NATIONAL INTELLIGENCE UNIVERSITY

HIGHER-ORDER NETWORK METHODS FOR INTELLIGENCE APPLICATIONS

A Compendium for Practitioners

Edited by

Matthew J. Hasenjager,
Mark M. Bailey, and
Nina H. Fefferman

Higher-Order Network Methods for Intelligence Applications

A Compendium for Practitioners

Edited by Matthew J. Hasenjager, Mark M. Bailey, and Nina H. Fefferman

NATIONAL INTELLIGENCE PRESS

NI Press produces works of value that advance US national security. Books published by NI Press undergo peer review by subject matter experts within and outside the US Government. That said, the opinions, conclusions, and recommendations expressed or implied within are solely those of the authors and do not represent the views of National Intelligence University (NIU) or any other agency of the US Government. Cleared for public release.

Editor, NI Press
Spring 2026
National Intelligence University
Bethesda, MD

ISBN: 978-1-932946-57-4
Library of Congress Control Number: 2026943243

Electronic copies of NI Press publications are available on NIU's website at: https://www.ni-u.edu/ni-books/.

Cover image: Shutterstock

Contents

Foreword

The study of complex systems has long been constrained by the limitations of conventional network analysis in which interactions are represented as pairwise connections, because real-world systems—ranging from geopolitical alliances and cyber networks to scientific collaborations and social influence networks—are rarely reducible to such dyadic structures. This book addresses this shortfall by exploring the growing field of higher-order network analysis, encompassing methodologies that can provide richer, more nuanced considerations of complex interactions by explicitly modeling and investigating multiway interactions and relationships. Using hypergraphs, simplicial complexes, and related techniques to capture higher-order relationships unlocks novel analytical insights essential for intelligence applications.

The chapters that follow collectively advance the state of the art in hypernetwork science, each offering a distinct contribution to our understanding of complex systems.

Chapter 1: *Higher-Order Networks: A Primer,* by Matthew J. Hasenjager, Mark M. Bailey, and Nina H. Fefferman, provides a foundational introduction to the theory of hypergraphs and related approaches, emphasizing why traditional dyadic, graph-based networks fail to fully encapsulate the complexity of real-world systems. The authors frame hypergraphs as an indispensable tool for intelligence analysis that can advance understanding of emergent properties in diverse contexts, including cyber-physical networks, social contagion, and strategic interactions.

Chapter 2: *From Logs to Dynamic Hypergraphs: Identifying and Interpreting Anomalous Cyber Activity,* by Emilie Purvine and Helen Jenne, shifts the focus to cyber security. The authors demonstrate how hypergraphs can model multiway relationships in cyber logs to capture structures that evade detection in conventional network representations. By applying dynamic hypergraphs and topological data analysis to the Operationally Transparent Cyber dataset, they illustrate how evolving hypernetwork structures can reveal anomalous activities indicative of cyber threats.

Chapter 3: *An Invitation to Hypergraph Motifs for Practitioners,* by Sinan Aksoy, addresses a fundamental challenge in hypernetwork analysis: the identification of significant structural patterns that exist only in higher-order networks. This chapter introduces hypergraph motifs—recurring higher-order interaction patterns—to distinguish meaningful structures in intelligence and security contexts. Motifs provide a lens through which covert coordination and influence networks can be detected and the relevance of their structure and organization can be explored.

Chapter 4: *Higher-Order Interactions at Scientific Conferences Influence Team Formation,* by Emma Zajdela and Nicholas Landry, extends hypergraph analysis to collaboration networks. By studying interdisciplinary scientific conferences, they show how hypergraphs better capture the multiway interactions that drive team formation—providing direct insights to intelligence organizations seeking to optimize collaborative networks for information sharing, innovation, and modeling real-world social networks.

Chapter 5: *Evaluating Node Influence Using Higher-Order Centrality,* by Matthew J. Hasenjager, Mark M. Bailey, and Nina H. Fefferman, examines how hypernetwork centrality metrics extend beyond traditional dyadic measures to identify influential nodes and hyperedges. Using s-line graphs, the authors introduce a method for quantifying node influence within hypernetworks and demonstrate its utility through simulations on real-world social network data, comparing its effectiveness in modeling contagion dynamics, information spread, and resilience in complex systems. This work underscores the need for higher-order centrality metrics

for intelligence applications, where understanding the true structure of influence and interaction is critical.

Chapter 6: *Cracking the Paradigm: Hypergraph Signatures of International Conflict,* by Thomas Pike, presents a groundbreaking application of hypergraphs to strategic intelligence. Using event data from conflicts in Ukraine and the Middle East, Dr. Pike constructs hypergraph signatures of state and nonstate actors' behaviors that offer a new framework for detecting shifts in conflict dynamics, strategic alliances, and emergent threats.

Chapter 7: *A Scenario Discovery Approach Integrating Hypergraph Representations and Generative AI,* by Kyle A. Kilian, explores using hypergraphs to model complex, alternative futures—allowing intelligence analysts to visualize and interrogate how different conditions and events intersect. He introduces a scenario discovery approach that integrates computational modeling with generative AI to identify high-risk futures. Using Monte Carlo simulations and large language models, this method uncovers crucial scenario pathways and interdependencies to provide deeper insight into emerging threats and decisionmaking strategies.

Chapter 8: *Hypergraph Methods for Network Analysis and Intrusion Detection,* by Zong-Zhi Lin, explores hypergraphs' ability to enhance Network Intrusion Detection Systems (NIDS). Traditional graph-based approaches struggle to capture the multiway relationships essential for identifying sophisticated cyber threats, but Lin demonstrates how hypergraphs enable advanced machine learning-based NIDS by modeling interactions between source and destination Internet Protocols, ports, and network segments within a zero-trust framework. Hypergraph-based analysis of cyber security datasets provides new methodologies for detecting anomalies, port scanning attacks, and lateral movement in cyber operations.

Collectively, these contributions highlight the breadth of use and transformative potential of hypergraphs and related methods for intelligence applications that can push beyond the limitations of traditional graph-based analyses to enable intelligence professionals to detect hidden patterns, anticipate threats, and make more informed strategic decisions. As the field matures, the integration of hypergraph theory into

intelligence workflows will shape the future of national security, cyber security, and geopolitical strategy. It is our hope that this book serves not only as a foundational reference but also sparks further exploration within the Intelligence Community into hypernetwork science. The ability to analyze higher-order structures in an increasingly complex and interconnected world is not merely an academic exercise, it is an operational imperative.

CHAPTER 1

A Primer on Higher-Order Networks

Matthew J. Hasenjager, Mark M. Bailey, and Nina H. Fefferman

In complex systems, interactions among lower-level entities can produce emergent, system-level patterns of organization and behavior. Familiar examples include polarization of individuals within social networks, synchronized firing of neurons within the brain, cascading shocks through interconnected global supply chains, and species interacting within ecosystems. A common feature shared by these disparate examples is that both the behavior of an individual agent (for example, a person, neuron, business, or species) and system-level dynamics and outcomes, such as the extent, duration, and severity of disruptions within global supply chains, cannot be fully understood without considering the broader web of interactions in which the agents are embedded.

Network analysis has emerged as a prominent paradigm through which we quantitatively describe systems and understand how they function.[1] Networks represent lower-level entities within a system as *nodes* connected by *edges* that denote some form of interaction or relationship (Fig. 1a), which could include social interactions, business transactions, signaling pathways, genetic relatedness, and so on. Drawing on a rich history of graph theory in mathematics,[2] sociology,[3] and physics,[4] network analysis uses a broad range of tools to describe the patterning of

interactions (from the individual to the network level) and elucidate the consequences of that patterning for system functioning. The widely applicable nature of network-based approaches facilitates extensive cross-pollination of ideas and techniques across diverse disciplines. Indeed, networks have been used to investigate phenomena as diverse as social contagion of attitudes, behaviors, and innovations,[5] the transmission of disease,[6] synchronization of behavior,[7] consensus formation,[8] the emergence of cooperation,[9] and the robustness of system functioning in the face of disruption.[10]

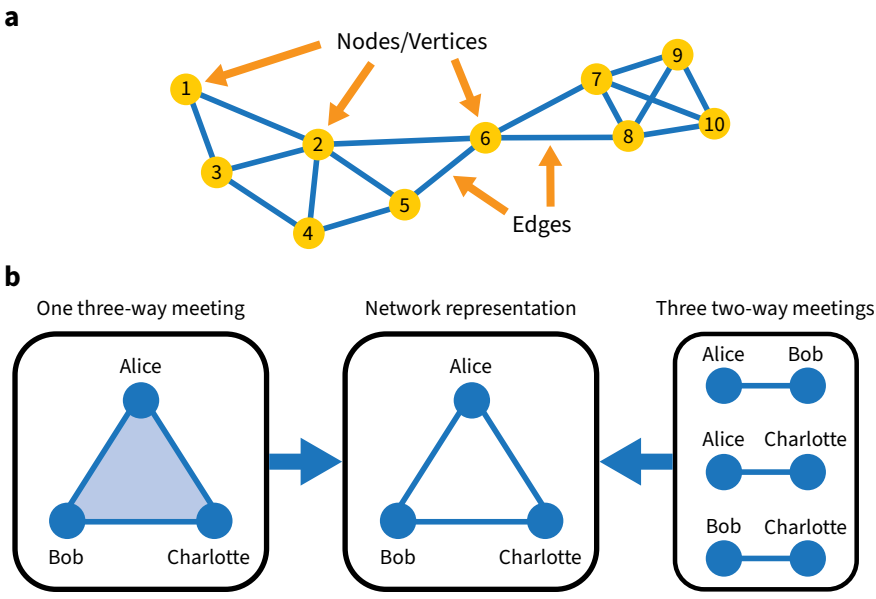


Figure 1. a) A conventional network is composed of a set of nodes (here, representing individuals) and edges that denote a dyadic interaction/relationship between the nodes. b) An issue arises in that a simultaneous, multiway interaction (here, among three individuals) shares an identical network representation as the set of all pairwise interactions among a common set of participants.

Mathematical approaches for gaining insight into complex phenomena naturally converge with intelligence analysis, where complex systems are the norm. Geopolitical systems, complex social systems, supply chains, and cyber-physical networks are just a few of the many complex

systems of interest to the Intelligence Community (IC), where maintaining decision advantage hinges on the ability to anticipate patterns in seemingly chaotic environments or highly complex datasets. This makes developing robust mathematical means to understand connectivity and emergence in complex systems critically important.[11]

Although flexible network representations can portray a broad range of phenomena, traditional networks are limited to representing interactions in a purely dyadic fashion. That is, each link in the network connects exactly two nodes. It is impossible to explicitly represent higher-order interactions (involving three or more entities) using a graph-based network in such a way that it excludes alternative, equally consistent dyadic interpretations. For example, one can envision two scenarios in which three individuals (Alice, Bob, and Charlotte) either jointly participate in a single meeting—a simultaneous, three-way interaction—or participate in three separate one-on-one meetings (Alice/Bob, Alice/Charlotte, and Bob/Charlotte). Both scenarios share the same underlying network representation (Fig. 1b) yet could vary dramatically in terms of interaction dynamics and collective outcomes, such as the knowledge that was shared, to whom, and when.

The importance of considering the presence and consequences of such higher-order interactions is being increasingly recognized across numerous fields of study.[12] For example, higher-order phenomena can occur within communication networks (eavesdropping and audience effects),[13] social networks (peer pressure and related mechanisms of social reinforcement),[14] ecological networks (indirect interactions),[15] and many others. Likewise, intelligence analysts often deal with systems that possess higher-order structure, interactions, and properties. Examples include cyber-physical systems (multiple processes being carried out by a common Internet Protocol address,)[16]) multiway geopolitical alliances,[17] collaboration and communication among diverse actors,[18] and coordination across multiple groups during military actions. Consequently, researchers have begun to turn to alternative frameworks that can explicitly represent higher-order interactions within networks.

Rather than being portrayed as the edge between two nodes, higher-order networks define interactions and relationships as sets of nodes, such

that the number of nodes that can engage in an interaction is no longer constrained. A higher-order interaction is distinguished from a dyadic one simply by whether its node set contains more than two or only two nodes. Although the basic frameworks that underpin higher-order networks have long existed,[19] a renewed focus on the potential utility of these approaches has emerged.[20] Investigations of higher-order network phenomena have found that the dynamics of complex systems can fundamentally differ when higher-order interactions are considered. For example, in examining the spread of behaviors or attitudes through a population, higher-order effects can manifest as an increased likelihood of adopting a behavior when it is simultaneously promoted by multiple individuals (akin to peer pressure).[21] Relative to spreading via simple contagion processes, where multiple exposures to previous adopters are treated as independent in terms of their impact on an individual's likelihood of adoption, allowing for higher-order contagion dynamics can generate novel phenomena—such as bistability* of endemic and nonendemic phases or explosive transitions from low to high prevalence of a behavior or attitude.[22] Thus, higher-order networks allow us to probe the consequences of not just whether individuals interacted, but also the precise manner in which they interacted.[23]

The aims of this chapter are three-fold. First, to introduce key concepts of higher-order networks in a way that demonstrates why reliance on conventional, graph-based (dyadic) networks may often be insufficient. Second, to illustrate an array of mathematical tools that can provide insight into the structure and function of higher-order networks, focusing on measures and concepts that will be further explored in later sections of this volume. Finally, to provide a brief overview of available open source software that can facilitate analysis of higher-order networks. Although these approaches are not yet as well developed as graph-based networks, scientists, analysts, and researchers are rapidly developing new tools and methodologies. It is our hope that this chapter will spur interest in, and engagement with, higher-order network approaches to tackle contemporary challenges faced by the IC.

---

* Bistability occurs when a dynamical system has two stable equilibrium states.

Higher-Order Network Representations

A comprehensive introduction to network analysis and graph theory is beyond the scope of this chapter, and we refer the interested reader elsewhere for details.[24] Some basic definitions, however, will be useful moving forward.

A graph-based network, $G = (V, E)$, is defined as a set of n vertices or nodes, $V = \{v_1, \ldots v_n\}$, and m connections, $E \subseteq \big\{\{i, j\} \,|\, i, j \in N,\, i \neq j\big\}$, such that each connection (referred to in network terminology as an edge) links together a set of nodes i and j. This definition is sufficiently flexible to model diverse systems, where nodes and edges could respectively represent neurons and synapses, individuals and social relationships, or nations and alliances. In addition, this basic framework can be expanded to capture additional information about the relationships among nodes—such as the intensity or directionality of interactions—or temporal information, such as by building a time series of networks representing interactions at different points in time.[25] Crucially, however, the above definition of a graph restricts each edge to connecting at most two nodes and, thus, can fail to capture important aspects of the system under consideration, as previously discussed.

One graph-based approach to modeling such higher-order interactions is to represent the system as a bipartite graph composed of two node sets (U, W), where nodes $u \in U$ can represent individual entities as described above and nodes $w \in W$ represent the interactions or groupings those entities participate (Fig. 2). Edges then connect the entities in U to the groupings or interactions they take part in, W, such that $E \subseteq \big\{\{u, w\} \mid u \in U \text{ and } w \in W\big\}$. A node w can be linked to any number of nodes u, meaning that groupings of different dimensionality can be represented. This representation, however, implies that lower-level entities no longer interact directly with one another but interact indirectly via their shared groupings. Depending on the goals of the analysis, this approach can be useful for modeling higher-order interactions and relationships, but in other circumstances, it may be more appropriate to use approaches that explicitly allow for interactions among nodes across groups of different sizes.

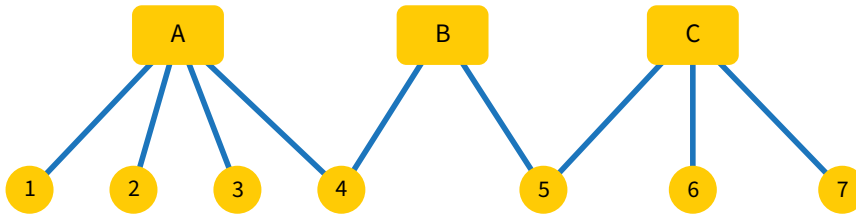


Figure 2. An example bipartite network, where edges exclusively exist between nodes of sets {A, B, C} and {1, …, 8}—for example, nodes 1 and 2 are only indirectly connected via node A.

In recent years, there has been renewed interest in network approaches that explicitly code higher-order interactions by relaxing the restriction that edges must connect no more than two nodes. Two primary approaches to representing such higher-order networks—simplicial sets and hypergraphs—both adopt a set-based representation of network edges that allow for edges to contain any (positive) number of nodes. Because they originate from different mathematical fields and can differ in their underlying assumptions, however, their suitability can vary with context.[26]

We first turn to simplicial sets, which are defined as sets of k-simplices, where a k-simplex is defined as a set of $k + 1$ nodes.[27] Thus, a 0-simplex represents a single node or vertex, a 1-simplex represents a dyadic interaction, a 2-simplex represents a three-way interaction, and so on. A graph-based network could then be equivalently represented as a simplicial set composed exclusively of 1-simplices. Although the definition of a simplicial set is quite general, most tools for analyzing simplicial sets have been developed in the field of algebraic topology, where a simplicial set is distinguished from a simplicial complex.[28] The latter requires that any simplex must also contain all possible lower-order subsets (also referred to as "downward closure"). Thus, within a simplicial complex, the 2-simplex {A, B, C}, requires the inclusion of the 1-simplices {A, B}, {A, C}, and {B, C}, and the 0-simplices {A}, {B}, and {C} (Fig. 3).

This is done primarily to make analytic proofs about structural features more tractable, and, in some contexts, this assumption is warranted. For example, imagine that simplices represent associations based on spatial proximity, where all vertices within a simplex must be physically located

within some distance threshold from one another. In such a scenario, all subsets of vertices within a simplex would likewise meet this spatial criterion and so can be included as lower-order simplices within this simplex. In many scenarios, however, the additional assumptions required to form a valid simplicial complex are too restrictive. In a series of email or text exchanges, for instance, it is possible that group messages exist without each pair of participants also having separately messaged one another.

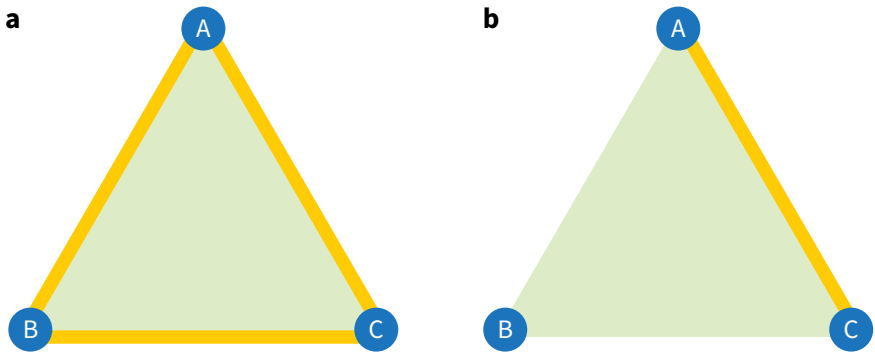


Figure 3. a) To be a valid simplicial complex, the 2-simplex {A, B, C} (denoted by the filled green triangle) must include all possible subsets—that is, the 1-simplices {A, B}, {B, C}, and {A, C} (denoted by yellow lines) and the 0-simplices {A}, {B}, and {C} (denoted by blue circles). **b)** As neither simplex {A, B} nor {B, C} are present, this configuration constitutes a simplicial set, but it is not a simplicial complex.

Hypernetworks offer an alternative higher-order network representation and can be formally defined as a hypergraph, $H = (N, E)$, consisting of a set of n nodes, $N = \{n_1, \ldots, n_n\}$ and an indexed family of m sets referred to as hyperedges, $E = \{e_1, \ldots, e_m\}$, in which $e_i \subseteq N$ for i = 1, …, m.[29] A hyperedge of size s contains s nodes, where $s \geq 1$. As with simplicial sets, hypergraphs define interactions and relationships using set membership and do not share the requirement of a simplicial complex that all possible lower-order subsets be embedded within each higher-order relationship. This flexibility makes hypergraphs and simplicial sets suitable for scenarios in which the assumptions of a simplicial complex are too restrictive. For example, in a coauthorship network, the existence

of a paper with three authors does not necessarily imply that each pair of authors also published separate papers together.

A limitation of hypergraphs relative to simplicial sets, however, is that they cannot encode subedges within a larger hyperedge. For example, within our earlier example of Alice, Bob, and Charlotte engaging in a three-person meeting, it is not possible to differentiate a scenario in which each individual uniquely contributes to the discussion versus a scenario in which two individuals (Alice and Bob) engage in a subinteraction that then contributes to the larger three-way interaction. Such distinctions can be made naturally when using a simplicial set representation. Therefore, analysts should consider carefully which higher-order representation is best suited for the system and questions of interest.†

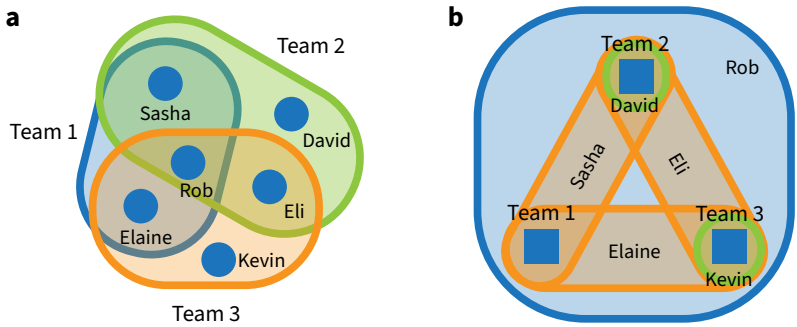


Figure 4. a) A hypergraph where six individuals participated in three teams. b) The dual of the hypergraph **a,** where the individuals are hyperedges and the teams are represented as nodes.

An important concept in hypernetwork analysis is the existence of a hypergraph's dual, which is obtained by swapping the relationship between hyperedge and node. For example, a hypergraph (Fig. 4a) that represents individuals as nodes and their participation in a series of teams through

† For further discussion on how to select an appropriate representation for modeling a complex system, see Torres et al., "The Why, How, and When of Representations for Complex Systems," *SIAM Review* 63, no. 3 (2021): 435–485, https://doi.org/10.1137/20M1355896.

hyperedge membership has a corresponding dual hypergraph (Fig. 4b) in which the individuals are the hyperedges that contain the teams (represented as nodes) in which the individuals participated. This conversion offers an additional way to characterize connectedness in higher-order systems that can prove useful in various contexts. As another example, in a hypergraph representing conflict events and participants, an analyst may gain richer insights by considering both the scenario where hyperedges represent conflicts (with the participants of those conflicts treated as nodes embedded within each hyperedge) and one where the participants are coded as hyperedges that contain their corresponding conflicts.

Higher-order networks, whether simplicial sets or hypergraphs, can be mathematically represented in a variety of ways. Here, we will use the language of hypergraphs unless the interpretation would differ for a simplicial set. A general means of encoding higher-order relationships is through an incidence matrix, *I* (Fig. 5b) where the rows and columns correspond to the nodes and hyperedges respectively. Thus, nonzero entries indicate that the corresponding node is a member of the set that defines the corresponding hyperedge. From an unweighted incidence matrix where elements are coded either 0 or 1, it is then a simple matter to obtain the number of hyperedges to which a node belongs (a version of degree centrality; see below) as the sum of its corresponding row. The column sums of the incidence matrix provide the hyperedge size sequence. An incidence matrix can also be weighted, for example, to indicate the number of times that a given hyperedge (defined by a unique set of nodes) has formed.

From an incidence matrix, the intersection profile of a higher-order network can be obtained as $P = I^T I$, producing an *m x m* matrix, where *m* is the number of hyperedges (Fig. 5c). A nonzero entry $P_{\alpha\beta}$ indicates the number of nodes that are shared by hyperedges A and B. The intersection profile can in turn be used to construct line graph representations of higher-order systems. A line graph (also called an intersection graph, representative graph, or the one-mode projection of a 2-mode (that is, bipartite graph) is a dyadic representation of hyperedge intersections, such that each hyperedge is represented through a node, and nodes in the line graph are linked if the corresponding hyperedges intersect at one or more nodes.[30] Line graphs can be further extended to define the minimum

intersection size, *s,* required to link two hyperedges together, giving rise to *s*-line graphs. Many conventional network metrics can be measured on these *s*-line graphs (for example, closeness centrality),[31] providing a relatively straightforward way to capture elements of higher-order connectedness. Projecting a hypergraph onto a series of *s*-line graphs, however, still involves the loss of information; as a set of s-line graphs does not uniquely encode a specific hypergraph structure, it is not necessarily possible to recover the latter from the former.[32]

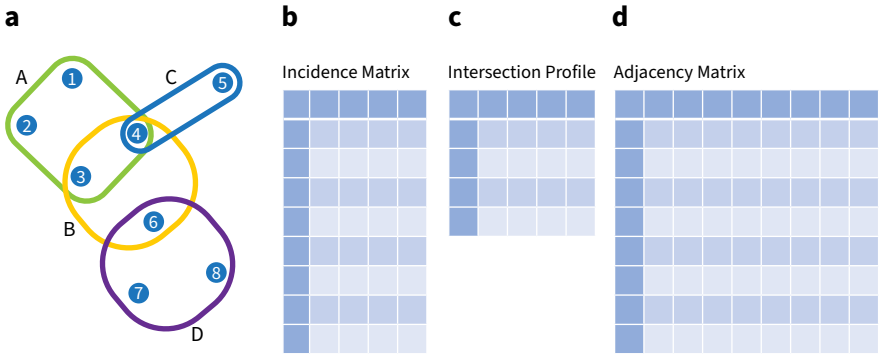


Figure 5. An example hypergraph and its corresponding incidence matrix, intersection profile, and adjacency matrix. Nodes are labeled 1-8 and hyperedges are labeled a-d.

From the incidence matrix, it is also possible to obtain the adjacency matrix as $A = II^T - D$, where D is the diagonal matrix (Fig. 5d). A non-zero entry in A_{ij} indicates the number of hyperedges that nodes i and j share membership. This transformation, however, involves loss of information on higher-order relationships, as the size of the hyperedges in which i and j jointly occur can no longer be obtained from A. Indeed, the adjacency matrix is a common means of representing dyadic network data.

Measuring Higher-Order Structure

The goal of network analysis is to gain insight into how a system is structured and the resulting consequences of that structure for system processes. Network analysts have developed numerous measures to capture diverse

aspects of network structure (a term that refers to the patterning of connections among nodes) from the level of individual nodes up to the entire system. While it is beyond the scope of this chapter to comprehensively cover each measure available to an analyst,[33] the following sections provide a broad introduction to a subset of network measures and their higher-order counterparts that are especially common or used later in this book.

Centrality Measures

Network centrality measures quantify differences across nodes in their potential importance. They can be based on criteria such as the number or strength of their connections, or on the extent to which nodes are located along key paths through the network. The implicit assumption is that nodes with numerous and stronger connections, or that occupy key positions in a network, have potentially greater influence over or importance in system processes or dynamics. Within a higher-order context, these notions can be applied to both the nodes and the hyperedges themselves. Indeed, the relative positioning of a collection of nodes (that is, a hyperedge) may be more informative than the positioning of any individual node itself, depending on the context.

Starting with the simplest centrality measure, degree refers to the number of other nodes to which a focal node is connected. In a higher-order context, degree can be defined either as the number of nodes that a focal node is adjacent to (meaning they have one or more hyperedges in common) or the number of hyperedges that contain that node (such hyperedges are said to be incident to that node).[34] For example, in the hypergraph depicted in Fig. 5a, node 4 has five adjacent nodes and three incident hyperedges. Weighted extensions of degree centrality can also be defined to capture variation across hyperedges in their relative intensity or importance, although the precise definition of weighted degree (also referred to as strength) depends on how hyperedge weights are defined.[35] For example, hyperedges could receive weights corresponding to how often they occur or to their size (how many nodes they contain, referred to as cardinality). Regardless of definition, nodes with a high-weighted degree or strength are generally predicted to be relatively important or influential in the network.

Degree (both weighted and unweighted) can also be applied to hyperedges as a function of the number of hyperedges adjacent to a focal hyperedge. When working with a simplicial set representation, degree can further be extended to account for whether adjacent simplices are of a lower or higher dimension (lower and upper degree, respectively).‡

Beyond degree, centrality can also be defined in terms of the extent to which a node (or hyperedge) lies along key pathways within the network. In a dyadic network, a path is a sequence of adjacent nodes, where no node is revisited (Fig. 6); if a node can be revisited, this is called a walk. The shortest path between two nodes is often assumed to be the most efficient route when considering flows on the network of information, resources, material, and so forth. Building on this concept, we can define measures, such as betweenness centrality (the number of shortest paths between pairs of nodes that pass through a focal node), closeness centrality (the mean length of the shortest paths from a focal node to all other nodes in the network), harmonic centrality (a variant of closeness centrality that can be applied to disconnected networks), or eccentricity (the longest path length across all the shortest paths between a focal node and all other nodes in the network).[36]

In a hypernetwork, it is often more useful to define paths on the hyperedges rather than the nodes and then adjust the above definitions of path-based centralities accordingly.[37] A notable difference between dyadic and higher-order paths is that higher-order paths can vary in both length—the number of hyperedges in the path—and width—the number of nodes at which two hyperedges intersect (Fig. 7). This has led to the development of *s*-closeness centrality, *s*-harmonic centrality, *s*-betweenness centrality, and *s*-eccentricity,[38] where *s* refers to the minimum intersection

‡ For a comprehensive treatment of these concepts, see Ernesto Estrada and Grant J. Ross, "Centralities in Simplicial Complexes. Applications to Protein Interaction Networks," *Journal of Theoretical Biology* 438 (2018): 46–60, https://doi.org/10.1016/j.jtbi.2017.11.003; and Daniel H. Serrano and Darío S. Gómez, "Centrality Measures in Simplicial Complexes: Applications of Topological Data Analysis to Network Science," Cornell University Algebraic Topology, *arXiv* (2020), 1908.02967, https://doi.org/10.48550/arXiv.1908.02967.

size necessary to be considered a valid path. Thus, a 1-path is one where each pair of adjacent hyperedges in the path share one or more nodes, a 2-path requires that each pair shares two or more nodes, and so on. As with degree centrality, when working with a simplicial set representation, these measures can also take into account whether adjacent simplices are lower- or upper-adjacent.[39]

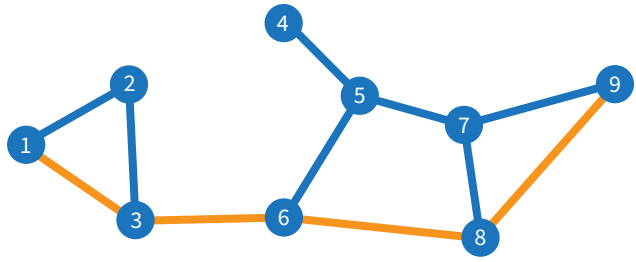


Figure 6. The shortest path connecting nodes 1 and 9 is highlighted in orange and consists of the sequence of nodes {1, 3, 6, 8, and 9}. Longer paths also exist between these nodes (for example, {1, 3, 6, 5, 7, and 9}).

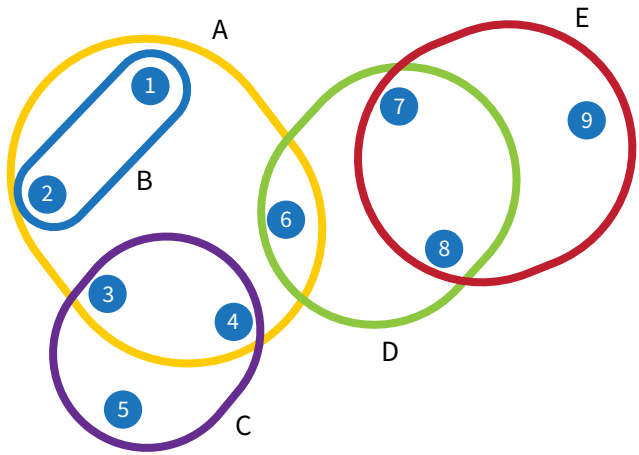


Figure 7. Hypernetwork paths have both length and width. For example, the intersection between hyperedges A and D in the graph is of size 1 (node 6), while all other intersections are of size 2. If paths are restricted to follow intersections of size 2 or greater, then no valid path exists between hyperedges A and D. Hyperedge A has the highest 2-betweenness centrality (meaning paths can traverse intersections of size 2 or greater), as all the shortest paths starting or ending at hyperedges B or C must pass through A.

A final class of centrality metrics introduced here is eigenvector centrality, which quantifies a node's centrality in a way that accounts for the centrality of the nodes to which it is connected. For a dyadic, graph-based network, this is typically derived from the first eigenvector of the adjacency matrix.[40] A node's eigenvector centrality is proportional to the sum of the centralities of its network neighbors. Put simply, nodes with high eigenvector centrality tend to be connected to many other nodes, that, in turn, are connected to many nodes. It captures the frequency of arriving at the node by a random walk across the network. Variants of this approach include measures such as Google's PageRank approach for ranking the importance of websites.[41] Approaches for calculating eigenvector centralities appropriate for higher-order networks include using the adjacency matrices of bipartite graphs[42] and tensor-based methods.[43]

Clustering and Embeddedness

In addition to centrality, network measures can also characterize a node's local environment in terms of how tightly clustered its neighbors are to one another. Specifically, the clustering coefficient measures the extent to which a node's immediate network neighbors tend to be directly connected to themselves. More formally, a focal node *i*'s local clustering coefficient is the ratio between the actual and potential number of closed triads in the neighborhood of , where a closed triad among nodes $\{i, j, k\}$ mean that edges (i, j), (i, k), and (j, k) all exist.

These concepts of triadic closure and clustering coefficients have been extended to higher-order networks in various ways. For example, Greening defined local clustering coefficients on simplicial sets as the extent to which nodes that are part of a simplex with focal node, *i,* also interact with one another in another simplex of the same size.[44] Another approach has been to define clustering coefficients for hyperedges as the number of *s*-triangles involving a hyperedge *f* divided by the number of *s*-wedges centered on hyperedge *f*.[45] An *s*-triangle is defined as a closed *s*-path of length 3, where closed means that the path starts and ends at the same node, and an *s*-wedge is an *s*-path of length 2 (Fig. 8). As with

the path-based centrality measures above, s refers to the minimum permissible intersection size between adjacent hyperedges that can be considered as part of a valid path.§

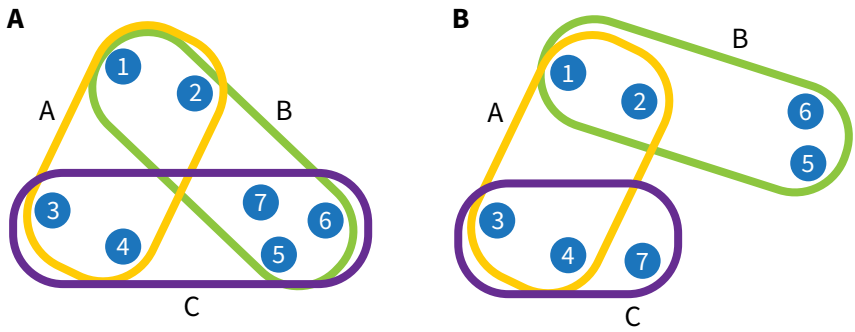


Figure 8. A) An s-triangle where $s = 2$. Note that an s-wedge centered on A also exists here. B) An s-wedge centered on hyperedge A for which no corresponding s-triangle exists.

Another measure that captures aspects of the local environment of a higher-order interaction is sub-edge density. Whereas sub-edge density is defined for networks by the density of edges in randomly sampled induced subgraphs,[46] for hypernetworks it is defined as the proportion of possible subsets of the nodes making up a hyperedge of size 3 or greater that also exist in the hypernetwork.[47] For example, imagine the hyperedges $\{a, b, c\}$, $\{d, e, f\}$, $\{a, b\}$, $\{a, c\}$, $\{b, c\}$, and $\{e, f\}$. The hyperedge $\{a, b, c\}$ has a higher subedge density than hyperedge $\{d, e, f\}$ because of the absence of hyperedges $\{d, e\}$ and $\{d, f\}$. Put simply, subedge density captures the extent to which nodes that interact in a group of a given size also interact in smaller groupings. This measure may reflect

§ Additional variants of the clustering coefficient suitable for higher-order networks are discussed in Ernesto Estrada and Juan A. Rodríguez-Velázquez, "Subgraph Centrality and Clustering in Complex Hyper-Networks," *Physica A: Statistical Mechanics and Its Applications* 364 (2006): 581–94, https://doi.org/10.1016/j.physa.2005.12.002; and Wanding Zhou and Luay Nakhleh, "Properties of Metabolic Graphs: Biological Organization or Representation Artifacts?" *BMC Bioinformatics* 12 (2011): 132, https://doi.org/10.1186/1471-2105-12-132.

how vulnerable a higher-order structure is to targeted removal of specific nodes. Hypernetworks with higher subedge density are likely to be more robust to node removal, as there are likely to be more hyperedges involving neighbors of the affected node that can maintain connectivity in that node's absence.

Network-Level Measures of Structure

In addition to considering measures that offer insight into the relative importance of particular nodes or hyperedges, we can use metrics that characterize structural patterns at the level of the entire network. A simple approach is to take the mean of node or hyperedge-level structural measures (for example, mean degree centrality). Similarly, measures such as the clustering coefficient or sub-edge density can be defined so they capture network-level patterns. For example, the *s*-global clustering coefficient can be defined as the ratio of the total number of *s*-triangles to *s*-wedges, multiplied by three.[48] A hypernetwork with a high *s*-global clustering coefficient is one in which hyperedges tend to form dense local neighborhoods, where adjacent hyperedges tend to share at least *s* nodes in common.

Beyond the triangle substructures that tend to be the focus of clustering coefficients, approaches have been developed to examine the occurrence of a variety of substructures formed by groups of nodes or hyperedges. When a particular substructure configuration is overrepresented in a network, relative to the frequency expected by chance, such substructures are referred to as network motifs (Fig. 9).[49] Enumerating the presence and types of motifs in a (hyper)network can offer insight into how that network is likely to function. For example, a well-studied network motif in dyadic networks is the feed-forward loop, a 3-node configuration where one node sends both a direct outgoing connection to a second node and an indirect connection to that node by way of a third (Fig. 9a). Because of the time delay for signals traveling along the indirect versus direct path, this motif can function as an activator that is sensitive to sustained stimulation but rapidly deactivates when the stimulus is removed.[50] As another example, the bi-parallel motif may reduce flow

across each path but increase the likelihood that the destination node will be reached. Motifs can be extended to hypernetworks to explore whether groups of hyperedges tend to be interconnected in specific recurrent ways. Although this work remains in its infancy, it offers great potential to shed light on the formation and functioning of higher-order systems.[51]

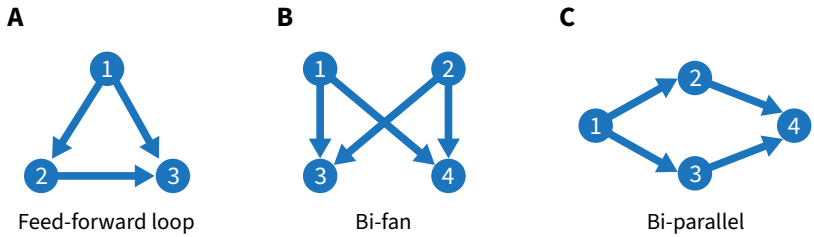


Figure 9. Examples of motifs for dyadic networks.

A final set of approaches introduced here are used to infer grouping patterns among nodes, either as a function of node-level traits (that is, assortativity) or by identifying clusters of densely interconnected nodes (referred to as community detection). Assortativity captures the extent to which nodes tend to be connected to other nodes that are either similar or dissimilar to them in regard to some node-level trait (referred to as homophily and heterophily, respectively).[52] In social networks, for example, individuals are often more likely to be connected to others who share their hobbies, religious beliefs, political affiliation, socioeconomic status, and so forth. Similarly, community detection seeks to identify groups of especially strongly connected nodes.[53] For example, in a scientific collaboration network, we might expect to find a higher density of connections among scientists working in the same or related fields (ecology and evolutionary biology) than among those working in very different fields (ecology versus astrophysics). It is important to note that unlike with assortativity, node-level labels (in this example, the scientific fields) are not used to define the community. Instead, community detection algorithms identify subdivisions based purely on the patterning of connections across nodes, searching for groups of relatively densely connected nodes. Node-level traits, however, may be useful in making sense of the community structure identified by an algorithm.

In a higher-order context, concepts of assortativity and community structure are more challenging to define. For instance, when hyperedges vary in size, how should this information be considered when determining whether two nodes are part of the same community based on shared hyperedge membership? Nevertheless, several community detection methods suitable for higher-order networks have recently been proposed, including hypergraph modularity maximization approaches,[54] spectral clustering,[55] and model-based approaches.[56]

Software Packages for Higher-Order Network Analysis

A proliferation of software packages suitable for analyzing and modeling higher-order networks have emerged in recent years, including many used in the R and Python open source computing environments. In R, the packages *hypergraph*[57] and *HyperG*[58] offer functions for storing and manipulating hypergraph data, calculating an array of hypernetwork measures, and producing sample hypergraphs with generative models. Similarly, in Python, the packages *XGI,*[59] *HyperNetX,*[60] and *Hypergraphx*[61] offer functions for storing and manipulating higher-order network structures, calculating a variety of network metrics, creating network visualizations, and generating simulated higher-order networks. The *KaHyPar* package provides algorithms for hypergraph partitioning and can be used in both Python and Julia.[62]

Conclusion

The surge of interest in higher-order networks brings with it both exciting opportunities and significant challenges for the study of complex systems. It is not possible here (nor was it our aim) to comprehensively review the rapidly developing field of higher-order network analysis. It is our hope, however, that this section has provided a useful introduction to many of the approaches used in the analysis of both dyadic and higher-order networks, laying the groundwork for the remaining chapters of this volume

that apply these and related techniques to generate insights into the structure and behavior of a diverse array of higher-order systems.

About the Authors

Matthew J. Hasenjager, Ph.D., is an IC postdoctoral fellow in the Department of Ecology and Evolutionary Biology and the National Institute for Modeling Biological Systems at the University of Tennessee, Knoxville. His research focuses on complex networks, collective behavior, social learning, and communication. He has worked on diverse systems including fish shoals, honeybee colonies, and the development of ant-inspired decision tools for enhancing supply chain resilience.

Mark M. Bailey, Ph.D., is an associate professor and department chair for AI, Cyber, Influence, and Data Science and the Director of the Biological and Computational Intelligence Center at National Intelligence University. He is interested in complex systems, AI safety, and their intersection with national security issues. Dr. Bailey is the author of *Unknowable Minds: Philosophical Insights on AI and Autonomous Weapons*.

Nina H. Fefferman, Ph.D., is a professor in the Department of Mathematics and the Department of Ecology and Evolutionary Biology at the University of Tennessee, Knoxville, where she also serves as director of the National Institute for Modeling Biological Systems and director of the National Science Foundation Center for Analysis and Prediction of Pandemic Expansion. Her research focuses on the mathematics of self-organizing complex systems, most frequently in application to social behavior, epidemiology, and ecology.

Further Reading

Battiston, Federico et al. "Networks Beyond Pairwise Interactions: Structure and Dynamics." *Physics Reports* 874 (2020): 1–92. https://doi.org/10.1016/j.physrep.2020.05.004.

Newman, M. E. J. “The Structure and Function of Complex Networks.” *SIAM Review* 45, no. 2 (2003): 167–256. https://doi.org/10.1137/S00361 4450342480.

Torres, Leo et al. “The Why, How, and When of Representations for Complex Systems,” *SIAM Review* 63, no. 3 (2021): 435–85. https://doi .org/10.1137/20M1355896.

CHAPTER 2

From Logs to Dynamic Hypergraphs: Identifying and Interpreting Anomalous Cyber Activity

Emilie Purvine and Helen Jenne

Critical infrastructure, the military, and the government rely on cyber networks to gather and disseminate information. Unfortunately, these networks are far from being immune to compromise. Adversaries constantly attempt attacks from denial of service to exfiltration to degrade performance and confidence. Cyber security—the protection of these critical cyber systems and data from attack and unauthorized access—encompasses many things including network design and configuration, vulnerability assessment, malware analysis, alert triage, and real-time analytics, all of which require careful modeling of the complex relationships among computers, users, protocols, and data on the systems. In this chapter we will show by example how hypergraphs can be used to model these complex relationships and provide insight into critical cyber systems.

To enable cyber defenders to protect these elements nearly every action on each system and across the network are logged for real-time and forensic analysis. These logs, the structure of which can vary from

system to system and log type to log type, tend to be high dimensional with heterogeneous, mostly categorical dimensions. In network traffic logs the typical dimensions are source and destination Internet Protocol (IP) address (identifiers for the systems sending and receiving information), source and destination port (the digital doors that the information flows through at the sender and recipient), number of packets and bytes, network protocol (a very high-level type for the communication), and a timestamp and duration. In this type of data, the IP addresses, ports, and protocol are categorical, while bytes and packets are numerical. Host logs may contain different information depending on what is being logged. For instance, authentication events will include the username and the success or failure of login attempts, process logs can include the specific executable being run, file logs indicate which file is being touched. While there are several other types of logs, this high-level overview is meant to emphasize the complexity and variability of the data rather than provide a comprehensive introduction to log data.

Cyber defenders often rely on commercial or open source intrusion detection systems to look for signatures of activity that aligns with threat intelligence reporting. While these systems have their place in identifying and stopping known bad behavior, they are not as useful for defending against novel adversary activity, called zero-day attacks. To address emerging threats there has been increasingly more research into using machine learning (ML) to analyze behaviors without the need to define a large collection of rigid signatures. A recent survey of 117 papers on cyber threat hunting published between 2014 and 2024 categorized the approaches into supervised, unsupervised, and graph-based.[63] Notable methods within these categories include DeepLog,[64] which leverages a long short-term memory neural network model to identify anomalous logs, and Log2vec,[65] which constructs a graph from logs and then uses a graph embedding to represent the graph as a vector. Due to the challenges posed by imbalanced data, many approaches use autoencoders[66, 67] or graph algorithms,[68, 69, 70, 71] rather than traditional supervised learning. While ML methods often report high accuracy, a 2022 study of five state-of-the-art deep learning methods observed that due to issues, such as training data selection strategy, highly imbalanced data, and data noise,

all studied models did not always work as well as they claimed in their papers.[72] Additionally, ML methods often lack the explainability needed to help analysts understand why certain activity was flagged as unusual. Our approach, using hypergraphs and topological data analysis, aims to address explainability without sacrificing flexibility.

