Models and Methods for Sparse (Hyper)Network Science in Business, Industry, and Government

Sinan G. Aksoy, Aric Hagberg, Cliff A. Joslyn, Bill Kay, Emilie Purvine, and Stephen J. Young

The authors of this piece are organizers of the AMS 2022 Mathematics Research Communities summer conference Models and Methods for Sparse (Hyper)Network Science, one of four topical research conferences offered this year that are focused on collaborative research and professional development for early-career mathematicians. Additional information can be found at `https://www.ams.org/programs/research-communities/2022MRC-HyperNet`. *Applications are open until February 15, 2022.*

Sinan G. Aksoy is a senior data scientist at Pacific Northwest National Laboratory. His email address is `sinan.aksoy@pnnl.gov`.

Aric Hagberg is the deputy division leader of the Computer, Computational, and Statistical Sciences Division at Los Alamos National Laboratory. His email address is `hagberg@lanl.gov`.

Cliff A. Joslyn is the Chief Knowledge Scientist and the Mathematics of Data Science Team Lead at Pacific Northwest National Laboratory. His email address is `cliff.joslyn@pnnl.gov`.

Bill Kay is a data scientist at Oak Ridge National Laboratory. His email address is `kaybw@ornl.gov`.

Emilie Purvine is a senior data scientist at Pacific Northwest National Laboratory. Her email address is `emilie.purvine@pnnl.gov`.

Stephen J. Young is a research mathematician and the Applied Mathematical Modeling Team Lead at Pacific Northwest National Laboratory. His email address is `stephen.young@pnnl.gov`.

DOI: https://doi.org/10.1090/noti2424

The authors are hosting an AMS sponsored Mathematics Research Community (MRC) focusing on two themes that have garnered intense attention in network models of complex relational data: (1) how to faithfully model *multi-way* relations in hypergraphs, rather than only pairwise interactions in graphs; and (2) challenges posed by modeling networks with extreme sparsity. Here we introduce and explore these two themes and their challenges. We hope to generate interest from researchers in pure and applied mathematics and computer science.

The Rise of Network Science

Graph theory has been driven by applied questions, from its apocryphal roots in the *Seven Bridges of Königsberg* problem to modern day network analysis. What had been cast in its infancy as a collection of recreational puzzles has evolved into an expansive and diverse discipline. Graph analyses are now common across nearly all areas of science. Accordingly, modern graph theory has evolved to engage methods from probability, topology, linear algebra, mathematical logic, computer science, and more.

Relational phenomena involving specific patterns of linkage between entities can afford natural representations as graphs. Prime examples are network systems, often massive, which arise in fields such as molecular biology, social systems, cyber systems, materials science, infrastructure modeling (e.g., Figure 1), and high performance computing. As elucidated by Chung in a 2010 *Notices* article

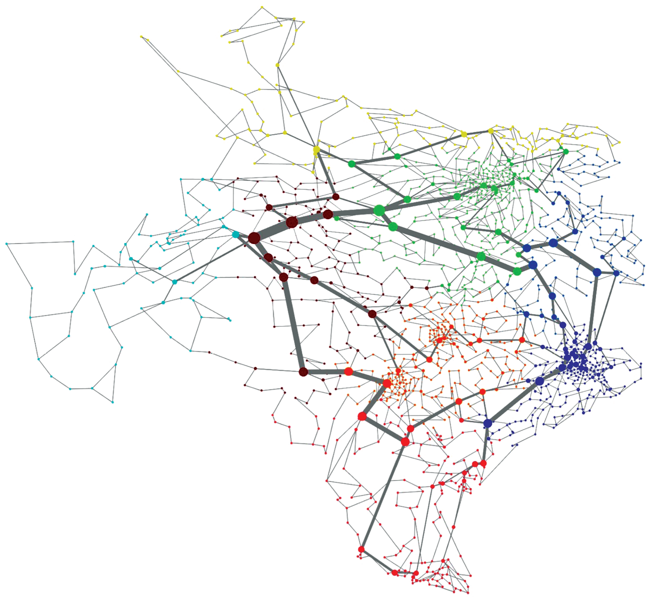

Figure 1. Synthetic Texas power transmission network generated from publicly available[1] test data.

