Relationships Between Node Degrees and Hyperedge Sizes in Empirical Hypergraphs

Bogumił Kamiński[1]*, Paweł Prałat[2]†, Aleksander Wojnarowicz[1]‡, Mateusz Zawisza[1]§

[1]Decision Analysis and Support Unit, SGH Warsaw School of Economics, Warsaw, Poland
[2]Department of Mathematics, Toronto Metropolitan University, Toronto, ON, Canada

Abstract

We investigate networks represented as hypergraphs and propose a generalizable statistical framework that captures the relationship between their node degrees and hyperedge sizes. We test the presence of such an association in 36 empirical hypergraphs from diverse domains, with a focus on social networks. Using nested model comparisons, we classify each such relationship as linear, monotonic, non-monotonic, or absent. Results reveal that true absence of this relationship is rare, while nearly half exhibit non-monotonic patterns. We evaluate three correlation measures of this association and find that Pearson correlation best aligns with relationship direction. We also consider three ways to capture this relationship (called: bipartite, node-centric or edge-centric) and show that the bipartite one yields most consistent results. Beyond providing empirical evidence, our results lay conceptual groundwork for linking static hypergraph structure to potential dynamical processes, positioning degree–size correlations as a structural bridge between hypergraph topology and function.

Keywords: empirical hypergraphs, node degree, hyperedge size, bipartite representation, Pearson correlation, non-monotonic relationship

1 Introduction

In recent years, hypergraphs have emerged as a powerful generalization of traditional pairwise graphs [3, 33, 78], particularly suited for modelling complex systems involving higher-order relationships [14, 72, 77, 104]. Unlike standard dyadic networks where edges connect pairs of nodes, hypergraphs allow hyperedges to connect any number of nodes, enabling a more expressive modelling framework [12, 15, 22]. This makes hypergraphs ideal for capturing group interactions found in diverse social networks, such as co-authorship networks [84, 73], affiliation or membership networks [31, 71, 107], social media [70, 6], social tagging [13], team sports [45, 83], but also ecological systems [47] and joint protein interactions in biological networks [75].

Due to their growing importance, many structural and statistical properties of empirical hypergraphs have been the focus of recent research [30, 65, 10]. Some of these measures, like node degree distribution [26], modularity [54, 53, 55, 27], clustering coefficients [35, 2] or the Bonacich eigenvector centrality [19] have analogues in traditional graphs [3, 76, 109, 17, 18, 36, 81, 80]. Others, such as hyperedge size distribution, node-hyperedge size correlation or the simplicial closure [84, 13] are unique to hypergraphs and open new questions for exploratory data analysis [1, 43]. Understanding these properties is essential for both descriptive purposes and for informing the design of algorithms and models tailored to higher-order data.

Descriptive analysis of hypergraph properties is also of practical significance. Structural characteristics—such as node degree distributions and hyperedge sizes—play a crucial role in shaping the dynamics of processes occurring on these networks [16]. For example, the spread of information, opinions, or infectious diseases can

*Email: bkamins@sgh.waw.pl
†Email: pralat@torontomu.ca
‡Email: alwojnarowicz@gmail.com
§Corresponding author: Mateusz Zawisza, SGH Warsaw School of Economics, Institute of Econometrics, ul. Madalińskiego 6/8, 02-513 Warsaw, Poland.
Tel.: +48 502 190 746.
Email: mzawisz@sgh.waw.pl

behave qualitatively differently depending on whether the system is modelled as a traditional graph or a hypergraph [63, 102, 117, 62]. Consequently, understanding and quantifying key structural features of hypergraphs is essential for developing accurate models of complex social systems.

In this paper, we propose and investigate a novel relationship in hypergraphs: between node degrees and their hyperedge sizes. While this relationship is non-existent in standard graphs, where all edges connect exactly two nodes, hypergraphs allow for nontrivial hyperedge size variability. This enables the study of correlations between how many hyperedges a node participates in (its degree) and how large those hyperedges tend to be. This important relationship has received limited attention in the literature.

From a theoretical perspective, we argue that degree–hyperedge size correlations constitute a previously underexplored structural property of hypergraphs that complements higher-order notions of assortativity and clustering. While classical assortativity describes how nodes connect to other nodes of similar activity, degree–size correlations capture how individual activity levels are coupled to the scale of collective interactions. As such, they provide a new lens for understanding participation, coordination, and resource allocation mechanisms in higher-order social systems. Several generative and random models for hypergraphs explicitly analyze the distributions of node degree and hyperedge size , often deriving these distributions in terms of model parameters. These models show that the mechanisms governing hyperedge formation, such as preferential attachment or copying, directly influence both node degree and hyperedge size distributions, and their interplay can lead to power-law or other heavy-tailed behaviours in real-world hypergraphs [11, 49, 89].