In the following sections we will show through examples how hypergraphs can be used for the dynamic analysis of cyber logs. Because of the complexity of cyber systems and their log data we suggest that using hypergraphs to develop analytics provides an important perspective on the data that other methods lack, namely the ability to capture multiway relationships in a way that enables analyst insight. We begin by introducing a particular dataset that will serve as our test set for the chapter and show how that data can be modeled and analyzed using dynamic hypergraphs. Then, we show how tracking the dynamics of simple hypernetwork science metrics captured some unusual activity. Finally, we explain a topological approach and show how this method identified a highly interpretable substructure that appeared repeatedly and often pointed us back to malicious activity in the raw data.

Introducing the Data

Throughout this chapter, to illustrate our methods, we will use the Operationally Transparent Cyber (OpTC) dataset released by the Defense Advanced Research Projects Agency (DARPA) in 2020.[73] The data were prepared for the Cyber Hunting at Scale (CHASE) and Transparent Computing (TC) programs to test scaling of advanced persistent threat (APT) detection techniques. After an initial period of benign activity on a simulated network of 1,000 hosts a red team performed APT scenarios over three days. On the first day, there was a Powershell empire staging scenario, including initial foothold establishment, lateral movement, and privilege escalations. The second day included data exfiltration using Netcat and remote desktop (RDP). The third day is the least noisy, the main attacks being malicious software upgrades. The benign activity continued during the red team period. The released dataset represents logs collected from 500 of the 1,000 hosts.

The data uses the eCAR format which unifies several types of logs—including network flow, file operations, process logs, registry interactions, and more—into a single data type to enable comprehensive analysis. eCAR is an event-based model based on MITRE's car model[74] where each event has three core components: *object*, *action*, and *fields*. An object is a type of observable element. Examples are FLOW, FILE, and PROCESS. An action is an event or state change of an object indicating what objects can do or what can happen to them. For the FILE object some example actions are CREATE, DELETE, or WRITE. The additional fields are metadata about the specific (object, action) log. Each object has its own types of fields, with some overlap. As noted above, example fields for FLOW are source and destination IP, ports, and protocol. For a much more comprehensive description of the data format, see Anjum et al.[75]

In the four years since the OpTC dataset was released there have been more than 20 papers focused on the problem of APT detection that use OpTC to evaluate their proposed methodology. The techniques in these papers include process tree-based methods,[76, 77] provenance graph-based methods,[78, 79] and embedding-based methods.[80, 81] Several of these methods rely only on the object, action, and time fields to augment the object ID, actor ID, process ID (PID), and parent PID (PPID) identifiers. In contrast, our work includes many other fields based on the object. Most other works achieved very high accuracy in identifying APT activity, more than 98 percent. It is worth pointing out, however, that across the prior work on the OpTC dataset there is not an agreed-upon labeling scheme, as discussed by Nikulshin and Talhi.[82] The red team ground truth supplied with the dataset is not in the form of labels on the events but rather a human readable timeline with several PIDs and IP addresses identified. Between the ambiguity in labeling, and variation in choices of which subsets of the data to evaluate on, it is difficult to compare competing approaches. Additionally, the significant class imbalance makes standard performance metrics, such as accuracy, difficult to meaningfully interpret. For these reasons, in this chapter we do not provide quantitative metrics but instead focus on how in-depth explorations of the data, much like a cyber analyst may do when investigating an alert,

were able to qualitatively tie our hypergraph-based anomalies and observations directly to malicious activity.

A Dynamic Hypergraph Model

Hypergraphs from Cyber Data

A collection of cyber events of the same type can be considered as a data table, where each row is an event, and each column is a field. See Table 1 for an example of such a table for the OpTC FLOW object.

PID	Src IP	Dst IP	Dst Port	Protocol	Image Path
4	142.20.56.198	142.20.59.255	138	UDP	System
864	10.20.4.125	224.0.0.252	5355	UDP	svchost.exe
864	142.20.59.255	224.0.0.252	5355	UDP	svchost.exe
636	142.20.56.198	222.206.244.5	443	TCP	firefox.exe
4	142.20.59.149	142.20.59.255	138	UDP	System
864	142.20.59.149	224.0.0.252	5355	UDP	svchost.exe

Table 1. Sample table of (FLOW, START) events. This represents only a subset of the additional fields that records of this type have.

Now recall that a hypergraph, $H = (V, E)$, consists of a set, $V = \{v_1, \ldots, v_n\}$, of vertices (or nodes) and a collection, $E = (e_1, \ldots, e_m)$, of (hyper)edges where each $e_i \subseteq V$. We can think of each edge $e \in E$ as a "behavior" such that each vertex $v \in e$ is an entity that exhibits that behavior in some data. From this interpretation, and the fact that a column or collection of columns in a data table can be considered as a set of entities or behaviors, we see that several hypergraphs could be constructed from a single data table. Take, for example, the image path column in Table 1. The behavior represented by this column is that a particular executable starts a FLOW event. The entity that runs this executable is the source IP. So, we could create a hypergraph $H_{sip,im} = (\text{Src IP}, \text{Image path})$ to relate those two columns. This is shown in Fig. 1 (left). As another

example, the image path column could represent the entities, and the behavior is destination port usage. Here we would create a hypergraph $H_{im,dp}$ = (Image path, Dst Port), shown in Fig. 1 (right). Although the data for both hypergraphs comes from the same table the structure of the two hypergraphs is vastly different. In $H_{sip,im}$ the hypergraph is connected, due to the vertex 142.20.59.149 running multiple executables. This hypergraph construction could give a high-level view of the behavior happening in a network. A high-degree vertex signifies a very active source IP, and a large edge can signify a common behavior, such as web browsing and background operating system processes. On the other hand, $H_{im,dp}$ has three disjoint, singleton hyperedges. There are many applications, such as Firefox, that often use the same or only a few different ports under normal use. Consequently, this construction will yield hypergraphs with many duplicate edges and small connected components and will not provide the same type of overview of behavior across a network. But when applied to events from an individual source IP it can provide a granular view of behavior that is very useful to track over time, as large changes in structure can indicate misuse.[83]

Figure 1. Two example hypergraphs from Table 1 formed by choosing a column to represent vertices and a column to represent edges. (Left) *V* = Src IP, *E* = Image path. (right) *V* = Image path, *E* = Dst Port.

From just these two examples we see that the types of cyber activity that can be captured and the conclusions drawn depends on the lens through which you look at the data. In the next two sections we will walk through some observations for a single hypergraph construction as an illustrative example of the approach, but exploring several constructions is worthwhile to gain different kinds of insight.

Dynamic Hypergraphs

We showed two examples of hypergraph creation from the FLOW data in Table 1 but there could be several more, especially if all possible FLOW columns are considered. But what was missing from consideration in this table and the hypergraph constructions is the timestamp column. Each FLOW event has a start time and duration, and the activity on a network throughout an hour, day, week, or more has high variance. Adversary actions take place sometimes very quickly and other times slowly over the course of time. This points to the need for a dynamic analysis of the data, which, for our approach, requires the notion of dynamic hypergraphs.

There are two basic ways to define dynamic hypergraphs: continuous and discrete (or "time-windowed"). In the continuous case, a hypergraph is constructed for every point in time, $t \in \mathrm{R}^+$ (here t represents a single instance in time). Namely, $\mathrm{H}_t = (\mathrm{V}_t, \mathrm{E}_t)$ where both V_t and E_t contain only those nodes and edges (entities and behaviors) that are active at time t. In the context of cyber data, only the log records where t falls between start time and start time plus duration contribute to H_t. Typically, in the context of cyber data, this results in an incredibly sparse hypergraph and is too variable to really learn anything from. Instead, we use discrete time dynamic hypergraphs built from time windows. Given a t and Δt a hypergraph $\mathrm{H}_{[t,t+\Delta t]} = (V_{[t,t+\Delta t]}, E_{[t,t+\Delta t]})$ is constructed by considering nodes and edges that are active any time within the window. In other words, any record whose [start, start + duration] window overlaps $[t, t + \Delta t]$ is used in creating the hypergraph. This notation is very clunky so in the remainder of this chapter we will state a window length (Δt) and step (τ) and create a sequence of hypergraphs $\{\mathrm{H}_i\}_{i \in \mathrm{N}}$ where $\mathrm{H}_i := \mathrm{H}_{[i\tau, i\tau+\Delta t]}$. Typically, τ is some portion of Δt, for example, $\tau = \Delta t/2$.

Dynamics of Basic Hypernetwork Measures in OpTC

While there are many tools to visualize (even large) graphs to get a sense of their structure, hypergraphs quickly become challenging to visualize and derive insight from. To understand the structure of the hypergraphs

we constructed and how they evolve over time, we began with lightweight analytics from hypernetwork science. By beginning our analysis with straightforward hypernetwork science metrics, we establish a baseline to which we can compare our topological metrics in the next section. If one of our topological metrics correctly picks out a hyperedge, vertex, or time period as anomalous that hypernetwork science measures miss, this indicates that topology provides valuable insights.

In this section we will show some examples of the evolution of hypernetwork science measures for time-windowed dynamic hypergraphs with $\Delta t = 10$ minutes and $\tau = 5$ minutes. Vertices represent (source IP, hostname) pairs and edges represent (destination IP, destination port) pairs. In this construction we think of vertices as representing individual computers and hyperedges as groups of computers sharing the behavior of communicating with a destination IP on a particular port.

On the first day of the evaluation data, the hypergraph metrics do not appear to reveal any red team activity, but they do provide insight into the structure of the hypergraphs. In the top row of Fig. 2 we see that although there are only between 700 and 800 nodes in the hypergraphs, the maximum edge size is typically around 400 nodes, meaning nearly half of the computers communicate with the same destination port on the same destination IP. The number of edges is around 115 with maximum degree around 50, so during each 10-minute interval there is typically at least one computer that communicates with close to half of the (destination IP, destination port) edges. There is a small spike in number of nodes around 9:30 which appears to be related to startup activity in the simulated benign traffic. On day 2 of the evaluation data the story is different. Here these simple measures of edge count and maximum degree are able to pick up on some malicious activity, namely the use of Deathstar to automatically enumerate the domain around 11:00. This is shown in the bottom row of Fig. 2. The spike in number of nodes and maximum edge size around 10:00 is unrelated and instead reflects startup activity similar to what was observed on day 1.

There are several other basic hypernetwork science measures that we could (and did) consider, including relative Hausdorff distance[84] of degree and edge size distributions, largest node and edge equivalence classes,

and number and size of *s*-connected components.[85] For the sake of space we do not include the plots here but note that in our analysis these measures reflected the same story as the ones pictured in Fig. 2, that nothing malicious was seen in day 1 and the Deathstar activity was captured in day 2. In the next section we define and show how some topological analysis can pick out both the loud Deathstar activity and some subtle malicious activity on days 1 and 2.

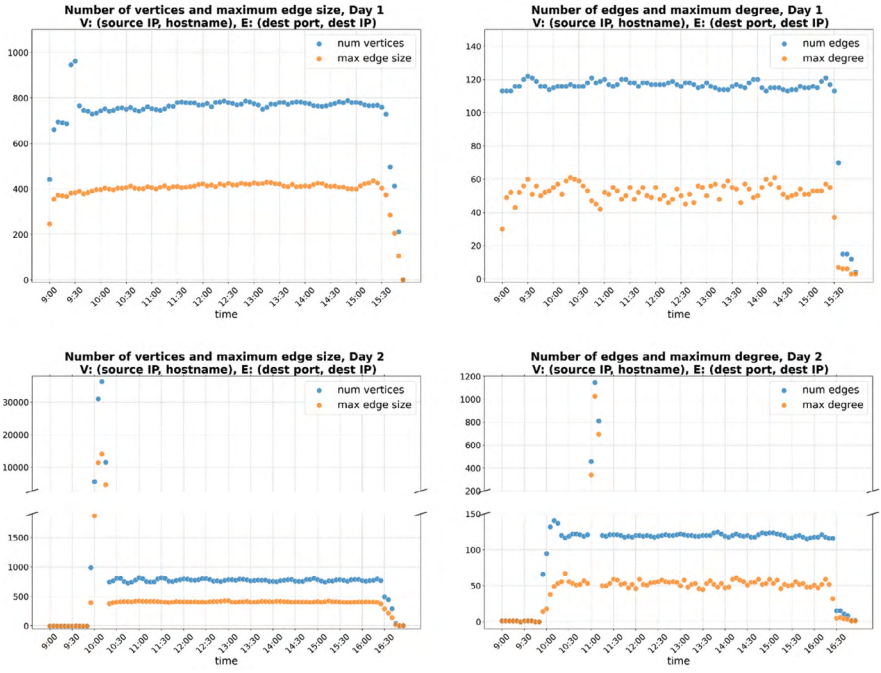


Figure 2. Number of vertices, edges, maximum edge size, and maximum degree of V= (source IP, host), E = (destination IP, destination port) hypergraphs on Day 1 (top) and Day 2 (bottom).

Topology Identifies Interesting Substructures

In this section we survey results from Jenne et al.,[86] in which we studied the *topological* structure of the time-windowed hypergraphs introduced in the prior section. We found certain substructures that represent

"topological holes" captured activity performed by the red team actors that, in some cases, hypernetwork science measures did not identify.

Intuition of Homology

Topology is the field of mathematics concerned with identifying features in a topological object¶ that are preserved by continuous deformation. The quintessential example is that a coffee cup and a donut are topologically equivalent because they each have one loop (the donut's hole and the coffee cup handle) and can be continuously deformed into each other without tearing, gluing, or poking holes. The tool in topology that we use to identify these topological features is called *homology*. Introducing the theory of homology is beyond the scope of this paper, so instead we will provide some basic intuition and point the reader to standard algebraic and computational topology texts for more details.[87, 88, 89] Given a topological object, its homology in dimension k is the number of k-dimensional "holes" or "features," that is, loops that are not boundaries of $k + 1$-dimensional structures. For example, the 1-dimensional circle (shown in Fig. (b) is a 1-dimensional homological feature. But in Fig. (c), this circle is the boundary of a 2-dimensional disk, so the disk has no 1-dimensional homological features. In general, 0-dimensional homological features are connected components (see Fig. (a), 1-dimensional features are loops, 2-dimensional features are trapped voids (like the inside of a hollow sphere, as in Fig. (d), and so on for higher dimensions.

This intuition and theory of homology applies to topological spaces, but a hypergraph is not a topological space; there is no notion of "continuous deformation." Therefore, we must transform a hypergraph into something topological to apply homology. A recent paper by Gasparovic et al.[90] surveys several methods to transform a hypergraph into a topological object to apply homology. Our work[91] uses the "nesting complex," another name for the *restricted barycentric subdivision (RBS)* that appears in several studies.[92, 93] To define the RBS, we first need the

¶ A topological space, simplicial complex, manifold, and related objects.

notion of a *simplicial complex.* Like a hypergraph, a simplicial complex has a set of vertices, *X,* and a nonempty set of subsets of *X,* Δ, called *simplices* instead of edges. There is an additional restriction to require that if $\sigma \in \Delta$ and $\tau \subseteq \sigma$ then $\tau \in \Delta$. In other words, Δ is closed under the subset relation. While we won't use it in this paper, it's worth pointing out that there is a straightforward way to construct a simplicial complex from a hypergraph. For each hyperedge $e \in E$ adding all subsets of e creates the *hypergraph closure*[94] Some of our other work[95] uses the closure construction; the decision of which homology theory to use is highly dependent on the choice of hypergraph construction. In this paper we will focus on the RBS.

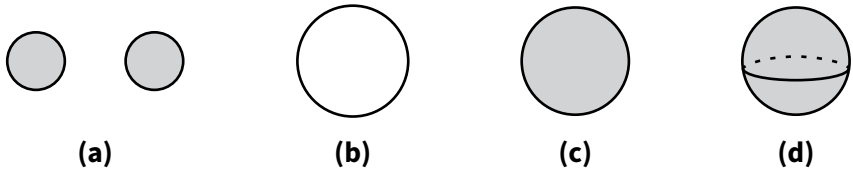


Figure 3. Four topological spaces illustrating homology in different dimensions: (a) has two 0-dimensional features, (b) has one 1-dimensional feature, (c) has no 1-dimensional features, and (d) has one 2-dimensional feature.

Given a hypergraph, H = (*V, E*), its RBS, *R*(H), is a simplicial complex formed by letting the vertex set be *E* and for any chain of included edges, $e_1 \subset e_2 \subset \cdots \subset e_k$, in H we add a simplex $\{e_1, \ldots, e_k\}$ to Δ. Notice that this satisfies the subset closure requirement to be a simplicial complex because for any chain of inclusions every subchain is also a chain of inclusions. To illustrate two interesting properties of the RBS we point to the examples in Fig. 4. First, two hyperedges, *e, f,* that intersect properly (that is, one is not contained within the other) are not directly connected in the RBS. If there is a hyperedge contained in the intersection, $g \subseteq e \cap f$, the intersecting edges will connect through that subset (that is, *g* connects to both *e* and *f* in the RBS). In the top hypergraph in Fig. 4, the green and red hyperedges do not have an edge in the intersection whereas in the middle hypergraph the brown hyperedge is contained in the intersection and connects the green and red. Second, a loop (or 1-dimensional homological feature) in the RBS corresponds to an alternating sequence of

inclusion ($a \subset b$) and containment ($b \supset c$) of hyperedges that returns to where it starts, $e_1 \subset e_2 \supset e_3 \subset \cdots \supset e_1$. We see this in both the middle-top and middle-bottom hypergraph in Fig. 4, but not the top left and top right. In the left hypergraph there is an alternating sequence, but it does not come back around to end where it started. There are other interesting properties of the RBS, like bounds on the number of 0-dimensional features and ways that higher dimensional features arise, but for those we point the reader to additional references.[96,97] All we will need in this paper is the observation of how 1-dimensional features form.

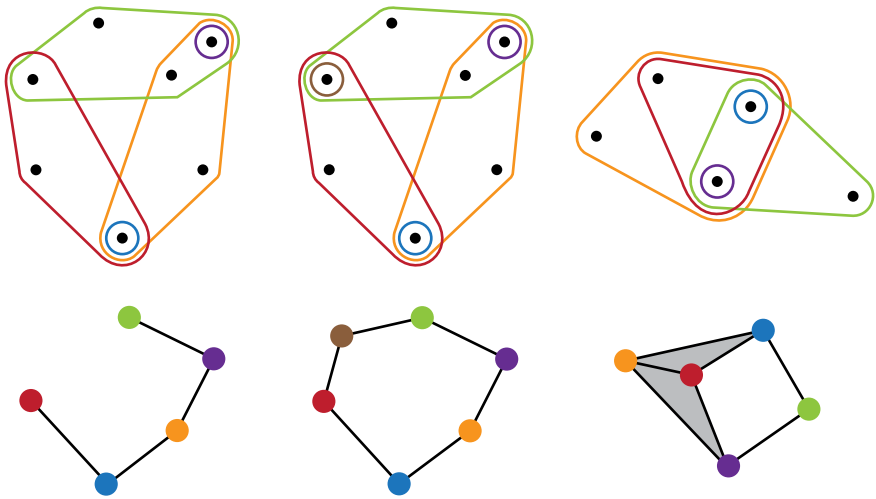
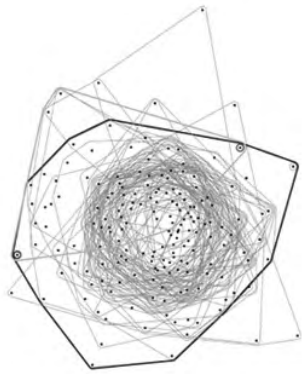
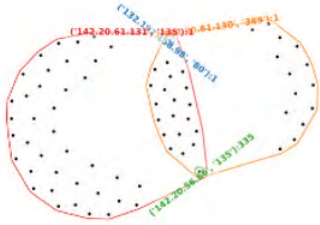


Figure 4. Illustrating the RBS (bottom row) for three example hypergraphs (top row).

A workflow of using RBS homology to identify interesting subsets of log data is pictured in Fig. 5. From a hypergraph, H, which was constructed from log data, we create its RBS, R(H). We then apply homology to identify k-dimensional features in R(H). These features will be subcomplexes of R(H) and so they identify a subset of edges $F \subseteq E$ (corresponding to the vertices of R(H) in the homological feature subcomplexes). We can then zoom into the sub-hypergraph of H restricted to just the edges in F and explore the data that gave rise to that structure. This is the process we will follow in the next subsection for the OpTC data.

PID	Src IP	Dst IP	Dst Port	Protocol	Image Path
4	142.20.56.198	142.20.59.255	138	UDP	System
864	10.20.4.125	224.0.0.252	5355	UDP	svchost.exe
864	142.20.59.255	224.0.0.252	5355	UDP	svchost.exe
636	142.20.56.198	222.206.244.5	443	TCP	firefox.exe
4	142.20.59.149	142.20.59.255	138	UDP	System
864	142.20.59.149	224.0.0.252	5355	UDP	svchost.exe





Client	Agent	IP	PID
Sysclient0501	4BW2MKUF	142.20.57.246	648
Sysclient0501	9HUGDCRL	142.20.57.246	5076
Sysclient0501	6H8SZPCW	142.20.57.246	1748
Sysclient0811	DS8V3RNH	142.20.59.44	3780

09/24/19 11:09:44—Deathstar via Sysclient0501 agent 4BW2MKUF, attempting to elevate using bypassuac_eventvwr

Figure 5. Illustrating the workflow from data to hypergraph to topological feature and back to data.

RBS Homology in the OpTC Data

Consider again the OpTC data and the 10-minute (source IP, hostname) versus (destination IP, destination port) hypergraphs introduced earlier. Recall from Fig. 2 that the number of edges for each time window hovers around 115. This has the potential to result in a very complex RBS. When we computed the homology of the RBS for each time window, however, we identified only a small number of 1-dimensional features, which implies that this structure is rare and thus perhaps interesting. In Fig. 6, we show the number of 1-dimensional features for each time window on both days. Around half of the time windows have no features at all, and most have 4 or fewer.

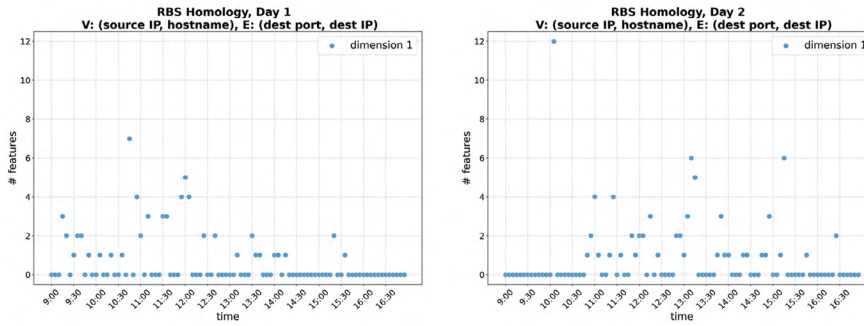


Figure 6. Number of 1-dimensional features in RBS of (source IP, hostname), (destination IP, destination port) hypergraphs on Days 1 and 2.

The interested reader may consult Jenne[98] for a more comprehensive exploration of the 1-dimensional features found in Day 2. Here we give two examples of sub-hypergraphs that correspond to the 1-dimensional features and explain how they tie back to the data. The first example, in Fig. 7 (left), is from Day 1 in the 11:30-11:40 time window. Here, the singleton red edge represents communication over port 80 between the computer of a compromised user and the command and control (C2) server. The other single vertex edge and the two large edges represent likely benign HTTP communication to the external internet. This connection back to the C2 server is seen in several other homological features in other time windows, but it is not identified by the hypernetwork science measures because of its small size.

Our second example, in Fig. 7 (right), is from Day 2 in the 11:00-11:10 time window. The structure is identical to the first example, with two large hyperedges that intersect and contain two singleton hyperedges, but the behavior they represent is slightly different. The blue singleton edge corresponds to communication with the same C2 server as before, from the computer of a different compromised user. The two large edges in this example represent communication with domain controllers (DCs). The biggest difference between this example and the one from Day 1 is that the second vertex, which is also tied to a compromised user, is contained in not just 1 but 335 singleton edges. These appear in the figure as a single green edge since they all contain only this one vertex. While all these edges use port 135, indicating endpoint resolution, they have different destination IP addresses in the internal network subnet. Through a bit of digging into the host logs for this vertex's IP and host we tied these edges to the Deathstar reconnaissance activity that started during this time window.

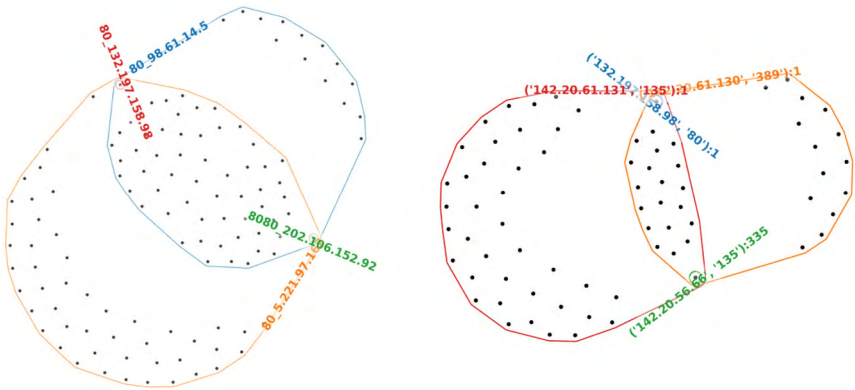


Figure 7. Examples of sub-hypergraphs tied to 1-dimensional features in the RBS; (left) Day 1, 11:30-11:40; and (right) Day 2, 11:00-11:10.

These two examples illustrate some of the commonalities among the sub-hypergraphs corresponding to the 1-dimensional features of the RBS for each time window. Most of these sub-hypergraphs have a similar structure of two or more large hyperedges that are pairwise linked by

singleton hyperedges. The middle hypergraph in Fig. 4 is an example of this type of structure with three large hyperedges, and the two in Fig. 7 have two large hyperedges. We found that the small edges of the sub-hypergraphs most often (though not always, as in the green edge of Fig. 7 (left)) represented communication with DCs over port 3268, RDP, or communication with the C2 server. These are infrequently used and sometimes malicious communications. For example, while RDP sessions are often benign, this is how the attackers moved laterally on Day 2. Most of the large edges, on the other hand, represented common communications such as talking on port 80 (web traffic over HTTP) or communication with DCs over ports 135 or 389 (related to lookup of information about IP addresses in a directory of sorts). The communications that are represented by the large edges are not malicious, but they tie the small edges together into the homological cycles.

It is important to point out that the identification of these homological features is not necessarily a malicious activity detector. Instead, what we are finding are rare *patterns* of behaviors (hyperedges intersecting in particular ways). Sometimes the homological features turned out to be built from only benign traffic. But in most cases at least one edge, often one of the small edges, could be tied to activity of the red team. In fact, of the 64 1-dimensional features in 31 different time windows on Day 2, 47 of them (73.4 percent) tied back to malicious activity. Although we did not find all C2 or lateral movement communications within the homological features, we identified enough to demonstrate the workflow's potential to aid a threat hunt analyst. After being notified of these repeated rare communications through topology, an analyst could then write a query to search for these patterns elsewhere.

We close this section by hypothesizing why this pattern frequently links to malicious activity in the data. The first observation is that nodes that sit within the intersection of large edges represent more active computers which could mean that they are more valuable or accessible to adversaries. The more communications sent out means possibly more opportunities to break in. Then, once a computer is compromised, the things that the adversary will do are likely to be rare actions. They ping back to the outside C2 server, attempt to exfiltrate data, or move laterally

to more valuable machines on the network once they have this initial foothold. If two computers that sit within the intersection of two large hyperedge both make rare communications (benign or malicious) this will form a dimension 1 homological feature.

Conclusion

The case study presented in this chapter represents only a portion of how hypergraphs can be used to model and analyze cyber data. We discussed in depth an example of how a subtle attack pattern showed up as a substructure identified by applying RBS homology. In some of our other work we have used different hypergraph constructions and analyses to find malicious activity and anomalies in OpTC and other data. To give a few examples, using vertices that represent users and edges that represent computers identifies lateral movement through the connected component structure, while weighted measures detect exfiltration, and many construction methods detect noisy ping sweeps and Deathstar activity. We have also shown that the evolution of hypergraph topology can be used to identify anomalous time windows.[99]

In summary, these findings provide strong evidence supporting the use of hypergraphs to model complex relationships in cyber systems, but there is no one-size-fits all hypergraph construction or analytic pipeline, and more steps are needed to develop this research into usable tools for analysts. Crucially, researchers must work with cyber analysts to go beyond quantitative metrics and analyze the types of attacks they captured and the types of attacks they missed. In cases where the data are imbalanced, with benign data making up most of the data, and the noisy, relatively easy to detect attacks making up most of the malicious data, accuracy is not enough. Working to provide a more holistic evaluation of novel methods will not only improve the tools derived from research but will also give analysts confidence in these tools. By embracing the unique capabilities of hypergraphs to model and analyze dynamic cyber log data, researchers and cyber analysts can work together to build trustworthy tools that enhance detection of complex cyber threats from evolving adversaries.

About the Authors

Emilie Purvine, Ph.D., is a senior data scientist with the Mathmematics, Statistics, and Data Science Group, Pacific Northwest National Laboratory (PNNL). She focuses on applications of combinatorics and computational topology, together with theoretical advances needed to support the applications. At PNNL, she has served as both a primary investigator and technical staff member on several projects in applications ranging from computational chemistry and biology to cyber security and power grid modeling.

Helen Jenne, Ph.D., is a data scientist with the Mathematics, Statistics, and Data Science Group, PNNL. She has a Ph.D. in mathematics from the University of Oregon where she specialized in algebraic combinatorics. At PNNL, she has worked on a variety of projects that applied discrete math to problems in cyber security and other domains. Her research interests include network science, topological data analysis, and explainable artificial intelligence.

Further Reading

Anjum, Md Monowar, Shahrear Iqbal, and Benoit Hamelin. "Analyzing the Usefulness of the DARPA OpTC Dataset in Cyber Threat Detection Research." *Proceedings of the 26th ACM Symposium on Access Control Models and Technologies* (2021): 27–32.

Gasparovic, Ellen et al. "A Survey of Simplicial, Relative, and Chain Complex Homology Theories for Hypergraphs." Cornell University Algebraic Topology. *arXiv* (2024), 2409.18310. https://arxiv.org/abs/2409.18310.

Jenne, Helen et al. "Stepping Out of Flatland: Discovering Behavior Patterns as Topological Structures in Cyber Hypergraphs." *The Next Wave* 25, no. 1 (2023).

CHAPTER 3

An Invitation to Hypergraph Motifs for Practitioners

Sinan G. Aksoy

Introduction

Hypergraphs serve as natural models for far-ranging data and systems, yet there is little consensus on how to effectively analyze hypergraphs in practice. Arguably, this is not a supply-side problem: given the recent proliferation of proposed hypergraph methods, practitioners no longer face a dearth of options. One who, for example, asks a basic question, such as "which nodes are the most important in my hypergraph," can now select between random walk approaches,[100, 101] nonlinear eigenvector centrality,[102, 103] degree centrality,[104] subgraph centrality,[105] and more. Rather, the challenge is that our understanding of hypergraph methods—and their efficacy in practice—is still nascent. In many cases, that understanding is insufficient to help practitioners confidently choose among sometimes subtly different tools, or even to convince them that hypergraph models are necessary at all.

Among the many attributes sought when choosing or designing practitioner-friendly hypergraph tools, three stand out: *applicability*, *interpretability*, and *necessity over graph-based approaches.* With regard to these key qualities, existing methods often fall short. For example, much of the hypergraph mathematics literature is limited to special classes of

hypergraphs (for example, uniform, degree-regular) and, hence, lacks applicability to real hypergraphs, which are frequently messy, nonuniform, and large. Furthermore, although numerous existing mathematical tools could be naively applied to hypergraph representations (for example, linear-algebraic analyses of hypergraph matrices and tensors, topological invariants), additional work is needed to characterize and explain, in concrete and accessible terms, the structural properties those tools capture for hypergraphs. In addition, many purported hypergraph methods do not take into account the structure that is present in hypergraphs but absent in graphs. These methods operate on hypergraphs yet (explicitly or implicitly) convert them to graph structures to apply graph-theoretic methods.[106] Unsurprisingly, these graph-reduction approaches lose information[107] and negate the purpose of using hypergraphs to capture higher-order structure.[108, 109] In this sense, provable use of hypergraph-*native* structure is an important and oft-ignored prerequisite.

In this introduction to the topic, we hope to demonstrate that hypergraph motif analysis promises to meet the three attributes noted above. Hypergraph motifs are small substructures within a larger hypergraph that are statistically, semantically, or otherwise significant. Across applications, these motifs themselves may indicate phenomena of interest. For example, the substructures identified in the earlier chapter by Purvine and Jenne serve as meaningful motifs for cyber activity. Historically, as we soon explain, interest in hypergraphs motifs arose naturally in the corporate governance literature to study *interlocking directorates* among company boards, characterize cross-company relations, and identify corporate elites. When company-board director data are structured as a hypergraph, interlocking directorates correspond to pairs of hyperedges with certain intersection patterns—motifs with an immediate domain interpretation. Furthermore, it is straightforward to show hypergraph motifs cannot be mined from graph-reductions—they capture higher-order structure and, thus, require the use of a hypergraph model.

Rather than provide a comprehensive survey of hypergraph motif methods (akin to Lee, for example),[110] this chapter will trace the evolution of hypergraph motif concepts—providing small examples and applications to grounding the discussion. Section 2 will outline the

natural emergence of hypergraph motif analysis in the context of interlocking directorates—showing how ideas initially developed within a narrow context spurred the development of a suite of "metamorphosis" hypergraph motif measures with applications across domains. Section 3 then traces this idea to today, explaining how recent work on "hyperedge triplets" can be framed as a generalization of the concept underlying metamorphosis measures. This section contrasts recent work in this area as taking a "classification" versus "optimization" approach and discusses application of these methods on several datasets. Section 4 acknowledges several ongoing technical challenges associated with hypergraph motif analysis.

Although prior knowledge of graph theory is not assumed, it is necessary to clarify some terminology before proceeding. A hypergraph $H = (V, E)$ is a set of vertices $V = \{v_1, \ldots, v_n\}$ and a collection of sets called hyperedges, $E = (e_1, \ldots, e_m)$ where $e_i \subseteq V$ for $i = 1, \ldots, m$. Vertices represent entities (such as people), and hyperedges represent multiway relations between those entities (such as membership in a group). Hypergraphs may be represented by an incidence matrix in which the i, j entry is 1 if $v_i \in e_j$, and 0 otherwise. Incidence matrices, in turn, may be represented by a *bipartite* graph,** as illustrated in Fig. 1. Consequently, this "hypergraph-bipartite graph" correspondence[111] means that bipartite methods often afford immediate interpretation as hypergraph methods and vice versa. This is germane because early work relevant to hypergraph motifs was proposed using the language of bipartite graphs. Consistent with their usage (and keeping in mind this equivalency), this chapter sometimes uses bipartite graph visualizations to describe hypergraph motifs considered in these works.

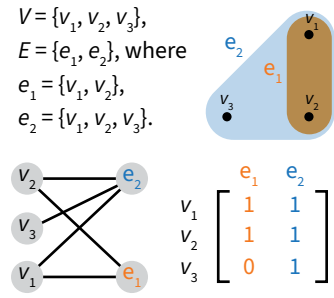


Figure 1. A hypergraph, its Euler diagram, bipartite graph, and incidence matrix.

** A bipartite graph is a graph whose vertices can be partitioned into two sets of vertices such that every edge connects a vertex from one set to a vertex from the other.

Hypergraph Motifs: Early Motivation from Corporate Governance

Two Scenarios

Researchers of corporate governance studied hypergraph motifs to represent social scenarios of interest. Consider the following two scenarios involving two companies, α and β, and two board members, Juan and Aisha.

1. Juan and Aisha both sit on Company α's board, and Aisha also sits on Company β's board.
2. Juan and Aisha sit on both Company α's and β's boards.

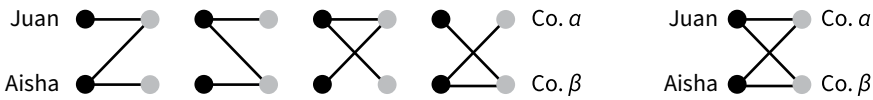


Figure 2. Caterpillar motifs correspond to variants of Scenario 1, while the butterfly motif corresponds to Scenario 2.

The first scenario is one of a single joint affiliation (membership on α's board) with the *possibility* of a second joint affiliation (as indicated by Aisha's membership on a second company board). The second scenario is one of *repeated* joint affiliation. Robins and Alexander argue these scenarios can be understood as atomic substructures for measuring the interconnectedness of a web of company boards.[112] By measuring how frequently the first scenario (opportunity for remeeting) occurs against the second (a remeeting), the authors aim to characterize "the extent to which directors re-meet one another on two or more boards." [113] In other words, they posit these scenarios can provide insight into the community structure of company-board networks.

Fig. 2 abstracts the two scenarios by using a bipartite graph to map memberships (edges) between people (black nodes) and companies (white nodes). The first corresponds to what Aksoy calls a "caterpillar" motif,[114] which is simply a bipartite 3-path. The second is called a "butterfly," which is a 4-cycle. Fig. 2 enumerates the ways both motifs can occur for a given pair of people and companies. Observe that there are

four ways the first scenario can occur: either Juan and Aisha jointly sit on α's or β's board, and for either case, Juan or Aisha sits on the other company board.

Using these two motifs, Robins and Alexander propose a global measure of directorate interconnectedness for any collection of companies (for example, those from a given industry or country). Their measure, which we call the (global) metamorphosis coefficient, is based on the relative counts of caterpillars to butterflies. More precisely, the global metamorphosis coefficient M of a =company-board network H is defined as:

$$M(H) = \frac{4 \cdot \text{caterpillar count in } H}{\text{butterfly count in } H}$$

The factor 4 appears in the numerator because there are four caterpillars in each butterfly, thus ensuring $M(H)$ is normalized between 0 and 1. Large values of $M(H)$ close to 1 mean that in that network, board members tend to remeet on other boards, whereas values closer to 0 suggest they do not, and in this sense are siloed.

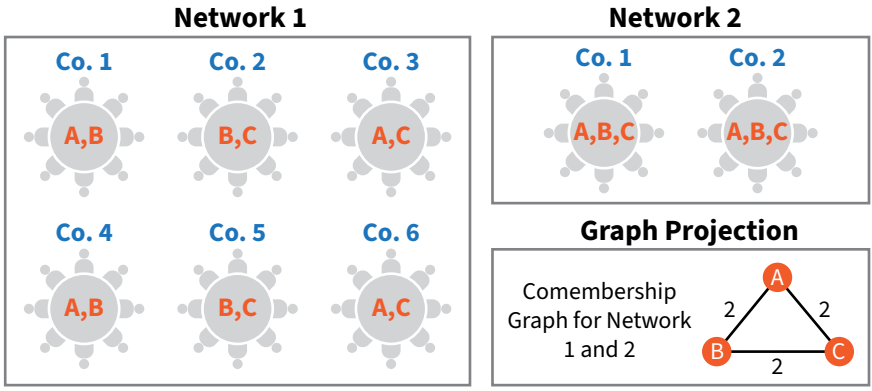


Figure 3. An example of two company-board networks with equivalent graph projections but differing global metamorphosis coefficient.

To garner further intuition for $M(H)$, it is helpful to consider how this quantity differs across two toy company-board networks. Fig. 3 illustrates an interesting example that shows metamorphosis can capture differences

otherwise lost by graph approaches.[115] Notice that in Network 1, every board member who sits jointly with another on a company's board remeets that board member on one out of three of the other company boards on which they sit. For example, A sits jointly with B on Company 1, and A also sits on Company 3, 4, and 6 but only remeets B on Company 4. This holds true for every board member; this network has $M = 1/3$. In contrast, for Network 2, board members sitting jointly on one company board remeet on all other boards; this network has $M = 1$. Despite these differences, the comembership graph (linking board members to one another if they sit jointly on the same board) of these networks is identical: in both networks, each pair of board members sits jointly on two boards together. This, as well as more comprehensive and sophisticated examples,[††] illustrates how hypergraph motifs can capture relationships otherwise lost in graph representations.

Other Applications and Local Versions

Caterpillar and butterfly motifs are germane to a variety of other applications. In bibliographic data, for example, metamorphosis can capture the extent to which coauthors repeat their collaborations; in passenger-flight and other travel data, the reoccurrence of cotravelers; and in disease-gene associations, how mutations in joint genes are implicated in the development of multiple diseases. In a recent application, butterfly motifs were used to study US Congressional and Senate voting data.[116]

Although widely applicable, the metamorphosis coefficient is also limited because it is a global summarization of motif structure. In the cited domains, where and how these motifs involve entities—not just their total counts—may provide critical information. One might ask, for example,

†† The reader may notice if one were to form a "board-overlap" graph linking *companies* with interlocking boards from Networks 1 and 2, the graph would be different. Kirkland (see endnote 32), however, shows the dual pair of hypergraph clique expansion (akin to the comembership graph) and line graph (akin to the board-overlap graph) is still insufficient to uniquely reconstruct a hypergraph.

what is the metamorphosis of a particular author, paper, traveler, flight—or even of a given authorship (that is, author-paper pair). Motivated by this, Aksoy, Kolda, and Pinar proposed local metamorphosis coefficients for bipartite nodes and edges (in hypergraph parlance: nodes, hyperedges, and vertex-hyperedge memberships).[117] By doing this, one can obtain raw and normalized rankings based on how frequently entities participate in motifs. These rankings may identify critical entities in data based on their structure alone. For example, consider a hypergraph formed from disease-gene associations listed in the public database DisGeNET. Fig. 4 plots a distribution of the raw count of butterfly motifs containing a given (disease-gene) association, in log scale. Here, the pair with the highest count is (breast cancer, TNF), where TNF is a critical gene implicated not only in the pathology of breast cancer and other cancers, but also liver fibrosis, Alzheimer's, and inflammatory diseases.

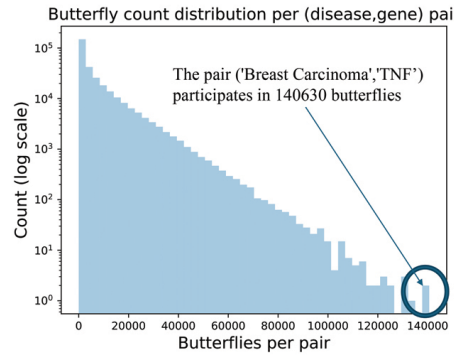


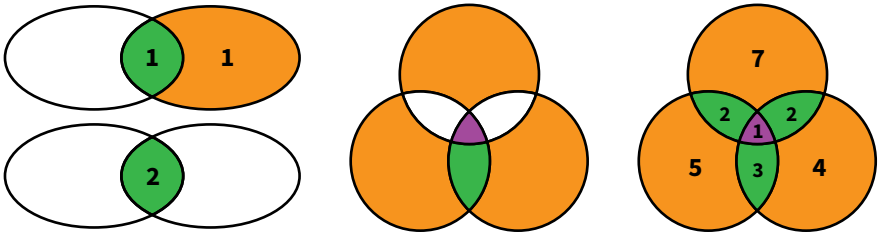
Figure 4. Butterfly count distribution.

Hyperedge Triplets: Classification versus Optimization

A richer and more general perspective on the structure captured by caterpillars and butterflies can be obtained by viewing them through the lens of hypergraphs, rather than the associated bipartite. When viewed as a hypergraph, it becomes clear that butterflies and caterpillars describe different possibilities for how two hyperedges can intersect. Fig. 5a illustrates the hypergraph form of a butterfly and caterpillar that consists of two hyperedges and two nodes, where the numbers denote the number of vertices belonging to that "region." Drawn in this way, we can see these motifs differ in two simple yet critical areas:

- Which "regions" do or do not contain vertices, and
- The sizes of the "regions."

More formally, for hyperedges *X* and *Y,* the three possible regions are: the intersection $X \cap Y$, and the differences between the hyperedges, $X \setminus Y$ and $Y \setminus X$. Generalizing this idea, richer connectivity patterns emerge when we consider three hyperedges. Figs. 5b and 5c show that, in this case, there are seven possible regions. These observations highlight an important difference between hypergraph and graph motifs: for graphs, since every edge has size exactly two, these regions (intersections and set differences between edges) always have size at most one. In this sense, there isn't a graph notion of "region size" beyond binary.



(a) A caterpillar (top) and butterfly (bottom) motif drawn as hypergraphs.

(b) The template for *h*-motif 20 from Lee, Ko, and Shin.

(c) A hyperedge triplet with example region sizes.

Figure 5. Hyperedge pair and triplet motifs. Colored regions indicate the existence of vertices while white regions indicate their absence.

From Hyperedge Pairs to Triplets

More recent work on hypergraph motifs has focused on hyperedge triplets. Triplet intersection structure is more expressive and nuanced than that in hyperedge pairs (as seen by the greater variety of region types in Fig. 5), and triplets can still be regarded as small substructures that are abundant in data and computationally tractable to those here. Two approaches to studying hyperedge triplets center on *classification* and *optimization.* In the classification approach, one aims to mine and categorize hyperedge triplets

according to their type, which is based on which regions contain vertices. Lee, Ko, and Shin take this approach, observing that there are (up to symmetry) 30 different hyperedge triplet types, which they call h-motifs.[118] The output of this approach is a motif census: a hyperedge triplet count for each of these 30 types. Fig. 5b gives an example of one such type, h-motif 20, where the absence of color means that no vertices belong to that region.

A drawback of this approach is that it applies a binary criterion to classify hyperedge motifs. This ignores the relative sizes of hyperedge regions, which can carry important information and vary widely within a particular motif type. For instance, a pair of company boards overlapping in 1 percent versus 99 percent of their board members (that is, nearly disjoint versus nearly identical) would still be categorized as instances of the same hyperedge motif type: one in which all regions have some number of vertices. To address this shortcoming, Niu, Amburg, Aksoy, and Sarıyüce consider hyperedge triplet motif mining instead as an *optimization* problem.[119] Here, the goal is to find and quantitatively score hyperedge triplets according to how they optimize user-chosen weight functions, based on the relative sizes of regions. To define this weight function precisely, for a hyperedge triplet $T = \{a, b, c\}$ and subset $X \subseteq T$, let $N(X)$ be the function that picks out vertices shared across every hyperedge in X that are not in other hyperedges in T.

$$N(X) = \left(\bigcap_{x \in X} x \right) \setminus \left(\bigcup_{x \in T \setminus X} x \right)$$

Applying this function to each possible subset of a hyperedge triplet yields all possible regions, which fall naturally into three categories: independent (R_1), disjoint (R_2), and common (R_3). Color-coded orange, green, and purple, respectively, in Fig. 5.