[Chu10], despite coming from disparate domains, such networks exhibit "amazing coherence" in their shared empirical properties. Such hallmarks include sparsity (the number of edges is linear in the number of vertices), the small world phenomenon (any two vertices are connected by a short path, local neighborhoods are typically dense), and heavy-tailed degree distributions (the number of degree k vertices is roughly $k^{-\beta}$). Researchers have addressed fundamental questions surrounding these networks—how they evolve, which structures are critical to their function, which graph invariants capture meaningful properties, and so on—in an area commonly referred to as "network science" [NBW06].

Since Chung's *Notices* article 12 years ago, the scope of network science has continued to grow beyond emphasizing small-world ubiquity, and into studying richer classes of mathematical structures that better reflect the nuances of real-world networks. In part due to the increasing widespread availability of complex relational data sets, researchers have coalesced around a new class of applied questions where the properties of the relational data are, in and of themselves, driving the questions.

Relations Beyond Graphs

Over the past several decades, there has been an increasing realization within network science that multi-way interactions can play a critical role in networked systems. For instance, as highlighted by COVID-19 spread, the interactions at group gatherings can have a cumulative effect that can be obscured when reduced to pairwise interactions. In order to faithfully capture these multi-way interactions, it is valuable to move beyond the standard graph structure consisting of vertices V and edges $E \subseteq \binom{V}{2}$, to the richer framework of *hypergraphs*, where the edge set E is a subset of 2^V, the power set of V.

Where graphs can represent only pairwise relations natively, hypergraphs naturally code for multi-way interactions. Nonetheless, it is routine to resort to analyzing systems and data exhibiting multi-way interactions via "auxiliary graphs" produced from multi-way data, such as the line graph (which encodes intersections between pairs of hyperedges) or the 2-section graph (which replaces hyperedges with graph cliques). However, as illustrated by Figure 2, two non-isomorphic hypergraphs may have the same line graph. Similarly, two non-isomorphic hypergraphs may have the same 2-sections as well. Simply put, these most natural encodings of hypergraphs by auxiliary graphs fail to retain some pertinent information. Despite hopes that incorporating weights into the auxiliary graphs would allow faithful representation of hypergraphs via graphs, recent work by Kirkland [Kir18] shows that this is not the case.

And while hypergraphs are bijective to bipartite graphs[2] in which one of the parts is labeled as vertices and the other as edges, naive deployment of graph methods against them will not necessarily reveal the "set"-valued properties of the original hypergraph. The resulting algorithms are at best cumbersome to phrase and study in this framework, and at worst simply recapitulate the corresponding hypergraph-native methods. Thus, it is apparent that hypergraphs require their own analytical tool set to avoid the information loss inherent to graph reduction or bipartite approaches.

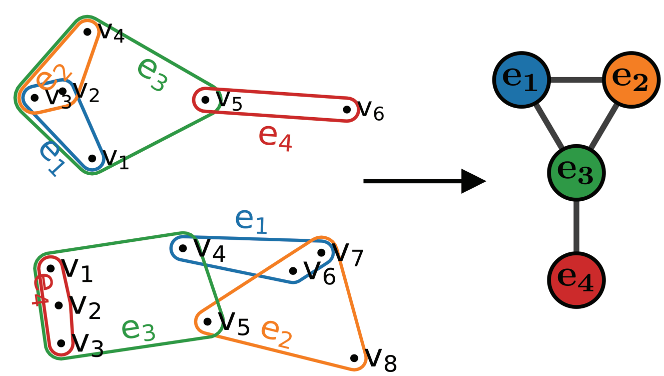

Figure 2. Non-isomorphic hypergraphs with the same line graph.

