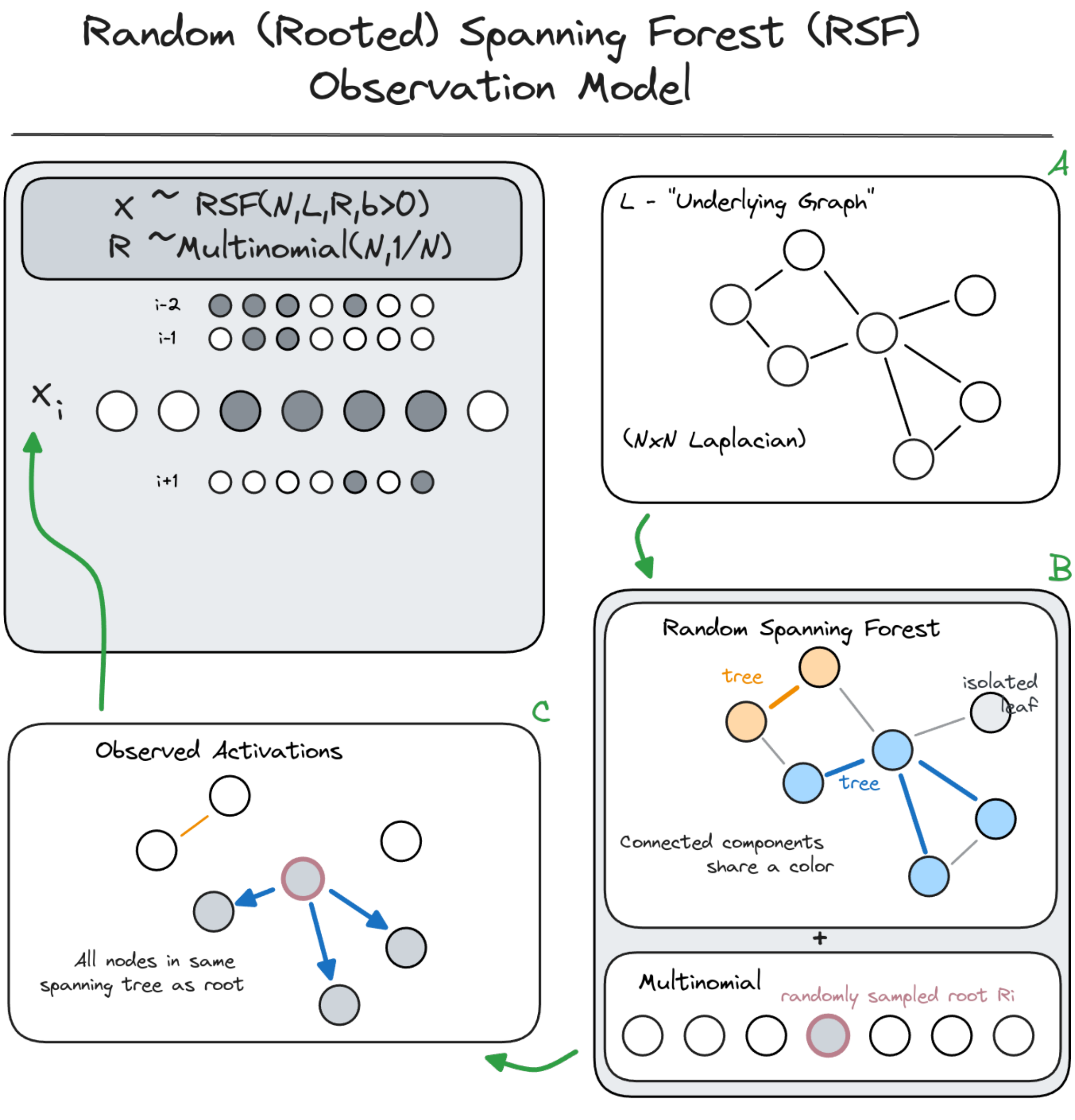

Figure 5.1: Explanation of the Random Spanning Forest generative model for binary activation observations

5.2 Forest Pursuit: Approximate Recovery in Near-linear Time

I.e. the PLOS paper (modified basis-pursuit via MSTs)

5.3 Bayesian Estimation by Gibbs Sampling

I.e. the unwritten paper, modifying technique by Duan and Dunson [2] for RSF instead of RSTs

Chapter 6: LFA: Latent Forest Allocation

Part III

Application and Future Development

Chapter 7: Qualitative Application of Relationship Recovery

Chapter 8: Recovery from Ordered Random-Walk Observation Sets

Like before, but with the added twist of *knowing* our nodes were activated with a particular partial order.

insert from [4, 5]

Bibliography

[1] L. Peel, T. P. Peixoto, and M. De Domenico, "Statistical inference links data and theory in network science," *Nature Communications*, vol. 13, no. 1, Nov. 2022, ISSN: 2041-1723. DOI: 10.1038/s41467-022-34267-9. [Online]. Available: https://www.nature.com/articles/s41467-022-34267-9.

[2] L. L. Duan and D. B. Dunson, "Bayesian spanning tree: Estimating the backbone of the dependence graph," arXiv, arXiv:2106.16120, Jun. 30, 2021, ZSCC: 0000001 type: article. DOI: 10.48550/arXiv.2106.16120. arXiv: 2106.16120.

[3] L. Torres, A. S. Blevins, D. Bassett, and T. Eliassi-Rad, "The why, how, and when of representations for complex systems," *SIAM Review*, vol. 63, no. 3, pp. 435–485, Jan. 2021, ISSN: 0036-1445. DOI: 10.1137/20M1355896. Accessed: Feb. 1, 2023.

[4] R. Sexton and M. Fuge, "Organizing tagged knowledge: Similarity measures and semantic fluency in structure mining," *Journal of Mechanical Design*, vol. 142, no. 3, Jan. 2020, ISSN: 1050-0472. DOI: 10.1115/1.4045686.

[5] R. Sexton and M. Fuge, "Using semantic fluency models improves network reconstruction accuracy of tacit engineering knowledge," in *Volume 2A: 45th Design Automation Conference*, American Society of Mechanical Engineers, Aug. 2019. DOI: 10.1115/detc2019-98429.

[6] T. P. Peixoto, “Reconstructing networks with unknown and heterogeneous errors,” *Physical Review X*, vol. 8, no. 4, p. 041 011, Oct. 16, 2018, ZSCC: 0000099. DOI: 10.1103/PhysRevX.8.041011. Accessed: Nov. 2, 2023.

[7] M. Gomez-Rodriguez, J. Leskovec, and A. Krause, “Inferring networks of diffusion and influence,” *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 5, no. 4, pp. 1–37, 2012, ISSN: 1556-472X. DOI: 10.1145/2086737.2086741.

[8] D. J. Wang, X. Shi, D. A. McFarland, and J. Leskovec, “Measurement error in network data: A re-classification,” *Social Networks*, vol. 34, no. 4, pp. 396–409, Oct. 2012, ISSN: 0378-8733. DOI: 10.1016/j.socnet.2012.01.003.