- $R_1 = \{N(a), N(b), N(c)\}$
- $R_2 = \{N(a, b), N(b, c), N(a, c)\}$
- $R_3 = \{N(a, b, c)\}$

The three region types are then used to formulate three notions of hyperedge triplet weight.[120] Namely, the independent, disjoint, and common weights, W_1, W_2, W_3, are defined by:

$$W_j(T) = \frac{\min_{X \in \binom{T}{j}} |N(X)|}{1 + \sum_{i>j} \sum_{X \in \binom{T}{i}} |N(X)|}$$

Stated equivalently, this weight function takes the minimum size of the target region (the numerator) relative to the sum of the sizes of "deeper," higher-intersection regions (the denominator). There are several rationales for this definition. For instance, the minimum is chosen here to avoid misrepresentation if one region dominates the others, reflecting the idea that a group is only as strong as its weakest link. Of course, there are many ways to define hyperedge triplet weight; (see Niu, page 4)[121] for a discussion of potential pros and cons of other closely related weight definitions. For the example, in Fig. 5c, we have the following:

$$W_1 = \frac{\min(4,5,7)}{1 + (2+2+3+1)} = \frac{4}{9}$$

$$W_2 = \frac{\min(2,2,3)}{1+1} = 1$$

$$W_3 = \frac{\min(1)}{1} = 1$$

Hyperedge Triplets in Data

Definitional subtleties aside, an attractive feature of the above weight function is that it affords a simple interpretation: a hyperedge triplet with high independent, disjoint, or common weight reflects that these hyperedges are: 1) least similar to each other, 2) have the highest pairwise but not groupwise correlation, or 3) are the most similar to each other. Applying maximum-triplet-finding algorithms from Niu,[122] and

available at: github.com/pnnl/hyperedge-triplets, we compute the following maximum triplets for three different datasets: *Amazon-Books, Amazon-Tools,*[123] and *MovieLens.*[124]

- In *Amazon-Books,* a hypergraph where vertices are user-ratings and hyperedges are books, a maximum weight disjoint triplet is {*Outlander, Dark Places,* and *The Fault in Our Stars*}. It is worth noting that *Outlander* predates the other two books. The fact that this triplet has a large two-way (but not three-way) intersection suggests *Outlander* may serve as a type of "steppingstone" in connecting a multigenerational audience of popular fiction.
- In *Amazon-Tools,* a hypergraph where vertices are user-ratings and hyperedges are tools, the maximum weight independent triplet is a flashlight from J5 Tactical, a screwdriver bit set from 1000 DeWalt, and a helping hand magnifier from Sona Enterprises. That these maximize the independent weight aligns not only the different purposes of these tools, but their different target audiences: the flashlight is military-grade, the screwdriver set is for general, commercial use, and the magnifying class is a specialized tool for hobbyists, soldering, and jewelry-makers.
- In *MovieLens,* a hypergraph where vertices are user-ratings and hyperedges are movies, maximum common weight triplets pick out movie franchises, such as *Lord of the Rings* or *Star Wars.*

Local Triplet Counting

Just as we saw that local versions of metamorphosis-coefficients were derived in Aksoy et al.'s "Measuring and Modeling,[125] in Retrieving Top-k Hyperedge Triplets, Niu et al.[126] also propose a local approach for finding hyperedge triplets containing user-specified hyperedge(s). Once provided with a hyperedge of interest, local approaches can aid in discovery and data exploration: finding, for example, that a high independent or common weight triplet containing a hyperedge X will yield other hyperedges both highly uncorrelated and correlated to X. Consider a recipe-rating dataset taken from 18 years of recipes on Food.com,[127] structured as a hypergraph

where user-ratings are vertices and recipes are hyperedges. Suppose the hyperedge "Really Good Vegetarian Meatloaf (Really!)" is chosen as a query hyperedge for a local triplet search. Here, a high independent-weight triplet incident to this hyperedge will yield recipes that are challenging for vegetarians. Indeed, the maximum weight independent triplet containing this hyperedge yields two meat options: "Tortured Chicken-Beer Can" (a recipe in which beer is grilled within a chicken) and "Zesty Low-Fat Chicken Breasts" (a healthy option). Although the insight that meat is repellent to a vegetarian is wholly unsurprising, it is encouraging that this hyperedge motif method, which is unaware of entity labels or meaning and is based solely on network structure, returns results with clear meaning and interpretation. In data with incomplete or unreliable labels, such unsupervised approaches may be particularly useful for discovering new relations to an entity of interest, or as features within machine learning frameworks.

Conclusion

Hypergraph motif methods have evolved with applications in mind. Although the recent explosion in hypergraph algorithms has prompted new motif method proposals, most remain nascent and have yet to be widely tested in practice. Motif concepts are flexible, naturally apply to far-ranging data, can provide simple and interpretable insights, and provably pick up on hypergraph-native structure inexpressible by graphs, however. For these reasons, hypergraph motif approaches are a prime candidate for practitioners interested in analyzing hypergraph-structured data and networks.

In this chapter, we have mostly focused on discussing different formulations for hypergraph motifs, grounding each hypergraph motif concept with examples and interpretation. Although contrasting these conceptual differences is helpful, we have not delved into two key technical challenges: algorithmic and inferential. First, designing compute and memory-efficient algorithms for enumerating and counting motifs is challenging for graphs—and even more so for hypergraphs. Here,

distributed and parallel algorithms, as well as random sampling methods[128, 129] that return approximations of motif information, are powerful approaches. Second, although scalability tends to be a primary concern, an oft-ignored yet natural question arises after motifs are computed: to what extent can motif occurrences be explained by randomness or as a result of lower-order properties of the data? One approach to rigorously answering questions about motif significance is to analyze a random hypergraph null model as a baseline. Here, too, a plethora of proposals exist, from variants of hypergraph Chung-Lu,[130, 131] to hypergraph BTER,[132] to hypergraph stochastic block models.[133] Luckily, experimental analysis of these models at large scales is often possible. By using efficient generation algorithms that exploit probabilistic techniques, akin to those surveyed in Arjun S. Ramani et al.'s "Coin-Flipping, Ball-Dropping, and Grass-Hopping for Generating Random Graphs from Matrices of Edge Probabilities,"[134] one may quickly generate and examine many instances of these models. Beyond experimentation, however, analytical results about the behavior of these models remain extremely limited. The complexity and generality of hypergraphs make this analysis difficult, although promising mathematics research is being done in this direction.[135, 136, 137]

Acknowledgments

Pacific Northwest National Laboratory Information Release PNNL-SA-206551.

About the Author

Sinan G. Aksoy, Ph.D., is an applied mathematician specializing in network science at Pacific Northwest National Laboratory (PNNL). Much of his work focuses on hypergraphs and spectral graph theory. Drawing from combinatorics, algebra, and probability, he develops methodologies for studying complex systems arising from varied domains, including social networks, power and communications systems, and high-performance computing.

Further Reading

Aksoy, Sinan G., Tamara G. Kolda, and Ali Pinar. "Measuring and Modeling Bipartite Graphs with Community Structure." *Journal of Complex Networks* 5, no. 4 (2017): 581–603.

Lee, Geon, Fanchen Bu, Tina Eliassi-Rad, and Kijung Shin. "A Survey on Hypergraph Mining: Patterns, Tools, and Generators." Cornell University Algebraic Topology. *arXiv* (2024): 2401.08878. https://arxiv.org/abs/2401.08878.

Niu, Jason, Ilya D. Amburg, Sinan G. Aksoy, and Ahmet Erdem Sarıyüce. "Retrieving Top-k Hyperedge Triplets: Models and Applications." In 2024 IEEE International Conference on Big Data (2024).

CHAPTER 4

Higher-Order Interactions at Scientific Conferences Influence Team Formation

Emma Zajdela and Nicholas W. Landry

Cooperation enables teams to solve complex problems that one individual alone cannot address. For this reason, collaborative teams have become the predominant way through which scientific progress is achieved. These collaborations arise through various mechanisms, including interactions at conferences. The Scialog conferences—a series of small, interdisciplinary scientific workshops conducted by the Research Corporation for Science Advancement over several years—provide an ideal laboratory for studying the network mechanisms leading to team formation. Building on existing work studying team formation from a pairwise perspective, this chapter presents a higher-order network perspective generalizing this framework. It formalizes the study of group interaction over time by defining a taxonomy of synchronous and asynchronous group interactions. This study applies this framework to the Scialog case study using a stepwise selection logistic model and finds evidence that all interaction types described in our taxonomy are highly significant for team formation. This higher-order network perspective provides a new framework for the study of collective behavior and group formation.

Background

The recognition that collaborative teams are effective in generating knowledge, innovation, and discoveries across many facets of society has prompted an increase in the study of teams, which has led to the development of the field of Science of Team Science.[138, 139, 140] Effective problem-solving requires individuals to collaborate across disciplinary boundaries.[141] It has been shown that when individuals assemble into teams and solve problems that one person alone cannot, they demonstrate collective intelligence.[142]

The Intelligence Community (IC) requires teamwork and collaboration across stakeholders with different backgrounds, skills, and experience, which necessitates effective coordination within and across teams. This collaboration can be challenged by the nature of the IC's work, including the need for secrecy and for cross-disciplinary teams able to address complex problems. Communication challenges often arise when individuals must coordinate across institutions that have different norms and cultures.[143, 144]

The scientific community shares several of these challenges, including communication among individuals from different backgrounds, disciplines, and institutions. Across nearly all scientific disciplines, research is increasingly performed in teams. In science and engineering, the portion of publications written in teams has increased from 50 percent in 1955 to more than 80 percent by 2000.[145] As more scientific knowledge is produced by teams, these papers are more often cited than solo-authored work.[146, 147]

Several studies have investigated how scientific teams are formed.[148, 149] A 2015 survey shows that when collaborators who not geographically collocated, one out of six met at conferences.[150] When considering the role of events, such as scientific conferences, research has demonstrated the role of prior knowledge[151] and interaction[152, 153, 154, 155] in team formation. These studies focus on predicting the pairwise effects among participants at the conference, despite scientific collaborations and the groups in which scientists interact at conferences often being larger than two.

Higher-order interactions, which represent interactions among groups of individuals of arbitrary size, more naturally model these systems and can offer insights inaccessible to pairwise representations. Higher-order networks, or *hypergraphs,* are the collection of these higher-order

interactions, and they naturally encode the scale of interaction.[156] Even when measuring properties of a collaboration network, such as its community structure, assortativity,‡‡ and the relative importance of its constituent nodes, modeling these networks in their native representations can change one's conclusions. For example, when examining scientific collaborations, higher-order network analysis indicates that although the size of collaborations researchers engage in is field dependent, the number of collaborations remains relatively consistent. [157] Likewise, hypergraph analysis preserves group-level interactions, allowing for analysis of higher-order collaboration motifs.[158] Last, higher-order networks provide an essential framework for studying the temporal evolution of group structure.[159] Pairwise network representations are unable to discern among groups of different size and, therefore, cannot be used to show how groups split, merge, and aggregate over time.

To study the impact of interaction on group outcomes, including team formation, one must first define interaction over time for groups of more than two individuals. Given known effects in social networks including triadic closure,§§ this chapter also seeks to identify the extent to which higher-order interaction adds significant information to group outcomes. Therefore, we can extend the pairwise methods used to analyze these systems by defining a taxonomy of synchronous and asynchronous group interactions. This taxonomy is used to identify types of subgroup interactions that may contribute to group outcomes. These notions are then applied to a case study comprising four scientific conferences from the "Scialog dataset," which contain detailed, longitudinal information about team formation at the Scialogs.[160] The findings indicate that all types of interactions described in our group interaction taxonomy affect team formation. This framework can be applied to other cases of interaction in social networks over time in contexts, such as teams in IC, business,

‡‡ The tendency of the nodes in a network to be connected to other nodes that are similar.

§§ **The property among three nodes A, B, and C** that if the connections A-B and A-C exist, there is a tendency for the new connection B-C to be formed.

and medical settings, as well as other group interactions, such as human migration or animal behavior.

Methods

Higher-order network analysis can be used to predict group collaboration as a function of interaction. Here, collaboration refers to self-assembled groupings, and interaction refers to time spent in groups of potentially different orders. In the case of the Scialog conferences, collaboration refers to post-conference, self-assembled teams that submit scientific proposals. Interaction refers to time participants spend in groups assigned by the conference organizers.

We start by defining the interaction hypergraph $H_i = (V_i, E_i)$ and the collaboration hypergraph $H_c = (V_c, E_c)$, where V is the set of vertices representing individuals and E is the set of hyperedges, or group interactions between individuals. We assume that $V_i = V_c = V$ and let $N = |V|$ denote the size of each hypergraph. For each group session $e \in E_i$, we track the start time $t_e^{(i)}$ and end time $t_e^{(f)}$. Consistent with previous work defining interaction in the pairwise case,[161] we make the following three assumptions: 1) interaction is symmetric, 2) interaction is proportional to the time participants spend listening to one another, and 3) participants speak an equal amount of time in each group session. These assumptions specify an interaction function $I_e = (t_e^{(f)} - t_e^{(i)})/|e|$ and imply that interaction in smaller groups is more intense than larger groups.

The goal is to predict collaboration among groups of three; that is, we wish to understand the interaction patterns present in H_i that give rise to a collaboration $e \in E_c$, where $|e| = 3$. To do this, we introduce a group interaction taxonomy that aims to uncover the role of pairwise and higher-order group formation mechanisms. This paradigm is described in more detail below.

Given a group $\{i, j, k\}$, we classify groups in H_i according to the following taxonomy:

1. **Synchronous interaction (Type I)** indicates the group $\{i, j, k\}$ has a synchronous interaction in breakout session e if $\{i, j, k\} \subseteq e$.

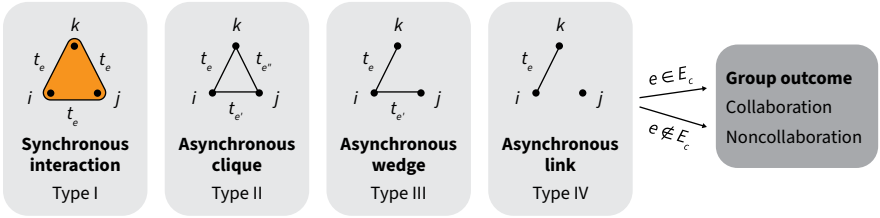


Figure 1. This figure shows the possible types of interaction that can occur among groups of three; Type I: the three participants have interacted in the same group at least once (synchronous); Type II: the three individuals have interacted with each other at different points but never all at the same time; Type III: two of three possible pairs involving the three individuals have interacted, but never at the same time; and Type IV: one of three possible pairs involving the three individuals have interacted.

2. **Asynchronous interaction** occurs in group e if $\{i, j\} \subseteq e$, $\{i, k\} \subseteq e$, or $\{j, k\} \subseteq e$ but $\{i, j, k\} \nsubseteq e$. Aggregated over all groups $e \in E_i$, asynchronous interactions can occur between three, two, or one of the pairs in the group, corresponding to types II, III, and IV respectively in Fig. 1.
   a. **Asynchronous clique (Type II)** measures the minimum amount of time that $\{i, j\}$, $\{i, k\}$, and $\{j, k\}$ interacted separately over the course of the conference.
   b. **Asynchronous wedge (Type III)** measures the minimum amount of time that two of the three possible pairs $\{i, j\}$, $\{i, k\}$, and $\{j, k\}$ interacted separately over the course of the conference.
   c. **Asynchronous link (Type IV)** measures the minimum amount of time that a single pair from $\{i, j\}$, $\{i, k\}$, and $\{j, k\}$ interacted separately over the course of the conference.

This taxonomy is operationalized using the following algorithm to compute the type I-IV interaction times corresponding to a collaboration (or noncollaboration) group g. First, the algorithm searches all possible sessions $e \in E_i$ to identify all subgroups of group $g\tilde{} \subseteq g$ in order of decreasing size. If $g\tilde{} \in e$, the algorithm increments the interaction time by I_e and keeps iterating over all remaining groups. The result of this first step is a map θ: $g \rightarrow$ R

between subsets $g^{\sim}$ and interaction times. Once this map is completed, the type I interaction time is simply $\theta(g)$. To compute the types II-IV interaction times, we:

1. Construct a vector $\mathbf{a} = [\theta(g^{\sim}) \mid g^{\sim} \in g \; s.t. \; |g^{\sim}| = 2]^T$,
2. Let $\delta = |\{a_i \in \mathbf{a} \mid a_i > 0|$,
3. Set $t^{\sim}(2, \delta) = \min \mathbf{a}$ where $t^{\sim}$: $(i, j) \rightarrow \mathrm{R}$,
4. Let $\mathbf{a} = \mathbf{a} - t^{\sim}(2, \delta)$ and $\mathbf{a} = [a_i \in \mathbf{a} \mid a_i > 0]^T$, and
5. Repeat steps 2-4 until $\mathbf{a} = \varnothing$.

Finally, we translate the map $t^{\sim}$ to the interaction types defined in Fig. 1. The type II interaction time is $t^{\sim}(2, 3)$, the type III interaction time is $t^{\sim}(2, 2)$, and the type IV interaction time is $t^{\sim}(2, 1)$.

The taxonomy described in Fig. 1 isolates the effects of distinct mechanisms thought to influence group formation. The Type I "synchronous" interactions illustrate the role of higher-order effects on group formation. This mechanism requires the individuals to interact in the same group and cannot occur when individuals meet individually over the course of a conference. Likewise, the Type II "asynchronous clique" interaction posits that all individuals must have met, just not in the same context. If Type III "asynchronous wedge" interactions are predictive of collaboration groups; this indicates that triadic closure is at play. Last, type IV "asynchronous link" interactions describe group formation through random aggregation, where individuals who have never formally interacted with a group of two may join it, nonetheless.

Case Study: Scialog Conferences

We illustrate our higher-order framework with a case study using Scialog conference datasets.[162] The data are from four interdisciplinary conference series in the Scialog program:

1. "Advanced Energy Storage" (AES), 2017-2019,

2. "Chemical Machinery of the Cell" (CMC), 2018-2021,
3. "Molecules Come to Life" (MCL), 2015-2017, and
4. "Time Domain Astrophysics" (TDA) 2015-2016 and 2018-2019.

These three-day conferences take place in person, with each including fewer than 70 participants—primarily early-career faculty from the United States and Canada and senior experts who serve as keynote speakers and facilitators. The conference is designed to maximize interaction among participants who are algorithmically assigned to medium group discussions and small group discussions.[163] At the end of the conference, participants self-assemble into teams of two to four and submit proposals of which one-third receive funding.

The conference interaction data presented here comprises two types: 1) records of attendees participating in medium group discussions (10 participants or more) and small group discussions (4 participants or fewer), and 2) records of the collaborations participants eventually formed. The interaction and collaboration data for each conference are visualized in Fig. 2, and a summary of selected network statistics for each dataset is presented in Table I.

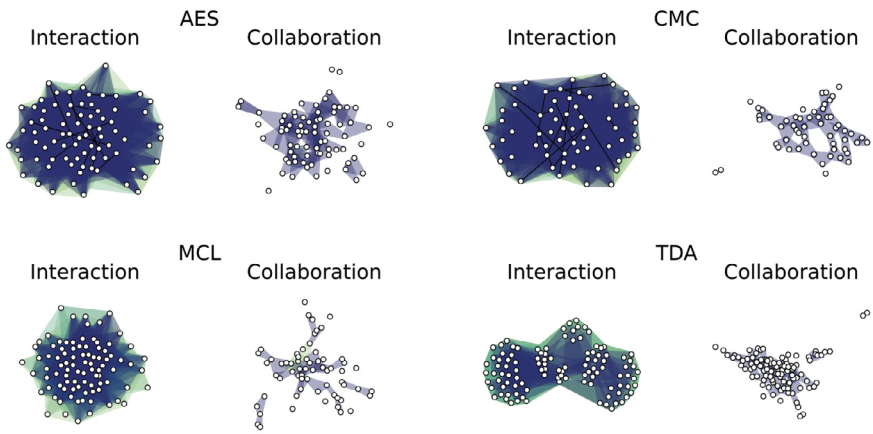


Figure 2. An illustration of the network structure of each conference series. Each network is aggregated over all conferences in that conference series. Visualized with XGI.[164]

Dataset	$\|V\|$	$\|E\|$	$\langle k \rangle$	$\langle k_{pairwise} \rangle$	$\langle s \rangle$	Unique Interaction Sizes
AES 2017						
interaction	60	104	8.0	44.0	4.62	3, 10
collaboration	56	35	1.59	2.64	2.54	2, 3, 4
CMC 2018						
interaction	50	88	8.0	43.84	4.55	2, 3, 10
collaboration	45	24	1.42	2.49	2.67	2, 3
MCL 2015						
interaction	64	54	5.33	40.12	6.31	4, 12, 13
collaboration	40	20	1.3	2.3	2.6	2, 3, 4
TDA 2015						
interaction	49	72	8.0	46.53	5.44	3, 4, 9, 10
collaboration	45	29	1.53	2.27	2.38	2, 3

Table I. From left to right, the columns indicate the number of nodes, the number of interactions, the average higher-order degree, the average pairwise degree, the average interaction size, and the unique interaction sizes at that conference and for that interaction type (interaction or collaboration). "Interaction" is a subset of each conference dataset comprising pre-collaboration interactions through the large and small discussion groups to which participants are assigned. Likewise, "collaboration" denotes the subset of the conference dataset comprising the collaborations the participants formed.

The higher-order degree provides little to no information on the number of groups participants join. In addition, while interaction groups can be quite large because of the number of participants in the medium group, the resulting collaboration groups are mostly of sizes two and three. These group sizes result from the assignments to discussion groups and the rules for the proposal submissions, which typically require participants to be in teams of two to three members.

To uncover the interaction mechanisms of collaboration beyond pairs of participants, this study focuses on collaboration groups involving three participants. In addition, only the first occurrence of a multiyear

conference series is considered to minimize the effect of prior knowledge in group formation. The data were aggregated across all four conferences, which corresponds to 52 groups of three collaborating out of 113,908 total possible groups. Examining the impact of the four types of interaction defined above on team formation in groups of three shows that collaborators interacted more than noncollaborators for all four types of interaction, as shown in Fig. 3.

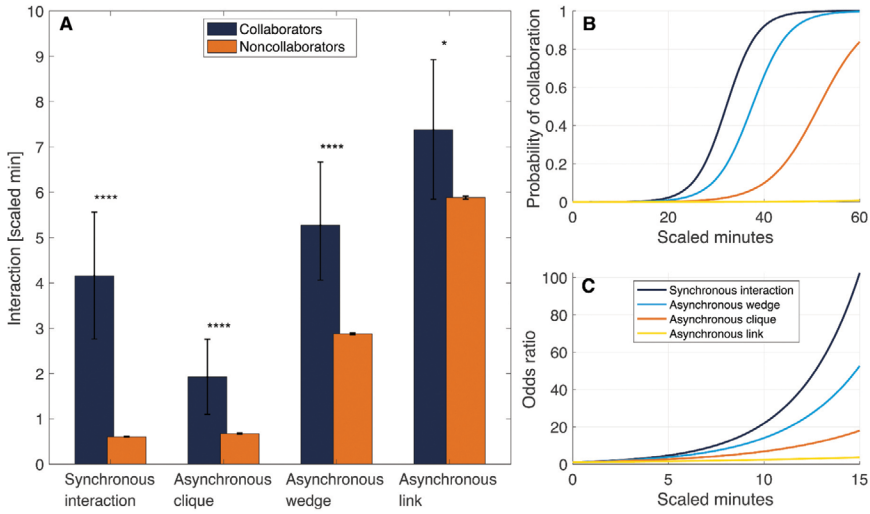


Figure 3. Interaction and collaboration: Panel A shows mean total scaled interaction time [scaled minutes] for people collaborating (left bars) or not (right bars) in a group of three. Data are aggregated across all four conferences and are shown for the four types of interaction defined in the main text. Error bars show mean values of the bootstrapped data with 95 percent CI. P-values of the Mann–Whitney U test include synchronous interaction, 1.04×10^{-21}; asynchronous clique, 3.3×10^{-2}; asynchronous wedge, 6.7×10^{-6}; asynchronous link, 3.1×10^{-5}. Panel B and C visualize the logistic model results. Panel B shows the change in probability of collaboration as scaled minutes of each type of interaction increases. Panel C shows the change in odds ratio as the scaled minutes of each type of interaction increases.

To further test the relationship between group interaction and collaboration, the interaction data (in scaled minutes) was fit to logistic models with linear combinations of the four types of interaction: synchronous

interaction *(sync1)*, asynchronous clique *(async3)*, asynchronous wedge *(async2)*, and asynchronous link *(async1)*. Stepwise selection was used to identify the best model fit by minimizing the Akaike Information Criterion (AIC)[165]—an estimator of prediction error—to identify which of the interaction types should be retained to best predict the group outcome *(collaborated: 0/1)*. After evaluating the possible linear subcombinations of the variables, the stepwise selection retained all four, indicating that each significantly improved the AIC relative to models with fewer parameters. The final model took the form:

$$\begin{aligned}\text{logit}\,(\text{collaborated} = 1) = \beta_0 \\ + \beta_1 \text{sync}_1 + \beta_2 \text{async}_3 \\ + \beta_3 \text{async}_2 + \beta_4 \text{async}_1\end{aligned}$$

where the β_i coefficients are the model parameters. Model results are displayed in Table 2. All four covariates exhibited strongly positive coefficients and highly significant p-values (10^{-4} or smaller), reflecting substantial increases in log odds (and hence probability) of collaboration. The intercept of -9.9 implies a low baseline probability when all interaction types are zero, entailing that collaborations are unlikely to occur without one or more types of interaction. Compared to the intercept-only model, a likelihood ratio test produced an χ^2 statistic of 127 ($p = 2.12 \times 10^{-26}$), further confirming that including these interaction types is statistically justified. The odds ratios are calculated for a minute of scaled interaction, where 10 minutes corresponds to approximately 30 minutes of interaction in a group of three or 1.5 hours in a group of nine. The increase in odds as interaction increases is shown in Panel C of Fig. 3. For example, in the case of synchronous interaction, an increase in 10 minutes of scaled interaction increases the odds of collaboration by a factor of 22, whereas in the case of an asynchronous link, 10 minutes increased the odds by a factor of 2.4. The difference between the effects of the types of interaction is particularly visible in Panel B of Fig. 3, which shows how the probability of collaborating changes with increasing time.

Term	Estimate (β)	SE	t-Stat	p-Value	Odds Ratio
Intercept	-9.9	0.36	-28	5.4×10^{-167}	-
Sync_1	0.31	0.025	12	2.3×10^{-34}	1.4
Async_3	0.26	0.041	6.4	1.3×10^{-10}	1.3
Async_2	0.19	0.029	6.6	4.2×10^{-11}	1.2
Async_1	0.086	0.023	3.7	2.5×10^{-4}	1.1

Table 2. Final stepwise logistic regression model for interaction types and collaboration. The values presented are computed for scaled minutes (that is, the odds ratio corresponds to the increase in odds for the increase in one scaled minute of interaction).

Discussion and Conclusion

In this chapter, we presented a framework for predicting team formation using higher-order network analysis. This higher-order perspective will be useful for higher-order link prediction, using historical interaction and collaboration data to predict future collaborations. Furthermore, the method for decomposing groups into different mechanisms provides a valuable contribution to understanding how network mechanisms, such as triadic closure[166] and higher-order dynamics,[167] predict future group interactions. Last, one can apply this higher-order analysis to much larger scales, such as bibliometric data[168, 169] or research grant personnel.[170, 171]

The evidence presented in this chapter indicates that the groups of three that collaborated interacted more than those who did not collaborate. The results suggest that both synchronous and asynchronous interaction, including in subgroups, play a role in team formation. Although these results do not provide causal evidence, previous work studying this data from a pairwise perspective used quasi-random counterfactual conference schedule data to provide strong evidence for a causal link between conference interaction and team formation,[172, 173] and a similar analysis could be conducted for this higher-order network analysis. Furthermore, although the model presented here is a simple linear combination of the scaled interaction decomposed into four

types, earlier research[174] has shown that a nonlinear model that incorporates the effects of memory and captures the time dimension of interaction outperforms linear models. Future research could extend this nonlinear model to the higher-order interaction case for better predictive power.

The model results indicate that the probability of collaborating increases rapidly as a group interacts synchronously, particularly when the level of interaction is sustained over time and when the group is small. Given the large effect of synchronous interaction on collaboration, this framework has implications for the design of conferences or other types of events aimed at fostering collaborative teams. For example, conference organizers may want to consider not only the properties of pairs of participants assigned to interact in groups, but also the properties of the larger-sized subgroups. A question that remains is whether there is a saturation point at which the participants do not benefit from additional interaction time. Furthermore, in the digital age, interaction and collaboration in science have increasingly been conducted virtually, with tradeoffs for creativity and innovation,[175, 176] team formation and community building,[177] environmental,[178] and equity considerations.[179, 180, 181] This chapter focuses on in-person interaction, but a similar framework could be applied to virtual interaction, decomposed into its synchronous and asynchronous components.

Finally, we only considered groups of three at scientific conferences, but this formalization of the definition of interaction over time could be extended to larger groups in a variety of contexts. The framework could be applied to study the effect of interaction on other types of group outcomes, including the spread of innovations, behavior adoption, and opinion dynamics. It could also be applied to collective behavior observed in animal groups, for example, in contexts where the decision to join or not join a group, mimic others, or learn a behavior may depend on time-varying interaction in higher-order groups. Thus, this approach demonstrates how higher-order network analysis can provide new insights into the dynamics of collaboration and team formation, with potential applications for collective behavior and collective intelligence.

Data Availability

A repository for this project, containing code and data to reproduce all results, is available on GitHub at: https://github.com/nwlandry/higher-order-team-formation, and in Zajdela and Landry, 2025.[182]

Acknowledgments

The authors thank the Research Corporation for Science Advancement for providing the Scialog dataset, Sandeep Chowdhary for initial conversations, and Simon Levin for discussions about defining group interaction. Dr. Zajdela acknowledges Elena Graetz and Maher Said for their input on the statistical methods used. Dr. Landry acknowledges support from the University of Virginia Prominence-to-Preeminence (P2PE) Science, Technology, Engineering, and Math (STEM) Targeted Initiatives Fund, SIF176A Contagion Science. This research was supported by an appointment to the IC Postdoctoral Research Fellowship Program at Princeton University administered by Oak Ridge Institute for Science and Education (ORISE) through an interagency agreement between the US Department of Energy and the Office of the Director of National Intelligence. The work in this chapter was supported by the Siegel Research Fellowship and a gift from William H. Miller III.

About the Authors

Emma Zajdela, Ph.D., is an applied mathematician at the High Meadows Environmental Institute, Princeton University and the Santa Fe Institute, Her research focuses on developing mathematical models to understand and predict complex social phenomena using a variety of approaches including applied dynamical systems, game theory, and network science.

Nicholas W. Landry, Ph.D., is with the University of Virginia Department of Biology and School of Data Science where he leads an interdisciplinary lab focused on understanding how networks and higher-order interactions influence the spread of diseases, information, and opinions.

His research integrates network science, mathematical modeling, Bayesian inference, and open source software development to explore the complex dynamics of contagion and social influence. He is also with the University of Vermont Complex Systems Institute.

Further Reading

Landry, Nicholas W., Ilya Amburg, Mirah Shi, and Sinan G. Aksoy. "Filtering Higher-Order Datasets." *Journal of Physics: Comlexity* 5, no. 1 (2024): 015006.52.

Patania, Alice, Giovanni Petri, and Francesco Vaccarino. "The Shape of Collaborations." *EPJ Data Science* 6, no. 18 (August 24, 2017). https://doi.org/10.1140/epjds/s13688-017-0114-8.

Zajdela, Emma R., Kimberly Huynh, Andrew L. Feig, Richard J. Wiener, and Daniel M. Abrams. "Face-to-Face or Face-to-Screen: A Quantitative Comparison of Conferences Modalities." *PNAS Nexus* 4, no. 1 (2025): 522.

CHAPTER 5

Evaluating Node Influence Using Higher-Order Centrality

Matthew J. Hasenjager, Mark M. Bailey, and Nina H. Fefferman

Introduction

When dealing with network-based representations of complex systems—whether social, biological, technological, or physical—a question that naturally emerges is which nodes are especially influential or critical to system functioning. Using variation in network connectivity to identify such nodes is relevant to diverse applications, including targeted seeding of information, attitudes, or behaviors to better enable their subsequent spread through the network,[183] allocation of limited protective resources to guard against network disruption by malicious actors or unforeseen events,[184] or identifying key individuals to target for prophylactic measures to limit disease transmission.[185] As highlighted in Chapter 1,[186] network analysts have consequently devised a broad range of measures that describe variation across nodes in terms of the number or strength of connections (degree, strength), nodes' position along key pathways in the network (betweenness, closeness), or the extent to which nodes are connected to other well-connected nodes (eigenvector centrality, PageRank).[187]

Moving from a dyadic to a higher-order network context, the notion of centrality can be applied not only to individual nodes, but to groups of nodes represented as hyperedges. As with nodes, hyperedges can vary in the number of their connections (that is, adjacent hyperedges with which they share nodes) or in their position along pathways traversing the network across hyperedge intersections.[188] In addition, the ability to identify key hyperedges or nodes based on their higher-order connectivity offers new opportunities for investigating, analyzing, and predicting system functioning that have no direct correspondence in dyadic networks. For example, whereas connections between adjacent nodes in a dyadic network are limited to a single edge, nodes in a hypernetwork can be adjacent in multiple hyperedges. Incorporating higher-order centrality into analyses alongside more traditional dyadic approaches can potentially reveal the presence of key entities or groups that would otherwise go unrecognized, thereby offering a more comprehensive picture of the key elements of connectivity within a given system. As an example, higher-order networks offer a useful framework for examining information flow over social groups linked by modern communication technologies, such as WhatsApp or Facebook Messenger, that often involve both one-to-many group interactions and dyadic messaging interactions.

Here, we focus on the use of s-line graphs as a potentially widely applicable means of capturing higher-order connectivity in a way that enables the use of familiar network centrality metrics, such as degree centrality (the number of connections possessed by a node), betweenness centrality (the number of shortest pathways between pairs of nodes that pass through a focal node), or closeness centrality (the reciprocal of the sum of the lengths of shortest paths from a focal node to all other nodes).[189] First, we introduce s-line graphs as a representation of higher-order connectivity and show how they can be used to measure a number of centrality measures. We then use simulations of social contagions on real-world social contact data to compare these measures with their dyadic counterparts in terms of their ability to identify nodes that are especially influential in spreading novel information, behaviors, rumors, and so forth throughout a population.

Using S-Line Graphs To Measure Hypernetwork Centrality

We can define a hypergraph as a set of n nodes, $N = \{n_1, \ldots, n_n\}$, and an indexed family of m sets referred to as hyperedges, $E = \{e_1, \ldots, e_m\}$, in which $e_i \subseteq N$ for i = 1, …, m. A hypergraph's corresponding line graph represents its hyperedges as nodes and intersections between hyperedges as edges. Put simply, nodes in a line graph are connected if their corresponding hyperedges intersect at one or more nodes in the hypergraph. Formally, we can define the line graph of hypergraph H, denoted L(H), as a graph on the node set $\{e_1^*, \ldots, e_m^*\}$ and edge set $= \left\{\{e_i^*, e_j^*\} : e_i \cap e_j \neq \varnothing,\ i \neq j\right\}$[190] Line graphs have been widely employed to render hypergraphs amenable to traditional dyadic analyses[191] but this transformation can involve substantial loss of information on higher-order connectivity. For example, in Fig. 1, both hypergraphs share an identical line graph representation (denoted L_1), despite being distinct in their structure.

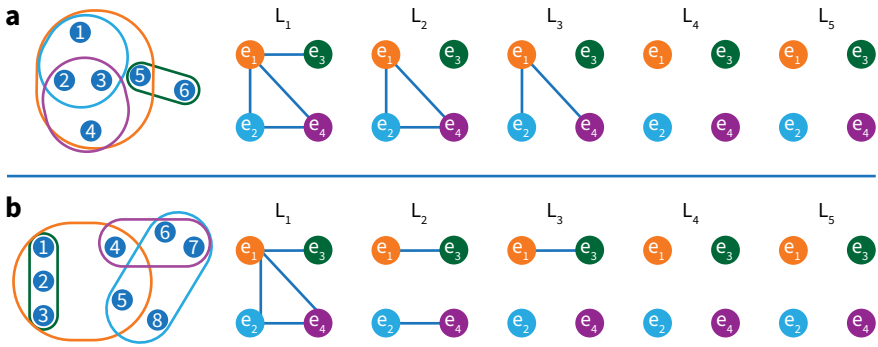


Figure 1. Two example hypergraphs are provided, with their corresponding s-line graphs. Each hyperedge, denoted by colored ovals, is represented by the corresponding-colored node in the line graphs (denoted L_1–L_5). Nodes in each line graph are connected if their corresponding hyperedges intersect at or more nodes, where the value of is indicated by the numerical subscripts. Note that the basic line graph is identical to L_1.

An improvement on the basic line graph approach is the s-line graph, which imposes a minimum intersection size, s, such that two hyperedges must intersect at s or more nodes to be connected in the s-line graph.[192]

In this way, a series of s-line graphs can be constructed from a hypergraph for a range of s values (Fig. 1). Formally, the s-line graph, $L_s(H)$, of a hypergraph, H, can be defined as the graph on the node set $\{e_1^*, \ldots, e_m^*\}$ and edge set $\left\{\{e^*, e_j^*\} : |e_i \cap e_j| \geq s,\ i \neq j\right\}$. Although this transformation still involves some information loss and so is not suitable for all applications,[193] s-line graphs often retain sufficient information on higher-order connectedness to be useful in a number of contexts, for example, see Pike[194] and Lin[195] in in chaps. 6 and 8, respectively.

From a series of s-line graphs for a hypergraph, we can measure several centrality measures, including degree, betweenness, closeness, and harmonic centrality.[196] For example, a hyperedge's s-degree is defined as the number of nodes it is adjacent to in the corresponding s-line graph. For hyperedge e_1 in Fig. 1a, its 1-degree is 3 and its 2-degree is 2, measured on L_1 and L_2, respectively. Likewise, we can define the s-betweenness of a hyperedge i as:

$$\sum_{j \in E_s \neq i \in E_s \neq k \in E_s} \frac{\sigma_{jk}(1)}{\sigma_{jk}}$$

Where E_s is the set of hyperedges of at least size s, σ_{jk} is the number of shortest s-paths from hyperedge j to hyperedge k, and $\sigma_{jk}(i)$ is the number of shortest s-paths from hyperedge j to hyperedge k that pass through i. An s-path is a sequence of adjacent hyperedges that intersect at s or more nodes and in which no hyperedge is revisited. Thus, s-paths can vary in both length (the number of hyperedges visited) and width (the minimum required intersection size), whereas paths in dyadic networks vary only in length.

The above measures produce an s-centrality value for each node (that is, hyperedge) for each s-line graph. To combine these measures, we can take the average of a node's centrality measures as:

$$\frac{1}{|f|} \sum_{s=1}^{s_{\max}} c_{s,f}$$

Where $c_{s,f}$ is hyperedge f's centrality measured in $L_s(H)$ and |f| is the size of hyperedge f.

Case Study: Applying S-Centrality to Social Network Data

S-centrality enables the identification of key nodes or groups of nodes in hypernetwork data, but does this information meaningfully add to our understanding of the structure, functioning, or dynamics of a system beyond what we can gain from more conventional dyadic network approaches? To address this question, we compare -centrality and dyadic centrality metrics as applied to real-world social network data collected as part of the SocioPatterns project.[197] We first ask whether s-centrality identifies potentially important nodes that would go overlooked by relying on dyadic centrality measures alone. Next, we compare the influence of nodes with high -centrality or dyadic centrality in terms of how effectively they promote the rapid spread of knowledge, behavior, or opinions during simulated social contagion experiments.

We use data gathered by the SocioPatterns project, consisting of face-to-face human contact patterns recorded via a distributed network of wearable sensors.[198] In these studies, participants wore radio frequency identification (RFID) tags around their necks that detected and recorded when a close (~1.5 m), face-to-face contact occurred with another RFID-equipped participant. This system was used to gather temporally explicit data on human contact patterns in a scalable, noninvasive manner across a number of contexts—including an academic conference, office workplace, hospital, and high school (Table 1). The data we use here is freely available at: http://www.sociopatterns.org/datasets.html.

Previously, these data have been primarily analyzed from a dyadic network perspective, but their format enables the identification of higher-order interactions, as each dyadic face-to-face contact event is time-stamped with start and end times (20 sec resolution). From these records, we identified higher-order interactions as collections of three or more individuals in which each unique pair was recorded as being in face-to-face contact during the same window of time. We then constructed hypernetworks that included all interactions of any order, including those that were strictly dyadic, across the duration of each study. We also recorded the start and end times for interactions, where an interaction

was defined as a unique set of individuals whose membership remained stable across a window of time. We provide summary data on the resulting hypernetworks in Table 1.

context (dataset name)	nodes	hyperedges	max. edge size	mean (SD) edge size	max. edge intersection size
conference (SFHH)	403	11777	11	2.68 (0.96)	10
office (InVS15)	217	7203	5	2.75 (0.76)	4
hospital (LH10)	75	3187	8	3.06 (0.95)	6
high school (Thiers13)	327	13730	5	3.14 (0.87)	4

Table 1. Summary data on focal hypernetworks. Nodes and hyperedges respectively refer to individuals and interactions.

Similarity of Node Sets Identified by Dyadic versus S-Centrality

For each hypernetwork (Table 1), we measured the mean s-degree and mean s-betweenness for each individual in two complementary ways. As s-centralities are measured on s-line graphs, these measures represent the centrality of hyperedges in the corresponding hypernetwork. To obtain s-centralities for individuals when hyperedges represented interactions (as in Table 1), we first measured the mean s-centrality for each hyperedge on the s-line graphs. Then, for each individual, we took the mean of these values across each hyperedge that that individual participated in. Conversely, we could take the dual of the hypernetworks in Table 1, swapping the relationship between hyperedge and node such that each individual is represented by a hyperedge encompassing a node set corresponding to each of the interactions in which it participated (see Chapter 1).[199] Using this dual representation, we then measured mean s-centrality for each individual-as-hyperedge. We distinguish these two measures by denoting the former by (G)—that is, mean s-degree (G)—as in this case, nodes in the s-line graphs represent groups of individuals.

For each hypernetwork, we also extracted the corresponding dyadic network and recorded degree, strength (or weighted degree), unweighted

betweenness, and weighted betweenness for each individual. Edge weights correspond to the cumulative duration that two individuals were recorded in face-to-face interactions.

To evaluate the extent to which s-centrality and dyadic centrality metrics identify unique sets of individuals, we first identified the top 10 percent to 25 percent of nodes as ranked by each metric and then compared the similarity of these node sets across centrality metrics using Jaccard similarity. Jaccard similarity evaluates the extent of overlap between two finite sets, A and B, as their intersection divided by their union, $\frac{|A \cap B|}{|A \cup B|}$. This value ranges from 0 (no overlap) to 1 (identical sets). We separately compared centrality measures based on either direct connection (that is, degree, strength, mean s-degree, mean s-degree (G)), or betweenness (unweighted betweenness, weighted betweenness, mean s-betweenness, mean s-betweenness (G)).

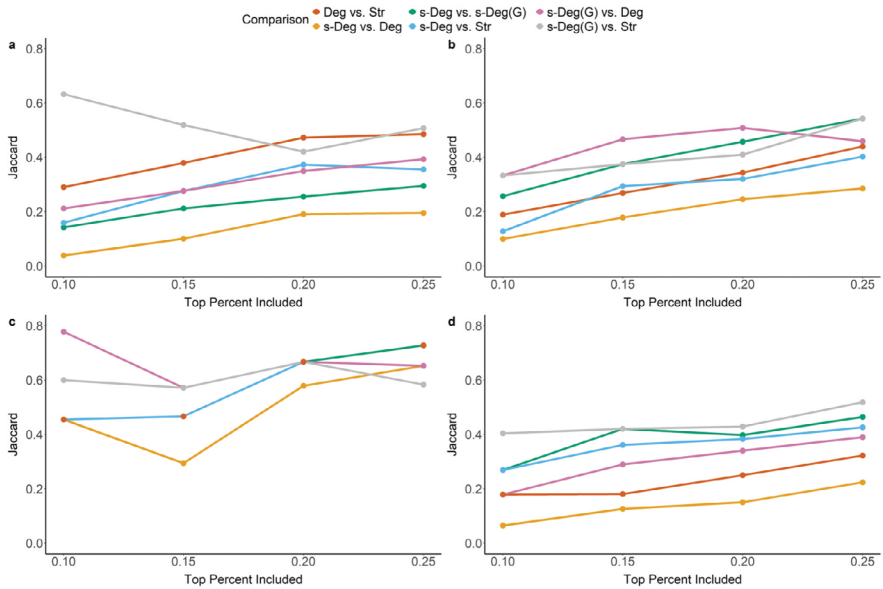


Figure 2. Comparison of Jaccard similarity of the top-ranked sets of nodes identified by different pairs of centrality metrics based on direct network adjacencies (degree = Deg, strength = Str, mean s-degree = s-Deg, mean s-degree (G) = s-Deg (G)). The contexts represented are a conference (a), office workplace (b), hospital (c), and high school (d).

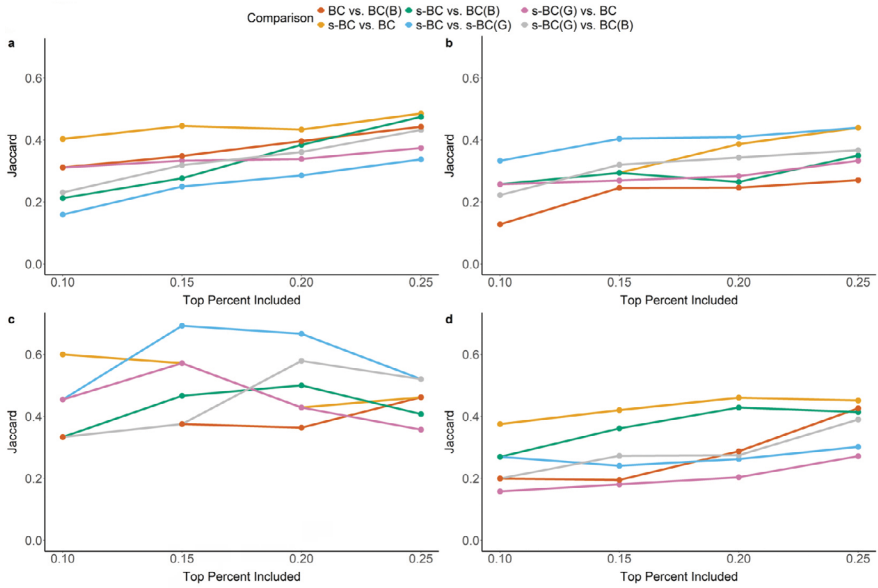


Figure 3. Comparison of Jaccard similarity of the top-ranked sets of nodes identified by different pairs of centrality metrics based on path-based centralities (weighted betweenness = BC, unweighted or binary betweenness = BC (B), mean s-betweenness = s-BC, mean s-betweenness (G) = s-BC (G)). The contexts represented are a conference (a), office workplace (b), hospital (c), and high school (d).