[1] *https://egriddata.org/dataset/activsg2000-2000-bus-synthetic-grid-geolocated-texas*

[2] *Technically "bicolored," there are caveats here in the case of disconnected hypergraphs.*

While shifting from modeling pairwise to multi-way interactions may seem like a subtle change, the implications are far-reaching and profound. For example, in a hypergraph the natural generalization of a walk of length k is a sequence $e_1, \ldots, e_k$ of hyperedges $e_i \in E$ such that for all $1 \leq i \leq k-1$ we have $e_i \cap e_{i+1} \neq \emptyset$. However, unlike in graphs these pairwise intersections have a non-trivial notion of "width," i.e., $|e_i \cap e_{i+1}|$. This allows the set of hypergraph walks to be partitioned by functions of their width, such as the minimum or mean width of intersections. In contrast with graphs, these width-based partitions induce non-trivial filtrations on a set of hypergraph walks. Since walks are foundational to defining many network science concepts, these filtrations in turn induce filtrations on component structure, connectivity, diameter, centrality, etc., which can be used to provide further insight into the network structure [AJM+20].

Building off this increased expressivity, a number of analytical tools have been developed to study hypergraphs, ranging from walk and centrality based methods [AJM+20, Ben19], motif and subgraph pattern analysis [LKS20], and dynamical processes on hypergraphs [dATM21, LR20]. Additionally, hypergraphs interact strongly with important structures from computational topology such as *abstract simplicial complexes* (hypergraphs that include all possible subedges), and there is a burgeoning movement to join network science to analytical approaches bridging to these higher-order mathematical fields [BGHS21, IPBL19].

Challenges of (Hyper)Analytics

Rather than attempt a methods survey, here we discuss thematic challenges associated with hypergraph spectral methods that reflect common issues facing hypernetwork science. Hypergraph Laplacians and associated spectral methods are commonly used to obtain embeddings, rank entities, and cluster data across domain areas, ranging from partitioning circuit netlists in VLSI, grouping term-document data in natural language processing, and performing image segmentation. How to optimally define hypergraph matrix and tensor representations to better enable such analyses has emerged as a central consideration.

Despite a plethora of proposals over the past several decades, there is little consensus as to which notion of hypergraph Laplacian is most appropriate. Furthermore, such proposals are starkly different, depending on whether or not one assumes uniformly sized hyperedges. For example, in the uniform case, Chung [Chu93] took a homological approach, Lu and Peng [LP11] introduced a so-called higher-order generalized Laplacian rooted in hypergraph random walks, while Cooper and Dutle [CD12] utilize multilinear-algebraic techniques to study multidimensional arrays they call hypermatrices. Unfortunately, there appears to be no obvious way to extend these notions to non-uniform hypergraphs. Accordingly, these methods likely have limited applicability to hypergraphs arising from real data, which are almost always naturally non-uniform.

While proposed non-uniform hypergraph Laplacians *are* applicable to real, messy hypergraph data, whether they effectively capture higher order structure present in hypergraphs but absent in graphs is disputed. As shown by Agarwal [ABB06], a number of non-uniform hypergraph Laplacians are related, via trivial shifts and scalings, to graph Laplacians associated with the auxiliary graphs mentioned above like the two-section (clique expansion). To mitigate the information loss inherent in such reductions [Kir18], one approach is to study hypergraph matrices associated with non-reversible random walks [CR19, HAPP20], while other recent work advocates non-uniform hypergraph adjacency tensors [BCM17]. However, these and other "hypergraph-native" approaches often come with caveats, underscoring the difficulty of devising practical yet faithful hypergraph methods: in this case, the former approach requires external weights to be effective, while the high-dimensionality of the tensor suggested in the latter poses computational challenges.

Modeling Sparsity

In addition to developing hypergraph analytic tools, network science is also grappling with how to develop models that capture the unusual combination of extreme properties exhibited by many complex networks. From the very first attempts to develop a robust theory of random graphs, it was recognized that the models being developed were, at best, imperfect representations of the real world. Indeed, Erdős and Rényi pointed this out in [ER61]:

> *"The evolution of random graphs may be considered as a (rather simplified) model of the evolution of certain real communication-nets, e.g. the railway-, road- or electric network system of a country or some other unit, or of the growth of structures of anorganic or organic matter, or even of the development of social relations. Of course, if one aims at describing such a real situation, our model of a random graph should be replaced by a more complicated but more realistic model."*

Since then numerous random graph models have been developed to capture various underlying structural or mechanistic properties; including approaches to capture the degree sequence (either exactly or probabilistically), network self-similarity, structural restrictions, network evolution based on preferences or biological mechanisms, and others.