A central objective of this work is to determine how correlations between node degree and hyperedge size should be defined and measured. This task is not straightforward, as these quantities are inherently associated with different types of objects: node degree is a vertex-level property, whereas hyperedge size is defined at the level of hyperedges. Therefore it requires mapping these quantities into a common analytical domain, and different such mappings lead to distinct interpretations serving different research questions. In some applications this choice will be predetermined by the research objectives. However, often the user might not have a preference. In our paper we address this by treating the choice of representation as a question to be evaluated empirically rather than assumed a priori. To this end, we define and compare three alternative hypergraph representations that bring node degrees and hyperedge sizes into a shared domain: node-centric, edge-centric, and bipartite representations. Using classical correlation coefficients (Pearson, Spearman, Kendall) applied within each representation, and 36 empirical hypergraphs from diverse domains, we assess the interpretability and semantic alignment of the resulting correlations, with the goal of identifying the representation that provides the most robust and informative characterization of degree–size correlations in empirical hypergraphs.

We aim to explore whether this relationship is consistently non-zero, and whether its strength and functional form varies across semantic categories of hypergraphs. Such variation could help distinguish between types of real-world systems and provides insight into the mechanisms governing participation and coordination in higher-order interactions. For example, positive correlations may emerge in co-authorship networks [84], whereas different patterns may arise in systems governed by other participation mechanisms.

We introduce a generalizable statistical framework for identifying and classifying node–hyperedge size relationships in empirical hypergraphs, filling a methodological gap in existing hypergraph analysis. Using nested model comparisons, we classify each such relationship as linear, monotonic, non-monotonic, or absent. Distinguishing between these dependencies is not only methodologically relevant, but also theoretically informative: non-monotonic relationships may signal saturation or dilution effects, where increasing participation does not translate into proportionally larger group interactions, with potential consequences for coordination and diffusion processes [108, 24].

Detecting systematic patterns in this relationship could inform the development of generative models of hypergraphs [26, 27].Many current generative frameworks such as h–ABCD [56, 58] assume zero correlation between node degree and hyperedge size. Incorporating flexible control over this correlation could improve the realism of synthetic hypergraph models.

A core motivation for this study stems from the role that structural correlations play in governing collective behaviour and dynamic processes. We argue that in higher-order systems, the relationship between an individual's activity level (node degree) and their interaction environment (hyperedge size) represents a foundational mechanism—essentially a missing dimension of higher-order assortativity and clustering. Unlike standard assortativity, however, this mechanism is intrinsically higher-order, as it characterizes the coupling between individual participation intensity and the size of the groups in which interactions occur.

Degree–size correlations may influence dynamical processes by shaping how individual activity is coupled to group interactions. Positive correlations can create highly efficient transmission environments by concentrating

activity in large groups, whereas negative or non-monotonic relationships may promote local reinforcement and coordination within smaller groups. Consequently, identifying these correlations is a necessary first step toward understanding their role in diffusion, contagion, and cooperation processes on hypergraphs. In addition to characterizing empirical degree–hyperedge size relationships, we extend our framework conceptually by linking these static structural correlations to their potential implications for dynamical processes on hypergraphs. Specifically, we formulate a testable hypothesis that connects observed structural dependencies to diffusion dynamics, thereby bridging empirical hypergraph structure with dynamical modelling.

In summary, this study pursues three objectives. First, from an empirical perspective, we aim to detect, quantify, and classify the relationship between node degree and hyperedge size across diverse real-world hypergraphs, and to assess how this relationship varies across semantic domains. Second, from a methodological perspective, we systematically compare alternative representations , together with multiple correlation measures, with the goal of identifying the representation that provides the most stable, interpretable, and semantically aligned characterization of degree–size correlations. Third, from a theoretical perspective, we highlight that non-zero degree–size correlations can influence dynamical processes on hypergraphs, thereby justifying their relevance for modelling higher-order social systems. A systematic investigation of the implications of these correlations for higher-order dynamics exceeds the scope of the present work. Together, these objectives position degree–size correlations as a meaningful structural property of empirical hypergraphs and provide practical guidance for their analysis.