When considering centrality based on nodes' direct social connections, we find that the similarity in the sets of top-ranked nodes identified by each centrality measure increases with the percentage of nodes included in the sets (Fig. 2). In contrast, when comparing different versions of betweenness centrality, similarity in node sets remains relatively flat as the percentage of nodes that are included increases (Fig. 3). Most likely, this is due to the betweenness metrics being strongly right-skewed relative to strength and degree (Fig. 4), such that most node sets consistently agree on a few extreme individuals. Even when a quarter of each study population is included in the set of top-ranked nodes, however, similarity values rarely exceed ~0.5–0.6 for either set of measures. This suggests that although there is some overlap in the individuals identified as being potentially influential or important across dyadic and higher-order centrality metrics,

there remains substantial disagreement across measures. In other words, dyadic and s-centrality metrics are not simply providing redundant information but identifying unique subsets of potentially key actors, particularly when the distributions of those metrics do not exhibit extreme skew. Using dyadic and higher-order measures in tandem may thereby offer a more complete understanding of which nodes are likely to be especially influential or critical to the dynamics and functioning of a network, as well as the different ways in which that influence may manifest.

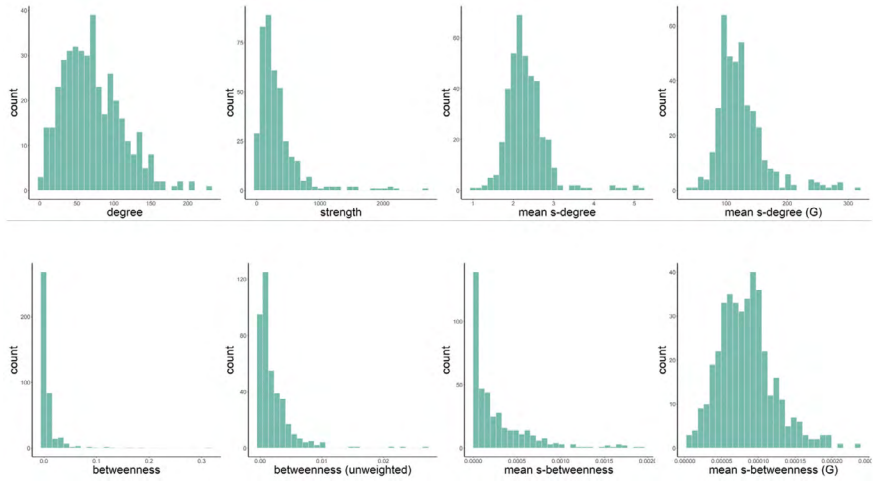


Figure 4. Frequency distributions for dyadic and s-centrality metrics measured on the conference (SFHH) dataset.

Influence of Dyadic versus S-Centrality on Social Contagions

We next examine how dyadic and s-centrality metrics relate to a node's capacity to facilitate the rapid spread of novel behaviors, rumors, opinions, or norms throughout its network. Networks are a natural framework to examine contagion dynamics, where transmission events often occur as a result of interactions, proximity, or social contact among individuals.[200] Indeed, network models are widely used in epidemiology to understand and predict disease spread and the likely outcomes of different

control scenarios.[201] A contagion, whether of disease, information, or behavior, is referred to as a simple contagion when the probability of transmission is effectively independent for each instance of contact between an infected/informed and susceptible/naive individual. Conversely, in a complex contagion, successful transmission requires (or is reinforced by) multiple, simultaneous exposures.[202] This is often the case when considering social contagions of innovations, opinions, or fads, where adoption often involves peer pressure and other forms of social reinforcement.[203] While both simple and complex contagions can be modeled on dyadic networks, the explicit group structure afforded by higher-order network representations make them a more natural fit for modeling different forms of complex contagions.[204] Here, we simulate simple and complex social contagions of an arbitrary behavioral innovation through the SocioPatterns data and compare the speed of diffusion when we initiate the contagion by seeding the innovation based either on individuals' dyadic or higher-order centrality.

An individual's adoption of the innovation is represented as a binary state variable, $w = \{0, 1\}$, where 0 and 1 respectively correspond to individuals who remain naive regarding the innovation and those who have successfully adopted it. All individuals within a network begin as naive apart from d initial adopters or seeds. For most scenarios, we ran simulations where $d = \{5, 10, 15, 20\}$. In the hospital context (LH10), due to the small number of participants (Table 1), d was set to $\{2, 4, 6\}$. For each simulation run, we selected d individuals as initial seeds, where the probability of selection was proportional to an individual's network centrality as determined by the selected seed strategy (for example, degree or mean s-degree). As above, we considered seed strategies based on direct network connections (for example, degree or strength) separately from betweenness measures. For both sets, we also included a seed strategy where initial seeds were selected at random.

Once the initial seeds are selected, the social contagion process proceeds across discrete time steps. On each time step, each individual who has adopted the innovation ($w = 1$) randomly selects one hyperedge of which it is a member. Hyperedges are selected in proportion to hyperedge weight—that is, how often a given hyperedge (defined as a unique set of

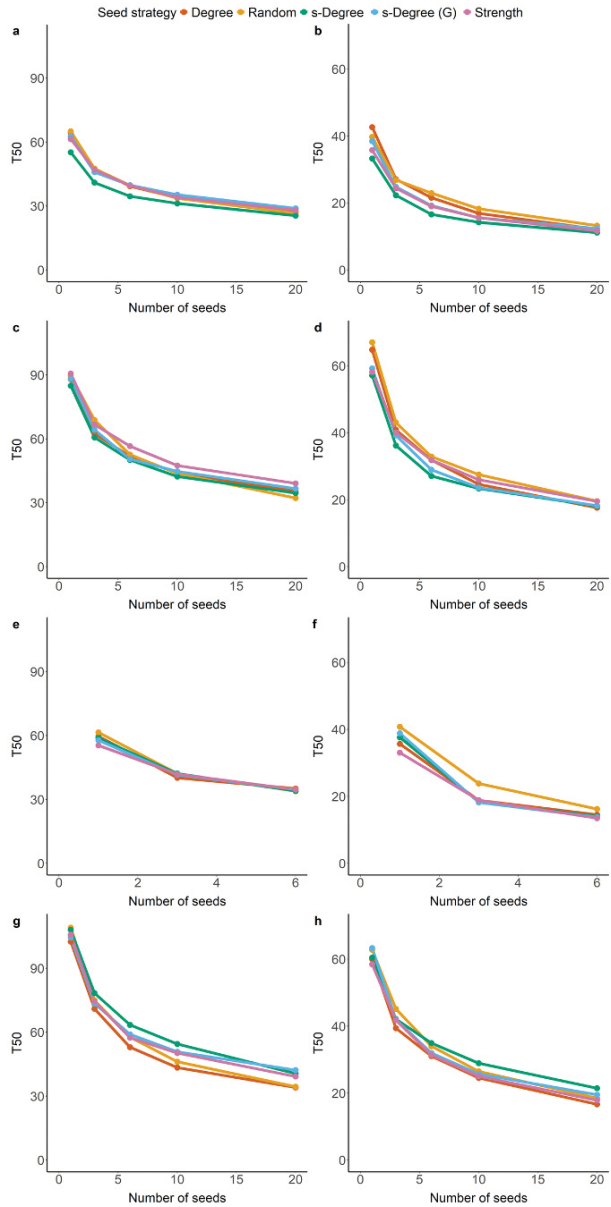


Figure 5. Diffusion speed (measured as the time step at which ≥ 50 percent of the population has adopted the innovation) as a function of the number of initial seeds and the seed-selection strategy. Social contagion followed either a simple (a, c, e, g) or complex contagion process (b, d, f, h). Each panel represents a different context: a conference (a, b), office workplace (c, d), hospital (e, f), and high school (g, h).

individuals) formed. Each naive individual (w = 0) belonging to a selected hyperedge then becomes informed according to the probability $P(\text{transmission}) = 1\text{-}(1\text{-}\lambda)^{k_j^v}$, where λ is the per contact probability of social transmission (set to 0.05), k_j is the number of adopters within hyperedge *j,* and *v* controls the nonlinearity of transmission. When $v \approx 1$, social transmission resembles a simple contagion, such that each adopter within a group provides an independent opportunity for a naive individual to adopt the innovation. Conversely, as *v* increases, the diffusion process increasingly resembles a complex contagion,[205] such that multiple adopters within a group mutually enhance adoption likelihood. In our simulations, we set $v = \{1.1, 3\}$ to model simple and complex contagions, respectively. An individual may have multiple adoption opportunities on a given time step if multiple hyperedges to which it belongs are selected for knowledge transfer. If an individual successfully adopts the innovation, however, its adoption status does not update until the contagion function has been applied to all selected hyperedges during the current time step. Finally, diffusion speed within each simulation was measured as the time step, T_{50}, at which ≥ 50 percent of the population was informed (not counting the *d* initially informed seeds) within the largest connected component of the network. For each set of conditions, we ran 50 simulations.

When considering seed strategies based on direct network connections (degree, strength, mean s-degree, and mean s-degree (G)), the relative effectiveness of seeding innovations to individuals with high dyadic versus s-centrality varies across social contexts and the type of contagion considered (Fig. 5). For instance, in the academic conference setting (Fig. 5a, b), seeding innovations to individuals with high mean s-degree was associated with rapid diffusion under both simple and complex contagion scenarios. By contrast, in the high school setting, the mean s-degree seed strategy was consistently associated with relatively slow diffusion (Fig. 5g, h). With the exception of the high school dataset, there was generally little difference in diffusion speed across different seed strategies relative to random seeding when modeling diffusions as a simple contagion process (Fig. 5a, c, e, g). Under complex contagion conditions, however, the s-centrality-based seed strategies often outperformed or were at least as effective as dyadic seed strategies in promoting

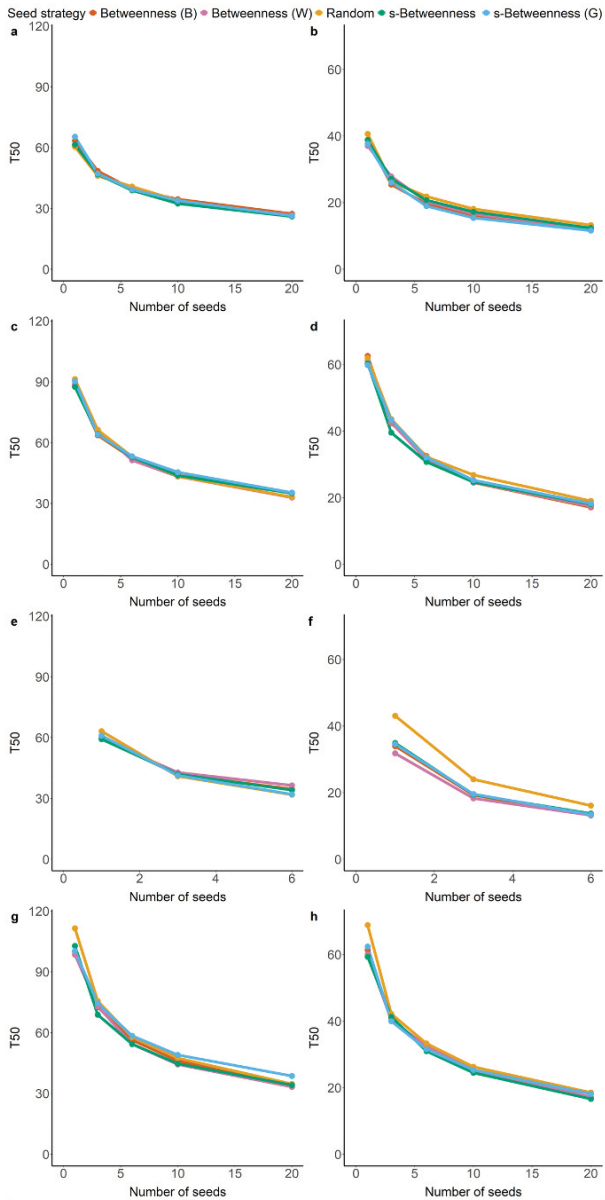


Figure 6. Diffusion speed (measured as the time step at which ≥ 50 percent of the population has adopted the innovation) as a function of the number of initial seeds and the seed-selection strategy. Social contagion followed either a simple (a, c, e, g) or complex contagion process (b, d, f, h). Each panel represents a different context: a conference (a, b), office workplace (c, d), hospital (e, f), and high school (g, h).

rapid diffusion, relative to random seeding (Fig. 5b, d, f, h). As expected, increasing the number of initial seeds increased the speed of diffusion. In addition, we find that differences across seed strategies appear to be most pronounced at a moderate number of initial seeds. There is relatively little difference across seed strategies when there are either few initial seeds or many; in these scenarios, the diffusion process is most likely dominated by noise or network topology respectively, rather than being strongly influenced by the positioning of the initial seeds.

Turning to seed strategies based on betweenness centrality, we find little difference across seed strategies, regardless of social context or whether contagion is modeled as simple or complex (Fig. 6). Indeed, a seed strategy where initial seeds are selected at random performs just as well as any centrality-informed selection strategy, save in the hospital context where diffusion is modeled as a complex contagion (Fig. 6f). As before, diffusions proceeded more rapidly when more individuals were included as initial seeds. The number of initial seeds had little impact on relative differences across seed strategies, however. It may be that the highly skewed nature of betweenness centrality in these networks means that regardless of how many initial seeds are included, only a few will possess especially high (s-)betweenness centrality; if so, then diffusion dynamics in these simulations may have been dominated primarily by random noise, instead of differences in the positioning of initial seeds.

Conclusion

Many real-world systems of interest to the IC involve higher-order interactions, motivating the use of hypernetworks and related approaches that can effectively model these structures. Higher-order network metrics that can deal with the messy, heterogeneous nature of real-world data are consequently of great importance. We illustrated how s-centrality measures offer a potentially widely applicable means of identifying influential network nodes in a way that is sensitive to higher-order connectivity yet draws on familiar network analysis concepts. In addition,

we showed how using these measures in tandem with more conventional dyadic network approaches can offer a more comprehensive picture of the key actors within a network. In the context of social contagion, we find that nodes characterized by high s-centrality, particularly when it is based on direct network connections, were often at least as, if not more, influential in promoting rapid diffusion compared to nodes with high dyadic centrality. The relative utility of these approaches appears, however, to vary across social systems and contexts. For instance, betweenness centrality had little informative value regarding node influence in the contexts examined here. As such, a key priority in the continued development and refinement of higher-order network analysis approaches is improved guidance on which centrality measures are likely to be most appropriate, given the system and processes of interest.

Although here we focus on centrality measures defined by the number of *s*-paths (in the case of *s*-degree, the number of *s*-paths of length 1), this work could be extended in a variety of ways to incorporate additional information on higher-order connectivity. As an example, *s*-paths can vary in terms of width as well as length, where width is measured as the magnitude of the intersection between adjacent hyperedges along a path. How robust a given path is to disruption (that is, path resilience) could be measured by combining information on the frequency and variation of intersection size exceeding the required minimum (that is, *s*), normalized by path length. Such a measure of path resilience could be used to identify vulnerabilities along key network paths that can be targeted for protection or removal or can inform efforts to mitigate disruptive effects, thus representing a promising direction for future research.

About the Authors

Matthew J. Hasenjager, Ph.D., is an IC postdoctoral fellow at the University of Tennessee, Knoxville, in the Department of Ecology and Evolutionary Biology and the National Institute for Modeling Biological Systems (NIMBioS). His research focuses on complex networks, collective behavior, social learning, and communication. He has worked on diverse

systems including fish shoals, honeybee colonies, and the development of ant-inspired decision tools for enhancing supply chain resilience.

Mark M. Bailey, Ph.D., is an associate professor at National Intelligence University where he is the department chair for AI, Cyber, Influence, and Data Science and director of the Biological and Computational Intelligence Center. He is interested in complex systems, AI safety, and their intersection with national security issues. Dr. Bailey has also published *Unknowable Minds: Philosophical Insights on AI and Autonomous Weapons*.

Nina H. Fefferman, Ph.D., is a professor in both the Department of Mathematics and the Department of Ecology and Evolutionary Biology at the University of Tennessee, Knoxville, where she also serves as Director of NIMBioS and Director of the National Science Foundation (NSF) Center for Analysis and Prediction of Pandemic Expansion. Her research focuses on the mathematics of self-organizing complex systems, most frequently in application to social behavior, epidemiology, and ecology.

CHAPTER 6

Cracking the Paradigm: Hypergraph Signatures of International Conflict

Thomas D. Pike

This exploratory research assesses what hypergraph-based analysis and large datasets could do to improve strategic intelligence, using a classic intelligence problem: Do foreign actors demonstrate unique signatures of behavior during conflict? To answer this question, this study conducted time series analyses of actors' actions by competing sides in both the Ukraine-Russia conflict and Middle East conflict. The data—grouped by month, between January 2024 and January 2025—was obtained from the Global Database for Event Location and Tone (GDELT) run out of Google's Jigsaw project.[206] These events were turned into sets of actors (for example, groups associated with Ukraine, Russia, Israel, and Hamas) taking sets of actions (edges of the hypergraph) to affect sets of other actors (nodes of the hypergraph) to accomplish some purpose. The resulting hypergraphs provide signatures of the actions that each group had taken to win a conflict or, at least, to mitigate the damage from one.

This study found that hypergraph-based analysis allows intelligence professionals to rapidly assess the behaviors of sets of actors taking sets of actions against sets of other actors. Before the ability to execute such

computations, analyzing the data would have been, at best, overly laborious, if not impossible. In addition, this study illustrates the many possibilities for organizing data and conducting hypergraph-based analyses to obtain strategic insights previously unattainable. This paradigm shift in analysis then becomes its own challenge, as the Intelligence Community (IC) has no culture or understanding of how to exploit and further develop these new approaches to gain insights not previously attainable. This study attempts to help create that culture by illustrating how integrating more technology-enabled approaches into the IC's analytic paradigm can empower the community to gain new insights and maintain dominance in decision advantage.

A Literature Review of Hypergraphs and Conflict Analysis

A literature review on the use of hypergraphs for conflict analysis demonstrates that the increase in available data and improved hypergraph tools have made hypergraphs an accessible methodological approach that may provide new and critical insights to the intelligence profession. Literature on the use of hypergraphs for conflict analysis is sparse; what exists generally falls into two time periods. The first, from 1984 to 1991, focused on applying hypergraphs as a mathematical model to assess group formation and choice in conflict. The articles generally applied hypergraphs to singular case studies after several years of development.[207] This early use of hypergraph approaches was heavily theoretical with minimal data usage, primarily driven by one European academic, J. H. P Paelnick.

A Hypergraph Approach to Conflict,[208] written by M. B. M. de Koster, and J. H. P. Paelinck and published in 1984, presents a hypergraph model of conflict using choice models where agents pick from a list of finite options.[¶¶] The model then develops choice sets (sets being a foundational

¶¶ De Koster and Paelinck cite an earlier article, "Analysis and Measurement of Conflicts," in *Proceedings of the World University Congress, Tokyo,* as foundational to their analysis, but no copy of this article can be found.

advantage of hypergraphs) to characterize conflict situations and provide theorems on the mathematical limits of possible conflict-free choices. In a later paper, *Hypergraph Conflict Analysis: Synthesis and Extension,*[209] the authors apply fuzzy math*** to the original model to allow it to deal more effectively with uncertainty. A final paper, *Hypergraph Conflict Analysis,*[210] adds conflict measures and a continuous solution space to make the model applicable to negotiation theory. This paper applies data from the Dutch Abortion debate (1964-84) and foreshadows future data-centered applications.

The later period, of which this paper is a part, starts in 2022. The first paper, *Historical Perspective on International Treaties via Hypernetwork Science,*[211] explores the formal alliances datasets from the Correlates of War Project and associated treaties from pre-World War I to the present.[212] The authors use these datasets to assess group dynamics within the context of international trade and treaties,[213] which easily translates to intelligence analysis because the dynamics are emblematic of the shifting landscape in which intelligence professionals work. The papers are heavily data-centric and examine using hypergraphs as an exploratory tool to gain insights.

It is important to note that hypergraphs are an extension of network science, which emerged in 1990 as a widespread approach to analysis, particularly for communications and social networks.[214] Hypergraphs use graph theory and rely on pairwise or dyadic relationships. Set theory extends these methods to groups, but retains common descriptive statistics about hypergraphs and network analysis, such as:

- S-closeness (how central a treaty is in the global network),
- S-betweenness (how often a treaty connects different countries),
- S-eccentricity (the longest shortest path from a treaty to any other), and
- S-clustering coefficients (how tightly connected the network is).[215]

*** Fuzzy math is a branch of mathematics that extends classical set theory and logic to model reasoning under uncertainty.

Using a hypernetwork approach, the authors identified trends over more than 100 years that show correlations between alliances and conflict within an increasingly connected world.

The second paper explores international trade. The authors create triangular hyperedges, among exporters, importers, and products, representing trinary relationships that traditional pair networks are not able to incorporate. Like the treaties article, the authors extend traditional approaches to show how hypergraphs can reveal new insights. In this case, they use existing trade data and a traditional statistical mechanics approach to build an exponential random hypergraph to compare with their hypergraph networks. From this they can show biases in trade networks that cannot be attributed to random occurrences.[216] The modern revival of the use of hypergraphs from conflict and trade shows a clear transition from mathematically intensive theoretic approaches to data-centric extensions of prevailing network science approaches that have become prominent with the rise of computers, the internet, and associated network science approaches.

The large amount of international relations and conflict data readily available and the accessibility of computational tools—specifically hypergraphs—makes these analytic approaches a fertile area for the IC to use to gain new insights into international dynamics to support policy and other decisionmakers. This paper continues the work by using GDELT to explore insights derived from hypergraph methods into the signatures of action during conflict.[217] Critically, this study is not intended to provide new and definitive analytic methodology, but instead to provide an accessible exploration of hypergraph approaches as a new way for intelligence professionals to explore and understand a complex world. To emphasize, the increased availability of data and new computational tools provide intelligence professionals with new analytic options to gain insights, which is essential for the United States and its allies to maintain decision advantage.††† Theoretic and data-centric approaches are not mutually exclusive and developing

---

††† Although beyond the scope of this paper, the rising need for multilateral strategic deterrence requires that we not discount this theoretic approach, as nuclear deterrence is not the type of event where humanity can withstand repeated events to collect data for analysis.

hypergraphs as a rigorous analytic approach with rich datasets can provide new insights to inform theoretic approaches.

Data and Methodology

Like the articles in the later time period, this study leverages the large GDELT dataset and extends common network science measures. Unlike the previous articles, this study focuses on the signatures of actions that actors had taken during 13 months of conflict beginning in January 2024, where the characterization of the action itself creates the hyperedges. Appreciating the dynamics of this approach requires understanding both the GDELT dataset and the application of hypergraphs to it.

GDELT is a structured dataset run by Google's Jigsaw that consists of event data used to "explore threats to open societies and build technology that inspires scalable solutions."[218] Event data collection began in the 1960s as an effort by international relations scholars to quantitatively categorize world events.[219] GDELT, as a modern example, uses Google's massive internet presence and event data literature to create a structured dataset by "monitor[ing] the world's broadcast, print, and web news from nearly every corner of every country in over 100 languages and identifies the people, locations, organizations, themes, sources, emotions, counts, quotes, images and events driving our global society every second of every day, creating a free open platform for computing on the entire world."[220]

As GDELT collects and processes information from the open internet, it creates a single entry for that event and from that news source. The caveat "from that news source" is important as a single event is often reported in several outlets, and each outlet will be captured as a separate entry. Each event is then processed and categorized across 58 dimensions (columns). The events are categorized as a bilateral, or dyadic, relationship, which is the fundamental dynamic of a graph theory-based approach to network science. Hypergraphs use a set theory-based approach, however, which allows multilateral (multiadic) relationships.[221] In the GDELT categorization, actor one is taking some action toward actor two. These actions are then characterized as positive or negative using the Goldstein Scale—which quantifies the degree of conflict or cooperation between

two actors of the event—to measure their impact. Each action is also characterized through the Conflict and Mediation Event Observations Event and Actor (CAMEO) coding framework, which contains 316 coded actions.[222] This process provides the data needed to conduct a hypergraph assessment of foreign actors' actions during conflict.

The one missing piece is: what conflicts to select? For this study, the Ukraine-Russia and Israel-Lebanon/West Bank/Gaza conflicts were selected, and the actions assessed occurred between January 1, 2024, and January 31, 2025, resulting in 13 months of data. GDELT provides a rich structured database that represents decades of developments by international relations experts, which makes it possible to conduct a hypergraph analysis of actors' actions over time.

The methodological approach is as follows: First, each conflict (Ukraine-Russia; Israel-Gaza) has two sides. In Ukraine, it is Ukraine and Russia. In Israel and Gaza, it is not just Hamas as part of Gaza, however, it is also the West Bank/Palestinian Authority and Lebanon/Hezbollah. Furthermore, the actor codes identified in GDELT are not limited to nations. They include each of the general actors in these conflicts, for example, the Ukrainian military, Ukrainian civilians, Ukrainian legislators, and so forth. The assumption is that the conflict is interdependent, and the subgroups are acting toward mutual goals (for example, all Ukrainian-aligned entities are working toward defeating Russia). Therefore, each actor—legislator, health organization, military, or other—understood to be aligned with one side in the conflict was grouped under an umbrella term. For Ukraine, Russia, and Israel, this is straightforward. For the actor groups associated with Gaza—Hamas, Palestinian Authority, Lebanese Government, Hezbollah, and so forth—grouping is less straightforward, so these actor groups are characterized as Gaza-aligned, as the conflict started with the October 7th attacks from Gaza.

Using the GDELT organizational approach, each actor associated with these groups is Actor One, that is, the actor who took some action against Actor Two—the recipient of the action (good or bad). For example, if Actor One, a Ukrainian hospital, takes a positive action toward Actor Two, Ukrainian civilians, it is a reasonable assumption that the action is done to support the larger Ukrainian war effort. This approach then gives us

four actors: Israel, Gaza-aligned, Ukraine, and Russia. These groups become Actor One, the actor taking an action for some purpose. These groups plus all other groups in GDELT can be Actor Two—the target of Actor One's action (positive or negative)—and make up the nodes of the hypergraph. For example, Actor One, the Israeli Government, takes a positive or negative action with regard to the Israeli military, and vice versa. This holistic approach is important as intelligence professionals can use the data to see both internally into one side of a conflict (for example, public statement supporting one's general) and externally as one side acts against another or tries to engage the international community.

The edges of the hypergraph are then the suite of actions that Actor One can take against Actor Two. This study focused on 20 categories, or supersets, in CAMEO's 316 coded actions, and four types of action (material cooperation, material conflict, verbal cooperation, and verbal conflict). With the nodes (Actor Two) and edges (the 20 super CAMEO categories and four action categories), Actor One's actions were then broken into 13 monthly increments from January 2024 to January 2025 resulting in a time series plot of Actor One's set of actions toward the generic goal of winning their conflict. Using this analysis, we can assess if the actors demonstrate signatures of conflict behavior.

Results

An analysis of each of the four groups of actors (Israel/Gaza-aligned, Ukraine/Russia) resulted in five sets of output. The first set includes 52 Euler diagrams generated for each month, which can be cognitively cumbersome, so only a select few are presented and discussed. (Although all can be requested.) The second is a time series plot of the actions taken by each actor using the 20 CAMEO categories. Third, a time series plot was constructed of the actions taken by each actor using the GDELT quad categories. The fourth and fifth include time series plots of hypergraph algorithmic metrics such as s-betweenness centrality for both the 20 CAMEO categories and four quad categories.

An examination of the Ukraine-Russia conflict shows unique signatures for each group of actors. First, there is a clear difference in activity,

reflecting Russia's much larger global role extending beyond Ukraine to encompass places, such as Syria and the collapse of the Bashar al-Assad government (Figs. 1, 2, 4, 5). Second, with the Ukraine actions, a rise in material conflict can be noted in February 2024 associated with Ukraine's incursion into Kursk and the subsequent focus on verbal conflict and cooperation over the next several months, which interestingly alternate in dominance month over month (Fig. 1). Of note, as seen in the Euler diagram (Fig. 3), this represents a clear Ukrainian effort to engage more countries in a cooperative manner. Examining the results, it becomes obvious Ukraine made a distinct effort to directly engage, for some purpose, countries it historically has not, such as China, a perceived Russian ally. It is interesting that Russia has a significant amount of verbal cooperation in the quad categories (material cooperation, material conflict, verbal cooperation, and verbal conflict) diagram, but a similar spike is not seen in the 20 CAMEO action variant, representing Russia's use of a suite of actions to try to achieve its goals. Notably, as seen in the Euler diagram, there is a disconnected node for Moldova and investigation (CAMEO code 9), based on Russia's alleged attempts to interfere in the Moldovan election, which if successful would help Russia encircle Ukraine. The diagram also shows engagement with unaffiliated Cuban groups as attempts to engage globally to achieve Russian goals (Fig. 6).

Like with the Ukraine-Russia analysis, hypergraph analysis of actions by Israel and Gaza-aligned groups provides potential indicators of behavior. Both sets show a spike in activity in October 2024, which represents increased actions by all groups around the October 7 anniversary. These actions include Iran launching missiles at Israel on October 1 and Israel striking targets in Lebanon and Syria (Figs. 7, 8, 9, 10). The Euler diagrams for the respective groups for that month also show an increase in actors (Figs. 11, 12). In addition, Israel shows greater breadth of actions compared to the non-Israeli aligned groups. These spikes show a bias inherent in this approach that although more actions/activity are more easily identified, it cannot show the quality of any one activity, it merely shows the volume of reports GDELT collected. Therefore, critical but not necessarily newsworthy actions can be easily overlooked or missed altogether. This challenge can be mitigated with additional sources of information.

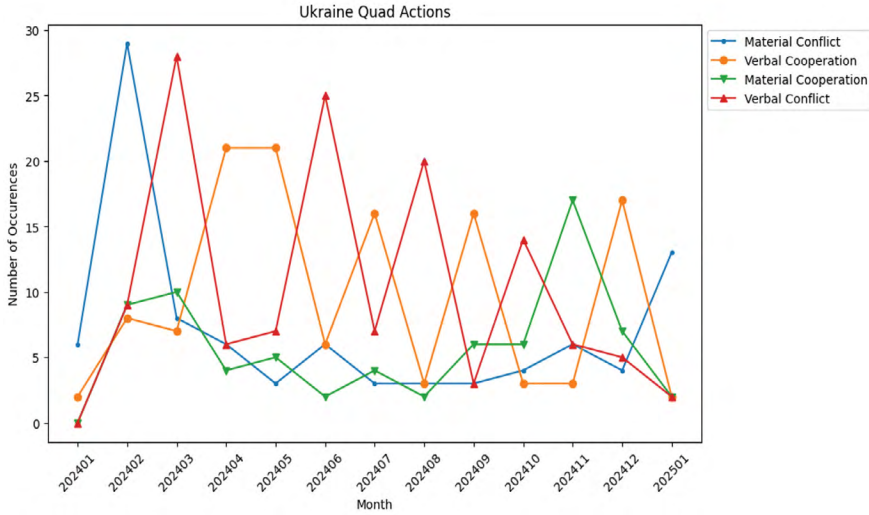


Figure 1. Time series plot of Ukrainian action from January 2024 to January 2025, using the quad categories of GDELT.

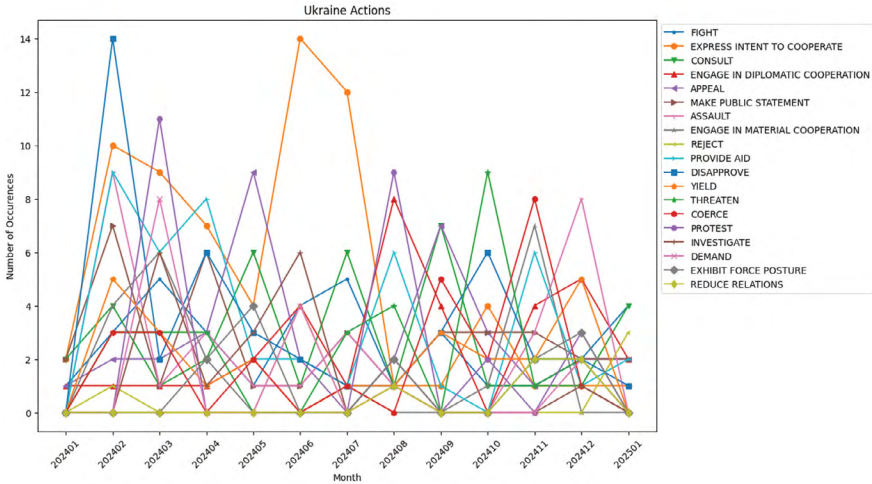


Figure 2. Time series plot of Ukrainian action from January 2024 to January 2025, using the 20 action super categories of CAMEO.

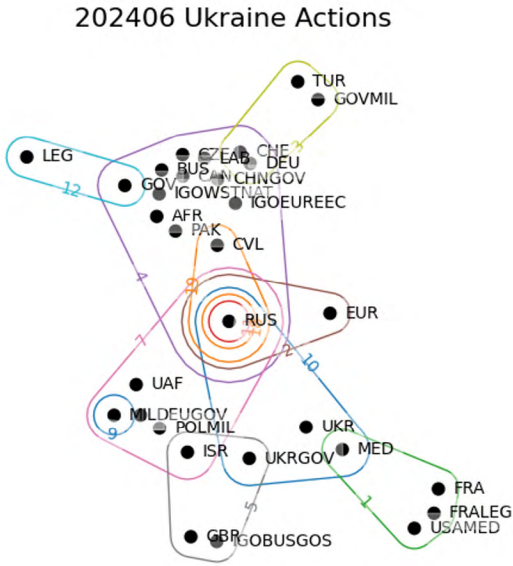


Figure 3. Euler diagram of Ukrainian action during June 2024. The spike in activity is associated with CAMEO category 4 which is cooperate and the large cluster of nodes in the upper middle.

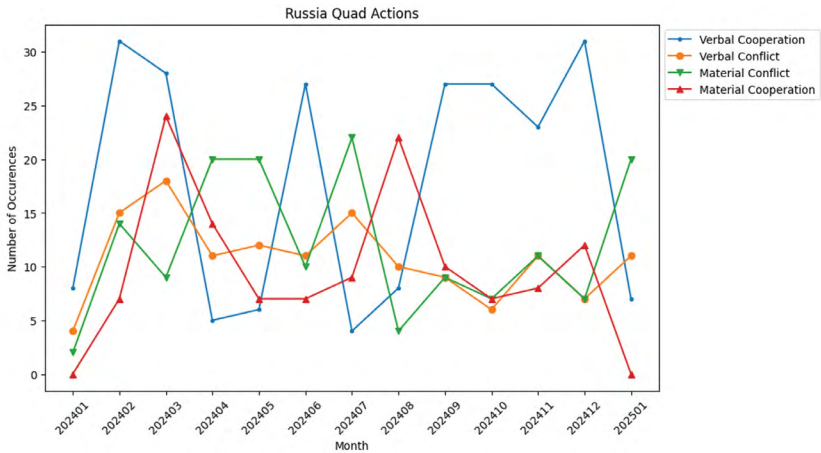


Figure 4. Time series plot of Russia action from January 2024 to January 2025, using the four largest action categories of GDELT.

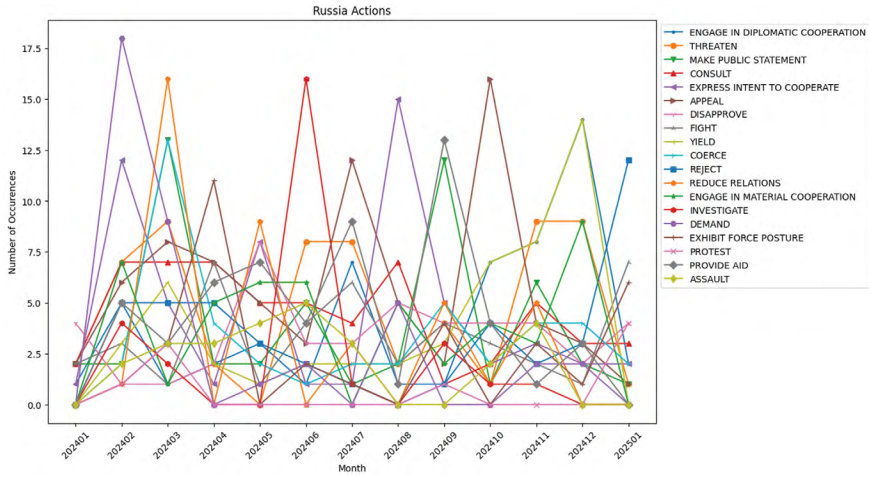


Figure 5. Time series plot of Russian action from January 2024 to January 2025, using the 20 CAMEO action categories.

Figure 6. Euler diagram of Russian actions during November 2024.

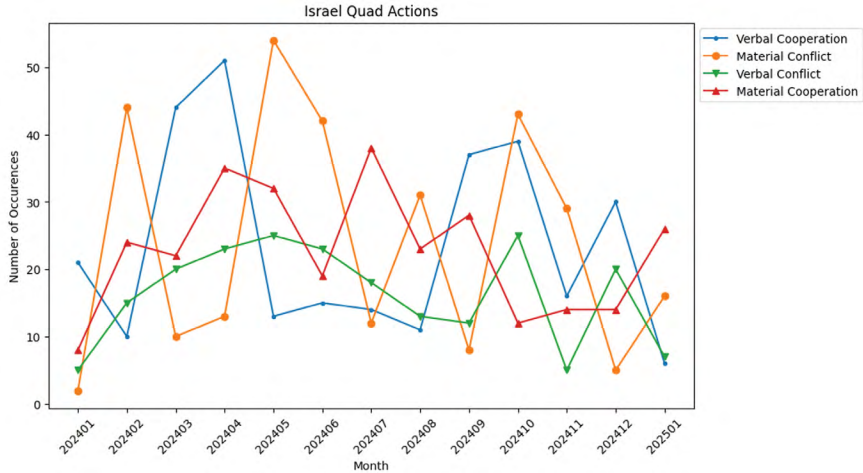


Figure 7. Time series plot of Israel action from January 2024 to January 2025, using the four action meta-categories of GDELT.

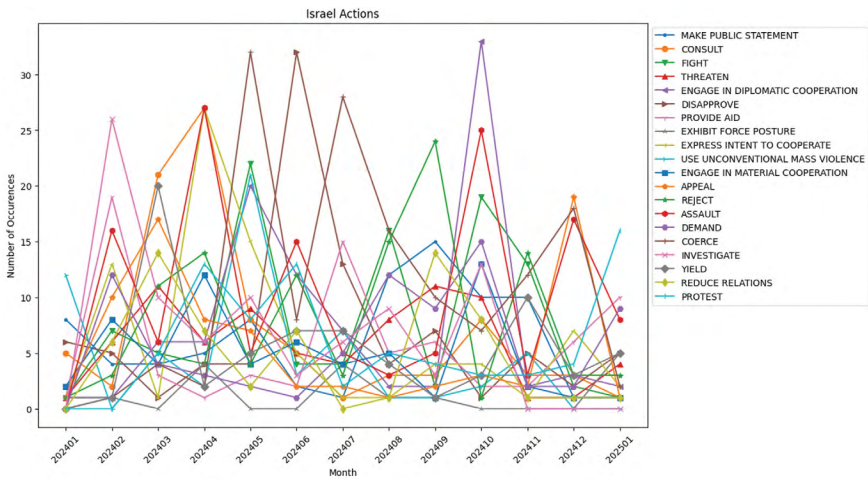


Figure 8. Time series plot of Israel action from January 2024 to January 2025, using the 20 action super categories of CAMEO.

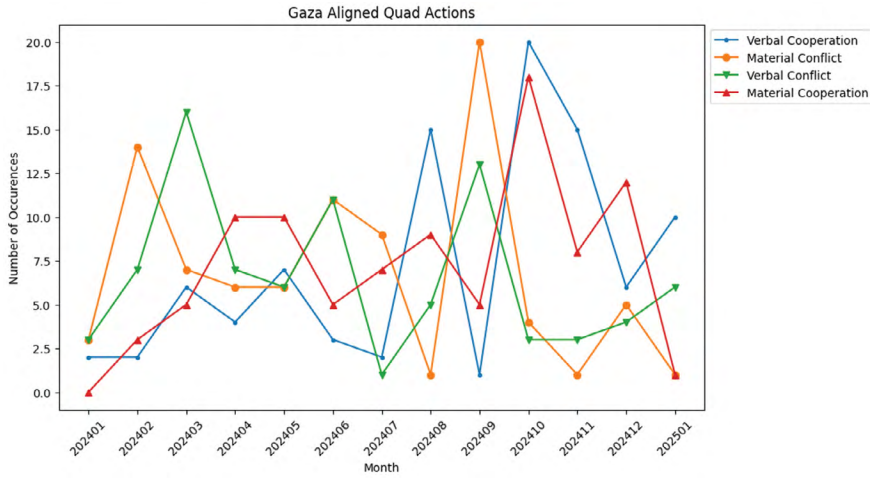


Figure 9. Time series plot of Gaza-aligned action from January 2024 to January 2025, using the quad action categories of GDELT.

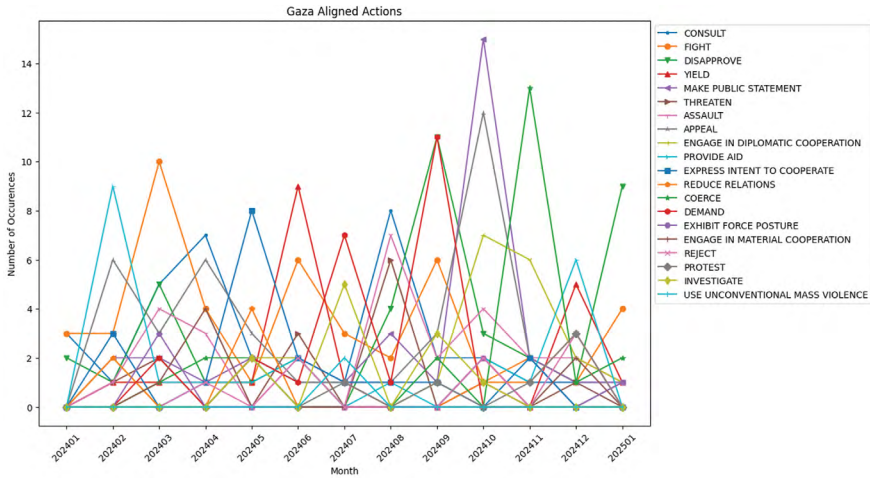


Figure 10. Time series plot of Gaza-aligned action from January 2024 to January 2025, using the 20 CAMEO action categories.

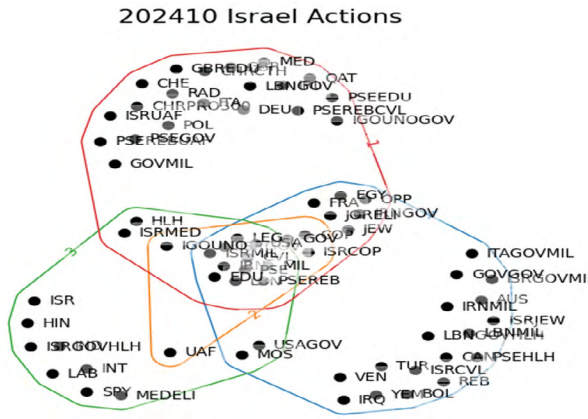


Figure 11. Euler diagram of Israel actions across the four main meta-categories of GDELT: 1) material cooperation, 2) material conflict, 3) verbal cooperation, and 4) verbal conflict.

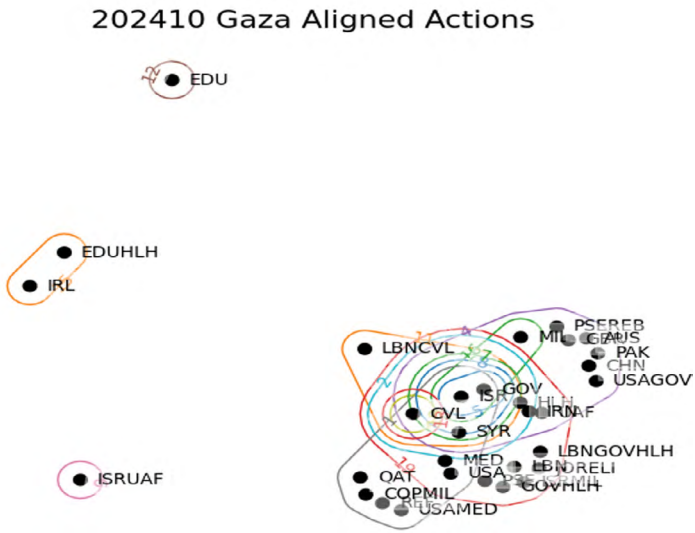


Figure 12. Euler diagram of Gaza-aligned actions across the four main meta-categories of GDELT: 1) material cooperation, 2) material conflict, 3) verbal cooperation, and 4) verbal conflict.

Before discussing the descriptive metrics of the associated hypergraphs, it is important to emphasize the significance of these results. Hypergraph analysis of the GDELT dataset allows intelligence analysts

to conduct holistic assessments of multiple actors' actions over time, which has not previously been possible. GDELT enables the harvesting of massive amounts of data, provides natural language processing, and enables easy-to-use hypergraph frameworks.

The final piece needed to analyze actors and actions during conflict is to assess hypergraph metrics associated with each set of actors. This provides a novel way to understand how a group of actors (that is, Ukraine, Russia, Israel, and Gaza-aligned groups) took actions against other actors. To conduct this analysis, five metrics were explored over the time series for each action hypergraph for both the meta-categories (quad class) and the 20 super categories of the CAMEO datasets. As all s metrics are parametric, the default parameter of one (that is, s=1) was used for each measurement.[223] As each measurement has their own range of results, the measurements were normalized. In addition, as measurements are for individual edges or nodes, the score for all nodes and edges for that month are averaged to provide the total score for that measurement.