Despite the proliferation of random graph models, there are significant structural features of data for business, industrial, and governmental (BIG) applications that still are not captured. For instance, while many of the networks important in BIG applications exhibit both connectivity and extreme sparsity, random graph models typically require an average degree of $\Omega(\log(n))$ (for models with more edge independence) or at least 3 (for models with less edge independence) in order to ensure connectivity. However, for systems such as the power grid (see Figure 1) or networks built from communication traces, connectivity is present a priori despite an average degree between 1 and 2.

In addition, many BIG applications are driven by experimental data that are essentially correlational in nature. Examples include correlated gene expression across multiple experimental conditions or macroscale structural properties of novel materials across a variety of microscale properties. These data sources are naturally represented in terms of a weighted hypergraph, and yet, many of the current analytical methods applied to these data sources rely on graph (as opposed to hypergraph) models. While there are many reasons for this discrepancy, one of the major contributing factors is a relative lack of random hypergraph models which can be meaningfully parameterized to be reflective of observed correlational data. While random bipartite graph models exist, they suffer from the problems described above. Between the need for connected random models exhibiting extreme sparsity, the increasing relevance of hypergraph data sources, and other peculiarities of BIG data sources, there is a significant opportunity to develop novel random hypergraph models driven by a new class of applications.

An Invitation

The authors of this article are organizing an AMS MRC in the summer of 2022 on these topics. We will be exploring the way that graphs and hypergraphs can be employed in real-world scenarios such as those in biology, computer science, social science, and power engineering. The goal of this collaborative workshop is to bring together researchers from multiple different domains including mathematics, computer science, and application domains to develop and extend graph-theoretical concepts that are rooted in problems of national significance, including:

- In critical infrastructure systems, such as the power grid or natural gas distribution system, it is often necessary to understand the combinatorial structure of the system to understand macroscale system behavior.
- Computer network data represents point to point information exchanges such as emails, network traffic, or process logs. This type of data is frequently modeled as a rapidly changing dynamic graph with the goal of discovering behavioral patterns and anomalies in the system.
- In the case of *-omics data from biology, much of the data is pairwise, or multi-way, rate of expression under various environmental conditions. This naturally leads to a variety of graphical structures, from directed hypergraphs to undirected graphs, depending on the choice of data representation.
- A key factor in the understanding of the behavior of microbial communities is the directed graph of reinforcing interactions, i.e., the presence of microbe A increases with the increase of microbe B.
- In blogging and social networks such as Twitter, users interact with external content by posting links, thereby forming a user-content hypergraph whose structure affects information spread.

We invite early-career applicants from all domains to join us. The most crucial characteristic of the applicants is the desire to build a community that is willing to teach and learn about other disciplines and to form true interdisciplinary teams. The organizers have identified several deep theoretical problems and will provide guidance and resources as participants tackle them. In addition to the technical collaborations, there will be opportunities to learn about research in the national laboratory system and in industry, expand networks, and participate in other professional development activities. We invite you to apply and join us in exploring this topic of theoretical interest and practical significance.

ACKNOWLEDGMENT. This manuscript has been authored in part by UT-Battelle, LLC under Contract No. DE-AC05-00OR22725 with the US Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes. The Department of Energy will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (`http://energy.gov/downloads/doe-public-access-plan`).

This work was supported in part by the US Department of Energy through the Los Alamos National Laboratory. Los Alamos National Laboratory is operated by Triad National Security, LLC, for the National Nuclear Security Administration of US Department of Energy (Contract No. 89233218CNA000001).

Pacific Northwest National Laboratory is a multi-program national laboratory operated for the US Department of Energy (DOE) by Battelle Memorial Institute under Contract No. DE-AC05-76RL01830.

References

[ABB06] Sameer Agarwal, Kristin Branson, and Serge Belongie, *Higher order learning with graphs*, Proceedings of the 23rd International Conference on Machine Learning, 2006, pp. 17–24.