The remainder of the article is organized as follows. In Section 2, we introduce the foundational concepts, including hypergraph notation and key properties. Section 2.1 presents three data preprocessing strategies designed to enable meaningful comparisons between node degree and hyperedge size. Section 2.2 introduces the Eta-squared (η^2) statistic for evaluating alignment with semantic groupings. In Section 2.3, we outline a nested statistical testing procedure to classify the relationship between structural quantities. Section 2.4 summarizes data sources and computing tools, while section 2.5 details the limitations and the scope of our study. The results are presented in three parts. Section 3.1 evaluates and justifies the choice of the optimal data preprocessing strategy. Section 3.2 investigates which correlation metric best captures global structural trends. Section 3.3 applies model-based tests to classify empirical hypergraphs by the type of relationship between node degree and hyperedge size. In Section 4 we provide a detailed discussion of results and explore the potential impact of the identified relationship on social dynamics, as well as discuss implications for generative hypergraph models. Finally, conclusions and directions for future research are offered in Section 5. The Appendix includes the proof of the covariance equivalence between preprocessing strategies (Subsection A.1) and provides detailed background on the empirical hypergraph datasets used in this study, including semantic segments, node and hyperedge interpretations, and descriptive statistics (Subsection A.2). It also documents implementation details and computational complexity (Subsection A.3), defines the correlation measures employed (Subsection A.4), and includes supplementary analyses referenced in the main text (Subsection A.5).

2 Notation, Methods & Data

In this section, we introduce the fundamental definitions, methods, and datasets that underpin our analysis of structural patterns in empirical hypergraphs. We begin by formalizing hypergraph notation and its incidence graph representation, which serves as the mathematical foundation for our computations (Section 2.1), together with the data preprocessing strategies designed to enable meaningful comparisons between node degree and hyperedge size. Subsequently, we present a method for evaluating the alignment between correlation values and semantic segments using the η^2 statistic (Section 2.2). We also describe a statistical procedure to classify the functional relationship between node degree and hyperedge size (Section 2.3). Finally, we give the overview of the empirical datasets used in this study and computational handling (Section 2.4), while the more detailed discussion of these topics is provided in Appendix (Sections A.2 and A.3) together with the review of correlation measures suitable for quantifying the statistical association (Section A.4). Lastly, limitations of the study is presented in Section 2.5.

2.1 Hypergraph Notation and Data Preprocessing Steps

Understanding the relationship between node degree and hyperedge size in a hypergraph requires precise definitions and careful data transformation. In this section, we formalize the notation used throughout the paper and

describe three alternative preprocessing strategies that enable meaningful correlation and relationship analysis between these two structural quantities.

Hypergraph and Its Bipartite Representation as an Incidence Graph A hypergraph is a generalization of a graph in which edges, called hyperedges, can connect any number of vertices. Formally, a hypergraph is defined as $H = (V, E)$, where $V = \{v_1, \ldots, v_n\}$ is the set of vertices, and $E = \{e_1, \ldots, e_m\}$ is the set of hyperedges, with each $e_j \subseteq V$ [57].

To facilitate analysis, we use the bipartite representation of a hypergraph, also known as its incidence graph. In social network analysis, bipartite graphs of this type are often referred to as *affiliation networks*, and are also known as *two-mode networks* or *membership networks*, where one node set represents actors and the other represents groups or events [20, 21, 64, 107]. This representation is information-preserving and equivalent to the original hypergraph structure. Specifically, we define the incidence graph $IG(H) = (V', E')$, where the vertex set $V' = V \cup E$ includes both the original vertices and hyperedges of H, and edges E' connect $v \in V$ to $e \in E$ if and only if $v \in e$. This structure naturally yields an incidence matrix $B = (b_{ij}) \in \{0, 1\}^{n \times m}$, where each entry $b_{ij} = 1$ if vertex v_i belongs to hyperedge e_j, and $b_{ij} = 0$ otherwise. Both H and $IG(H)$ are fully recoverable from this matrix.

Hypergraph Data Preprocessing Steps Let the degree of a vertex $v_i \in V$ be defined as the number of hyperedges that include it, i.e.,

$$k_i \;=\; \deg(v_i) \;=\; \big|\{\, e \in E \;:\; v_i \in e \,\}\big|. \tag{1}$$

Similarly, the size of a hyperedge $e_j \in E$ is given by

$$s_j \;=\; |e_j| \;=\; \big|\{\, v \in V \;:\; v \in e_j \,\}\big|. \tag{2}$$

To quantify the relationship between node degrees and hyperedge sizes, these measures must be defined in the same domain. However, node degree is inherently a vertex-level property, while hyperedge size is defined at the hyperedge level. Thus, we construct data pre-processing transformations to bring both into a common domain for meaningful comparison.