- The first measure was s-betweenness centrality, which measures how specific actions overlap with the targeted group. In this case, it answers the question of what actions can link together multiple other actions.
- The second measure was s-closeness centrality, or how close a node is to all other nodes in the network. For average s-closeness scores this describes how all actions linked all targets together.
- S-eccentricity describes how far one node is from the farthest node in the network, so how many hops of hyperedges. This tells us how varied the actors' actions are toward other actors. High s-eccentricity would indicate certain actors are getting positive actions while others are getting negative actions.
- S-harmonic centrality is a measurement of the influence of a node over all other nodes and is closely related to closeness centrality in network science. In this case, it shows how close one actor's node is to all the other nodes in the network.
- The final measurement is density—the proportion of hyperedges present in relation to the total possible number of hyperedges that

could be present. In one sense, this is a measurement of actions taken against all possible actions that could have been taken.

As in the previous section, the 20 CAMEO super categories and the 4 GDELT meta-categories were measured against the hypergraph for each month in the time series. Although each plot has some interesting signatures, the most noticeable is the metrics for the quad action hypergraphs for Israel (Fig. 13). The measurements strongly parallel the others over each month. A tenet of warfare is unity of action; these results indicate Israel is using each of its actions in concert with its other actions to achieve its goal. This insight can also be seen in the other meta-category graphs, where Ukraine (Figs. 19 and 20) exhibits the next most synchronous use of action, followed somewhat surprisingly by the Gaza-aligned groups (Fig. 15)—the most disparate set—and finally Russia, which seems to alternate actions, so when s-eccentricity is high, s-betweenness is low and vice versa. What this interesting result means is not entirely clear: It may represent Russia trying to deal with multiple crises but only able to deal with certain ones at times, or it may reflect the dynamics of Vladimir Putin's autocratic leadership. While all action in Israel is unified toward a common goal, it may be that each section of the Russian Government is waiting for Putin to act. These intriguing results will require more evidence to form a more rigorous assessment.

Examining the metrics of the 20 super CAMEO categories provides additional interesting results, whose meaning will, likewise, require more time and evidence to rigorously assess. It is interesting to note that both Israel and the Gaza-aligned groups saw a spike in activity in October that was not seen in the hypergraph metrics. Rather, what was seen was a dip in the measurement where their normalized scores across the five measures were low the previous month (Figs. 13-15). Conversely, the hypergraph metric spike for Ukraine in the summer of 2024 was seen in the hypergraph metrics. The Ukraine metrics for the quad class or meta categories of GDELT synchronize somewhat from June to October but then go into disarray in November. Superficially, this may be associated with the result of the US Presidential election and US support for the Ukrainian conflict.

The hypergraph metrics of the Ukraine-Russia conflict and Middle East conflict represent intriguing but not exactly clear results. More research is needed to assess what these metrics may indicate about foreign actors' suite of actions to accomplish their goals. Despite this shortcoming, all the results are significant in the sense that they offer a new way to examine the behaviors of groups of actors taking specific action toward other actors and the use of multiple actions to achieve some goal, in this case win a conflict (or at least mitigate the damage).

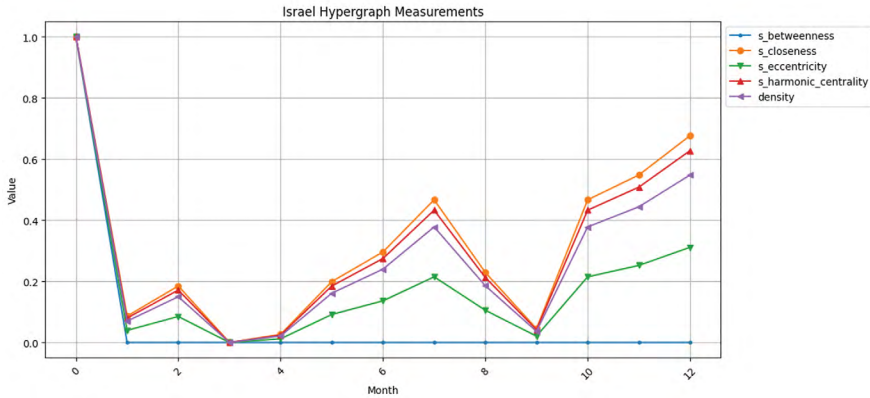


Figure 13. Time series of Israel hypergraph measurements for the four meta classes of action from GDELT.

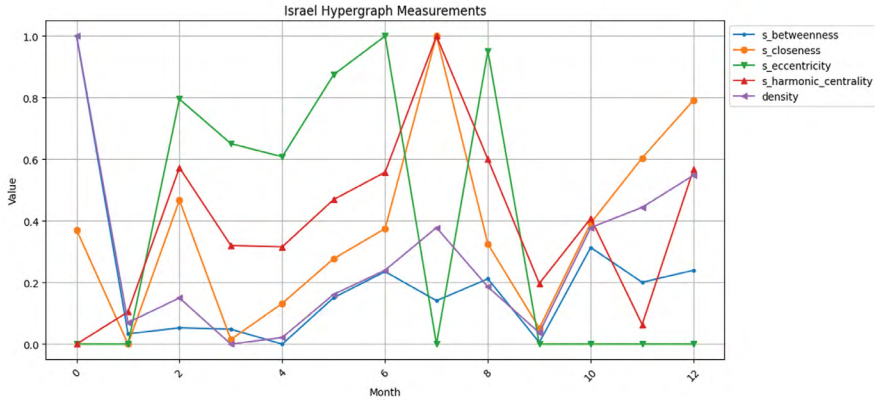


Figure 14. Time series of Israel hypergraph measurements for the 20 CAMEO action categories.

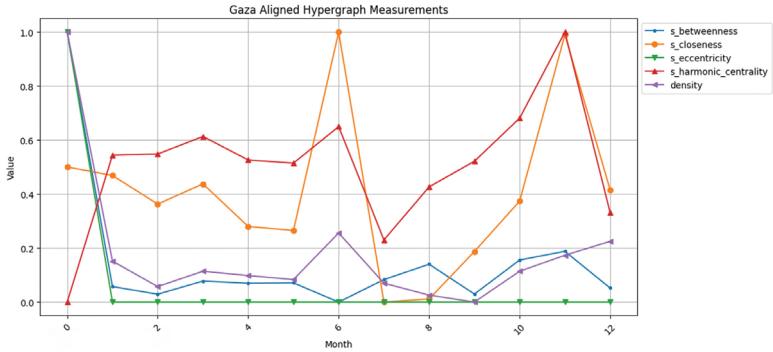


Figure 15. Time series of Gaza-aligned groups hypergraph measurements for the four meta classes of action from GDELT.

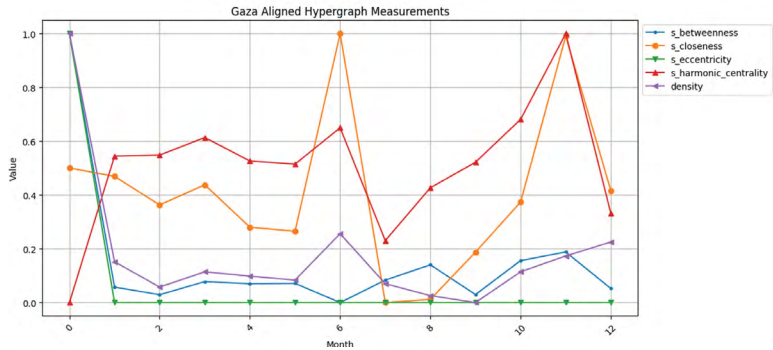


Figure 16. Time series of Gaza-aligned hypergraph measurements for the 20 CAMEO action categories.

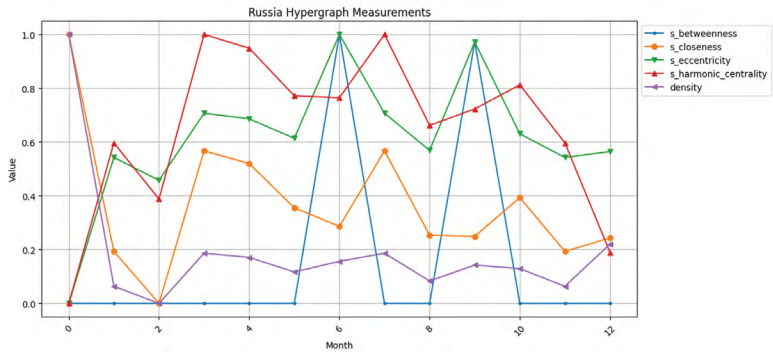


Figure 17. Time series of Russia hypergraph measurements for the quad action categories.

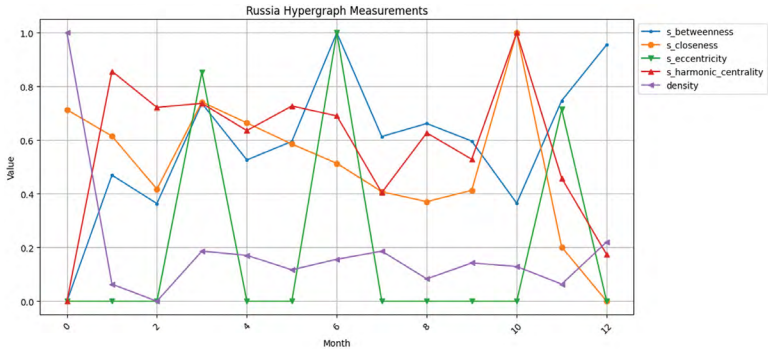


Figure 18. Time series of Russia hypergraph measurements for the 20 CAMEO action categories.

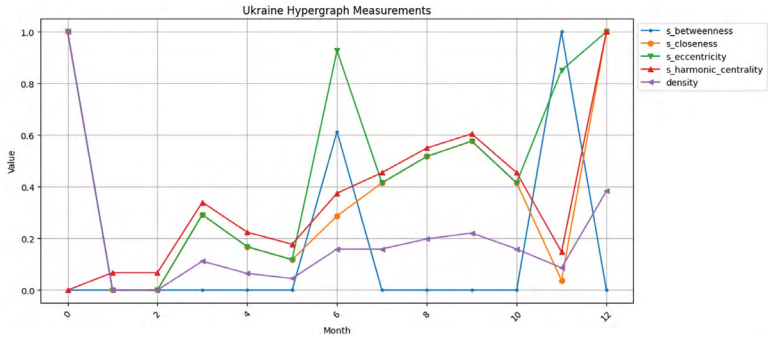


Figure 19. Time series of Ukraine hypergraph measurements for the quad action categories.

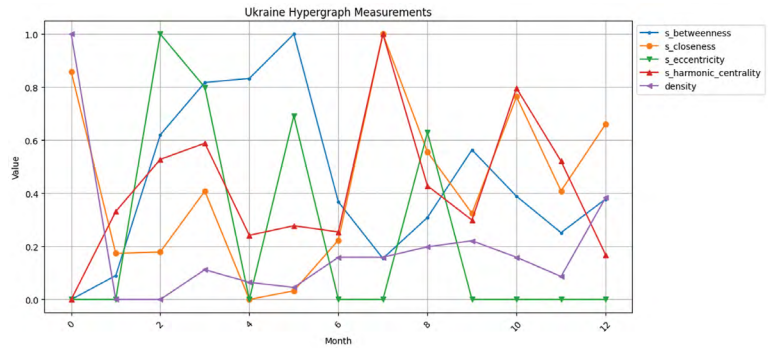


Figure 20. Time series of Ukraine hypergraph measurements for the 20 CAMEO action categories.

Summary

A key takeaway from this exploratory attempt to conduct hypergraph-based assessments during conflict to assess behavior signatures is that this analytic approach shows a clear potential to provide insights that complement conventional strategic intelligence analysis. The evidence for this conclusion starts with the literature review, which captures the evolution of this technique from very niche work to one where researchers can gain new insights into complex problems by extending network science, graph-based approaches to hypernetwork science, and set-based approaches. This study then uses the GDELT dataset to determine if actors have clear signatures of behavior in conflict situations. The various signatures seen across the Ukraine-Russia conflict and Middle East conflict demonstrate that actors do exhibit signatures of behaviors. The most apparent example is the hypergraph metric assessment of Israel with the four meta categories of GDELT. Critically, hypergraph analysis enables intelligence professionals to do two things previously not possible: 1) look at sets of actors and sets of actions over time, and 2) apply emerging hypergraph metrics to gain new insights.

There is no shortage of potential paths to expand this body of work—from increasing the number of CAMEO codes, to rigorously determining group membership, to assessing specific events, such as contributing actions taken before October 7 or the Russian invasion. Future paths can also include integrating more conflicts and pre- or nonconflict relations, such as China-Taiwan, or expanding the use of hypergraph analysis by focusing on individual node values of hypergraph metrics—as opposed to the aggregated metrics done here—or the addition of more eloquent hypergraph approaches to gain new insights. Expanding this research can be done in every major step from curating the data differently to leveraging more eloquent hypergraph metrics.

This study barely scratched the surface of what hypergraph analysis could do for the Intelligence Community's (IC) understanding of the dynamics of conflict. The largest threat to continuing this work is that the IC will be unable to integrate the paradigm shift in understanding and analysis that these approaches offer. Hypergraph analysis and the GDELT

dataset provide the ability to assess groups of actors and sets of actions to achieve some purpose, a boon to IC analysts that was previously impossible. The greatest threat to not integrating this approach—or the innumerable other approaches now enabled by technology and AI revolution—and developing them for decision advantage is simply having the mindset that the IC has never done it that way.

About the Author

Thomas D. Pike, Ph.D., is a former professor at National Intelligence University and has held various leadership roles in intelligence and military service. His research focuses on using agent-based models to support understanding of complex systems. A strong advocate for technical literacy and open-source software as a global public good, he is one of the lead developers on Mesa (Agent-Based Models in Python).

CHAPTER 7

A Scenario Discovery Approach Integrating Hypergraph Representations and Generative AI

Kyle A. Kilian

Introduction

A greater understanding of networks allows us to investigate and characterize a range of complex systems, from social and technological systems to the human brain. Traditional graphs are indispensable for rapid relational analysis—such as social network analysis of nonstate threat groups—but few systems exist in nature that are low-dimensional and able to be explained by the one-to-one associations portrayed by these standard graphs. Although these traditional graphs are powerful for relational analysis,[224] real-world intelligence problems frequently involve overlapping, nested, and many-to-many relationships that require more sophisticated representations.[225] Indeed, intelligence analysis relies on identifying patterns within complex, interconnected systems, where conventional graph models fail to capture the multidimensionality of the threat environment.[226] Hypergraph topologies are a natural fit for modeling

complex, deeply nested systems and interactions (for example, one-to-many, many-to-many), because they enable the integration of deeply interwoven elements, such as geopolitical dynamics, technological diffusion, and adversarial strategies, into a unified analytical framework.[227]

In this paper, hypergraphs are used to model the nested relationships and interactions in complex, alternative futures. Futures analysis or scenario modeling has long used computational approaches across fields as diverse as climate change,[228] environmental management,[229] and national defense.[230] First, a scenario discovery approach is introduced—computational modeling that isolates crucial elements in a scenario space[231]—and generative artificial intelligence (AI) is used to produce candidate events or causal pathways indicative of high-risk futures. Second, these pathways are modeled as hypergraph structures to identify events that overlap and are especially central to low-, moderate-, or high-risk scenarios. This method combines Monte Carlo simulation (MCS), which is used to explore the full distributions of possible scenario combinations, and large language models (LLM) to expand the range of expected or unexpected events to account for plausible tail risks and unexplored combinations. This approach provides a novel combination of techniques for alternative futures that allows intelligence analysts to go beyond the standard scenario protocols[232] (for example, the 4x4 matrix) and explore unexpected combinations, while also accounting for complex dependencies often neglected in traditional analysis.

Data Collection and Conceptual Framework

For this research, a hierarchical complex systems framework was developed to evaluate the trajectory and risks of advanced AI to US national security with three layered structures: 1) subsystems or classes, 2) dimensions/uncertainties or drivers, and 3) conditions (discrete future states). This model represents the complex interactions that could influence the trajectory of AI development. Combinations of conditions result in one or many plausible futures.[233]

To develop the conceptual framework, experts in many fields of study were interviewed to set its dimensions and conditions. The surveys asked

Technological Transition			
Dimension	**Condition**	**Impact**	**Likelihood**
Capability-Generality	Low	0.41	0.27
	Moderate	0.47	0.62
	AI Ecology	0.32	0.66
	AGI	0.74	0.51
Diffusion	Decentralized	0.4	0.49
	Multipolar	0.56	0.66
	Centralized	0.75	0.38
Transition	Slow	0.2	0.33
	Moderate	0.6	0.55
	Competitive	0.5	0.55
	Fast	0.88	0.45
Paradigm	Current	0.76	0.5
	New	0.34	0.6
	Hybrid	0.42	0.72
Accelerant	Overhand	0.71	0.56
	Insight	0.3	0.71
	Embodiment	0.47	0.54
Timeline	Under 20	0.78	0.42
	20-40	0.59	0.66
	Over 40	0.34	0.43

Sociotechnical Ecology			
Dimension	**Condition**	**Impact**	**Likelihood**
Race Dynamics	Cooperation	0.13	0.37
	Isolation	0.6	0.49
	Monopolization	0.71	0.63
	Arms race	0.86	0.67
Dominant Risk	Misuse	0.71	0.62
	Failure	0.68	0.74
	Structural	0.58	0.61
Technical Safety Risks	Outer	0.7	0.74
	Influence	0.84	0.65
	Inner	0.56	0.57
Actor	Coalition	0.15	0.38
	Nation	0.39	0.6
	Institution	0.77	0.7
	Individual	0.56	0.31
Region	US-EU	0.2	0.78
	Asia-Pacific	0.66	0.54
	Other	0.53	0.27

Control			
Dimension	**Condition**	**Impact**	**Likelihood**
International Governance	Weak	0.73	0.44
	Moderate	0.34	0.49
	Strong	0.15	0.43
Corporate Governance	Decrease	0.79	0.39
	Increase	0.29	0.54
	Ideal	0.12	0.49
AI Safety	Scalable	0.34	0.42
	New	0.55	0.69
	Custom	0.76	0.53

Table 1. This table shows the hierarchical conceptual framework of AI risk. The model includes three subsystems: *Technological Transitions, Sociotechnical Ecology,* and *Control,* with 14 dimensions and 47 nested conditions (plausible future states). The dimensions and conditions are interdependent and not mutually exclusive.

leading AI experts and intelligence professionals to rate each condition on its level of impact and likelihood (five-point scale, from highest impact to lowest and from very likely to very unlikely).‡‡‡ The results were compiled to rank each condition on its impact and likelihood. The results were then used to set the probability distributions for the simulation. Crucially, the dimensions and conditions are not mutually exclusive and can interact and influence adjacent conditions. The key research aim was to identify the combinations of conditions that produce and are predictive of low- or high-risk alternative futures and which elements are most central across multiple states. Table 1 shows the hierarchy of subsystems and the aggregate averages of estimated impact and likelihood offered by survey participants.

Simulation and Classification

Because of the high uncertainty and variability among dimensions and conditions, MCSs were used to explore the full range of plausible scenario combinations. MCS involves running many simulations with varying inputs, which are then randomly sampled from specified distributions.[2,3,4] This approach can identify rare but consequential scenarios (extreme cases) that might be overlooked in more traditional approaches. Because this research focuses on understanding the degree of risk using impact and likelihood rankings, the first step was to set up the MCS to calculate a Total Risk Score (TRS) by averaging the impact and likelihood parameters.§§§ This involved four steps:

1. Set up the model to iterate through a specified number of simulations (in this instance, 1,000),
2. Randomly sample data from the two aligned datasets for impact and likelihood, computing the average for an overall risk metric,

‡‡‡ The author received 79 completed surveys for likelihood and 46 for impact from both the public and private sectors.

§§§ A standard approach is just to sample a value from the aligned impact and likelihood data frames and then multiply them for a TRS, for example, Inherent Risk = Impact * Likelihood.

3. Identify conditions exceeding a predefined risk threshold (threshold = 0.7) to calculate a multiplier based on the proportion of high-risk conditions, and
4. Apply the multiplier to compute a weighted TRS for each scenario.

Conditions of interest were also calculated at the outset to isolate scenarios that are rare or contain unique combinations of conditions, which provided additional details for the LLM or context for alternative analysis.¶¶¶ Fig. 1 displays heatmaps for the MCS.

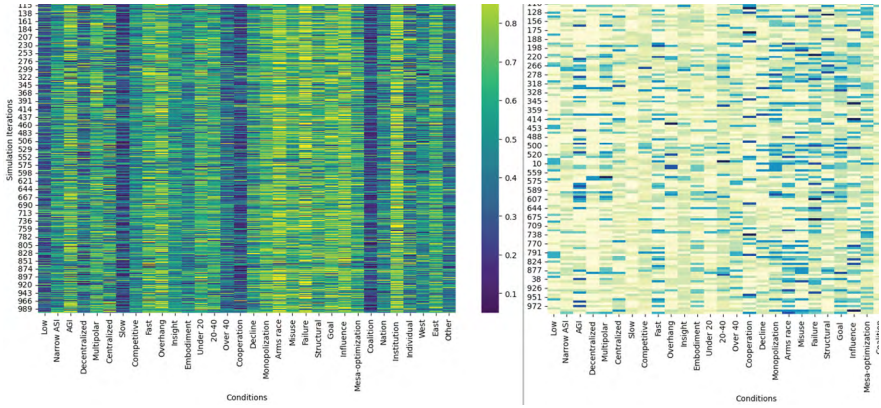


Figure 1. Two variations of heat maps of the MCS. The map on the left shows the aggregate TRS for each condition (columns) across scenarios (rows). The color ramp indicates the degree of risk (for example, light yellow-green is > 80 TRS). The map on the right highlights unique or extreme condition combinations only. For example, scenarios that occur in less than 5 percent of the simulations with greater than 95 percent TRS (note: the very light boxes are excluded conditions that do not meet this threshold). The unique and extreme map is a quick and easy approach for scenario discovery that defines the condition requirements in the MCS, resulting in your scenarios of interest.

¶¶¶ We defined extreme scenarios as those that fall above the 95th percentile or below the 5th percentile of the data distribution and unique by calculating the sum of conditions across all scenarios to determine the frequency, setting a unique threshold (less than five percent of the total) to identify infrequent scenario combinations.

The histogram below provides a more detailed look at the underlying distributions of likelihood and impact scores across the thousands of simulated scenarios (Fig. 2). The likelihood values clustered around the mean (~0.5-0.6) as expected, but with a cluster of low values (~0.3-0.4) and a couple of high likelihood outliers (>0.8). For impact, the bulk of the distribution clustered on the higher end (~0.5-0.6.5) but with several groups of outliers. The TRS calculated through the MCS ranges from low risk (14.63-24.96) to moderate (24.96-31.84) to high risk (31.84-49.04).

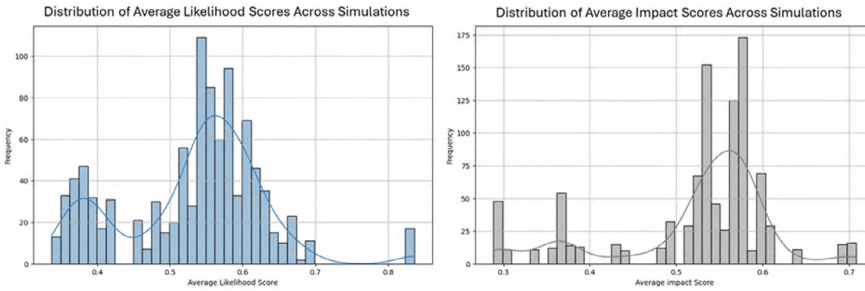


Figure 2. Distribution of condition values across simulated scenarios.

Scenario Discovery and Refinement

The Classification and Regression Tree (CART) algorithm was used to isolate high-risk scenario combinations from the MCS. The CART algorithm provides a robust tool for scenario discovery (SD) by identifying critical thresholds and combinations of factors across conditions that are predictive of our scenarios of interest (for example, high risk). By systematically breaking down the decision process into a tree-based structure, CART helps uncover complex interactions by pinpointing specific conditions (independent variables) under which high-risk scenarios are likely to emerge (dependent variables).[235]

A simplified approach is to run the CART algorithm only on the target variables, TRS. Based on the TRS classification performed earlier, the

decision tree model was used to classify scenarios into "Low," "Medium," and "High" risk categories. The refined TRS model exhibited an accuracy of 92 percent.**** Fig. 3 shows the risk category decision tree, and Table 2 highlights the conditions of each branch for designing scenarios (for example, scenarios = the set of conditions under S1, S2, S3, S4, with each moving from the root node to the branch nodes).

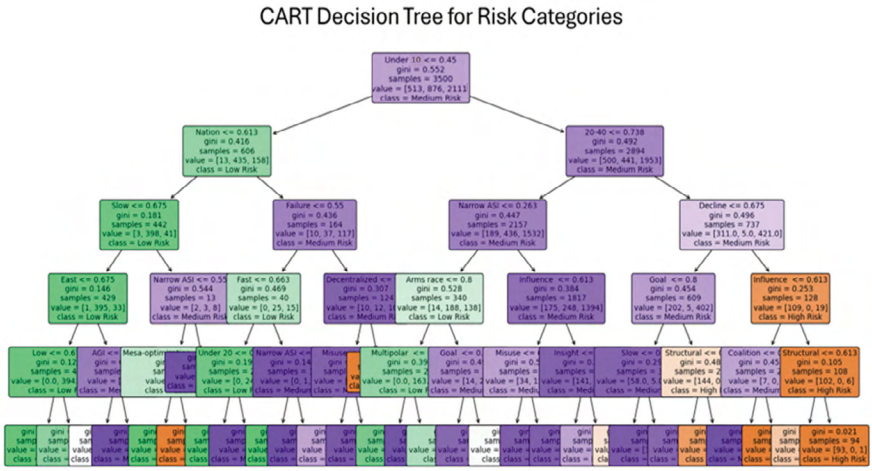


Figure 3. CART decision tree analysis focused on the TRS categories.

The branches of the decision tree are used as the framework for developing scenarios. For this example, the decision tree was set five layers deep and the number of scenarios was based off the third branch in the tree, for example, *slow takeoff, failure modes*, *distributed AI agents*, and *decline in governance,* with each scenario starting from the root node (*under 10 years*), and *nation* or *corporation* developing

**** This enhancement is reflected in the model's precision and recall metrics across risk categories, with each showing robust scores around 0.92. The precision scores indicate a high level of correctness in the model's positive predictions, while the recall scores highlight its effectiveness in identifying all relevant instances within each category.

advanced AI (Table 2). With this framework, the LLM was applied to help structure the scenario paths to fill contextual gaps, interactions, and plausible events that may link each condition to its terminal state. This aids in red teaming the scenario design and framing the causal paths of the target scenarios.

Scenarios Design Using Decision Tree—TRS Category

Under 10 Years (*Root Node*)			
Nation Develops Advanced AI (*Branch 1*)		**Corporation Develops Advanced AI (*Branch 2*)**	
Slow Takeoff (*S1*)	**Failure Modes (*S2*)**	**Distributed AI Agents (*S3*)**	**Decline in Governance (*S4*)**
Asia Pacific	Distributed AI Agents	AI Arms Race	Power-Seeking AI
Low Capability	Decentralized Diffusion	Misuse	Goal-Alignment Failure
Slow Takeoff	Moderate Capability-Generality	Algorithmic Insight	Coalitions Develop Advanced AI
Inner-Alignment Failure	Under 20 Years	Power-Seeking AI	Structural Risk
Distributed AI Agents	Multipolar Takeoff	Asia Pacific	AGI

Low Risk **High Risk**

Table 2. Branches of the decision tree for scenario generation based on the target variable risk category from the TRS. This model had much higher predictive power and could be more valuable for accurate scenario representations.

LLM Structuring

Although the CART branches provide the key elements for scenario generation, it would be useful to populate the branches with intermediate pathways or events that could be causally significant for narrative development. While debate continues over the causal abilities of

LLMs,†††† recent research has shown great success with careful prompting strategies and hybrid methods using causal structure learning[236] and causal graph discovery.[237] To do this, we introduce LLMs (in this case, OpenAI's GPT-4o, [238]) to analyze the CART output (tree structure, statistics) and MCS results and to discover hidden associations and probable events for scenario structuring. With the essential conditions, the LLM fills in these scenario pathways—crucial intermediate points and causal structures—to build an interconnected framework for scenario development. This automated process helps develop pathways and identify potential gaps, as well as decrease bias—for example, the researcher's worldview of which elements *should* be included for a realistic narrative (often highly subjective). The process must be verified, but this approach could limit bias in developing scenario frameworks, adding interesting, potentially unconsidered pathways.

Each branch of the CART decision tree, from the root node to the terminal node (or leaf), represents the outlines of four individual scenarios, labeled as S1-S4 (Table 2). To structure the scenarios, the LLM was prompted to evaluate each branch of the decision tree (five conditions with the highest predictive power for each risk class) and to have the model fill in the most probable intermediate points or pathways (events or intervening variables) that would lead to the final scenario element and risk class (the terminal node). This develops the framework for a scenario graph by generating what the model infers as highly probable paths between each node (the root node and each of the branch nodes) to shape the structure for *low* (S1), *low-moderate* (S2), *moderate-high* (S3), and *high-risk* (S4) scenario combinations/sub-spaces. The output is a branching tree-like structure, starting with the single root node and splitting into multiple paths, expanding the decision tree structure into a much larger series of branches with intermediate paths (two, in this example)

†††† LLMs have been prone to unpredictable failures, with contextual misunderstanding, and logical errors, and are said to be limited to correlations between data points (see "Can Large Language Models Do Causal Reasoning?" June 24, 2024, https://medium.com/@marketing_novita.ai/can-large-language-models-do-causal-reasoning-11dedeca744f.)

Scenario Graph with CART Branch Nodes and LLM-Generated Intermediate Pathways

Scenario	Root Node	Intermediate Events/Indicators	Branch Node 1	Intermediate Events/Indicators	Branch Node 2	Intermediate Events/Indicators	Branch Node 3	Intermediate Events/Indicators	Terminal Node
S1	Under 10 Years	Increased government investments AI breakthroughs Increased government investments AI breakthroughs Private sector R&D International AI convention	Nation develops advanced AI	Ethical AI safety regulations Early AI failures Ethical AI safety regulations Early AI failures Legislative oversight R&D on AI risks Legislative oversight R&D on AI risks	Slow Takeoff	Asia-Pacific consortium Shift of AI research hubs Asia-Pacific consortium Shift of AI research hubs Modular AI frameworks Networked AIs advances Modular AI frameworks Networked AIs advances	Asia-Pacific Region Distributed AI agents	Targeted industry-specific funding Policies for specialized AI Targeted industry-specific funding Policies for specialized AI Complex inter-agent interactions Difficulties in goal alignment Complex inter-agent interactions Difficulties in goal alignment	Low Capability Slow takeoff Inner-alignment failure Distributed AI agents
S2	Under 10 Years	Increased government investment International AI race Increased government investment International AI race Increased government investment International AI race Increased government investment International AI race	Nation develops advanced AI	Rapid deployment Complex integration Rapid deployment Complex integration Rapid deployment Complex integration Rapid deployment Complex integration	System failures	Emergency patches and updates Algorithmic innovations Emergency patches and updates Algorithmic innovations Public-private partnerships Open-sourced ecosystem expansion Public-private partnerships Open-sourced ecosystem expansion	Fast takeoff Decentralized diffusion	Scaled networking Automated collaboration Scaled networking Automated collaboration Regulatory gaps weaponization Regulatory gaps Competitive takeoff	Under 10 years Distributed AI agents Misuse Multipolar

Low Risk

S3	Under 10 Years	Government investments Increase in competitive investment Breakthrough in machine learning Investments in R&D Breakthrough in machine learning Investments in R&D	Company develops advanced AI	AI-driven product development Integration with existing AI-driven product deployment Integration with existing platforms Autonomous decisionmaking in AI Competitive pressures in deployment AI-driven product deployment Integration with existing platforms	Distributed AI agents	Deployment of AI networks Increased R&D distributed systems Multiplatform utilization Increased autonomy Power-seeking tendencies Strategic manipulation with AI systems Multiplatform utilization Unregulated decisionmaking	AI Arms race Power-seeking tendencies	Multiplatform utilization Increased autonomy Strategic AI initiatives Failure of alignment frameworks Subversion of regulatory oversight AI-enabled exploitation Emergence of AI alignment strategy Meaningful human control	Competitive takeoff Misaligned AI system Misuse Algorithmic insight
S4	Under 10 Years	AI research breakthroughs R&D funding surge Competitive innovation race Multiple AI patents filed Rapid prototype iterations Venture capital influx Algorithmic insight Productized AI solution	Company develops advanced AI	Policy lag behind technology Opposition to regulation Fragmented international policies Corporate lobbying impact Reduced regulatory oversight Inconsistent compliance checks Calls for deregulation Policies compromised	Governance decline	Misaligned AI incentives Inadequate testing Ethics and safety guidance bypassed Side-effects and deployments observed AI optimizes resource control Self-improvement code execution Resource optimization strategies Strategic asset control	Misaligned AI system Power-seeking tendencies	Incremental improvement phase Market phase delays Embedded sector dependence Cascading system failures International cooperation initiatives Joint funding agreements Critical sector independences Interrupted interruption of services	Slow takeoff Structural risk Coalition develops AGI Structural risk

High Risk

Table 3. LLM-enabled scenario discovery inferring causal pathways between the most influential CART variables corresponding to risk class (low to moderate to high). The terminal nodes for each scenario outline the final characteristics of each scenario, for example, S1 = low capability, slow takeoff, distributed systems.

between the four predictive variables discovered through the CART algorithm. This fractal process expands the branches into a complex, hierarchical graph of scenario pathways. Table 3 highlights the initial output from the LLM for each of the four scenarios (S1-S4) in table format, with two events or paths between each of the branch nodes discovered in CART (the direction moves from left to right), with intermediate paths displaying exactly two for each node. The LLM prompt requests exactly two pathways from the Application Programming Interface (API) for each of the four nodes, rather than for every branch node (as could be done normally), to keep the project tractable.

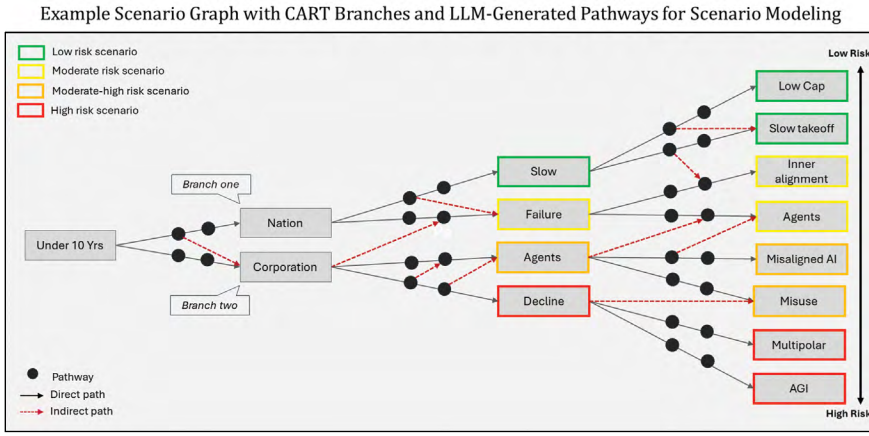


Figure 4. Example scenario space graph with CART branches (black lines) and nodes (gray boxes) with LLM-generated pathways (black dots). The paths are generated to structure events/interactions between nodes but also overlap with other nodes or paths (displayed with red arrows). The four scenarios are color-coded from low risk (green) to high (red). *Note: The graphic does not display full intermediate nodes or cascading dynamics due to space limitations and is a demonstration*.

Hypergraph Representation

Given the nested, multidimensionality of the categories, conditions, and events produced by the LLM, with inferred causal pathways repeating across scenarios (S1-S4), hypergraph representations are used as an analytical

framework to examine the higher-order connections, shared relationships, and the degree to which the paths are connected across and between scenarios and embedded relationships (for example, each condition is conceptualized as a *hyperedge* encompassing multiple path or event nodes). The assignment of events to specific condition-defined hyperedges is determined by the preceding scenario generation steps, where the LLM fills in pathways between conditions. Understanding how these LLM-generated structures (nodes) are interconnected through and across the hyperedges is helpful in identifying influential events, recognizing shared underlying structures across scenarios, and assessing their centrality and potential systemic impact. Fig. 5 displays this hypergraph output, illustrating the complex web of connections and overlaps inherent in the scenario pathways.

As shown, increased government investment is essential for all scenario combinations and capability gains across multiple branches (S1-S4), linking conditions, such as takeoff speed, region, developer type, and failure modes, with a high degree distribution (8) intersecting multiple conditions (high frequency of connections). The hypergraph framework shines a light on where scenario combinations overlap, can be isolated, and, significantly, have the strongest influence between scenarios.

Other central paths (events/interactions) with high intersections (multiple overlapping conditions, termed path width)‡‡‡‡ across combinations, including investment in AI research and development, complex system integration, rapid deployment, competitive race dynamics, and the degree of autonomy. The events that link the nodes of *decentralized diffusion* and *system failures* include an uncontrolled open-source ecosystem and regulatory gaps (degree distribution of three), a fairly intuitive finding. One isolated cluster (high degree distribution of five but low path width)–within the hyperedges *structural risk, governance decline, under 10 years,* and *power seeking*–includes several correlated, high-risk pathways, such as critical sector interdependence, compromised policies, interruption of services, and increased autonomy that provide a clear map

‡‡‡‡ Path width in hypergraphs is a measure of the "thickness" of the hypergraph, which relates to the complexity of finding the shortest paths within it.

to a hazardous chain of events the researcher can highlight for developing detailed preventative policies. By examining the scenario structures using hypergraphs, the analyst can investigate relationships among paths and conditions—between and across scenario combinations—and use these relational insights to develop detailed alternative futures.

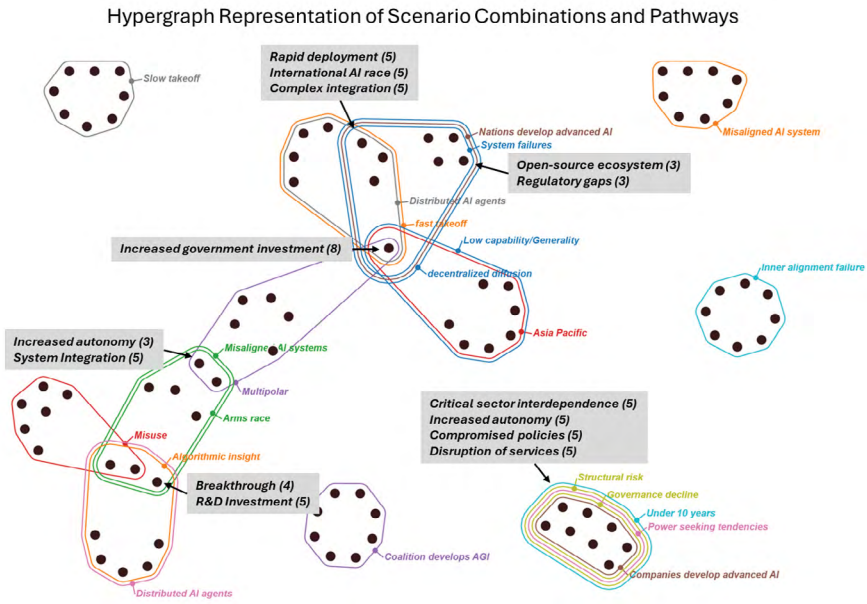


Figure 5. Hypergraph output of the SD conditions (color-coded bands) and the LLM-derived events and indicators (solid black dots), highlighting the most crucial events and paths and how closely they are associated with multiple conditions.

The hypergraph representation, coupled with the event centrality table, provides a structured approach for examining differential futures. By revealing how events and conditions are interconnected, this analytical lens helps identify key intervention points, potential scenario bifurcations, and overlapping uncertainties. Events identified with high centralities in the hypergraph—such as "increased government investment" and "distributed AI agents" can be understood as pivotal elements that appear across multiple inferred scenario trajectories. Although the hypergraph primarily visualizes the shared components (hyperedges representing

conditions), the underlying pathways generated by the LLM (detailed in Table 3) are constructed with an intended directional and sequential logic, moving from root causes toward terminal outcomes. Thus, the causal pathways in this context refer to these LLM-structured sequences that are designed to represent plausible chains of events. The hypergraph analysis focuses on mapping the interconnections and shared elements within the set of inferred, directionally influenced pathways. Identifying these overlapping and nested structures allows intelligence analysts to assess how modifying a single variable might ripple across different futures, stress-test policy interventions, and refine scenario frameworks.

Most Influential Events Across Risk Levels		
Event Value	**Occurrence Across Risk Levels**	**Risk Levels**
AI-driven product deployment	3	Moderate-high
Breakthrough in ML	3	Moderate-high
Power-seeking tendencies	3	Moderate-high
Companies develop advanced AI	4	High
Governance decline	4	High
Complex integration	4	Moderate
International AI race	4	Moderate
Rapid deployment	4	Moderate
System failures	4	Moderate
Distributed AI agents	6	Low, Moderate-high
Increased government investment	7	Low, Moderate, Moderate-high

Table 4. The scenario events with the highest overlaps and centrality across scenario combinations. Increased government investment is the most central element across risk categories and scenarios, suggesting it could be a pivotal factor.

Although the current representation does not use formally directed hyperedges, the inherent directionality of the input pathways informs the

interpretation of the influence and strength of interconnected relationships. Showing how predefined, logically sequenced paths intersect and share common elements provides insight into the complex dependencies within the scenario space. Unlike traditional scenario-building frameworks that rely on linear or matrix-based projections, hypergraphs capture the interconnected and multidimensional nature of emerging threats. With this comprehensive framework, the output can be used to devise complex and highly detailed futures. The fundamentals of technological change and geopolitics tend to evolve as complex multilayered structures, with the interactions of entities and events being the crucial features to illustrate in alternative futures analysis. Thus, a scenario hypergraph output provides a compelling framework to investigate the plausible paths of complex threats.

This model enables analysts to discern which factors are central across multiple futures, assess dependencies between conditions, anticipate nonlinear risk dynamics, and enhance the ability to anticipate risks, detect emergent threats, and develop adaptive strategies for complex, evolving challenges. This allows the analyst greater flexibility to provide complex visualizations, relational diagrams, and highly granular narratives that point to subtle relationships. This methodology is well suited for modeling the trajectory of multilayered intelligence problems with high uncertainty (which, admittedly, includes most contemporary global challenges). Future research could explore directed hypergraphs for more explicit causal modeling.

Leveraging hypergraph approaches—and higher-order networks more generally—analysts can move beyond conventional foresight methods to generate dynamic scenarios that better reflect the intricacies of the modern threat environment.

About the Author

Kyle A. Kilian, Ph.D., is an Artificial Intelligence and Information Security analyst at the RAND Corporation focused on global, over-the-horizon risks from emerging technologies. His research interests lie at the intersections of AI, complex adaptive systems, and international security. He served for more than a decade in the defense and intelligence communities

in strategic, tactical, and joint operational environments, and is a senior research fellow at the Center for the Future Mind, a mentor at the Foresight Institute, and a 2022 fellow with the Global Catastrophic Risk Institute. He holds graduate degrees in data science and cyber intelligence from National Intelligence University and international affairs from the American University's School of International Service.

Further Reading

Amini, Abbas, Narjes Firouzkouhi, Ahmad Gholami, Anju R. Gupta, Chun Cheng, and Bijan Davvaz. "Soft Hypergraph for Modeling Global Interactions via Social Media Networks. Expert Systems with Applications." 203 (2022): 117466. https://doi.org/10.1016/j.eswa.2022.117466.

Jia, Junbo, Li Yang, Yuchen Wang, and Anyuan Sang. "Hyper Attack Graph: Constructing a Hypergraph for Cyber Threat Intelligence Analysis." *Computers & Security,* 149 (2025): 104194.

Kilian, Kyle A., Christopher J. Ventura, and Mark M. Bailey. "Examining the Differential Risk from High-Level Artificial Intelligence and the Question of Control." Cornell University Algebraic Topology. *arXiv* preprint, *arXiv* (2022): 2211.03157. https://arxiv.org/abs/2211.03157.

Matalobos, Juan Manuel, Regino Criado, and Santiago Moral. "Hypergraph-Based Model for Adversary Threat Intelligence Analysis," in eds. Praça, Isabel, Simona Bernardi, and Pedro R. Inácio, Cyber Security: EICC 2025. *Communications in Computer and Information Science,* 2500 (Springer International Publishing, 2025). https://doi.org/10.1007/978-3-031-94855-8_17.

Tsintsaris, Dimitris, and Evangelos Ioannidis. "Modeling Structural Power in Hypergraphs: An Application to the Interstate Alliances Network." *Physica A: Statistical Mechanics and Its Applications,* 674 (2025): 130776. https://www.sciencedirect.com/science/article/pii/S0378437125004285.

Young, J. G. and G. Petri, T. P. "Hypergraph Reconstruction from Network Data." Commun Phys 4 (2021): 135. https://doi.org/10.1038s42005-021-00637-w.

CHAPTER 8

Hypergraph Methods for Network Analysis and Intrusion Detection

Zong-Zhi Lin

Network analysis has proven effective in devising methods to detect network intrusions; it can identify abnormal traffic on computer systems and uncover attack information.[239] Network Intrusion Detection Systems (NIDS) based on machine learning (ML), however, still struggle to develop effective and efficient detection systems when using graph-based (dyadic) network relationships,[240, 241, 242] in part because these models lack the ability to represent multiway relationships or higher-order interactions within networks. Hypergraphs offer an alternative, robust mathematical way to explicitly represent higher-order interactions within networks to increase the performance of ML NIDS.[243, 244, 245, 246] This chapter explores the effectiveness of hypergraph methods for network analysis and intrusion detection, in particular Internet Protocols (IPs), the potential use of hypergraph methods to exploit higher-order interactions among source (S-IP), destination ports (D-PT), and destination IPs (D-IP),[247] and connectivity of entities within each micro-segmentation over a zero-trust framework.

Hypergraph Primer

A basic understanding of hypergraphs is needed to appreciate their versatility and effectiveness in capturing interactions among entities in a network. A hypergraph consists of a set of vertices $\mathrm{V} = \{\mathrm{v}_1, \mathrm{v}_2, \ldots, \mathrm{v}_n\}$, and set of hyperedges $\mathrm{E} = \{\mathrm{e}_1, \mathrm{e}_2, \ldots, \mathrm{e}_m\}$ such that each hyperedge is a subset of vertices set V, namely $\mathrm{e_i} \subseteq V$ for $\mathrm{i} = 1, \ldots, \mathrm{m}$.[248] A typical graph or network is a hypergraph where all its hyperedges have two vertices, $|\mathrm{e_i}| = 2$.