[AJM$^+$20] Sinan G. Aksoy, Cliff Joslyn, Carlos Ortiz Marrero, Brenda Praggastis, and Emilie Purvine, *Hypernetwork science via high-order hypergraph walks*, EPJ Data Science **9** (2020), no. 1, 16.

[BCM17] Anirban Banerjee, Arnab Char, and Bibhash Mondal, *Spectra of general hypergraphs*, Linear Algebra Appl. **518** (2017), 14–30, DOI 10.1016/j.laa.2016.12.022. MR3598572

[Ben19] Austin R. Benson, *Three hypergraph eigenvector centralities*, SIAM J. Math. Data Sci. **1** (2019), no. 2, 293–312, DOI 10.1137/18M1203031. MR3953463

[BGHS21] Christian Bick, Elizabeth Gross, Heather A. Harrington, and Michael T. Schaub, *What are higher-order networks?*, arXiv:2104.11329 (2021).

[CD12] Joshua Cooper and Aaron Dutle, *Spectra of uniform hypergraphs*, Linear Algebra Appl. **436** (2012), no. 9, 3268–3292, DOI 10.1016/j.laa.2011.11.018. MR2900714

[Chu10] Fan Chung, *Graph theory in the information age*, Notices Amer. Math. Soc. **57** (2010), no. 6, 726–732. MR2674816

[Chu93] Fan R. K. Chung, *The Laplacian of a hypergraph*, Expanding graphs (Princeton, NJ, 1992), DIMACS Ser. Discrete Math. Theoret. Comput. Sci., vol. 10, Amer. Math. Soc., Providence, RI, 1993, pp. 21–36. MR1235565

[CR19] Uthsav Chitra and Benjamin Raphael, *Random walks on hypergraphs with edge-dependent vertex weights*, International Conference on Machine Learning, 2019, pp. 1172–1181.

[dATM21] Guilherme Ferraz de Arruda, Michele Tizzani, and Yamir Moreno, *Phase transitions and stability of dynamical processes on hypergraphs*, Communications Physics **4** (2021), no. 1, 1–9.

[ER61] P. Erdős and A. Rényi, *On the evolution of random graphs* (English, with French summary), Bull. Inst. Internat. Statist. **38** (1961), 343–347. MR148055

[HAPP20] Koby Hayashi, Sinan G. Aksoy, Cheong Hee Park, and Haesun Park, *Hypergraph random walks, laplacians, and clustering*, Proceedings of the 29th ACM International Conference on Information & Knowledge Management, 2020, pp. 495–504.

[IPBL19] Iacopo Iacopini, Giovanni Petri, Alain Barrat, and Vito Latora, *Simplicial models of social contagion*, Nature Communications **10** (2019), 2485.

[Kir18] Steve Kirkland, *Two-mode networks exhibiting data loss*, J. Complex Netw. **6** (2018), no. 2, 297–316, DOI 10.1093/comnet/cnx039. MR3801727

[LKS20] Geon Lee, Jihoon Ko, and Kijung Shin, *Hypergraph motifs: Concepts, algorithms, and discoveries*, arXiv:2003.01853 (2020).

[LP11] Linyuan Lu and Xing Peng, *High-ordered random walks and generalized Laplacians on hypergraphs*, Algorithms and models for the web graph, Lecture Notes in Comput. Sci., vol. 6732, Springer, Heidelberg, 2011, pp. 14–25, DOI 10.1007/978-3-642-21286-4_2. MR2842309

[LR20] Nicholas Landry and Juan G. Restrepo, *The effect of heterogeneity on hypergraph contagion models*, Chaos **30** (2020), no. 10, 3117. https://doi.org/10.1063/5.0020034.

[NBW06] Mark Newman, Albert-László Barabási, and Duncan J. Watts (eds.), *The structure and dynamics of networks*, Princeton Studies in Complexity, Princeton University Press, Princeton, NJ, 2006. MR2352222

Credits

Figures 1 and 2 were created by Sinan Aksoy.