To define a node-centric counterpart to hyperedge size, we compute for each node v_i the average size of hyperedges in which it participates:

$$\bar{s}_i \;=\; \frac{1}{k_i} \sum_{j:\, b_{ij}=1} s_j \;=\; \frac{\sum_j b_{ij}\, s_j}{k_i} \;=\; \frac{\sum_j b_{ij} \sum_k b_{kj}}{\sum_j b_{ij}}. \tag{3}$$

Algebraically, this corresponds to summing over the columns of the incidence matrix B for those hyperedges e_j that contain node v_i, and dividing by v_i's degree. This defines the node-centric preprocessing step, which produces a dataset of the form $\{(k_i, \bar{s}_i)\}_{i=1}^{n}$. One may then compute, for example, the Pearson correlation between node degree and average hyperedge size. This correlation reflects the expected hyperedge size for a randomly chosen node (uniformly at random) that has above-average degree.

Analogously, we define a hyperedge-centric counterpart to node degree by computing, for each hyperedge e_j, the average degree of its participating nodes:

$$\bar{k}_j \;=\; \frac{1}{s_j} \sum_{i:\, b_{ij}=1} k_i \;=\; \frac{\sum_i b_{ij}\, k_i}{s_j} \;=\; \frac{\sum_i b_{ij} \sum_k b_{ik}}{\sum_i b_{ij}}. \tag{4}$$

This amounts to summing over the rows of the incidence matrix B corresponding to nodes v_i in hyperedge e_j, and dividing by the hyperedge size. This defines the edge-centric[1] preprocessing step resulting in a dataset $\{(\bar{k}_j, s_j)\}_{j=1}^{m}$. The Pearson correlation between these quantities captures the expected degree of nodes involved in a hyperedge with size above the average.

A third approach is the bipartite representation preprocessing step, based on the incidence graph $IG(H)$ of the hypergraph H, where nodes and hyperedges form two disjoint sets of vertices. This representation leads to a dataset $\{(k_i, s_j) : b_{ij} = 1\}$ defined over the edges of the bipartite graph. Here, one can compute the assortativity

[1] A more precise name would be *hyperedge-centric preprocessing step*, but for conciseness, we refer to it simply as *edge-centric* throughout the text.

coefficient as the Pearson correlation of the degrees of incident vertex pairs in the bipartite graph [79, 57]. This statistic reflects the expected hyperedge size for a randomly sampled node–hyperedge pair, conditional on the node having above-average degree.

2.2 Assessing the Alignment between Correlation and Hypergraph Segment

To evaluate how well different correlation designs align with semantic distinctions in hypergraph structure, we employ the Eta-squared statistic (η^2). This metric quantifies the proportion of variance in a continuous variable that is explained by a categorical predictor, and is commonly interpreted as a measure of effect size in analysis of variance (ANOVA) [93]. In our case, η^2 captures how well a given correlation coefficient (e.g., between hyperedge size and node degree) can be predicted based on the hypergraph segment to which it belongs.

Formally, let Y_i denote the correlation value computed for the i-th hypergraph, and let $\mathcal{S}_i$ be its segment label (e.g., *email*, *tag-question*, etc.). Then, η^2 is defined as:

$$\eta^2 = \frac{\mathrm{SS}_{\text{between}}}{\mathrm{SS}_{\text{total}}} = \frac{\sum_{g=1}^{G} n_g(\bar{Y}_g - \bar{Y})^2}{\sum_{i=1}^{N}(Y_i - \bar{Y})^2}, \tag{5}$$

where G is the number of groups (segments), n_g is the number of observations in group g, $\bar{Y}_g$ is the mean correlation in group g, and $\bar{Y}$ is the overall mean. The numerator is the between-group sum of squares ($\mathrm{SS}_{\text{between}}$), and the denominator is the total sum of squares ($\mathrm{SS}_{\text{total}}$).

We consider nine correlation configurations, defined by the combination of three data preprocessing strategies (edge-centric, node-centric, and bipartite representation) with three correlation measures (Pearson, Spearman, and Kendall). For each configuration, a single correlation value is computed for each of the 12 hypergraph segments under study, which include: *user-answer*, *physical contact*, *part-whole*, *diseases and genes*, *email*, *person-place*, *political*, *participant-conference*, *user-review*, *drugs*, *tag-question*, and *user-thread*.

To compute η^2, we fit a linear model where the response variable is the correlation value and the predictor is a one-hot encoded vector representing the hypergraph segment. The resulting η^2 is equivalent to the R^2 of this model and indicates the proportion of variance in the correlation values that can be attributed to segment identity. All η^2 values reported in this paper were calculated in `R` using the `etaSquared()` function from the `lsr` package [74].