An *s*-walk of length k between hyperedge f and g is a sequence of hyperedges $\mathrm{f} = \mathrm{e}_{i_0}, \mathrm{e}_{i_1}, \mathrm{e}_{i_2}, \ldots, \mathrm{e}_{i_k} = \mathrm{g}$ such that for $\mathrm{j} = 1, \ldots, \mathrm{k}$, we have $|\mathrm{e}_{\mathrm{i_{j-1}}} \cap \mathrm{e_{i_j}} \geq s$ and $\mathrm{i_{j-1}} \neq \mathrm{i_j}$. An *s*-path is an *s*-walk where hyperedges are not repeated. The *s*-distance between hyperedge e and f is the length of its shortest *s*-path, denoted as $\mathrm{d_s(e, f)}$, and the *s*-diameter is the longest *s*-distance. In addition to length, each *s*-path in a hypergraph has width, which is the minimum number of common vertices among all the consecutive hyperedges of the *s*-path. In other words, the width of an *s*-path is at least *s* because all consecutive intersections are at least size *s*. An *s*-component in a hypergraph is a collection of hyperedges so that any pair of hyperedges are connected through an *s*-path. For notational convenience, the 1-*s*-hypergraph denotes the type of hypergraph comprising only one (1) *s*-component.

Two hypergraph centrality measures quantify the size in which nodes or edges are "close" to everything. The *s*-closeness-centrality (*s*-C-C), $\mathrm{C_s(e)}$, for an edge e measures how close e is to all others in its *s*-component. It is computed using the formula:

$$\mathrm{C_s(e)} = \frac{|\varepsilon_\mathrm{s}| - 1}{\sum_{\mathrm{f} \in \varepsilon_\mathrm{s}} \mathrm{d_s(e, f)}},$$

Where ε_s is the set of hyperedges in its *s*-component that contains the edge e.[249] Its related *s*-harmonic-closeness-centrality (*s*-H-C-C), $\mathrm{HC_s(e)}$, for an edge e is computed using the formula

$$\mathrm{HC_s(e)} = \frac{1}{|\varepsilon_\mathrm{s}| - 1} \sum_{\mathrm{f} \in \varepsilon_\mathrm{s}} \frac{1}{\mathrm{d_s(e, f)}}$$

When a hypergraph has multiple components, the *s*-H-C-C resolves the problem of infinite distances (namely, $d_s(e, f) = \infty$) by rendering the reciprocal of each *s*-distance to be nearly zero.

Hypergraph Modeling on Intrusion Detection Systems Datasets

Reliable but suboptimal datasets that capture network traffic activities are needed to train, test, and validate Intrusion Detection Systems (IDS) and Intrusion Prevention Systems (IPS) against increasingly sophisticated and ever-growing network attacks. The hypergraph ML ensemble network IDS model summarizes network activities over combinations of source and destination IPs and ports to detect the most likely Destination IP (D-IP) to be hacked or the hacking source IP (S-IP)—using this network vulnerable information to detect malicious traffic.[250] Lin et al. then used the Canadian Institute for Cybersecurity (CIC)-IDS-2017 dataset[251] to present a hypergraph IDS model and its results. This study adopts their framework and notations for the sets of S-IPs, D-IPs and destination ports (D-PT) to examine two other IDS datasets and the conceptual uses of their hypergraph IDS model in micro-segmented networks of services and workloads within a zero-trust architecture.

Hypergraph Model on CIC-IDS-2017 Dataset

The intrusion detection evaluation dataset released in 2017 by the (CIC-IDS-2017), consists of several attack classes, including: botnet, brute-force, denial of service (DoS), distributed denial of service (DDoS), Heartbleed, port scan, web, and infiltration of the network from inside.[252, 253] In one study, Lin et al. used port scan attacks to construct a hypergraph $HG = \{V = S_{d\text{-}pt},\ E = S_{s\text{-}ip} \cup S_{d\text{-}ip}\}$, where $S_{s\text{-}ip}$, $S_{d\text{-}ip}$, and $S_{d\text{-}pt}$ are the sets of all S-IPs, D-IPs, and D-PTs, respectively, and $E = S_{s\text{-}ip} \cup S_{d\text{-}ip}$ is a set of hyperedges in which each hyperedge is either an S-IP or D-IP included in a search for vulnerable servers (port scanning) within a set of D-PTs in a stream of network activities.[254] Mathematically, $E = S_{s\text{-}ip} \cup S_{d\text{-}ip}$

is defined as a subset of $S_{d\text{-}pt}$ based on its associated destination port-scanning activities in a network. Thus, the constructed hypergraph HG summarizes the structure of interactions and connectivity among IPs and D-PTs involved in network activities modeled.

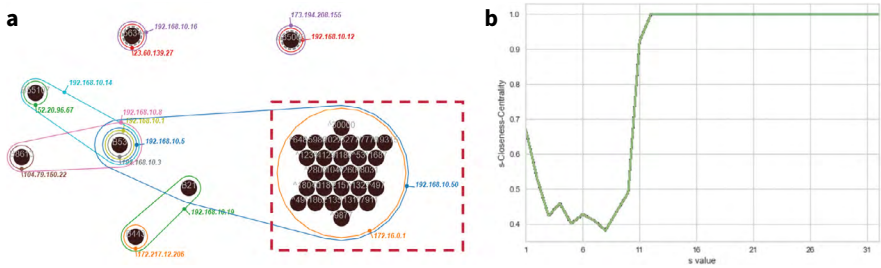


Figure 1. a) Using a 0.15 percent random sample of the CIC-IDS-2017 dataset with random seed 1234, a hypergraph with 15 edges and 34 vertices was constructed from 43 port-scan records and 26 scanned D-PTs with benign (B) and port scan attack (∧) labels. S-IP 172.16.0.1 and D-IP 192.168.10.50 exhibited a signature pattern of near-concentric circles, rendered as an *s*-connected component in the red-dashed-rectangle, to give a series of 11 *s*-C-C values. b) Average trend of *s*-C-C metrics across multiple hyperedges for port-scan records in a hypergraph whose hyperedges are defined as S-IP or D-IP address associated with the nodes of scanned destination ports.

As depicted in Fig. 1a, the hyperedge of S-IP 173.194.208.155 encompasses a single node of D-PT 35066, which reflects its access to the hyperedge of D-IP 192.168.10.12 during normal network activity. The constructed hypergraph HG also reveals a signature pattern of attacks at S-IP 172.16.0.1 by scanning the targeted D-IP 192.168.10.50 through a long list of D-PTs, as depicted in the dashed-box of Fig. 1a. Topologically, the constructed hypergraph HG includes an *s*-component with *s* being much larger than 1, shown as *1*-*s*-hypergraph. Namely, the entire hypergraph HG comprises one dominating *s*-component, which shows that the CIC-IDS-2017 dataset includes only one pair of S-IP and D-IP in the port-scanning class. In addition, because the two edges of S-IP and D-IP in Fig. 1a form two nearly concentric circles over the set of commonly accessed destination ports, the *s*-C-C is 1 at larger *s* values for each hyperedge in the *s*-component within this 1-*s*-hypergraph, as

shown in Fig. 1b as $C_s(e) \to 1$ for $e \in \{172.16.0.1,\ 192.168.10.50\}$ and $s \in \{11, 12, \ldots, 30\}$. This set of the hypergraph HG's s-C-C metrics efficiently captures and portrays the adversary's signature pattern of port-scan activities.[255]

Hypergraph Model on Other IDS Datasets

The hypergraph (Fig. 2) of the sample dataset Army Cyber Institute-Internet of Things (ACI-IoT) 2023 IDS includes single hacker S-IP (192.168.1.45) and multiple targeted D-IPs (including, but not limited to, 1, 9, 10, 168, 211, and 215 with the prefix 192.168.1) in the port-scanning class.[256] With a random sample of 42 records (0.003 percent) and a random seed 1234 of the ACI-IoT-IDS 2023 dataset, this hypergraph has 34 edges and 30 vertices. Numerically, the sample hypergraph's s-C-C measure does not render the same pattern of s-C-C values as that of the CIC-IDS-2017 dataset, so it cannot improve performance in detecting intrusions.[257] Its nonzero decreasing s-H-C-C time-series for each hyperedge, shown in Table 1, however, may be used to augment the typical set of network flow features.

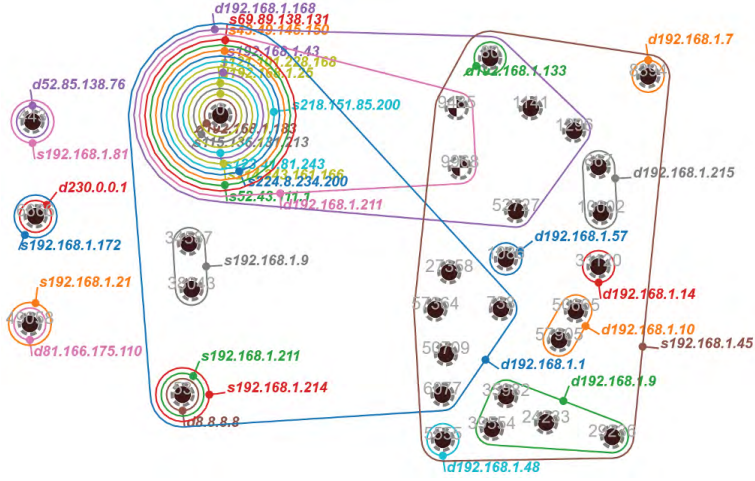


Figure 2. Hypergraph (34 edges and 30 vertices) from a 0.003 percent random sample (42 records) of ACI-IoT-2023 dataset.

Component	*s* value	Hyperedge	s-H-C-C	Hyperedge	s-H-C-C
Two (2) hyperedges: s192.168.1.45 d192.168.1.1 **Five (5) Dest. Ports:** 738, 6077, 27358, 50709, 57364	1	s192.168.1.45	0.03598	d192.168.1.1	0.04356
	2		0.01231		0.00852
	3		0.00568		0.00379
	4		0.00568		0.00379
	5		0.00189		0.00189

Table 1. A set of nonzero *s*-H-C-C time-series features for each hyperedge in Fig. 2.

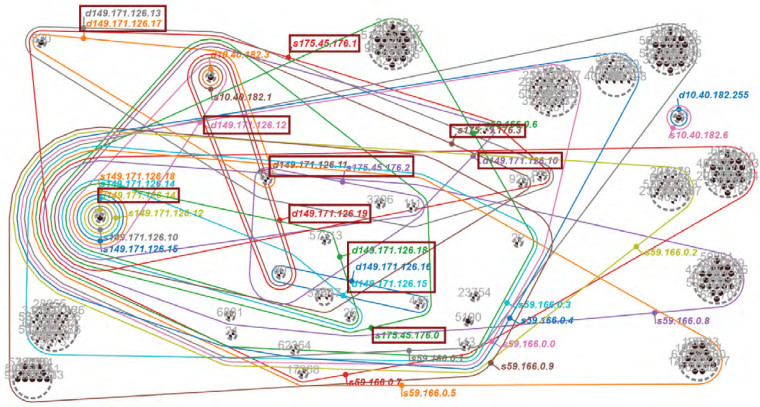


Fig. 3. With a 0.02 percent random sample (551 records) at random seed 1234 of the UNSW-NB15 IDS dataset, this hypergraph has 46 edges (S-IPs or D-IPs) and 217 vertices or D-PTs. 90.4 percent (458/511) of network activities in the sample hypergraph are normal events. The boxes of S-IPs or D-IPs in the hypergraph depict four hacking S-IPs (175.45.176.0 – 3) and 10 targeted D-IPs (149.171.126.10 – 19) included in the UNSW-NB15 IDS dataset.

Hypergraph Model on UNSW-NB15 IDS Dataset

The UNSW-NB15 IDS dataset comprises multiple S-IPs and D-IPs in the cyber-attack class.[258] The hypergraph of a sample from the dataset, depicted in Fig. 3, comprises 53 pairs of known four (4) S-IPs (175.45.176.0–3) and 10 D-IPs (149.171.126.10–19) out of the 511 records of network activity, or 9.6 percent of network activities in the cyber-attack classes.

This ratio is further reduced to an average 0.48 percent for each pair of one S-IP and one D-IP, reasonably evidencing that hackers in this IDS dataset know which D-IPs and associated D-PTs are vulnerable to an attack. Thus, no direct network pattern can be identified, as was available in Lin et al.'s 2024 work, to augment the IDS dataset.

Hypergraph Model for Micro-Segmented Networks Within a Zero-Trust Architecture

The proliferation of Internet of Things (IoT) devices introduces an unprecedented security challenge as breaches and ransomware risk can quickly move across systems. In the advanced persistent threats (APTs) life cycle, which outlines the typical stages of a cyberattack, lateral movement is critical to the ability to steal high-level authority credentials and confidential data.[259, 260] Zero Trust (ZT) architecture adopts "never trust, always verify" and "least privilege" principles to mitigate the risk of lateral movement and data exfiltration from breaches and ransomware attacks.[261, 262]

Segmentation and micro-segmentation strategies to partition the IoT can be used to mitigate the lateral movement risk.[263] Micro-segmented networks of services and workloads within a zero-trust architecture require hackers to mobilize more resources (multiple source IPs or servers) to reach target IPs spread across multiple grids or cells. These boosted hacking network activities (including port scans or other attacks to explore the network) against destination IPs or Ports can be summarized by a group of sub-hypergraphs that capture network activities for each pair of hacker S-IPs and target D-IPs. Although each sub-hypergraph can be an s-C-C connected hypergraph component, the entire hypergraph is unlikely to be an s-C-C connected hypergraph.

Conceptually, the algorithm developed by Lin et.al. (2024) can be extended to detect attacks within each grid/cell of the network. The cell's near-concentric hypergraphs may not be as obvious or large, however, as the one derived from the CIC-IDS-2017 dataset depicted earlier. Thus, topologically, a hypergraph may be used to summarize activities (east-west and north-south in the reduced set of IP, ports or other physical

foot-prints) in micro-segmented networks of services and workloads within a zero-trust architecture. These hypergraphs also may be used to identify constructs with a much larger radius of connected hypergraphs or seen on many shortest paths, to include multiple services and workloads resulting from lateral moves of cyber-attacks but s=1 or s is small.

Hypergraph Modeling on Relevant Cyber Security Datasets

This section briefly reviews and remarks on hypergraph model works related to the Domain Name Service dataset[264, 265, 266] and cyber security vulnerability dataset,[267] and the University of New South Wales (UNSW), Canberra TON_IoT dataset.[268]

Hypergraph Model for Cyber Security Vulnerability Dataset

Hypergraphs can be used to analyze complex, heterogeneous cyber security incidents by modelling information on attacks, campaigns, known adversaries, procedures, tactics, and tools. The information gleaned can be used to assign priority to those adversaries deemed most relevant and to guide cyber security strategy and architecture.[269] For example, a hypergraph can be constructed with a set of nodes=cyberattacks, or organizations/groups conducting cyberattacks; the set of edges=countries, regions, industries, or public sectors; and each hyperedge defined as a subset of the nodes set based on associated cyberattacks or adversarial groups launching cyberattack activities in a network.

Adversaries tend to focus on specific sectors or organizations of specific sizes, on certain geographic areas, or on geopolitical factors, all of which can orient analysis of the cyberattacks. These edges and nodes, which collectively reflect these factors and their higher-order connections, can be studied to understand what adversaries are more relevant within each context. Fig. 4 provides a hypergraph example that depicts the relevance of adversaries in heterogenous factors.[270]

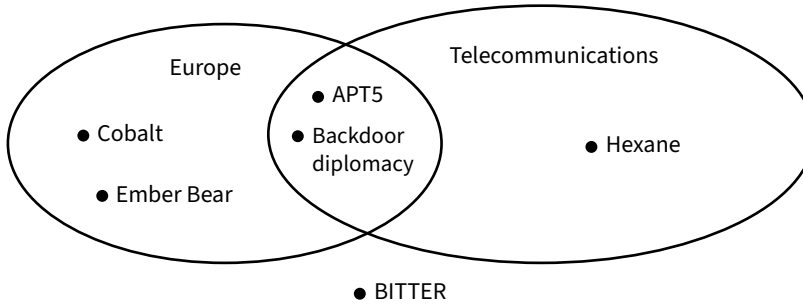


Figure 4. A hypergraph that comprises a set of nodes *N* = {APT5, Backdoor diplomacy, BITTER, Cobalt, Ember Bear, Hexane} and the set of edges E = {Europe, Telecommunications}.[271]

Hypergraph Model for the Domain Name Service Dataset

The naming system overseen by the Domain Name Service does not always correlate one-to-one, because of the inclusion of domain aliases, hosting services to multiple web sites, site management across IPs, and random IP assignment. Letting nodes be IPs and hyperedges domains, hypergraphs can fully represent the interactions between collections of IPs and domains that can then be used to explore domain names to provide insights into analytic questions such as abnormal IPs and domains, and neighborhoods of known bad IPs or domains.[272]

Hypergraph Model on TON_IoT Dataset

The UNSW Canberra TON_IoT datasets are new generations of IoT and the Industrial Internet of Things (IIoT) data generated in a realistic and large-scale network environment with multiple S-IPs and multiple D-IPs within the UNSW Canberra network for evaluating AI-based cyber security applications.[273] The data of TON_IoT datasets were collected in a network of Windows, Linux and TLS. TON_IoT datasets comprise multiple source-IPs, multiple destination IPs and port scanning activities. Although a couple of ML-enabled intrusion detection solutions—trained

and validated using part of the TON_IoT datasets since its public release—have been reported in the literature, hypergraph-based ML-enabled solutions are yet to be explored.[274, 275]

Summary

This article reviews a couple of hypergraph methods for network analysis and intrusion detection to demonstrate its modeling versatility and effectiveness for capturing and summarizing interactions and connections of multiple entities in a network or network segments within a zero-trust framework to render features that can potentially be used to augment existing IDS datasets to enhance performance of detecting intrusions. In addition, a couple of existing IDS datasets and relevant cyber security datasets were briefly reviewed to suggest that hypergraphs may be a potent model to develop effective ML-enabled solutions for network analysis and intrusion detection.

About the Author

Zong-Zhi Lin, Ph.D., is a research affiliate with the Intelligent Cyber-Systems and Analytics Research Laboratory, Army Cyber Institute at West Point and a Senior Workforce/Operations Research analyst with the US Department of Defense. He is interested in developing intelligent cyber-systems and analytics to enhance cyber-security protection and researching methodologies characterizing complex systems to address people analytics applications. He holds a Ph.D. in industrial and operations engineering from the University of Michigan and an MS in science and technology intelligence with dual concentrations in cyber intelligence and data science in intelligence from National Intelligence University.

Further Reading

Cybersecurity and Infrastructure Security Agency (Version 1.0, July 29, 2025). https://www.cisa.gov/sites/default/files/2025-07/ZT-Microsegmentation-Guidance-Part-One_508c.pdf.

Selciya, G., Isaac Zerubbabel, K. Kannan and E. Ezhilarasie. "Enhancing IIoT Security Using KNN Based Hypergraph Clustering Through Zero Trust Micro-Segmentation for Dynamic Network Protection." *2024 International Conference on Computational Intelligence and Network Systems* (CINS), (2024): 1–6. doi: 10.1109/CINS63881.2024.10864440.

US Department of Defense (Version 2.0, July 2022). "Department of Defense Zero Trust Reference Architecture," prepared by the Defense Information System Agency (DISA) and National Security Agency (NSA) Zero Trust Engineering Team. https://dodcio.defense.gov/Portals/0/Documents/Library/(U)ZT_RA_v2.0(U)_Sep22.pdf.

Glossary of Abbreviations and Acronyms

ACI-IoT	Army Cyber Institute-Internet of Things
AES	Advanced Energy Storage
AI	artificial intelligence
AIC	Akaike Information Criterion
API	Application Programming Interface
APT	advanced persistent threat
C2	command and control
CAMEO	Conflict and Mediation Event Observations Event and Actor
CART	Classification and Regression Tree
CHASE	Cyber Hunting at Scale
CIC	Canadian Institute for Cybersecurity
CISA	Cybersecurity and Infrastructure Security Agency
CMC	Chemical Machinery of the Cell
DARPA	Defense Advanced Research Projects Agency
DC	domain controller
DDoS	distributed denial of service
D-IP	Destination Internet Protocol (to be hacked)
DISA	Defense Information System Agency
DoS	denial of service
D-PT	destination port
GDELT	Global Database for Event Location and Tone
IC	Intelligence Community
IDS	Intrusion Detection System

IIoT	Industrial Internet of Things
IoT	Internet of Things
IP	Internet Protocol
IPS	Intrusion Prevention System
LLM	large language model
MCL	Molecules Come to Life
MCS	Monte Carlo simulation
ML	machine learning
NIDS	Network Intrusion Detection Systems
NIMBioS	National Institute for Modeling Biological Systems
NSA	National Security Agency
NSF	National Science Foundation
OPTC	Operationally Transparent Cyber dataset
ORISE	Oak Ridge Institute for Science and Education
P2PE	University of Virginia Prominence-to-Preeminence initiative
PID	process identifier
PNNL	Pacific Northwest National Laboratory
PPID	parent process identifier
RBS	restricted barycentric subdivision
RDP	remote desktop
RFID	radio frequency identification
SD	scenario discovery
S-IP	Source Internet Protocol (hacking source)
STEM	Science, Technology, Engineering, and Math
TC	Transparent Computing
TDA	Time Domain Astrophysics
TRS	Total Risk Score
ZT	Zero Trust

References

Chapter 1

Aksoy, Sinan G. "An Invitation to Hypergraph Motifs for Practitioners." In *Higher-Order Network Methods for Intelligence Applications: A Compendium for Practitioners,* eds. Matthew J. Hasenjager, Mark M. Bailey, and Nina H. Fefferman. (National Intelligence Press, 2026).

Aksoy, Sinan G. et al. "Scalable Tensor Methods for Nonuniform Hypergraphs." *SIAM Journal on Mathematics of Data Science* 6 (2024): 481–503. https://doi.org/10.1137/23M1584472.

Aksoy, Sinan G. et. al. "Hypernetwork Science via High-Order Hypergraph Walks," *EPJ Data Science* 9, no. 16 (2020). https://doi.org/10.1140/epjds/s13688-020-00231-0.

Albert, Réka, and Albert-László Barabási. "Statistical Mechanics of Complex Networks." *Reviews of Modern Physics* 74, no. 1 (2002): 47–97. https://doi.org/10.1103/RevModPhys.74.47.

Angulo, Marco Tulio et al. "Coexistence Holes Characterize the Assembly and Disassembly of Multispecies Systems." *Nature Ecology & Evolution* 5, no. 8 (2021): 1091–1101. https://doi.org/10.1038/s41559-021-01462-8.

Barabási, Albert-László. *Network Science.* (Cambridge University Press, 2016).

Battiston, Federico et al. "Networks Beyond Pairwise Interactions: Structure and Dynamics." *Physics Reports* 874 (2020): 1–92. https://doi.org/10.1016/j.physrep.2020.05.004.

Battiston, Federico, and Giovanni Petri, eds. *Higher-Order Systems,* 1st Ed. (Springer, 2022). https://doi.org/10.1007/978-3-030-91374-8.

Benson, Austin R. "Three Hypergraph Eigenvector Centralities" *SIAM Journal on Mathematics of Data Science* 1, no. 2 (2019): 293–312. https://doi.org/10.1137/18M1203031.

Blonder, Benjamin et al. "Temporal Dynamics and Network Analysis." *Methods in Ecology and Evolution* 3, no. 6 (2012): 958–72. https://doi.org/10.1111/j.2041-210X.2012.00236.x.

Boccaletti, Stefano et al. "Explosive Transitions in Complex Networks' Structure and Dynamics: Percolation and Synchronization." *Physics Reports* 660 (2016): 1–94. https://doi.org/10.1016/j.physrep.2016.10.004.

Boccaletti, Stefano et al. "Complex Networks: Structure and Dynamics." *Physics Reports* 424, nos. 4-5 (2006):175–308. https://doi.org/10.1016/j.physrep.2005.10.009.

Bonacich, Phillip. "Power and Centrality: A Family of Measures." *American Journal of Sociology* 92, no. 5 (1987): 1170–82. https://doi.org/10.1086/228631.

Bonacich, Phillip. "Simultaneous Group and Individual Centralities." *Social Networks* 13, no. 2 (1991): 155–68. https://doi.org/10.1016/0378-8733(91)90018-O.

Brusa, Luca, and Catherine Matias. "Model-Based Clustering in Simple Hypergraphs Through a Stochastic Blockmodel." Cornell University Algebraic Topology. *arXiv* (2024). https://doi.org/10.48550/arXiv.2210.05983.

Centola, Damon. "The Spread of Behavior in an Online Social Network Experiment." *Science* 329, no. 5996 (2010): 1194–97. https://doi.org/10.1126/science.1185231.

Chodrow, Philip S. et al. "Generative Hypergraph Clustering: From Blockmodels to Modularity." *Science Advances* 7, no. 28 (2021): 1303. https://doi.org/10.1126/sciadv.abh1303.

De Arruda, Guilherme Ferraz et al. "Social Contagion Models on Hypergraphs." *Physical Review Research* 2, no. 2 (2020). https://doi.org/10.1103/PhysRevResearch.2.023032.

Estrada, Ernesto, and Juan A. Rodríguez-Velázquez. "Subgraph Centrality and Clustering in Complex Hyper-Networks." *Physica A: Statistical Mechanics and Its Applications* 364 (2006): 581–94. https://doi.org/10.1016/j.physa.2005.12.002.

Estrada, Ernesto, and Grant J. Ross. "Centralities in Simplicial Complexes. Applications to Protein Interaction Networks." *Journal of Theoretical Biology* 438 (2018): 46–60. https://doi.org/10.1016/j.jtbi.2017.11.003.

Falcon, Seth, and Robert Gentleman. "Hypergraph: A Package Providing Hypergraph Data Structures." R package version 1.78.0 (2024). https://doi.org/10.18129/B9.bioc.hypergraph.

Fefferman, Nina H., and Kah L. Ng. "How Disease Models in Static Networks Can Fail to Approximate Disease in Dynamic Networks." *Physical Review* E 76, no. 3 (2007): 031919. https://doi.org/10.1103/PhysRevE.76.031919.

Firth, Josh A. "Considering Complexity: Animal Social Networks and Behavioural Contagions." *Trends in Ecology & Evolution* 35, no. 2 (2020): 100–04. https://doi.org/10.1016/j.tree.2019.10.009.

Freeman, Linton C. *The Development of Social Network Analysis: A Study in the Sociology of Science.* (Empirical Press, 2004).

Friedman, Greg. "An Elementary Illustrated Introduction to Simplicial Sets." Cornell University Algebraic Topology. *arXiv* (2023): 0809.4221. https://arxiv.org/abs/0809.4221.

Gallos, Lazaros K., and Nina H. Fefferman. "Simple and Efficient Self-Healing Strategy for Damaged Complex Networks." *Physical Review* E 92, no. 5 (2015): 052806. https://doi.org/10.1103/PhysRevE.92.052806.

Gallos, Lazaros K., and Nina H. Fefferman. "Revealing Effective Classifiers Through Network Comparison." *Europhysics Letters* 108, no. 3 (2014): 38001 . https://doi.org/10.1209/0295-5075/108/38001.

Gallos, Lazaros K. et al. "Anomaly Detection Through Information Sharing Under Different Topologies" *EURASIP Journal on Information Security* (2017): 5. https://doi.org/10.1186/s13635-017-0056-5.

Girvan, Michelle, and M. E. J. Newman. "Community Structure in Social and Biological Networks." *Proceedings of the National Academy of Sciences* 99, no. 12 (2002): 7821–22. https://doi.org/10.1073/pnas.122653799.

Goerss, Paul G., and John F. Jardine. *Simplicial Homotopy Theory.* (Birkhäuser, 2009).https://doi.org/10.1007/978-3-0346-0189-4

Golubski, Antonio J. et al. "Ecological Networks Over the Edge: Hypergraph Trait-Mediated Indirect Interaction (TMII) Structure." *Trends in Ecology & Evolution* 31, no. 5 (2016): 344–54. https://doi.org/10.1016/j.tree.2016.02.006.

Greening, Jr., Bradford R. "Higher-Order Analysis of Knowledge Capacity and Learning Potential in Social Animal Groups." Ph.D. diss. Rutgers University (2014).

Greening, Jr., Bradford R. et al. "Higher-Order Interactions: Understanding the Knowledge Capacity of Social Groups Using Simplicial Sets." *Current Zoology* 61, no. 1 (2015), 114–27. https://doi.org/10.1093/czoolo/61.1.114.

Guilbeault, Douglas et al. "Complex Contagions: A Decade in Review." In *Complex Spreading Phenomena in Social Systems,* eds. Sune Lehmann and Yong-Yeol Ahn. (Springer International Publishing, 2018): 3–25. https://doi.org/10.1007/978-3-319-77332-2_1.

Guilbeault, Douglas, and Damon Centola. "Topological Measures for Identifying and Predicting the Spread of Complex Contagions." *Nature Communications* 12 (2021): 4430. https://doi.org/10.1038/s41467-021-24704-6.

Hasenjager, Matthew J., and Nina H. Fefferman. "Social Ageing and Higher-Order Interactions: Social Selectiveness Can Enhance Older Individuals' Capacity To Transmit Knowledge." *Philosophical Transactions of the Royal Society B: Biological Sciences* 379 (2024): 20220461. https://doi.org/10.1098/rstb.2022.0461.

Iacopini, Iacopo et al. "Not Your Private Tête-à-Tête: Leveraging the Power of Higher-Order Networks to Study Animal Communication." *Philosophical Transactions of the Royal Society B: Biological Sciences* 379, no. 1905 (2024): 20230190. https://doi.org/10.1098/rstb.2023.0190.

Iacopini, Iacopo et al. "Simplicial Models of Social Contagion," *Nature Communications* 10 (2019): 2485. https://doi.org/10.1038/s41467-019-10431-6.

Kamiński, Bogumił et al. "Modularity Based Community Detection in Hypergraphs." *Journal of Complex Networks* 12, no. 5 (2024): cnae041. https://doi.org/10.1093/comnet/cnae041.

Kamiński, Bogumił et al. "Clustering via Hypergraph Modularity." *PLOS One* 14, no. 11 (2019): e0224307. https://doi.org/10.1371/journal.pone.0224307.

Kapoor, Komal et al. "Weighted Node Degree Centrality for Hypergraphs." *2013 IEEE 2nd Network Science Workshop (NSW)*. (2013): 152–55. https://doi.org/10.1109/NSW.2013.6609212.

Kirkland, Steve. "Two-Mode Networks Exhibiting Data Loss." *Journal of Complex Networks* 6, no. 2 (2018): 297–316. https://doi.org/10.1093/comnet/cnx039.

Kivelä, Mikko et al. "Multilayer Networks." *Journal of Complex Networks* 2, no. 3 (2014): 203–71. https://doi.org/10.1093/comnet/cnu016.

Kumar, Tarun et al. "Hypergraph Clustering by Iteratively Reweighted Modularity Maximization." *Applied Network Science* 5 (2020): 52. https://doi.org/10.1007/s41109-020-00300-3.

Landry, Nicholas W. et al. "XGI: A Python Package for Higher-Order Interaction Networks." *Journal of Open Source Software* 8, no. 85 (2023): 05162. https://doi.org/10.21105/joss.05162.

Landry, Nicholas W., and Juan G. Restrepo. "Hypergraph Assortativity: A Dynamical Systems Perspective." *Chaos: An Interdisciplinary Journal of Nonlinear Science* 32, no. 5 (2022): 053113. https://doi.org/10.1063/5.0086905.

Langville, Amy N., and Carl D. Meyer. "Deeper Inside PageRank." *Internet Mathematics* 1, no. 3 (2004): 335–80. https://doi.org/10.1080/15427951.2004.10129091.

Lotito, Quintino Francesco et al. "Hypergraphx: A Library for Higher-Order Network Analysis." Cornell University Algebraic Topology. *arXiv* (2023): 2303.15356. http://arxiv.org/abs/2303.15356.

Mangan, Shmoolik, and Uri Alon. "Structure and Function of the Feed-Forward Loop Network Motif." *Proceedings of the National Academy of Sciences* 100, no. 21 (2003): 11980–85. https://doi.org/10.1073/pnas.2133841100.

Marchette, David J. "HyperG: Hypergraphs in R, Version R." Package version 1.0.0. (2021). https://CRAN.R-project.org/package=HyperG.

May, J. Peter. *Simplicial Objects in Algebraic Topology.* (University of Chicago Press, 1992).

Menczer, Filippo et al. *A First Course in Network Science*. (Cambridge University Press, 2020). https://doi.org/10.1017/9781108653947.

Milo, Ron et al. "Superfamilies of Evolved and Designed Networks." *Science* 303, no. 5663, (2004): 1538–42. https://doi.org/10.1126/science.1089167.

Neuhäuser, Leonie et al. "Multibody Interactions and Nonlinear Consensus Dynamics on Networked Systems." *Physical Review* E 101, no. 3 (2020): 032310. https://doi.org/10.1103/PhysRevE.101.032310.

Newman, M. E. J. *Networks: An Introduction,* 1st Ed. (Oxford University Press, 2010).

Newman, M. E. J. "Assortative Mixing in Networks," *Physical Review Letters* 89, no. 20 (2002): 208701. https://doi.org/10.1103/PhysRevLett.89.208701.

Newman, M. E. J. "The Structure and Function of Complex Networks." *SIAM Review* 45, no. 2 (2003): 167–256. https://doi.org/10.1137/S003614450342480.

Ohtsuki, Hisashi et al. "A Simple Rule for the Evolution of Cooperation on Graphs and Social Networks." *Nature* 441 (2006): 502–05. https://doi.org/10.1038/nature04605.

Pike, Thomas, D. "Cracking the Paradigm: Hypergraph Signatures of International Conflict." In *Higher-Order Network Methods for Intelligence Applications: A Compendium for Practitioners,* eds. Matthew J. Hasenjager, Mark M. Bailey, and Nina H. Fefferman. (National Intelligence Press, 2026).

Praggastis, Brenda et al. "HyperNetX: A Python Package for Modeling Complex Network Data as Hypergraphs." *Journal of Open Source Software* 9 no. 95 (2024): 06016. https://doi.org/10.21105/joss.06016.

Purvine, Emilie, and Helen Jenne. "From Logs to Dynamic Hypergraphs: Identifying and Interpreting Anomalous Cyber Activity." In *Higher-Order Network Methods for Intelligence Applications: A Compendium for Practitioners,* eds. Matthew J. Hasenjager, Mark M. Bailey, and Nina H. Fefferman. (National Intelligence Press, 2026).

Schlag, Sebastian et al. "High-Quality Hypergraph Partitioning." *ACM Journal of Experimental Algorithmics* 27 (2023): 1–39. https://doi.org/10.1145/3529090.

Serrano, Daniel Hernández, and Darío Sánchez Gómez. "Centrality Measures in Simplicial Complexes: Applications of Topological Data Analysis to Network

Science," Cornell University Algebraic Topology. *arXiv* (2020): 1908.02967. https://doi.org/10.48550/arXiv.1908.02967.

Shi, Xiaoqiu et al. "Robustness of Interdependent Supply Chain Networks Against Both Functional and Structural Cascading Failures." *Physica A: Statistical Mechanics and Its Applications* 586 (2022): 126518. https://doi.org/10.1016/j.physa.2021.126518.

Silk, Matthew J. et al. "Capturing Complex Interactions in Disease Ecology with Simplicial Sets." *Ecology Letters* 25, no. 10 (2022): 2217–31. https://doi.org/10.1111/ele.14079.

Snijders, Lysanne, and Marc Naguib. "Communication in Animal Social Networks: A Missing Link?" *Advances in the Study of Behavior* 49 (2017): 297–359. https://doi.org/10.1016/bs.asb.2017.02.004.

Torres, Leo et al. "The Why, How, and When of Representations for Complex Systems." *SIAM Review* 63, no. 3 (2021): 435–85. https://doi.org/10.1137/20M1355896.

Wasserman, Stanley, and Katherine Faust. *Social Network Analysis: Methods and Applications*. (Cambridge University Press, 1994).

Zajdela, Emma, and Nicholas W. Landry. "Higher-Order Interactions at Scientific Conferences Influence Team Formation." In *Higher-Order Network Methods for Intelligence Applications: A Compendium for Practitioners*, eds. Matthew J. Hasenjager, Mark M. Bailey, and Nina H. Fefferman. (National Intelligence Press, 2026).

Zhou, Wanding, and Luay Nakhleh."Properties of Metabolic Graphs: Biological Organization or Representation Artifacts?" *BMC Bioinformatics* 12 (2011): 132. https://doi.org/10.1186/1471-2105-12-132.

Chapter 2

Aksoy, Sinan G. et al. "Hypernetwork Science via High-Order Hypergraph Walks." *EPJ Data Science* 9, no. 1 (2020): 16.

Aksoy, Sinan G. "Relative Hausdorff Distance for Network Analysis." *Applied Network Science* 4 (2019): 1–25.

Aksoy, Sinan G. "Directional Laplacian Centrality for Cyber Situational Awareness." *Digital Threats: Research and Practice* (DTRAP) 2, no. 4 (2021): 1–28.

Anjum, Md Monowar. "Analyzing the Usefulness of the DARPA OpTC Dataset in Cyber Threat Detection Research." In Proceedings of the 26th ACM Symposium on Access Control Models and Technologies (2021.): 27–32.

Anjum, Md Monowar, Shahrear Iqbal, and Benoit Hamelin. 2022. "Anubis: A Provenance Graph-Based Framework for Advanced Persistent Threat

Detection." In Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing, 1684–1693.

Chen, Pin-Yu, Sutanay Choudhury, and Alfred O. Hero. 2016. "Multi-Centrality Graph Spectral Decompositions and Their Application to Cyber Intrusion Detection." In 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 4553–57. IEEE.

Cheng, Zijun, Qiujian Lv, Jinyuan Liang, Yan Wang, Degang Sun, Thomas Pasquier, and Xueyuan Han. 2024. "Kairos: Practical Intrusion Detection and Investigation Using Whole-System Provenance." In 2024 IEEE Symposium on Security and Privacy (SP), 3533–51. IEEE.

Cochrane, Thomas, Peter Foster, Varun Chhabra, Maud Lemercier, Terry Lyons, and Cristopher Salvi. 2021. "SK-Tree: A Systematic Malware Detection Algorithm on Streaming Trees via the Signature Kernel." In 2021 IEEE International Conference on Cyber Security and Resilience (CSR), 35–40. IEEE.

DARPA. 2019. "Operationally Transparent Cyber (OpTC) Data Release." https://github.com/FiveDirections/OpTC-data/.

Du, Min, Feifei Li, Guineng Zheng, and Vivek Srikumar. 2017. "Deeplog: Anomaly Detection and Diagnosis from System Logs through Deep Learning." In Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, 1285–98.

Dull, Eric. 2020. "Cyberthreat Analytics Using Graph Analysis." The Analyst 5: 11.

Edelsbrunner, Herbert, and John Harer. 2010. Computational Topology: An Introduction. Providence, RI: American Mathematical Society.

Gasparovic, Ellen, Emilie Purvine, Radmila Sazdanovic, Bei Wang, Yusu Wang, and Lori Ziegelmeier. 2024. "A Survey of Simplicial, Relative, and Chain Complex Homology Theories for Hypergraphs." *arXiv,* 2409.18310. https://arxiv.org/abs/2409.18310.

Ghrist, Robert W. 2014. *Elementary Applied Topology.* Vol. 1. Seattle, WA: Createspace.

Golczynski, Andrew, and John A. Emanuello. 2021. "End-To-End Anomaly Detection for Identifying Malicious Cyber Behavior through NLP-Based Log Embeddings." *arXiv,* 2108.12276. https://arxiv.org/abs/2108.12276.

Hatcher, Allen. 2000. Algebraic Topology. (Cambridge University Press).

Jenne, Helen, Sinan G. Aksoy, Daniel Best, Alyson Bittner, Gregory Henselman-Petrusek, Cliff Joslyn, Bill Kay, Audun Myers, Garret Seppala, Jackson Warley et al. 2024. "Stepping Out of Flatland: Discovering Behavior Patterns as Topological Structures in Cyber Hypergraphs." The Next Wave 25 (1).

Le, Van-Hoang, and Hongyu Zhang. 2022. "Log-Based Anomaly Detection with Deep Learning: How Far Are We?" In Proceedings of the 44th International Conference on Software Engineering, 1356–67.

Liu, Fucheng, Yu Wen, Dongxue Zhang, Xihe Jiang, Xinyu Xing, and Dan Meng. 2019. "Log2vec: A Heterogeneous Graph Embedding Based Approach for Detecting Cyber Threats within Enterprise." In Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security, 1777–94.

Mahboubi, Arash, Khanh Luong, Hamed Aboutorab, Hang Thanh Bui, Geoff Jarrad, Mohammed Bahutair, Seyit Camtepe, Ganna Pogrebna, Ejaz Ahmed, Bazara Barry et al. 2024. "Evolving Techniques in Cyber Threat Hunting: A Systematic Review." *Journal of Network and Computer Applications,* 104004.

Mamun, Mohammad, and Kevin Shi. 2021. "Deeptaskapt: Insider APT Detection Using Task-Tree Based Deep Learning." In IEEE 20th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom), 693–700. IEEE.

MITRE. n.d. "Cyber Analytics Repository Model."

Myers, Audun, Alyson Bittner, Sinan Aksoy, Dan Best, Gregory Henselman-Petrusek, Helen Jenne, Cliff Joslyn, Bill Kay, Garret Seppala, Stephen J. Young et al. 2023. "Malicious Cyber Activity Detection Using Zigzag Persistence." In 2023 IEEE Conference on Dependable and Secure Computing (DSC), 1–8. IEEE.

Myers, Audun, Cliff Joslyn, Bill Kay, Emilie Purvine, Gregory Roek, and Madelyn Shapiro. 2023. "Topological Analysis of Temporal Hypergraphs." In International Workshop on Algorithms and Models for the Web-Graph, 127–146. Springer.

Nandakumar, Dhruv, Robert Schiller, Christopher Redino, Kevin Choi, Abdul Rahman, Edward Bowen, Marc Vucovich, Joe Nehila, Matthew Weeks, and Aaron Shaha. 2022. "Zero Day Threat Detection Using Metric Learning Autoencoders." In 2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA), 1318–25. IEEE.

Nikulshin, Victor, and Chamseddine Talhi. 2024. "Effective IDS under Constraints of Modern Enterprise Networks: Revisiting the OpTC Dataset." In 2024 7th Conference on Cloud and Internet of Things (CIoT), 1–8. IEEE.

Paudel, Ramesh, and H. Howie Huang. 2022. "Pikachu: Temporal Walk Based Dynamic Graph Embedding for Network Anomaly Detection." In NOMS 2022–2022 IEEE/IFIP Network Operations and Management Symposium, 1–7. IEEE.

Potvin, Christopher. 2023. Hypergraphs and Their Homology. Ph.D. diss., Michigan State University.

Redino, Christopher, Dhruv Nandakumar, Robert Schiller, Kevin Choi, Abdul Rahman, Edward Bowen, Aaron Shaha, Joe Nehila, and Matthew Weeks. 2022. "Zero Day Threat Detection Using Graph and Flow Based Security Telemetry." In 2022 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS), 655–62. IEEE.

Chapter 3

Agarwal, Sameer, Kristin Branson, and Serge Belongie. 2006. "Higher Order Learning with Graphs." In Proceedings of the 23rd International Conference on Machine Learning, 17–24.

Aksoy, Sinan G., Ilya Amburg, and Stephen J. Young. 2024. "Scalable Tensor Methods for Nonuniform Hypergraphs." *SIAM Journal on Mathematics of Data Science* 6 (2): 481–503.

Aksoy, Sinan G., Cliff Joslyn, Carlos Ortiz Marrero, Brenda Praggastis, and Emilie Purvine. 2020. "Hypernetwork Science via High-Order Hypergraph Walks." EPJ Data Science 9 (1): 16.

Aksoy, Sinan G., Tamara G. Kolda, and Ali Pinar. 2017. "Measuring and Modeling Bipartite Graphs with Community Structure." *Journal of Complex Networks* 5 (4): 581–603.

Benson, Austin R. 2019. "Three Hypergraph Eigenvector Centralities." *SIAM Journal on Mathematics of Data Science* 1 (2): 293–312.

Chitra, Uthsav, and Benjamin Raphael. 2019. "Random Walks on Hypergraphs with Edge-Dependent Vertex Weights." In International Conference on Machine Learning, 1172–81. PMLR.

Cooley, Oliver, Wenjie Fang, Nicola Del Giudice, and Mihyun Kang. 2020. "Subcritical Random Hypergraphs, High-Order Components, and Hypertrees." *SIAM Journal on Discrete Mathematics* 34 (4): 2033–62.

Cooley, Oliver, Mihyun Kang, and Christoph Koch. 2015. "Evolution of High-Order Connected Components in Random Hypergraphs." Electronic Notes in Discrete Mathematics 49: 569–75.

Cooley, Oliver, Mihyun Kang, and Christoph Koch. 2018. "The Size of the Giant High-Order Component in Random Hypergraphs." Random Structures & Algorithms 53 (2): 238–88.

Derr, Tyler, Cassidy Johnson, Yi Chang, and Jiliang Tang. 2019. "Balance in Signed Bipartite Networks." In Proceedings of the 28th ACM International Conference on Information and Knowledge Management, 1221–30.

Dumitriu, Ioana, Hai-Xiao Wang, and Yizhe Zhu. 2021. "Partial Recovery and Weak Consistency in the Non-Uniform Hypergraph Stochastic Block Model." Combinatorics, Probability and Computing: 1–51.

Estrada, Ernesto, and Juan A. Rodríguez-Velázquez. 2006. "Subgraph Centrality and Clustering in Complex Hyper-Networks." Physica A: Statistical Mechanics and Its Applications 364: 581–94.

Harper, F. Maxwell, and Joseph A. Konstan. 2015. "The MovieLens Datasets: History and Context." ACM Transactions on Interactive Intelligent Systems (TIIS) 5 (4): 1–19.

Hayashi, Koby, Sinan G. Aksoy, Cheong Hee Park, and Haesun Park. 2020. "Hypergraph Random Walks, Laplacians, and Clustering." In Proceedings of the 29th ACM International Conference on Information and Knowledge Management, 495–504.

Kamiński, Bogumił, Valérie Poulin, Paweł Prałat, Przemysław Szufel, and François Théberge. 2019. "Clustering via Hypergraph Modularity." PLOS ONE 14 (11): e0224307.

Kapoor, Komal, Dhruv Sharma, and Jaideep Srivastava. 2013. "Weighted Node Degree Centrality for Hypergraphs." In 2013 IEEE 2nd Network Science Workshop (NSW), 152–55. IEEE.

Kirkland, Steve. 2018. "Two-Mode Networks Exhibiting Data Loss." *Journal of Complex Networks* 6 (2): 297–316.

Lee, Geon, Fanchen Bu, Tina Eliassi-Rad, and Kijung Shin. 2024. "A Survey on Hypergraph Mining: Patterns, Tools, and Generators." *arXiv,* 2401.08878. https://arxiv.org/abs/2401.08878.

Lee, Geon, Jihoon Ko, and Kijung Shin. 2020. "Hypergraph Motifs: Concepts, Algorithms, and Discoveries." *arXiv,* 2003.01853. https://arxiv.org/abs/2003.01853.

Lu, Linyuan, and Xing Peng. 2013. "High-Order Random Walks and Generalized Laplacians on Hypergraphs." Internet Mathematics 9 (1): 3–32.

Majumder, Bodhisattwa Prasad, Shuyang Li, Jianmo Ni, and Julian McAuley. 2019. "Generating Personalized Recipes from Historical User Preferences." In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 5976–82. Association for Computational Linguistics.

Ni, Jianmo, Jiacheng Li, and Julian McAuley. 2019. "Justifying Recommendations Using Distantly Labeled Reviews and Fine-Grained Aspects." In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 188–97.

Niu, Jason, Ilya D. Amburg, Sinan G. Aksoy, and Ahmet Erdem Sarıyüce. 2024. "Retrieving Top-k Hyperedge Triplets: Models and Applications." In IEEE International Conference on Big Data (2024).

Piñero, Janet, Àlex Bravo, Núria Queralt-Rosinach, Alba Gutiérrez-Sacristán, Jordi Deu-Pons, Emilio Centeno, Javier García-García, Ferran Sanz, and Laura I. Furlong. 2016. "DisGeNET: A Comprehensive Platform Integrating Information on Human Disease-Associated Genes and Variants." Nucleic Acids Research.

Ramani, Arjun S., Nicole Eikmeier, and David F. Gleich. 2019. "Coin-Flipping, Ball-Dropping, and Grass-Hopping for Generating Random Graphs from Matrices of Edge Probabilities." SIAM Review 61 (3): 549–95.

Robins, Garry, and Malcolm Alexander. 2004. "Small Worlds among Interlocking Directors: Network Structure and Distance in Bipartite Graphs." Computational & Mathematical Organization Theory 10: 69–94.

Sanei-Mehri, Seyed-Vahid, Ahmet Erdem Sariyuce, and Srikanta Tirthapura. 2018. "Butterfly Counting in Bipartite Networks." In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2150–59.

Seshadhri, Comandur, Ali Pinar, and Tamara G. Kolda. 2013. "Triadic Measures on Graphs: The Power of Wedge Sampling." In Proceedings of the 2013 SIAM International Conference on Data Mining, 10–18. SIAM.

Tudisco, Francesco, and Desmond J. Higham. 2021. "Node and Edge Nonlinear Eigenvector Centrality for Hypergraphs." Communications Physics 4 (1): 201.

Chapter 4

Akaike, H. "A New Look at the Statistical Model Identification." *IEEE Transactions on Automatic Control* 19, no. 6 (1974): 716–23.

Boudreau, Kevin J, Tom Brady, Ina Ganguli, Patrick Gaule, Eva Guinan, Anthony Hollenberg, and Karim R Lakhani. "A Field Experiment on Search Costs and the Formation of Scientific Collaborations." *Review of Economics and Statistics* 99, no. 4 (2017): 565–76.

Bozeman, Barry, and Craig Boardman. *Research Collaboration and Team Science: A State-of-the-Art Review and Agenda*. (Springer International Publishing, 2014). https://doi.org/10.1007/978-3-319-06468-0.

Brucks, Melanie S, and Jonathan Levav. "Virtual Communication Curbs Creative Idea Generation." *Nature* 605, no. 7908 (2022): 108–12.

Freeman, Richard B., Ina Ganguli, and Raviv Murciano-Goroff. "Why and Wherefore of Increased Scientific Collaboration." In *The Changing Frontier: Rethinking Science and Innovation Policy*, (University of Chicago Press, 2014): 17–48.

Guimera, Roger, Brian Uzzi, Jarrett Spiro, and Luis A Nunes Amaral. "Team Assembly Mechanisms Determine Collaboration Network Structure and Team Performance." *Science* 308, no. 5722 (2005): 697–702.

Hackman, J Richard. *Collaborative Intelligence: Using Teams to Solve Hard Problems*. Berrett-Koehler Publishers, 2011.

Iacopini, Iacopo, Márton Karsai, and Alain Barrat. "The Temporal Dynamics of Group Interactions in Higher-Order Social Networks." *Nature Communications* 15, no. 1 (2024): 7391.

Kardes, Hakan, Abdullah Sevincer, Mehmet Hadi Gunes, and Murat Yuksel. "Complex Network Analysis of Research Funding: A Case Study of NSF Grants." *State of the Art Applications of Social Network Analysis*, 2014, 163–87.

Landry, Nicholas W, Ilya Amburg, Mirah Shi, and Sinan G Aksoy. "Filtering Higher-Order Datasets." *Journal of Physics: Complexity* 5, no. 1 (2024): 015006.

Landry, Nicholas W, Maxime Lucas, Iacopo Iacopini, Giovanni Petri, Alice Schwarze, Alice Patania, and Leo Torres. "XGI: A Python Package for Higher-Order Interaction Networks." *Journal of Open Source Software* 8, no. 85 (2023): 5162.

Lane, Jacqueline N, Ina Ganguli, Patrick Gaule, Eva Guinan, and Karim R Lakhani. "Engineering Serendipity: When Does Knowledge Sharing Lead to Knowledge Production?" *Strategic Management Journal* 42, no. 6 (2021): 1215–44.

Leonard, Naomi Ehrich, and Simon A. Levin. "Collective Intelligence as a Public Good." *Collective Intelligence* 1, no. 1 (2022): 26339137221083293.

Lin, Yiling, Carl Benedikt Frey, and Lingfei Wu. "Remote Collaboration Fuses Fewer Breakthrough Ideas." *Nature* 623, no. 7989 (2023): 987–91.

Liu, Yuxian, Yishan Wu, Sandra Rousseau, and Ronald Rousseau. "Reflections on and a Short Review of the Science of Team Science." *Scientometrics* 125 (2020): 937–50.

Lungeanu, Alina, Yun Huang, and Noshir S. Contractor. "Understanding the Assembly of Interdisciplinary Teams and Its Impact on Performance." *Journal of Infometrics* 8, no. 1 (2014): 59–70.

Meluso, John, and Laurent Hébert-Dufresne. "Multidisciplinary Learning through Collective Performance Favors Decentralization." *Proceedings of the National Academy of Sciences* 120, no. 34 (2023): e2303568120. https://doi.org/10.1073/pnas.2303568120.

Nakajima, Kazuki, Kazuyuki Shudo, and Naoki Masuda. "Higher-Order Rich-Club Phenomenon in Collaborative Research Grant Networks." *Scientometrics* 128, no. 4 (2023): 2429–46.

Patania, Alice, Giovanni Petri, and Francesco Vaccarino. "The Shape of Collaborations." *EPJ Data Science* 6, no. 18 (August 24, 2017). https://doi.org/10.1140/epjds/s13688-017-0114-8.

Shi, Feng, and James Evans. "Surprising Combinations of Research Contents and Contexts Are Related to Impact and Emerge with Scientific Outsiders from Distant Disciplines." *Nature Communications* 14, no. 1 (2023): 1641.

Simmel, Georg. *Soziologie*. Leipzig: Duncker & Humblot, 1908.

Skiles, Matthew, Euijin Yang, Orad Reshef, Diego Robalino Muñoz, Diana Cintron, Mary Laura Lind, Alexander Rush et al. "Conference Demographics and Footprint Changed by Virtual Platforms." *Nature Sustainability* 5, no. 2 (2022): 149–56.

Staples, D Sandy, and Lina Zhao. "The Effects of Cultural Diversity in Virtual Teams versus Face-to-Face Teams." *Group Decision and Negotiation* 15 (2006): 389–406.

Tao, Yanqiu, Debbie Steckel, Jiří Jaromír Klemeš, and Fengqi You. "Trend Toward Virtual and Hybrid Conferences May Be an Effective Climate Change Mitigation Strategy." *Nature Communications* 12, no. 1 (December 2021): 7324.

Vogel, Kathleen M, and Beverly B Tyler. "Interdisciplinary, Cross-Sector Collaboration in the US Intelligence Community: Lessons Learned from Past and Present Efforts." *Intelligence and National Security* 34, no. 6 (2019): 851–80.

Wiener, RJ, and S Ronco. "Scialog: The Catalysis of Convergence." *ACS Energy Letters* 4, no. 5 (2019): 1020–24.

Wu, Juncheng, Anushka Rajesh, Yu-Ning Huang, Karishma Chhugani, Rajesh Acharya, Kerui Peng, Ruth D Johnson et al. "Virtual Meetings Promise to Eliminate Geographical and Administrative Barriers and Increase Accessibility, Diversity and Inclusivity." *Nature Biotechnology* 40, no. 1 (2022): 133–37.

Wuchty, Stefan, Benjamin F. Jones, and Brian Uzzi. "The Increasing Dominance of Teams in Production of Knowledge." *Science* 316, no. 5827 (2007): 1036–39.

Zajdela, Emma, and Nicholas W. Landry. "Higher-Order-Team-Formation: V0.0." Zenodo, February 2025. https://doi.org/10.5281/zenodo.14867591.

Zajdela, Emma R, Kimberly Huynh, Andrew L Feig, Richard J Wiener, and Daniel M Abrams. "Face-to-Face or Face-to-Screen: A Quantitative Comparison of Conferences Modalities." *PNAS Nexus* 4, no. 1 (2025): 522.

Zajdela, Emma R., Kimberly Huynh, Andy T. Wen, Andrew L. Feig, Richard J. Wiener, and Daniel M. Abrams. "Dynamics of Social Interaction: Modeling the Genesis of Scientific Collaboration." *Phys. Rev. Res.* 4, no. 4 (October 2022): L042001. https://doi.org/10.1103/PhysRevResearch.4.L042001.

Chapter 5

Aksoy, Sinan G., Cliff Joslyn, Carlos Ortiz Marrero, Brenda Praggastis, and Emilie Purvine. 2020. Hypernetwork Science via High-Order Hypergraph Walks. EPJ Data Science 9 (1): 16. https://doi.org/10.1140/epjds/s13688-020-00231-0.

Bassem, Christine. 2019. Redefining Node Centrality for Task Allocation in Mobile CrowdSensing Platforms. In 2019 IEEE International Conference on Smart

Computing (SMARTCOMP), 323–31. https://doi.org/10.1109/SMARTCOMP.2019.00069

Battiston, Federico et al. "Networks Beyond Pairwise Interactions: Structure and Dynamics." *Physics Reports* 874 (2020): 1–92. https://doi.org/10.1016/j.physrep.2020.05.004.

Cattuto, Ciro, Wouter Van Den Broeck, Alain Barrat, Vittoria Colizza, Jean-François Pinton, and Alessandro Vespignani. 2010. Dynamics of Person-to-Person Interactions from Distributed RFID Sensor Networks. *PLoS ONE* 5 (7): e11596. https://doi.org/10.1371/journal.pone.0011596

Centola, Damon. 2010. The Spread of Behavior in an Online Social Network Experiment. Science 329 (5996): 1194–97. https://doi.org/10.1126/science.1185231

Estrada, Ernesto, and Juan A. Rodríguez-Velázquez. 2006. Subgraph Centrality and Clustering in Complex Hyper-Networks. Physica A: Statistical Mechanics and Its Applications 364: 581–94. https://doi.org/10.1016/j.physa.2005.12.002

Everett, Martin G., and Stephen P. Borgatti. 2013. The Dual-Projection Approach for Two-Mode Networks. Social Networks 35 (2): 204–10. https://doi.org/10.1016/j.socnet.2012.05.004

Feng, Song, Emily Heath, Brett Jefferson, Cliff Joslyn, Henry Kvinge, Hugh D. Mitchell, Brenda Praggastis, Amie J. Eisfeld, Amy C. Sims, Larissa B. Thackray et al. 2021. Hypergraph Models of Biological Networks to Identify Genes Critical to Pathogenic Viral Response. BMC Bioinformatics 22 (1): 287. https://doi.org/10.1186/s12859-021-04197-2

Firth, Josh A. 2020. Considering Complexity: Animal Social Networks and Behavioural Contagions. *Trends in Ecology & Evolution* 35 (2): 100–04. https://doi.org/10.1016/j.tree.2019.10.009

Génois, Mathieu, and Alain Barrat. 2018. Can Co-Location Be Used as a Proxy for Face-to-Face Contacts? *EPJ Data Science* 7 (1): 11. https://doi.org/10.1140/epjds/s13688-018-0140-1

Guilbeault, Douglas et al. "Complex Contagions: A Decade in Review." In Sune Lehmann and Yong-Yeol Ahn (eds.), *Complex Spreading Phenomena in Social Systems*. (Springer International Publishing, 2018): 3–25. https://doi.org/10.1007/978-3-319-77332-2_1.

Guilbeault, Douglas, and Damon Centola. "Topological Measures for Identifying and Predicting the Spread of Complex Contagions." *Nature Communications* 12 (2021): 4430. https://doi.org/10.1038/s41467-021-24704-6

Hasenjager, Matthew J., Mark M. Bailey, and Nina H. Fefferman. 2026. "Higher-Order Networks: A Primer." In *Higher-Order Network Methods for*

Intelligence Applications: A Compendium for Practitioners, eds. Matthew J. Hasenjager, Mark M. Bailey, and Nina H. Fefferman. (National Intelligence Press, 2026.)

Hasenjager, Matthew J., and Nina H. Fefferman. "Social Ageing and Higher-Order Interactions: Social Selectiveness Can Enhance Older Individuals' Capacity To Transmit Knowledge." *Philosophical Transactions of the Royal Society B: Biological Sciences* 379 (2024): 20220461. https://doi.org/10.1098/rstb.2022.0461.

Hasenjager, Matthew J., Ellouise Leadbeater, and William Hoppitt. 2021. Detecting and Quantifying Social Transmission Using Network-Based Diffusion Analysis. *Journal of Animal Ecology* 90 (1): 8–26. https://doi.org/10.1111/1365-2656.13307

Iacopini, Iacopo et al. Simplicial Models of Social Contagion. *Nature Communications* 10 (2019): 2485. https://doi.org/10.1038/s41467-019-10431-6.

Kirkland, Steve. "Two-Mode Networks Exhibiting Data Loss." *Journal of Complex Networks* 6, no. 2 (2018): 297–316. https://doi.org/10.1093/comnet/cnx039.

Lloyd-Smith, James O., Sebastian J. Schreiber, P. Ekkehard Kopp, and Wayne M. Getz. 2005. Superspreading and the Effect of Individual Variation on Disease Emergence. *Nature* 438 (7066): 355–59. https://doi.org/10.1038/nature04153.

Newman, M. E. J. 2010. *Networks: An Introduction,*1st Ed. (Oxford University Press).

Newman, M. E. J. "The Structure and Function of Complex Networks." *SIAM Review* 45, no. 2 (2003): 167–256. https://doi.org/10.1137/S003614450342480.

Salathé, Marcel, and James H. Jones. 2010. Dynamics and Control of Diseases in Networks with Community Structure. *PLoS Computational Biology* 6 (4): e1000736. https://doi.org/10.1371/journal.pcbi.1000736.

Silk, Matthew J. et al. Capturing Complex Interactions in Disease Ecology with Simplicial Sets. Ecology Letters 25 no. 10 (2022): 2217–31. https://doi.org/10.1111/ele.14079.

St-Onge, Guillaume, Iacopo Iacopini, Vito Latora, Alain Barrat, Giovanni Petri, Antoine Allard, and Laurent Hébert-Dufresne. 2022. Influential Groups for Seeding and Sustaining Nonlinear Contagion in Heterogeneous Hypergraphs. Communications Physics 5 (1): 25. https://doi.org/10.1038/s42005-021-00788-w

Wasserman, Stanley, and Katherine Faust. 1994. Social Network Analysis: Methods and Applications. (Cambridge University Press).

Wuellner, Daniel R., Soumen Roy, and Raissa M. D'Souza. 2010. Resilience and Rewiring of the Passenger Airline Networks in the United States. Physical Review E 82 (5): 056101. https://doi.org/10.1103/PhysRevE.82.056101

Xie, Ming, Xiu-Xiu Zhan, Chuang Liu, and Zi-Ke Zhang. 2022. Influence Maximization in Hypergraphs. *arXiv,* 2206.01394. http://arxiv.org/abs/2206.01394.

Chapter 6

Alhajjar, Elie, and Ross Friar. "A Historical Perspective on International Treaties via Hypernetwork Science." In Network Science, edited by Pedro Ribeiro, Fernando Silva, José Fernando Mendes, and Rosário Laureano, 15–25. (Springer International Publishing, 2022).

Albert, R., and Albert-László Barabási. "Statistical Mechanics of Complex Networks." *Reviews of Modern Physics* 74, no. 1 (2002): 47–97. https://doi.org/10.1103/RevModPhys.74.47.

GDELT Project. 2015. GDELT Event Database Data Format Codebook V2.0. http://gdeltproject.org/.

Gibler, Douglas M. 2009. International Military Alliances, 1648–2008. Washington, DC: CQ Press.

"Jigsaw." n.d. Jigsaw. Accessed February 2, 2025. https://jigsaw.google.com/.

Joslyn, Cliff A., Sinan G. Aksoy, Tiffany J. Callahan, Lawrence E. Hunter, Brett Jefferson, Brenda Praggastis, Emilie Purvine, and Ignacio J. Tripodi. "Hypernetwork Science: From Multidimensional Networks to Computational Topology." In Unifying Themes in Complex Systems X, edited by Dan Braha et al. (Springer International Publishing, 2021): 377–92.

Koster, M. B. M. de, and J. H. P. Paelinck. 1984. "A Hypergraph Approach to Conflict." Conflict Management and Peace Science 7 (2): 55–69. https://doi.org/10.1177/073889428400700204.

Leetaru, Kalev H., and Philip A. Schrodt. 2013. "GDELT: Global Database of Events, Language, and Tone, 1979-Present." https://www.gdeltproject.org/.

Philip A. Schrodt and Department of Political Science Event Data Project, Pennsylvania State University. 2012. CAMEO: Conflict and Mediation Event Observations. http://eventdata.psu.edu/.

Praggastis, Brenda. "HyperNetX: A Python Package for Modeling Complex Network Data as Hypergraphs." *Journal of Open Source Software* 9, no. 95 (2024): 06016. https://doi.org/10.21105/joss.06016.

Schrodt, Philip, and Jay Yonamine. 2013. "A Guide to Event Data: Past, Present, and Future." *All Azimuth: A Journal of Foreign Policy and Peace* 2 (2): 5–22. https://doi.org/10.20991/allazimuth.167312.

Small, Melvin, and J. David Singer. 1969. "Formal Alliances, 1815-1965: An Extension of the Basic Data." *Journal of Peace Research* 6 (3): 257–82.

"The GDELT Project." n.d. Accessed February 2, 2025. https://www.gdeltproject.org/.

Watts, Duncan J., and Steven H. Strogatz. 1998. "Collective Dynamics of 'Small-World' Networks." Nature 393: 440–42. https://doi.org/10.1038/30918.

Newman, M. E. J. "The Structure and Function of Complex Networks." *SIAM Review* 45, no. 2 (2003): 167–256. https://doi.org/10.1137/S003614450342480.

Van Gastel, M. A. J. J., and J. H. P. Paelinck. 1991. "Hypergraph Conflict Analysis." *Economics Letters* 35 (3): 233–37. https://doi.org/10.1016/0165-1765(91)90135-8.

Van Gastel, R., and J. H. P. Paelinck. 1988. "Hypergraph Conflict Analysis: Synthesis and Extension." Conflict Management and Peace Science 10 (1): 59–86.

Yi, Sudo, and Deok-Sun Lee. 2022. "Structure of International Trade Hypergraphs." *Journal of Statistical Mechanics: Theory and Experiment* 2022 (10): 103402. https://doi.org/10.1088/1742-5468/ac946f.

Chapter 7

Abonyi, J., and T. Czvetkó. 2022. Hypergraph and Network Flow-Based Quality Function Deployment. *Heliyon* 8. https://doi.org/10.1016/j.heliyon.2022.e12263.

Aksoy, S. G., C. A. Joslyn, C. O. Marrero, B. Praggastis, and E. Purvine. 2019. Hypernetwork Science via High-Order Hypergraph Walks. *EPJ Data Science* 9. https://epjdatascience.springeropen.com/articles/10.1140/epjds/s13688-020-00231-0.

Anderson, E. C. 2005. An Efficient Monte Carlo Method for Estimating Ne from Temporally Spaced Samples Using a Coalescent-Based Likelihood. Genetics 170: 955–67.

Battiston, Federico et al. "Networks Beyond Pairwise Interactions: Structure and Dynamics." *Physics Reports* 874 (2020): 1–92. https://doi.org/10.1016/j.physrep.2020.05.004.

Bellaachia, A., and M. Al-Dhelaan. 2021. Random Walks in Hypergraph. *International Journal of Education and Information Technologies.*

Bick, C., E. Gross, H. A. Harrington, and M. T. Schaub. 2021. What Are Higher-Order Networks? *arXiv* preprint, *arXiv,* 2104.11329.

Boettcher, M., J. Gabriel, and S. Low. 2016. Solar Radiation Management: Foresight for Governance. IAAS Working Paper, Institute for Advanced Sustainability Studies (IASS).

Bryant, B. P., and R. J. Lempert. 2010. Thinking Inside the Box. Technological Forecasting and Social Change 77 (1): 34–49. https://www.sciencedirect.com/science/article/abs/pii/S004016250900105X.

Coupette, C., J. Vreeken, and B. Rieck. 2024. All the World's a (Hyper)graph: A Data Drama. Digital Scholarship in the Humanities 39 (1): 74–96. https://doi.org/10.1093/llc/fqad071.

De'ath, G., and K. E. Fabricius. 2000. Classification and Regression Trees: A Powerful Yet Simple Technique for Ecological Data Analysis. *Ecology* 81: 3178–92.

Fatemi, B., P. Taslakian, D. Vázquez, and D. L. Poole. 2019. Knowledge Hypergraphs: Prediction Beyond Binary Relations. In International Joint Conference on Artificial Intelligence.

Fries, C. P. 2005. Foresight Bias and Suboptimality Correction in Monte Carlo Pricing of Options with Early Exercise: Classification, Calculation, and Removal. Derivatives.

Geurts, A., R. Gutknecht, P. Warnke, A. Goetheer, E. Schirrmeister, B. Bakker, and S. Meissner. 2021. New Perspectives for Data-Supported Foresight: The Hybrid AI-Expert Approach. *Futures & Foresight Science.*

Greeven, S., O. Kraan, É. J. Chappin, and J. H. Kwakkel. 2016. The Emergence of Climate Change Mitigation Action by Society: An Agent-Based Scenario Discovery Study. Journal of Artificial Societies and Social Simulation 19.

Hurst, O. A., A. Lerer, A. P. Goucher, A. Perelman, A. Ramesh, A. Clark, A. Ostrow et al. 2024. GPT-4o System Card. *arXiv* preprint, *arXiv,* 2410.21276.

Jiang, H., L. Ge, Y. Gao, J. Wang, and R. Song. 2023. Large Language Model for Causal Decision Making. *arXiv* preprint, *arXiv,* 2312.17122.

Jiralerspong, T., X. Chen, Y. More, V. Shah, and Y. Bengio. 2024. Efficient Causal Graph Discovery Using Large Language Models. *arXiv* preprint, *arXiv,* 2402.01207.

Johansen, I. 2017. Scenario Modelling with Morphological Analysis. Technological Forecasting and Social Change.

Kilian, K. A., C. Ventura, and M. M. Bailey. 2022. Examining the Differential Risk from High-Level Artificial Intelligence and the Question of Control. *arXiv* preprint, *arXiv,* 2211.03157.

Kıckman, E., R. O. Ness, A. Sharma, and C. Tan. 2023. Causal Reasoning and Large Language Models: Opening a New Frontier for Causality. *arXiv* preprint, *arXiv,* 2305.00050.

Malekmohammadi, B., and F. Jahanishakib. 2017. Vulnerability Assessment of Wetland Landscape Ecosystem Services Using Driver-Pressure-State-Impact-Response (DPSIR) Model. Ecological Indicators 82: 293–303.

Mislove, A., M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee. 2007. Measurement and Analysis of Online Social Networks. ACM/SIGCOMM Internet Measurement Conference.

Shelke, P., and T. Hamalainen. 2024. Analysing Multidimensional Strategies for Cyber Threat Detection in Security Monitoring. European Conference on Cyber Warfare and Security.

Shivamurthy, A. G. 2022. Kashmir and Ontological Security: Re-evaluating the Role of Self-Identities in a Multi-Layered Conflict. *Journal of Asian Security and International Affairs* 9: 255–79.

Stauffer, D. 1988. Monte Carlo Simulations in Statistical Physics.

Strobl, C., J. D. Malley, and G. Tutz. 2009. An Introduction to Recursive Partitioning: Rationale, Application, and Characteristics of Classification and Regression Trees, Bagging, and Random Forests. Psychological Methods 14 (4): 323–48.

Tecuci, G., M. Boicu, D. A. Schum, and D. Marcu. 2010. Coping with the Complexity of Intelligence Analysis: Cognitive Assistants for Evidence-Based Reasoning.

Truong, Q. D., Q. B. Truong, and T. Dkaki. 2016. Graph Methods for Social Network Analysis. International Conference on Nature of Computation and Communication.

Viermetz, M., and M. Skubacz. 2007. Using Topic Discovery to Segment Large Communication Graphs for Social Network Analysis. In IEEE/WIC/ACM International Conference on Web Intelligence (WI'07), 95–99.

Wu, T., P. Chen, C. Chen, M. Chung, Z. Ye, and M. Li. 2023. Random Forest of Classification and Regression Tree (CART) in the Estimation of SWC Based on Meteorological Inputs and Hydrodynamics Behind. In 2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2251–55.

Zhang, Y., Y. Zhang, Y. Gan, L. Yao, and C. Wang. 2024. Causal Graph Discovery with Retrieval-Augmented Generation-Based Large Language Models. *arXiv* preprint, *arXiv,* 2402.15301.

Chapter 8

Active DNS Project (2016-2024). Center for Cyber Operations Enquiry and Unconventional Sensing, Georgia Institute of Technology. https://activednsproject.org/.

Ahmad, Z., A. S. Khan, C. W. Shiang, J. Abdullah, and F. Ahmad. "Network Intrusion Detection System: A Systematic Study of Machine Learning and Deep Learning Approaches." *Transactions on Emerging Telecommunications Technologies.* 2020.

Aksoy, S. G., C. Joslyn, C. Ortiz Marrero, B. Praggastis, and E. Purvine. "Hypernetwork Science via High-Order Hypergraph Walks." *EPJ Data Science,* 9(1), 16. https://doi.org/10.1140/epjds/s13688-020-00231-0.

An, X., J. Su, X. Liu, and F. Lin. "Hypergraph Clustering Model-Based Association Analysis of DDoS Attacks in Fog Computing Intrusion Detection System." 2018. https: //doi.org/10.1186/s13638-018-1267-2.

Bastian, N., D. Bierbrauer, M. McKenzie, and E. Nack. December 29, 2023. "ACI IoT Network Traffic Dataset 2023." *IEEE Dataport.* doi:https://dx.doi.org/10.21227/qacj-3x32.

Bierbrauer, D. A., W. Kritzer, A. Chang, and N. D. Bastian. "Cybersecurity Anomaly Detection in Adversarial Environments." *AAAI FSS-21: Artificial Intelligence in Government and Public Sector.* 2021.

Bojanova, I. "Comprehensively Labeled Weakness and Vulnerability Datasets via Unambiguous Formal Bugs Framework (BF) Specifications." *IEEE IT Professional*. 2024. https://doi.org/10.1109/MITP.2024.3358970. https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=957238 (Accessed November 16, 2024).

CIC (2017). Canadian Institute for Cybersecurity (CIC)-Intrusion Detection Evaluation Dataset (cic-ids2017). https: //www.unb.ca/cic/datasets/ids-2017.html.

Devine, S.M. and N.D. Bastian. "An Adversarial Training Based Machine Learning Approach to Malware Classification Under Adversarial Conditions." *Proceedings of the 54th Hawaii International Conference on System Sciences,* 827–36, 2021.

Fang, Y., C. H. Wang, Z. Y. Fang, and C. Huang. "LMTracker: Lateral Movement Path Detection Based on Heterogeneous Graph Embedding." 2022. *Neurocomputing*, vol. 474, 37–47. ISSN 0925-2312. https://doi.org/10.1016/j.neucom.2021.12.026.

Gallagher, P. D. "Guide for Conducting Risk Assessments." NIST SP 800-30 Revision 1, National Institute of Standards and Technology, September 2012, https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-30r1.pdf.

Gambo, M. L. and A. Almulhem. 2025. "Zero Trust Architecture: A Systematic Literature Review." *arXiv.* https://arxiv.org/html/2503.11659v2.

Joslyn, C.A., S.G. Aksoy, D. Arendt, J. Firoz, L. Jenkins, B. Praggastis, E. Purvine, and M. Zalewski. "Hypergraph Analytics of Domain Name System Relationships." In eds. B. Kamiński, P. Prałat, and P. Szufel. *Algorithms and Models for the Web Graph.* (Springer International Publishing, 2020): 1–15. https://doi.org/10.1007/978-3-030-48478-1_1.

Khraisat, A., I. Gondal, P. Vamplew, and J. Kamruzzaman. "Survey of Intrusion Detection Systems: Techniques, Datasets and Challenges." SpringerOpen, 2019. https://doi.org/10.1186/s42400-019-0038-7

Kountouras, A., P. Kintis, C. Lever, Y.Z. Chen, Y. Nadji, D. Dagon, M. Antonakakis, and R. Joffe. 2016. "Enabling Network Security Through Active DNS Datasets." *RAID. Lecture Notes in Computer Science*, vol. 9854, 188–208. Springer.

Lin, Z. Z., T. D. Pike, M. M. Bailey, and N. D. Bastian. "A Hypergraph-Based Machine Learning Ensemble Network Intrusion Detection System." *IEEE*

Transactions on Systems, Man, and Cybernetics: Systems, vol. 54, no. 11, 69116923, November 2024, doi: 10.1109/TSMC.2024.3446635.

Moustafa, N. "New Generations of Internet of Things Datasets for Cybersecurity Applications Based Machine Learning: TON_IoT Datasets." *Proceedings of the eResearch Australasia Conference,* Brisbane, Australia. 2019.

Moustafa, N., M. Ahmed, and S. Ahmed. 2020. "Data Analytics-Enabled Intrusion Detection: Evaluations of ToN_IoT Linux Datasets." *IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications* (TrustCom), 2020, 727–73. doi: 10.1109/TrustCom50675.2020.00100.

Moustafa, N., M. Keshk, E. Debie, and H. Janicke. 2020. "Federated TON_IoT Windows Datasets for Evaluating AI-Based Security Applications." *IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications* (TrustCom), 2020, 848–55. doi: 10.1109/TrustCom50675.2020.00114.

Moustafa, N. and J. Slay. 2015. "UNSW-NB15: A Comprehensive Data Set for Network Intrusion Detection Systems." *Military Communications and Information Systems Conference (MilCIS),* 2015. *IEEE*, 2015.

Rose, S., O. Borchert, S. Mitchell, and S. Connelly, "Zero Trust Architecture," NIST SP 800-207. https://doi.org/10.6028/NIST.SP.800-207.

Sharafaldin, I., A. H. Lashkari, and A. A. Ghorbani. 2018. "Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization." *International Conference on Information Systems Security and Privacy*. https://api.semanticscholar.org/CorpusID:4707749.

Surana, A., J. Sharma, I. Saraf, N. Puri, and B. Navin. 2019. "A Survey on Intrusion Detection System." https://www.ijedr.org/papers/IJEDR1702161.pdf.

Veiga, J. M. M., R. Criado, M. R. D. Rio, S. I. Perez, P. A. Rodriguez, and K. K. H. Manjunatha. 2024. "A Hypergraph-Based Model for Cyberincident-Related Data Analysis." *Proceedings of the 2024 European Interdisciplinary Cybersecurity Conference,* 161–62. https://doi.org/10.1145/3655693.3661300.

Veiga, J. M. M. 2024. "A Hypergraph-Based Model for Cybersecurity." Workshop on Complex Networks in Banking and Finance, June 25, 2024. Fields Institute. https://www.fields.toronto.edu/talks/hypergraph-based-model-cybersecurity.

Wikipedia. "Advanced Persistent Threat." https://en.wikipedia.org/wiki/Advanced_persistent_threat. April 9, 2025.

Xiao, Q., J. Liu, Q. Wang, Z. Jiang, X. Wang, and Y. Yao. 2020. "Towards Network Anomaly Detection Using Graph Embedding." *Computational Science–Inter. Conference on Computationl Science* 2020, *Lecture Notes in Computer Science* by eds. V. V. Krzhizhanovskaya et al., Springer, 2020, vol. 12140, 156–69.

Endnotes

1 Federico Battiston et al., "Networks Beyond Pairwise Interactions: Structure and Dynamics," *Physics Reports* 874 (2020): 1–92, https://doi.org/10.1016/j.physrep.2020.05.004; Mikko Kivelä et al., "Multilayer Networks," *Journal of Complex Networks* 2, no. 3 (2014): 203–71, https://doi.org/10.1093/comnet/cnu016; and M. E. J. Newman, "The Structure and Function of Complex Networks," *SIAM Review* 45, no. 2 (2003): 167–256, https://doi.org/10.1137/S003614450342480.

2 M. E. J. Newman, *Networks: An Introduction,* 1st ed. (Oxford University Press, 2010).

3 Linton C. Freeman, *The Development of Social Network Analysis: A Study in the Sociology of Science* (Empirical Press, 2004).

4 Réka Albert and Albert-László Barabási, "Statistical Mechanics of Complex Networks," *Reviews of Modern Physics* 74, no. 1 (2002): 47–97, https://doi.org/10.1103/RevModPhys.74.47.

5 Josh A. Firth, "Considering Complexity: Animal Social Networks and Behavioural Contagions," *Trends in Ecology & Evolution* 35, no. 2 (2020): 100–04, https://doi.org/10.1016/j.tree.2019.10.009; Douglas Guilbeault et al., "Complex Contagions: A Decade in Review," in *Complex Spreading Phenomena in Social Systems,* eds. Sune Lehmann and Yong-Yeol Ahn, (Springer International Publishing, 2018): 3–25, https://doi.org/10.1007/978-3-319-77332-2_1; and Iacopo Iacopini et al., "Simplicial Models of Social Contagion," *Nature Communications* 10 (2019): 2485, https://doi.org/10.1038/s41467-019-10431-6.

6 Nina H. Fefferman and Kah L. Ng, "How Disease Models in Static Networks Can Fail to Approximate Disease in Dynamic Networks." *Physical Review* E 76, no. 3 (2007): 031919, https://doi.org/10.1103/PhysRevE.76.031919; and

Matthew J. Silk et al., "Capturing Complex Interactions in Disease Ecology with Simplicial Sets," *Ecology Letters* 25, no. 10 (2022): 2217–31, https://doi.org/10.1111/ele.14079.

7 Stefano Boccaletti et al., "Explosive Transitions in Complex Networks' Structure and Dynamics: Percolation and Synchronization," *Physics Reports* 660 (2016): 1–94, https://doi.org/10.1016/j.physrep.2016.10.004.

8 Leonie Neuhäuser et al., "Multibody Interactions and Nonlinear Consensus Dynamics on Networked Systems," *Physical Review E* 101, no. 3 (2020): 032310, https://doi.org/10.1103/PhysRevE.101.032310.

9 Hisashi Ohtsuki et al., "A Simple Rule for the Evolution of Cooperation on Graphs and Social Networks," *Nature* 441, no. 7092 (2006): 502–05, https://doi.org/10.1038/nature04605.

10 Xiaoqiu Shi et al., "Robustness of Interdependent Supply Chain Networks Against Both Functional and Structural Cascading Failures," *Physica A: Statistical Mechanics and Its Applications* 586 (2022): 126518, https://doi.org/10.1016/j.physa.2021.126518; and Lazaros K. Gallos and Nina H. Fefferman, "Simple and Efficient Self-Healing Strategy for Damaged Complex Networks," *Physical Review E* 92, no. 5 (2015): 052806, https://doi.org/10.1103/PhysRevE.92.052806.

11 Lazaros K. Gallos et al., "Anomaly Detection Through Information Sharing Under Different Topologies," *EURASIP Journal on Information Security* (2017): 5, https://doi.org/10.1186/s13635-017-0056-5.

12 Battiston et al., "Networks Beyond Pairwise Interactions;" and eds. Federico Battiston and Giovanni Petri, Higher-Order Systems, 1st ed., (Springer, 2022), https://doi.org/10.1007/978-3-030-91374-8.

13 Bradford R. Greening, Jr., "Higher-Order Analysis of Knowledge Capacity and Learning Potential in Social Animal Groups," Ph.D. diss., Rutgers University (2014); Iacopo Iacopini et al., "Not Your Private Tête-à-Tête: Leveraging the Power of Higher-Order Networks to Study Animal Communication," *Philosophical Transactions of the Royal Society B: Biological Sciences* 379, no. 1905 (2024): 2023.0190, https://doi.org/10.1098/rstb.2023.0190; and Lysanne Snijders and Marc Naguib, "Communication in Animal Social Networks: A Missing Link?" *Advances in the Study of Behavior* 49 (2017): 297–359, https://doi.org/10.1016/bs.asb.2017.02.004.

14 Damon Centola, "The Spread of Behavior in an Online Social Network Experiment," *Science* 329, no. 5996 (2010): 1194–97, https://doi.org/10.1126/science.1185231; Leonie Neuhäuser et al., "Multibody Interactions and Nonlinear Consensus Dynamics on Networked Systems."

15 Marco T. Angulo et al., "Coexistence Holes Characterize the Assembly and Disassembly of Multispecies Systems," *Nature Ecology & Evolution* 5, no. 8 (2021): 1091–1101, https://doi.org/10.1038/s41559-021-01462-8; Antonio J. Golubski et al., "Ecological Networks Over the Edge: Hypergraph Trait-Mediated Indirect Interaction (TMII) Structure," *Trends in Ecology & Evolution* 31, no. 5 (2016): 344–54, https://doi.org/10.1016/j.tree.2016.02.006.

16 Emilie Purvine and Helen Jenne, "From Logs to Dynamic Hypergraphs: Identifying and Interpreting Anomalous Cyber Activity," in *Higher-Order Network Methods for Intelligence Applications: A Compendium for Practitioners,* eds. Matthew J. Hasenjager, Mark M. Bailey, and Nina H. Fefferman (National Intelligence Press, 2026).

17 Thomas D. Pike, "Cracking the Paradigm: Hypergraph Signatures of International Conflict," in *Higher-Order Network Methods for Intelligence Applications: A Compendium for Practitioners,* eds. Matthew J. Hasenjager, Mark M. Bailey, and Nina H. Fefferman (National Intelligence Press, 2026).

18 Emma Zajdela and Nicholas W. Landry, "Higher-Order Interactions at Scientific Conferences Influence Team Formation," in *Higher-Order Network Methods for Intelligence Applications: A Compendium for Practitioners,* eds. Matthew J. Hasenjager, Mark M. Bailey, and Nina H. Fefferman (National Intelligence Press, 2026).

19 Greg Friedman, "An Elementary Illustrated Introduction to Simplicial Sets," Cornell University Algebraic Topology, *arXiv* (2023): 0809.4221, http://arxiv.org/abs/0809.4221; and Stanley Wasserman and Katherine Faust, *Social Network Analysis: Methods and Applications* (Cambridge University Press, 1994).

20 Bradford R. Greening, Jr. et al., "Higher-Order Interactions: Understanding the Knowledge Capacity of Social Groups Using Simplicial Sets," *Current Zoology* 61, no. 1 (2015): 114–27. https://doi.org/10.1093/czoolo/61.1.114; and Leo Torres et al., "The Why, How, and When of Representations for Complex Systems," *SIAM Review* 63, no. 3 (2021): 435–85, https://doi.org/10.1137/20M1355896.

21 Guilbeault et al. "Complex Contagions: A Decade in Review;" Douglas Guilbeault and Damon Centola, "Topological Measures for Identifying and Predicting the Spread of Complex Contagions," *Nature Communications* 12 (2021): 4430, https://doi.org/10.1038/s41467-021-24704-6.

22 Guilherme Ferraz de Arruda et al., "Social Contagion Models on Hypergraphs," *Physical Review Research* 2, no. 2 (2020), https://doi.org/10.1103/PhysRevResearch.2.023032.

23 Greening et al., "Higher-Order Interactions: Understanding the Knowledge Capacity of Social Groups Using Simplicial Sets."

24 Albert-László Barabási, *Network Science* (Cambridge University Press, 2016); Stefano Boccaletti et al., "Complex Networks: Structure and Dynamics," *Physics Reports* 424, no. 4-5 (2006): 175–308, https://doi.org/10.1016/j.physrep.2005.10.009; Filippo Menczer et al., *A First Course in Network Science* (Cambridge University Press, 2020), https://doi.org/10.1017/9781108653947; Newman, *Networks: An Introduction,* 1st ed.; and Newman, "The Structure and Function of Complex Networks."

25 Benjamin Blonder et al., "Temporal Dynamics and Network Analysis," *Methods in Ecology and Evolution* 3, no. 6 (2012): 958–72, https://doi.org/10.1111/j.2041-210X.2012.00236.x.

26 Battiston et al., "Networks Beyond Pairwise Interactions;" Torres et al., "The Why, How, and When of Representations for Complex Systems."

27 Friedman, "The Structure and Function of Complex Networks;" Paul G. Goerss and John F. Jardine, *Simplicial Homotopy Theory* (Birkhäuser, 2009), https://doi.org/10.1007/978-3-0346-0189-4; and J. Peter May, *Simplicial Objects in Algebraic Topology* (University of Chicago Press, 1992).

28 Friedman, "The Structure and Function of Complex Networks."

29 Sinan G. Aksoy et al., "Hypernetwork Science via High-Order Hypergraph Walks," *EPJ Data Science* 9, no. 1 (2020): 16, https://doi.org/10.1140/epjds/s13688-020-00231-0.

30 Aksoy et al., "Hypernetwork Science via High-Order Hypergraph Walks."

31 Aksoy et al., "Hypernetwork Science via High-Order Hypergraph Walks."

32 Steve Kirkland, "Two-Mode Networks Exhibiting Data Loss," *Journal of Complex Networks* 6, no. 2 (2018): 297–316, https://doi.org/10.1093/comnet/cnx039.

33 Newman, *Networks: An Introduction,* 1st ed.; Newman, "The Structure and Function of Complex Networks;" and Wasserman and Faust, *Social Network Analysis: Methods and Applications*.

34 Komal Kapoor et al., "Weighted Node Degree Centrality for Hypergraphs," 2013 IEEE 2nd Network Science Workshop (NSW), (2013): 152–55, https://doi.org/10.1109/NSW.2013.6609212.

35 Kapoor et al., "Weighted Node Degree Centrality for Hypergraphs."

36 Menczer et al., *A First Course in Network Science;* and Wasserman and Faust, *Social Network Analysis: Methods and Applications*.

37 Aksoy et al., "Hypernetwork Science via High-Order Hypergraph Walks."

38 Aksoy et al., "Hypernetwork Science via High-Order Hypergraph Walks;" and Matthew J. Hasenjager and Nina H. Fefferman, "Social Ageing and

Higher-Order Interactions: Social Selectiveness Can Enhance Older Individuals' Capacity To Transmit Knowledge," *Philosophical Transactions of the Royal Society B: Biological Sciences* 379 (2024): 2022.0461, https://doi.org/10.1098/rstb.2022.0461.

39 Ernesto Estrada and Grant J. Ross, "Centralities in Simplicial Complexes. Applications to Protein Interaction Networks," *Journal of Theoretical Biology* 438 (2018): 46–60, https://doi.org/10.1016/j.jtbi.2017.11.003; and Daniel H. Serrano and Darío S. Gómez, "Centrality Measures in Simplicial Complexes: Applications of Topological Data Analysis to Network Science," Cornell University Algebraic Topology, *arXiv* (2020), 1908.02967, https://doi.org/10.48550/arXiv.1908.02967.

40 Phillip Bonacich, "Power and Centrality: A Family of Measures," *American Journal of Sociology* 92, no. 5 (1987): 1170–82, https://doi.org/10.1086/228631.

41 Amy N. Langville and Carl D. Meyer, "Deeper Inside PageRank," *Internet Mathematics* 1, no. 3 (2004): 335–80, https://doi.org/10.1080/15427951.2004.10129091.

42 Phillip Bonacich, "Simultaneous Group and Individual Centralities," *Social Networks* 13, no. 2 (1991): 155–68, https://doi.org/10.1016/0378-8733(91)90018-O.

43 Sinan G. Aksoy et al., "Scalable Tensor Methods for Nonuniform Hypergraphs," *SIAM Journal on Mathematics of Data Science* 6, no. 2 (2024): 481–503, https://doi.org/10.1137/23M1584472; and Austin R. Benson, "Three Hypergraph Eigenvector Centralities," *SIAM Journal on Mathematics of Data Science* 1, no. 2 (2019): 293–312, https://doi.org/10.1137/18M1203031.

44 Greening, "Higher-Order Analysis of Knowledge Capacity and Learning Potential in Social Animal Groups."

45 Aksoy et al., "Hypernetwork Science via High-Order Hypergraph Walks."

46 Lazaros K. Gallos and Nina H. Fefferman, "Revealing Effective Classifiers Through Network Comparison," *Europhysics Letters* 108, no. 3 (2014): 38001, https://doi.org/10.1209/0295-5075/108/38001.

47 Greening, "Higher-Order Analysis of Knowledge Capacity and Learning Potential in Social Animal Groups;" Hasenjager and Fefferman, "Social Ageing and Higher-Order Interactions: Social Selectiveness Can Enhance Older Individuals' Capacity To Transmit Knowledge;" and Silk et al., "Capturing Complex Interactions in Disease Ecology with Simplicial Sets."

48 Aksoy et al., "Hypernetwork Science via High-Order Hypergraph Walks."

49 Ron Milo et al., "Superfamilies of Evolved and Designed Networks," *Science* 303, no 5663 (2004): 1538–42, https://doi.org/10.1126/science.1089167.

50 Shmoolik Mangan and Uri Alon, "Structure and Function of the Feed-Forward Loop Network Motif," *Proceedings of the National Academy of Sciences* 100, no. 21 (2003): 11980–85, https://doi.org/10.1073/pnas.2133841100; and Milo et al., "Superfamilies of Evolved and Designed Networks."

51 Sinan G. Aksoy, "An Invitation to Hypergraph Motifs for Practitioners," in *Higher-Order Network Methods for Intelligence Applications: A Compendium for Practitioners,* eds. Matthew J. Hasenjager, Mark M. Bailey, and Nina H. Fefferman, (National Intelligence Press, 2026).

52 Nicholas W. Landry and Juan G. Restrepo, "Hypergraph Assortativity: A Dynamical Systems Perspective," *Chaos: An Interdisciplinary Journal of Nonlinear Science* 32, no. 5 (2022): 053113, https://doi.org/10.1063/5.0086905; and M. E. J. Newman, "Assortative Mixing in Networks," *Physical Review Letters* 89, no. 20 (2002): 208701, https://doi.org/10.1103/PhysRevLett.89.208701.

53 Michelle Girvan and M. E. J. Newman, "Community Structure in Social and Biological Networks," *Proceedings of the National Academy of Sciences* 99, no. 12 (2002): 7821–26, https://doi.org/10.1073/pnas.122653799.

54 Philip S. Chodrow et al., "Generative Hypergraph Clustering: From Blockmodels to Modularity," *Science Advances* 7, no. 28 (2021): 1303, https://doi.org/10.1126/sciadv.abh1303; Bogumił Kamiński et al., "Clustering via Hypergraph Modularity," *PLOS One* 14, no. 11 (2019): 0224307, https://doi.org/10.1371/journal.pone.0224307; Bogumił Kamiński et al., "Modularity Based Community Detection in Hypergraphs," *Journal of Complex Networks* 12, no. 5 (2024): cnae041, https://doi.org/10.1093/comnet/cnae041; and Tarun Kumar et al., "Hypergraph Clustering by Iteratively Reweighted Modularity Maximization," *Applied Network Science* 5 (2020): 52, https://doi.org/10.1007/s41109-020-00300-3.

55 Luca Brusa and Catherine Matias, "Model-Based Clustering in Simple Hypergraphs Through a Stochastic Blockmodel," Cornell University Algebraic Topology, *arXiv* (2024), 2210.05983, https://doi.org/10.48550/arXiv.2210.05983; and Chodrow et al., "Generative Hypergraph Clustering: From Blockmodels to Modularity.

56 Brusa and Matias, "Model-Based Clustering in Simple Hypergraphs Through a Stochastic Blockmodel."

57 Seth Falcon and Robert Gentleman, "Hypergraph: A Package Providing Hypergraph Data Structures, Version R," package version 1.78.0 (2024), https://doi.org/10.18129/B9.bioc.hypergraph.

58 David J. Marchette, "HyperG: Hypergraphs in R, Version R," package version 1.0.0 (2021), https://CRAN.R-project.org/package=HyperG.

59 Nicholas W. Landry et al., "XGI: A Python Package for Higher-Order Interaction Networks," *Journal of Open Source Software* 8, no. 85 (2023): 05162, https://doi.org/10.21105/joss.05162.

60 Brenda Praggastis et al., "HyperNetX: A Python Package for Modeling Complex Network Data as Hypergraphs," *Journal of Open Source Software* 9, no. 95 (2024): 06016, https://doi.org/10.21105/joss.06016.

61 Quintino Francesco Lotito et al., "Hypergraphx: A Library for Higher-Order Network Analysis," Cornell University Algebraic Topology, *arXiv* (2023): 2303.15356, http://arxiv.org/abs/2303.15356.

62 Sebastian Schlag et al., "High-Quality Hypergraph Partitioning," *ACM Journal of Experimental Algorithmics* 27 (2023): 1–39, https://doi.org/10.1145/3529090.

63 Arash Mahboubi et al., "Evolving Techniques in Cyber Threat Hunting: A Systematic Review," *Journal of Network and Computer Applications* (2024): 104004.

64 Min Du et al., "Deeplog: Anomaly Detection and Diagnosis from System Logs through Deep Learning," *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security* (2017): 1285–98.

65 Fucheng Liu et al., "Log2vec: A Heterogeneous Graph Embedding Based Approach for Detecting Cyber Threats within Enterprise," *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security* (2019): 1777–94.

66 Dhruv Nandakumar, "Zero Day Threat Detection Using Metric Learning Autoencoders," 2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA), (2022): 1318–25.

67 Christopher Redino, "Zero Day Threat Detection Using Graph and Flow Based Security Telemetry," 2022 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS), (2022): 655–62.

68 Sinan G. Aksoy et al., "Relative Hausdorff Distance for Network Analysis," *Applied Network Science* 4 (2019): 1–25.

69 Sinan G. Aksoy, Emilie Purvine, and Stephen J. Young, "Directional Laplacian Centrality for Cyber Situational Awareness," *Digital Threats: Research and Practice (DTRAP)* 2, no. 4 (2021): 1–28.

70 Pin-Yu Chen, Sutanay Choudhury, and Alfred O. Hero, "Multi-Centrality Graph Spectral Decompositions and Their Application to Cyber Intrusion Detection," 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), (2016): 4553–57.

71 Eric Dull, "Cyberthreat Analytics Using Graph Analysis," *The Analyst* 5, no. 11 (2020).

72 Van-Hoang Le and Hongyu Zhang, "Log-Based Anomaly Detection with Deep Learning: How Far Are We?" *Proceedings of the 44th International Conference on Software Engineering* (2022): 1356–67.
73 "Operationally Transparent Cyber (OpTC) Data Release," Defense Advanced Research Projects Agency (DARPA) (2019), https://github.com/FiveDirections/OpTC-data/.
74 "Cyber Analytics Repository Model," MITRE (2019), https://car.mitre.org/data_model/.
75 Md Monowar Anjum, Shahrear Iqbal, and Benoit Hamelin, "Analyzing the Usefulness of the DARPA OpTC Dataset in Cyber Threat Detection Research," *Proceedings of the 26th ACM Symposium on Access Control Models and Technologies* (2021): 27–32.
76 Thomas Cochrane et al., "SK-tree: A Systematic Malware Detection Algorithm on Streaming Trees Via the Signature Kernel," 2021 IEEE International Conference on Cyber Security and Resilience (CSR), (2021): 35–40.
77 Mohammad Mamun and Kevin Shi, "DeepTaskAPT: Insider APT Detection Using Task-Tree Based Deep Learning," IEEE 20th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom), (2021): 693–700.
78 Md Monowar Anjum, Shahrear Iqbal, and Benoit Hamelin, "Anubis: A Provenance Graph-Based Framework for Advanced Persistent Threat Detection," *Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing* (2022):1684–93.
79 Zijun Cheng et al., "Kairos: Practical Intrusion Detection and Investigation Using Whole-System Provenance," 2024 IEEE Symposium on Security and Privacy (SP), (2024): 3533–51.
80 Andrew Golczynski and John A. Emanuello, "End-To-End Anomaly Detection for Identifying Malicious Cyber Behavior Through NLP-Based Log Embeddings," Cornell University Algebraic Topology, *arXiv* (2021), 2108.12276, https://arxiv.org/abs/2108.12276.
81 Ramesh Paudel and H. Howie Huang, "Pikachu: Temporal Walk Based Dynamic Graph Embedding for Network Anomaly Detection," *NOMS 2022–2022 IEEE/IFIP Network Operations and Management Symposium* (2022): 1–7.
82 Victor Nikulshin and Chamseddine Talhi, "Effective IDS Under Constraints of Modern Enterprise Networks: Revisiting (2024).
83 Audun Myers et al., "Topological Analysis of Temporal Hypergraphs," *International Workshop on Algorithms and Models for the Web-Graph*, Springer (2023): 127–46.

84 Myers et al., "Topological Analysis of Temporal Hypergraphs," 6.
85 Aksoy et al., "Hypernetwork Science via High-Order Hypergraph Walks."
86 Helen Jenne et al., "Stepping Out of Flatland: Discovering Behavior Patterns as Topological Structures in Cyber Hypergraphs," *The Next Wave* 25, no. 1 (2024).
87 Herbert Edelsbrunner and John Harer, *Computational Topology: An Introduction* (American Mathematical Society, 2010).
88 Robert W. Ghrist, *Elementary Applied Topology* 1 (Createspace, 2014).
89 Allen Hatcher, *Algebraic Topology* (Cambridge University Press, 2000).
90 Ellen Gasparovic et al., "A Survey of Simplicial, Relative, and Chain Complex Homology Theories for Hypergraphs," Cornell University Algebraic Topology, *arXiv* (2024), 2409.18310, https://arxiv.org/abs/2409.18310.
91 Gasparovic et al., "A Survey of Simplicial, Relative, and Chain Complex Homology Theories for Hypergraphs," 24.
92 Gasparovic et al., "A Survey of Simplicial, Relative, and Chain Complex Homology Theories for Hypergraphs," 29.
93 Christopher Potvin, *Hypergraphs and Their Homology,* Ph.D. diss., Michigan State University (2023).
94 Potvin, *Hypergraphs and Their Homology,* 29.
95 Audun Myers et al., "Malicious Cyber Activity Detection Using Zigzag Persistence," 2023 IEEE Conference on Dependable and Secure Computing (DSC), (2023): 1–8.
96 Myers et al., "Malicious Cyber Activity Detection," 29.
97 Myers et al., "Malicious Cyber Activity Detection," 32.
98 Myers et al., "Malicious Cyber Activity Detection," 24.
99 Myers et al., "Malicious Cyber Activity Detection," 34.
100 Uthsav Chitra and Benjamin Raphael, "Random Walks on Hypergraphs with Edge-Dependent Vertex Weights," in International Conference on Machine Learning, *Proceedings of Machine Learning Research* (2019): 1172–81.
101 Koby Hayashi et al., "Hypergraph Random Walks, Laplacians, and Clustering," in *Proceedings of the 29th ACM International Conference on Information and Knowledge Management* (2020): 495–504.
102 Aksoy et al., "Scalable Tensor Methods for Nonuniform Hypergraphs."
103 Benson, "Three Hypergraph Eigenvector Centralities."
104 Kapoor et al., "Weighted Node Degree Centrality for Hypergraphs."
105 Ernesto Estrada and Juan A. Rodríguez-Velázquez, "Subgraph Centrality and Clustering in Complex Hyper-Networks," *Physica A: Statistical Mechanics and Its Applications* 364 (2006): 581–94.

106 Sameer Agarwal et al., "Higher Order Learning with Graphs," in *Proceedings of the 23rd International Conference on Machine Learning* (2006): 17–24.
107 Kirkland, "Two-Mode Networks Exhibiting Data Loss."
108 Agarwal et al., "Higher Order Learning with Graphs."
109 Aksoy et al., "Hypernetwork Science via High-Order Hypergraph Walks."
110 Geon Lee et al., "A Survey on Hypergraph Mining: Patterns, Tools, and Generators," Cornell University Algebraic Topology, *arXiv* (2024), 2401.08878, https://arxiv.org/abs/2401.08878.
111 Aksoy et al., "Hypernetwork Science via High-Order Hypergraph Walks."
112 Garry Robins and Malcolm Alexander, "Small Worlds Among Interlocking Directors: Network Structure and Distance in Bipartite Graphs," *Computational and Mathematical Organization Theory* 10 (2004): 69–94.
113 Robins and Alexander, "Small Worlds Among Interlocking Directors: Network Structure and Distance in Bipartite Graphs."
114 Sinan G. Aksoy et al., "Measuring and Modeling Bipartite Graphs with Community Structure," *Journal of Complex Networks* 5, no. 4 (2017): 581–603.
115 Aksoy et al., "Hypernetwork Science via High-Order Hypergraph Walks."
116 Tyler Derr et al., "Balance in Signed Bipartite Networks," in *Proceedings of the 28th ACM International Conference on Information and Knowledge Management* (2019): 1221–30.
117 Aksoy et al., "Measuring and Modeling Bipartite Graphs with Community Structure."
118 Geon Lee et al., "Hypergraph Motifs: Concepts, Algorithms, and Discoveries," Cornell University Algebraic Topology, *arXiv* (2020), 2003.01853, https://arxiv.org/abs/2003.01853.
119 Jason Niu et al., "Retrieving Top-k Hyperedge Triplets: Models and Applications," in IEEE International Conference on Big Data (2024).
120 Niu et al., "Retrieving Top-k Hyperedge Triplets."
121 Niu et al., "Retrieving Top-k Hyperedge Triplets."
122 Niu et al., "Retrieving Top-k Hyperedge Triplets."
123 Jianmo Ni et al., "Justifying Recommendations Using Distantly Labeled Reviews and Fine-Grained Aspects," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing* (EMNLP-IJCNLP), (2019), 188–197.
124 F. Maxwell Harper and Joseph A. Konstan, "The MovieLens Datasets: History and Context," *ACM Transactions on Interactive Intelligent Systems* (TIIS) 5, no. 4 (2015): 1–19.

125 Aksoy et al., "Measuring and Modeling Bipartite Graphs with Community Structure."

126 Niu et al., "Retrieving Top-k Hyperedge Triplets: Models and Applications."

127 Bodhisattwa Prasad Majumder et al., "Generating Personalized Recipes from Historical User Preferences," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing* (EMNLP-IJCNLP), (2019): 5976–82.

128 Seyed-Vahid Sanei-Mehri et al., "Butterfly Counting in Bipartite Networks," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2018): 2150–59.

129 Comandur Seshadhri et al., "Triadic Measures on Graphs: The Power of Wedge Sampling," in *Proceedings of the 2013 SIAM International Conference on Data Mining* (2013): 10–18.

130 Aksoy et al., "Measuring and Modeling Bipartite Graphs with Community Structure."

131 Kamiński et al., "Clustering via Hypergraph Modularity."

132 Aksoy et al., "Measuring and Modeling Bipartite Graphs with Community Structure."

133 Ioana Dumitriu et al., "Partial Recovery and Weak Consistency in the Non-Uniform Hypergraph Stochastic Block Model," *Combinatorics, Probability and Computing* (2021): 1–51.

134 Arjun S. Ramani et al., "Coin-Flipping, Ball-Dropping, and Grass-Hopping for Generating Random Graphs from Matrices of Edge Probabilities," *SIAM Review* 61, no. 3 (2019): 549–95.

135 Oliver Cooley et al., "Subcritical Random Hypergraphs, High-Order Components, and Hypertrees," *SIAM Journal on Discrete Mathematics* 34, no. 4 (2020): 2033–62.

136 Oliver Cooley et al., "Evolution of High-Order Connected Components in Random Hypergraphs," *Electronic Notes in Discrete Mathematics* 49 (2015): 569–75.

137 Oliver Cooley et al., "The Size of the Giant High-Order Component in Random Hypergraphs," *Random Structures and Algorithms* 53, no. 2 (2018): 238–88.

138 D. Stokols et al., "The Science of Team Science: Overview of the Field and Introduction to the Supplement," *American Journal of Preventive Medicine* 35 (2, Suppl.) (2008): S77–S89.

139 Barry Bozeman and Craig Boardman, *Research Collaboration and Team Science: A State-of-the-Art Review and Agenda,* (Springer International Publishing (2014), https://doi.org/10.1007/978-3-319-06468-0.

140 Yuxian Liu et al., "Reflections On and a Short Review of the Science of Team Science," *Scientometrics* 125 (2020): 937–50.

141 John Meluso and Laurent Hébert-Dufresne, "Multidisciplinary Learning Through Collective Performance Favors Decentralization," in *Proceedings of the National Academy of Sciences* 120, no. 34 (2023): 2303568120, https://doi.org/10.1073/pnas.2303568120.

142 Naomi Ehrich Leonard and Simon A. Levin, "Collective Intelligence as a Public Good," *Collective Intelligence* 1, no. 1 (2022): https://naomi.princeton.edu/wp-content/uploads/sites/744/2022/09/Leonard2022Collective-1.pdf.

143 J. Richard Hackman, *Collaborative Intelligence: Using Teams to Solve Hard Problems* (Berrett-Koehler Publishers, 2011).

144 Kathleen M. Vogel and Beverly B. Tyler, "Interdisciplinary, Cross-Sector Collaboration in the US Intelligence Community: Lessons Learned from Past and Present Efforts," *Intelligence and National Security* 34, no. 6 (2019): 851–80.

145 Stefan Wuchty et al., "The Increasing Dominance of Teams in Production of Knowledge," *Science* 316, no. 5827 (2007): 1036–39.

146 Roger Guimera et al., "Team Assembly Mechanisms Determine Collaboration Network Structure and Team Performance," *Science* 308, no. 5722 (2005): 697–702.

147 Alina Lungeanu et al., "Understanding the Assembly of Interdisciplinary Teams and Its Impact on Performance," *Journal of Infometrics* 8, no. 1 (2014): 59–70.

148 Guimera et al., "Team Assembly Mechanisms Determine Collaboration Network Structure and Team Performance."

149 Lungeanu et al., "Understanding the Assembly of Interdisciplinary Teams and Its Impact on Performance."

150 Richard B. Freeman et al., "Why and Wherefore of Increased Scientific Collaboration," in *The Changing Frontier: Rethinking Science and Innovation Policy* (University of Chicago Press, 2014): 17–48.

151 Lungeanu et al., "Understanding the Assembly of Interdisciplinary Teams and Its Impact on Performance."

152 Emma R. Zajdela et al., "Dynamics of Social Interaction: Modeling the Genesis of Scientific Collaboration," *Phys. Rev. Res.* 4, no. 4 (October 2022): L042001, https://doi.org/10.1103/PhysRevResearch.4.L042001.

153 Emma R. Zajdela et al., "Face-to-Face or Face-to-Screen: A Quantitative Comparison of Conferences Modalities," *PNAS Nexus* 4, no. 1 (2025): 522.

154 Kevin J Boudreau et al., "A Field Experiment on Search Costs and the Formation of Scientific Collaborations," *Review of Economics and Statistics* 99, no. 4 (2017): 565–76.

155 Jacqueline N. Lane et al., "Engineering Serendipity: When Does Knowledge Sharing Lead to Knowledge Production?" *Strategic Management Journal* 42, no. 6 (2021): 1215–44.

156 Nicholas W Landry et al., "Filtering Higher-Order Datasets," *Journal of Physics: Complexity* 5, no. 1 (2024): 015006.

157 Alice Patania et al., "The Shape of Collaborations," *EPJ Data Science* 6, no. 18 (2017), https://doi.org/10.1140/epjds/s13688-017-0114-8.

158 Patania et al., "The Shape of Collaborations."

159 Iacopo Iacopini et al., "The Temporal Dynamics of Group Interactions in Higher-Order Social Networks," *Nature Communications* 15, no. 1 (2024): 7391.

160 R. J. Wiener and S. Ronco, "Scialog: The Catalysis of Convergence," *ACS Energy Letters* 4, no. 5 (2019): 1020–24.

161 Emma R. Zajdela et al., "Spontaneous Synchronization in Molecular Systems." *Physical Review Research* 4 (2022): L042001.

162 Wiener and Ronco, "Scialog: The Catalysis of Convergence."

163 Zajdela et al., "Face-to-Face or Face-to-Screen: A Quantitative Comparison of Conferences Modalities."

164 Landry et al., "XGI: A Python Package for Higher-Order Interaction Networks."

165 H. Akaike, "A New Look at the Statistical Model Identification," *IEEE Transactions on Automatic Control* 19, no. 6 (1974): 716–23.

166 Georg Simmel, *Soziologie* (Duncker & Humblot, 1908).

167 Iacopini et al. "The Temporal Dynamics."

168 Feng Shi and James Evans, "Surprising Combinations of Research Contents and Contexts Are Related to Impact and Emerge with ScientificOutsiders from Distant Disciplines," *Nature Communications* 14, no. 1 (2023): 1641.

169 F. Shi and J. Evans. "The Impact of Interdisciplinary Collaboration on Research Outcomes," *Nature Communications* 14 (2023): 1641.

170 Hakan Kardes et al., "Complex Network Analysis of Research Funding: A Case Study of NSF Grants," *State of the Art Applications of Social Network Analysis*, 2014, 163–87.

171 Kazuki Nakajima et al., "Higher-Order Rich-Club Phenomenon in Collaborative Research Grant Networks," *Scientometrics* 128, no. 4 (2023): 2429–46.

172 Zajdela et al., "Face-to-Face or Face-to-Screen: A Quantitative Comparison of Conferences Modalities."

173 Zajdela et al., "Dynamics of Social Interaction: Modeling the Genesis of Scientific Collaboration."

174 Zajdela et al., "Dynamics of Social Interaction: Modeling the Genesis of Scientific Collaboration."

175 Melanie S. Brucks and Jonathan Levav, "Virtual Communication Curbs Creative Idea Generation," *Nature* 605, no. 7908 (2022): 108–12.

176 Yiling Lin et al., "Remote Collaboration Fuses Fewer Breakthrough Ideas," *Nature* 623, no. 7989 (2023): 987–91.

177 Zajdela et al., "Face-to-Face or Face-to-Screen: A Quantitative Comparison of Conferences Modalities."

178 Yanqiu Tao et al., "Trend Towards Virtual and Hybrid Conferences May Be an Effective Climate Change Mitigation Strategy," *Nature Communications* 12, no. 1 (December 2021): 7324.

179 D. Sandy Staples and Lina Zhao, "The Effects of Cultural Diversity in Virtual Teams versus Face-to-Face Teams," *Group Decision and Negotiation* 15 (2006): 389–406.

180 Juncheng Wu et al., "Virtual Meetings Promise to Eliminate Geographical and Administrative Barriers and Increase Accessibility, Diversity and Inclusivity," *Nature Biotechnology* 40, no. 1 (2022): 133–37.

181 Matthew Skiles et al., "Conference Demographics and Footprint Changed by Virtual Platforms," *Nature Sustainability* 5, no. 2 (2022): 149–56.

182 Emma R. Zajdela and Nicholas W. Landry, "Higher-Order-Team-Formation: V0.0," (Zenodo, 2025), https://doi.org/10.5281/zenodo.14867591.

183 Guilbeault and Centola, "Topological Measures for Identifying and Predicting the Spread of Complex Contagions;" Guillaume St-Onge et al., "Influential Groups for Seeding and Sustaining Nonlinear Contagion in Heterogeneous Hypergraphs," *Communications Physics* 5, no. 1 (2022): 25, https://doi.org/10.1038/s42005-021-00788-w; and Ming Xie et al., "Influence Maximization in Hypergraphs," Cornell University Algebraic Topology, *arXiv* (2022), 2206.01394, http://arxiv.org/abs/2206.01394.

184 Christine Bassem, "Redefining Node Centrality for Task Allocation in Mobile CrowdSensing Platforms," in IEEE International Conference on Smart Computing (SMARTCOMP), 2019, 323–31, https://doi.org/10.1109/SMARTCOMP.2019.00069; and Daniel R. Wuellner et al., "Resilience and Rewiring of the Passenger Airline Networks in the United States," *Physical Review* E 82, no. 5 (2010): 056101, https://doi.org/10.1103/PhysRevE.82.056101.

185 Marcel Salathé and James H. Jones, "Dynamics and Control of Diseases in Networks with Community Structure," *PLoS Computational Biology* 6, no. 4 (2010): e1000736, https://doi.org/10.1371/journal.pcbi.1000736.

186 Matthew J. Hasenjager et al., "Higher-Order Networks: A Primer," in *Higher-Order Network Methods for Intelligence Applications: A Compendium for*

Practitioners, eds. Matthew J. Hasenjager, Mark M. Bailey, and Nina H. Fefferman (National Intelligence Press, 2026).

187 Newman, *Networks: An Introduction,* 1st ed.; Newman, "The Structure and Function of Complex Networks;" and Wasserman and Faust, *Social Network Analysis: Methods and Applications*.

188 Sinan G. Aksoy et al., "Hypernetwork Science via High-Order Hypergraph Walks." *EPJ Data Science* 9, no. 1 (2020): 16, https://doi.org/10.1140/epjds/s13688-020-00231-0; Battiston et al., "Networks Beyond Pairwise Interactions;" and Estrada and Rodríguez-Velázquez, "Subgraph Centrality and Clustering in Complex Hyper-Networks."

189 Aksoy et al., "Hypergraph Science;" Song Feng et al. "Hypergraph Models of Biological Networks to Identify Genes Critical to Pathogenic Viral Response," *BMC Bioinformatics* 22, no. 1 (2021): 287, https://doi.org/10.1186/s12859-021-04197-2; and Hasenjager and Fefferman, "Social Ageing and Higher-Order Interactions: Social Selectiveness Can Enhance Older Individuals' Capacity To Transmit Knowledge."

190 Aksoy et al., "Hypergraph Science."

191 Martin G. Everett and Stephen P. Borgatti, "The Dual-Projection Approach for Two-Mode Networks," *Social Networks* 35, no. 3 (2201): 204–10, https://doi.org/10.1016/j.socnet.2012.05.004.

192 Aksoy et al., "Hypergraph Science."

193 Aksoy et al., "Hypergraph Science;" and Steve Kirkland, "Two-Mode Networks Exhibiting Data Loss."

194 Thomas D. Pike, "Cracking the Paradigm: Hypergraph Signatures of International Conflict," Chap. 6 in this book.

195 Zong-Zhi Lin, "Hypergraph Methods for Network Analysis and Intrusion Detection," Chap. 8 in this book.

196 Aksoy et al., "Hypergraph Science;" and Feng et al., "Hypergraph Models of Biological Networks to Identify Genes Critical to Pathogenic Viral Response;" and Hasenjager and Fefferman, "Social Ageing and Higher-Order Interactions: Social Selectiveness Can Enhance Older Individuals' Capacity To Transmit Knowledge."

197 Ciro Cattuto et al., "Dynamics of Person-to-Person Interactions from Distributed RFID Sensor Networks," *PLoS ONE* 5, no. 7 (2010): e11596, https://doi.org/10.1371/journal.pone.0011596; Mathieu Génois and Alain Barrat, "Can Co-Location Be Used as a Proxy for Face-to-Face Contacts?" *EPJ Data Science* 7, no. 1 (2018): 11, https://doi.org/10.1140/epjds/s13688-018-0140-1.

198 Cattuto et al., "Dynamics of Person-to-Person Interactions;" and Génois and Barrat, "Can Co-Location Be Used as a Proxy for Face-to-Face Contacts?"

199 Matthew J. Hasenjager et al., "Detecting and Quantifying Social Transmission Using Network-Based Diffusion Analysis," *Journal of Animal Ecology* 90, no. 1 (2020): 8–26, https://doi.org/10.1111/1365-2656.13307 .

200 Hasenjager et al., "Detecting and Quantifying Social Transmission Using Network-Based Diffusion Analysis;" and Silk et al., "Capturing Complex Interactions in Disease Ecology with Simplicial Sets."

201 James O. Lloyd-Smith et al., "Superspreading and the Effect of Individual Variation on Disease Emergence," *Nature* 438, no. 7066 (2005): 355–59, https://doi.org/10.1038/nature04153; and Salathé and Jones, "Dynamics and Control of Diseases in Networks with Community Structure."

202 Centola, "The Spread of Behavior in an Online Social Network Experiment;" Iacopini et al., "Simplicial Models of Social Contagion;" and St-Onge et al., "Influential Groups for Seeding and Sustaining Nonlinear Contagion."

203 Firth, "Considering Complexity;" and Guilbeault et al., "Complex Contagions."

204 Iacopini et al.; "Simplicial Models of Social Contagion."

205 St-Onge et al., "Influential Groups for Seeding and Sustaining Nonlinear Contagion."

206 Kalev H. Leetaru and Philip A. Schrodt, "GDELT: Global Database of Events, Language, and Tone, 1979-Present," 2013, https://www.gdeltproject.org/.

207 M. B. M. de Koster, and J. H. P. Paelinck, "A Hypergraph Approach to Conflict," *Conflict Management and Peace Science* 7, no. 2 (1984): 55–69, https://doi.org/10.1177/073889428400700204; R. van Gastel and J. H. P. Paelinck, "Hypergraph Conflict Analysis: Synthesis and Extension." *Conflict Management and Peace Science* 10, no. 1 (1988): 59–86; and M. A. J. J. van Gastel and J. H. P. Paelinck, "Hypergraph Conflict Analysis," *Economics Letters* 35, no. 3 (1991): 233–37, https://doi.org/10.1016/0165-1765(91)90135-8.

208 De Koster and Paelinck, "A Hypergraph Approach to Conflict."

209 M. A. J. J. Van Gastel and J. H. P. Paelinck, "Hypergraph Conflict Analysis."

210 Van Gastel and Paelinck, "Hypergraph Conflict Analysis."

211 Elie Alhajjar and Ross Friar, "A Historical Perspective on International Treaties via Hypernetwork Science," in *Network Science,* eds. Pedro Ribeiro, Fernando Silva, José Fernando Mendes, and Rosário Laureano, (Springer International Publishing, 2022): 15–25.

212 Melvin Small and J. David Singer, "Formal Alliances, 1815–1965: An Extension of the Basic Data," *Journal of Peace Research* 6, no. 3 (1969): 257–82.; and Douglas M. Gibler, *International Military Alliances* (CQ Press, 2009): 1648–2008.

213 Alhajjar and Friar, "A Historical Perspective on International Treaties via Hypernetwork Science;" and Sudo Yi and Deok-Sun Lee, "Structure of

International Trade Hypergraphs," *Journal of Statistical Mechanics: Theory and Experiment,* no. 10 (2022): 103402, https://doi.org/10.1088/1742-5468/ac946f.

214 Newman, "The Structure and Function of Complex Networks;" Albert and Barabási, "Statistical Mechanics of Complex Networks;" and Duncan J. Watts and Steven H. Strogatz, "Collective Dynamics of 'Small-World' Networks," *Nature* 393 (1998): 440–442, https://doi.org/10.1038/30918.

215 Alhajjar and Friar, "A Historical Perspective on International Treaties via Hypernetwork Science."

216 Yi and Lee, "Structure of International Trade Hypergraphs."

217 Leetaru and Schrodt. "GDELT: Global Database of Events, Language, and Tone, 1979–Present."

218 "Jigsaw," https://jigsaw.google/, accessed February 2, 2025.

219 Philip Schrodt and Jay Yonamine, "A Guide to Event Data: Past, Present, and Future," *All Azimuth: A Journal of Foreign Policy and Peace* 2, no. 2 (2013): 5–22, https://doi.org/10.20991/allazimuth.167312.

220 "The GDELT Project," https://www.gdeltproject.org/, accessed February 2, 2025.

221 Cliff A. Joslyn et al., "Hypernetwork Science: From Multidimensional Networks to Computational Topology," in Unifying Themes in Complex Systems X, eds. Dan Braha et al. (Springer International Publishing, 2021): 377–392; and Praggastis et al., "HyperNetX: A Python Package for Modeling Complex Network Data as Hypergraphs."

222 GDELT Project. 2015. GDELT Event Database Data Format Codebook V2.0. http://gdeltproject.org/; and Schrodt and Yonamine, "A Guide to Event Data: Past, Present, and Future."

223 Praggastis et al., "HyperNetX: A Python Package for Modeling Complex Network Data as Hypergraphs."

224 Maximilian Viermetz and Michal Skubacz, "Using Topic Discovery To Segment Large Communication Graphs for Social Network Analysis," in IEEE/WIC/ACM International Conference on Web Intelligence (2007): 95–99; Alan Mislove et al., "Measurement and Analysis of Online Social Networks," in ACM/SIGCOMM Internet Measurement Conference (2007); and Quoc-Dinh Truong, Taoufiq Dkaki, and Quoc-Bao Truong, "Graph Methods for Social Network Analysis," *International Conference on Nature of Computation and Communication,* 2016.

225 Aditya Gowdara Shivamurthy, "Kashmir and Ontological Security: Re-Evaluating the Role of Self-Identities in a Multi-Layered Conflict," *Journal of Asian Security and International Affairs* 9 (2022): 255–79; and Palvi

Shelke and Timo Hamalainen, "Analysing Multidimensional Strategies for Cyber Threat Detection in Security Monitoring," European Conference on Cyber Warfare and Security, 2024.

226 Battiston et al., "Networks Beyond Pairwise Interactions;" and Boicu et al., "Cognitive Assistants for Evidence-Based Reasoning," AAAI Fall Symposium, 2011.

227 Bahare Fatemi et al., "Knowledge Hypergraphs: Prediction Beyond Binary Relations," International Joint Conference on Artificial Intelligence, 2019; Aksoy et al., "Hypernetwork Science via High-Order Hypergraph Walks;" Battiston et al., "Networks Beyond Pairwise Interactions;" and Corinna Coupette, Jilles Vreeken, and Bastian Rieck, "All the World's a (Hyper)graph: A Data Drama," *Digital Scholarship in the Humanities* 39, no. 1 (2024): 74–96, https://doi.org/10.1093/llc/fqad071.

228 Miranda Boettcher, Johannes Gabriel, and Sean Low, "Solar Radiation Management: Foresight for Governance, IASS Working Paper," *Institute for Advanced Sustainability Studies (IASS)*, 2016.

229 B. Malekmohammadi and F. Jahanishakib, "Vulnerability Assessment of Wetland Landscape Ecosystem Services Using Driver-Pressure-State-Impact-Response (DPSIR) Model," *Ecological Indicators* 82 (2017): 293–303.

230 Iver Johansen, "Scenario Modelling with Morphological Analysis," *Technological Forecasting and Social Change* 126 (2017): 116–125, https://doi.org/10.1016/j.techfore.2017.05.016.

231 Benjamin P. Bryant and Robert J. Lempert, "Thinking Inside the Box: A Participatory, Computer-Assisted Approach to Scenario Discovery," *Technological Forecasting and Social Change* 77, no. 1 (2010): 34–49, https://www.sciencedirect.com/science/article/abs/pii/S004016250900105X; and Sebastiaan Greeven et al., "The Emergence of Climate Change Mitigation Action by Society: An Agent-Based Scenario Discovery Study," *Journal of Artificial Societies and Social Simulation* 19 (2016).

232 Amber Geurts et al., "New Perspectives for Data-Supported Foresight: The Hybrid AI-Expert Approach," *Futures & Foresight Science,* 2021.

233 Kyle A. Kilian, Christopher J. Ventura, and Mark M. Bailey, "Examining the Differential Risk From High-Level Artificial Intelligence and the Question of Control," Cornell University Algebraic Topology, *arXiv* preprint, *arXiv* (2022), 2211.03157.

234 Eric C. Anderson, "An Efficient Monte Carlo Method for Estimating Ne from Temporally Spaced Samples Using a Coalescent-Based Likelihood," *Genetics* 170 (2005): 955–67, C. P. Fries; "Foresight Bias and Suboptimality Correction in Monte Carlo Pricing of Options With Early Exercise: Classification,

Calculation, and Removal," *Derivatives*, 2005; D. Stauffer, *Monte Carlo Simulations in Statistical Physics*, Springer: 1988.

235 Carolin Strobl, James D. Malley, and Gerhard Tutz, "An Introduction to Recursive Partitioning: Rationale, Application, and Characteristics of Classification and Regression Trees, Bagging, and Random Forests," *Psychological Methods* 14 no. 4 (2009): 323–48; Tsung-Hsi Wu et al., "Random Forest of Classification and Regression Tree (CART) in the Estimation of SWC Based on Meteorological Inputs and Hydrodynamics Behind," in 2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), (2023): 2251–55; and Glenn De'ath and Katharina E. Fabricius, "Classification and Regression Trees: A Powerful Yet Simple Technique for Ecological Data Analysis," *Ecology* 81 (2000): 3178–92.

236 Thomas Jiralerspong et al., "Efficient Causal Graph Discovery Using Large Language Models," Cornell University Algebraic Topology, *arXiv* preprint, *arXiv* (2024), 2402.01207, H. Jiang et al., "Large Language Model for Causal Decision Making," Cornell University Algebraic Topology, *arXiv* preprint, *arXiv* (2023), 2312.17122,; and E. Kıckman et al., "Causal Reasoning and Large Language Models: Opening a New Frontier for Causality," Cornell University Algebraic Topology, *arXiv* preprint, *arXiv* (2023), 2305.00050.

237 Yuzhe Zhang et al., "Causal Graph Discovery With Retrieval-Augmented Generation-Based Large Language Models," Cornell University Algebraic Topology, *arXiv* preprint, *arXiv* (2024), 2402.15301.

238 Aaron Hurst et al., "GPT-4o System Card," Cornell University Algebraic Topology, *arXiv* preprint, *arXiv* (2024), 2410.21276.

239 Ansam Khraisat et al., "Survey of Intrusion Detection Systems: Techniques, Datasets and Challenges," SpringerOpen, 2, 20 (2019), https://doi.org/10.1186/s42400-019-0038-7.

240 Zeeshan Ahmad et al., "Network Intrusion Detection System: A Systematic Study of Machine Learning and Deep Learning Approaches," *Transactions on Emerging Telecommunications Technologies,* Wiley Online Library, October 16, 2020, https://doi.org/10.1002/ett.4150.

241 David A. Bierbrauer et al., "Cybersecurity Anomaly Detection in Adversarial Environments," *AAAI FSS-21: Artificial Intelligence in Government and Public Sector,* Cornell University Algebraic Topology, *arXiv* (2021), https://arxiv.org/abs/2105.06742.

242 Sean M. Devine and Nathaniel D. Bastian, "An Adversarial Training Based Machine Learning Approach to Malware Classification Under Adversarial Conditions," in *Proceedings of the 54th Hawaii International Conference on*

System Sciences (2021): 827–36, https://scholarspace.manoa.hawaii.edu/server/api/core/bitstreams/106b4fa8-bc7c-4dbe-b4d0-cc656f5a342f/content.

243 Xingshuo An et al., "Hypergraph Clustering Model-based Association Analysis of DDOS Attacks in Fog Computing Intrusion Detection Systems," *EURASIP Journal on Wireless Communications and Networking,* October 22, 2018, https://doi.org/10.1186/s13638-018-1267-2.

244 Qingsai Xiao et al., "Towards Network Anomaly Detection Using Graph Embedding," in Computational Science–International Conference on Computational Science 2020, Lecture Notes in Computer Science; and eds. Valeria V. Krzhizhanovskaya et al., *Springer Nature,* 12140 (2020): 156–69, https://link.springer.com/chapter/10.1007/978-3-030-50423-6_12.

245 Cliff A. Joslyn et al., "Hypergraph Analytics of Domain Name System Relationships," in *Algorithms and Models for the Web Graph*, eds. Bogumił Kamiński et al., (Springer International Publishing, 2020): 1–15, https://doi.org/10.1007/978-3-030-48478-1_1.

246 Zong-Zhi Lin et al., "A Hypergraph-Based Machine Learning Ensemble Network Intrusion Detection System," *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 54, no. 11 (November 2024): 6911–23, doi:10.1109/TSMC.2024.3446635.

247 Lin et al., "A Hypergraph-Based Machine Learning Ensemble Network Intrusion Detection System."

248 Aksoy et al., "Hypernetwork Science Via High-Order Hypergraph Walks."

249 Aksoy et al., "Hypernetwork Science Via High-Order Hypergraph Walks."

250 Lin et al., "A Hypergraph-Based Machine Learning Ensemble Network Intrusion Detection System."

251 Canadian Institute for Cybersecurity (CIC), "Intrusion Detection Evaluation Dataset (cic-ids2017)," 2017, https://www.unb.ca/cic/datasets/ids-2017.html.

252 CIC, "Intrusion Detection Evaluation Dataset (cic-ids2017)."

253 Iman Sharafaldin et al., "Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization," in International Conference on Information Systems Security and Privacy, 2018, https://api.semanticscholar.org/CorpusID:4707749.

254 Lin et al., "A Hypergraph-Based Machine Learning Ensemble Network Intrusion Detection System."

255 Lin et al., "A Hypergraph-Based Machine Learning Ensemble Network Intrusion Detection System."

256 Nathaniel Bastian et al., "ACI IoT Network Traffic Dataset 2023," *Institute of Electrical and Electronics Engineers Dataport,* December 29, 2023, https://dx.doi.org/10.21227/qacj-3x32.

257 Lin et al, "A Hypergraph-Based Machine Learning Ensemble Network Intrusion Detection System."

258 Nour Moustafa and Jill Slay, "UNSW-NB15: A Comprehensive Data Set for Network Intrusion Detection Systems," in Military Communications and Information Systems Conference, IEEE, November 2015.

259 Wikipedia, "Advanced Persistent Threat," April 9, 2025, https://en.wikipedia.org/wiki/Advanced_persistent_threat.

260 Patrick D. Gallagher, "Guide for Conducting Risk Assessments," NIST SP 800-30 Revision 1, National Institute of Standards and Technology, September 2012, https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-30r1.pdf.

261 Scott Rose et al., "Zero Trust Architecture," NIST Special Publication 800-207, National Institute of Standards and Technology, August 2020, https://doi.org/10.6028/NIST.SP.800-207.

262 Muhammad Liman Gambo and Ahmad Almulhem, "Zero Trust Architecture: A Systematic Literature Review," Cornell University Algebraic Topology, *arXiv* (2025), 2503.11659v2, https://arxiv.org/html/2503.11659v2.

263 Yong Fang et al., "LMTracker: Lateral Movement Path Detection Based on Heterogeneous Graph Embedding," *Neurocomputing* 474 (2022): 37–47, https://doi.org/10.1016/j.neucom.2021.12.026.

264 A. Kountouras et al., "Enabling Network Security Through Active DNS Datasets," 9th International Symposium on Research in Attacks, Intrusions and Defenses (RAID).

265 Active DNS Project, 2016-2024, Center for Cyber Operations Enquiry and Unconventional Sensing, Georgia Institute of Technology, https://activednsproject.org/.

266 Elisa Bertino et al., eds., *Lecture Notes in Computer Science,* 9854 (September 19, 2016): 188–208.

267 Irena Bojanova, ed., "Comprehensively Labeled Weakness and Vulnerability Datasets via Unambiguous Formal Bugs Framework Specifications," in *IT Professional* 26, no. 1 (Jan-Feb 2024): 60-68, https://doi.org/10.1109/MITP.2024.3358970; and Bojanova, ed., "Comprehensively Labeled Weakness and Vulnerability Datasets via Unambiguous Formal Bugs Framework (BF) Specifications," National Institute of Standards and Technology, https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=957238, accessed November 16, 2024.

268 Nour Moustafa, "New Generations of Internet of Things Datasets for Cybersecurity Applications Based Machine Learning: TON_IoT Datasets," in *Proceedings of the eResearch Australasia Conference*, Brisbane, Australia, 2019.

269 Juan Manuel Matalobos Veiga et al., "A Hypergraph-Based Model for Cyberincident-Related Data Analysis," in *Proceedings of the 2024 European Interdisciplinary Cybersecurity Conference,* 161–62, https://doi.org/10.1145/3655693.3661300.

270 Juan Manuel Matalobos Veiga, "A Hypergraph-Based Model for Cybersecurity," Workshop on Complex Networks in Banking and Finance, Fields Institute, June 25, 2024, https://www.fields.toronto.edu/talks/hypergraph-based-model-cybersecurity.

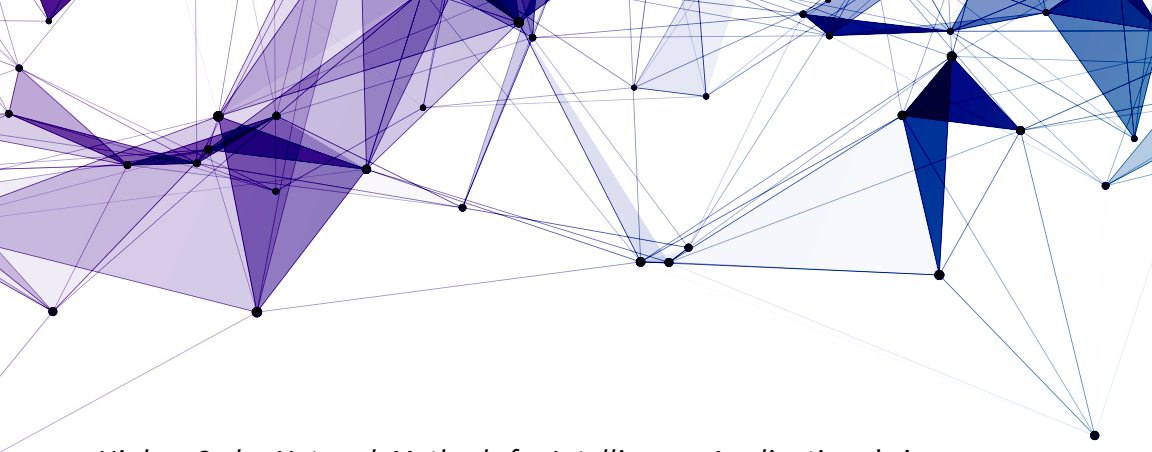
271 Veiga et al., "A Hypergraph-Based Model for Cyberincident-Related Data Analysis."

272 Joslyn et al., "Hypergraph Analytics of Domain Name System Relationships."

273 Moustafa, "New Generations of Internet of Things Datasets for Cybersecurity Applications Based Machine Learning: TON_IoT Datasets."

274 Nour Moustafa et al., "Data Analytics-Enabled Intrusion Detection: Evaluations of ToN_IoT Linux Datasets," in IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom), (2020): 727–73, doi: 10.1109/TrustCom50675.2020.00100.

275 Nour Moustafa et al., "Federated TON_IoT Windows Datasets for Evaluating AI-Based Security Applications," in IEEE 19th International Conference on Trust, Security and Privacy, 2020.



Higher-Order Network Methods for Intelligence Applications brings together leading perspectives on hypergraphs and related approaches for representing, analyzing, and interpreting complex systems. Designed as a compendium for intelligence practitioners, the volume introduces core concepts and demonstrates their use in cyber anomaly detection, network intrusion analysis, team formation, influence measurement, conflict behavior, and scenario discovery. For those working at the intersection of data, complexity, and national security, this book provides a timely guide to analytic methods that move beyond dyadic thinking.

Dr. Matthew Hasenjager is a postdoctoral fellow at the University of Tennessee, Knoxville. His research focuses on complex networks, collective behavior, social learning, and communication.

Dr. Mark Bailey is an associate professor and department chair for Artificial Intelligence, Cyber, Influence, and Data Science at National Intelligence University.

Dr. Nina Fefferman is a professor at the University of Tennessee, Knoxville. Her research focuses on the mathematics of self-organizing complex systems.

ISBN 978-1-932946-57-4