A higher η^2 signifies a stronger alignment between correlations and hypergraph segments. Thus, we use η^2 as a criterion for selecting the optimal design of data preprocessing method and correlation measure that yields the most segment-sensitive correlation values.

2.3 Statistical Identification of Relationship Type

This subsection outlines the statistical procedure used to classify the relationship between hyperedge size and node degree into one of four categories: non-monotonic, monotonic (but not linear), linear, or no relationship. This classification is later employed in the results Subsection 3.3. The identification procedure is model-based and involves fitting three types of models to each dataset: (i) an unrestricted Generalised Additive Model (GAM) [48, 112], (ii) a monotonic GAM, and (iii) a simple linear regression model estimated using ordinary least squares (OLS) [37].

In this work, we define a monotonic GAM as a special case of a shape-constrained additive model (SCAM) in which the smooth term is restricted to be either monotonically increasing or decreasing. To implement this, we fit two SCAMs with a monotonic increasing constraint and another with a decreasing constraint. Then our procedure selects the model with the lower residual sum of squares as the representative monotonic GAM. This approach follows the methodology introduced by [91] and is implemented using the `mgcv` and `scam` packages in R [112]. For general background on generalized additive models, we refer readers to [48].

The classification is conducted via a sequence of statistical comparisons using ANOVA tests for nested models[2] [39]. As illustrated in Figure 1, the procedure begins by comparing the unrestricted GAM to the best-fitting monotonic GAM using an F-test. The null hypothesis (H_0) states that the monotonic GAM sufficiently explains the data; rejecting it (at $\alpha < 10^{-5}$) implies the presence of significant non-monotonicity, and the

[2]Alternative model-selection procedures based on information criteria, particularly BIC, could also be considered within this sequential framework. However, we prefer nested hypothesis testing because our primary objective is to identify the underlying relationship type rather than optimize predictive performance, for which criteria such as AIC are more commonly used.

relationship is classified as non-monotonic. The choice of such a stringent significance threshold is motivated by the large size of most empirical hypergraphs and the need to limit false discovery in favour of simpler models unless the data strongly supports added complexity.

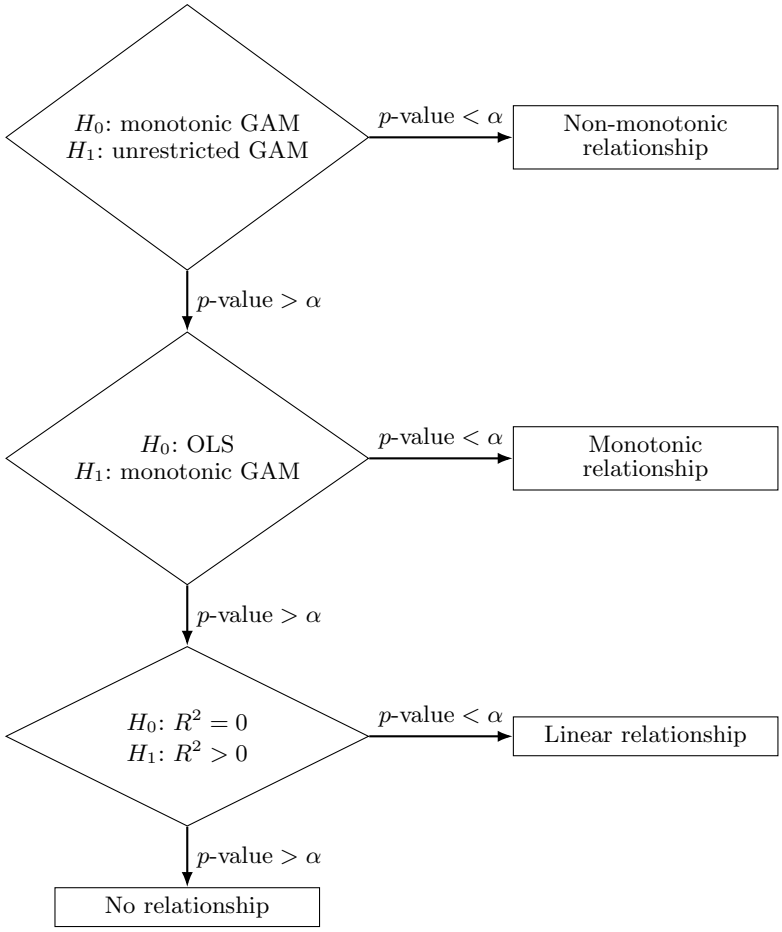


Figure 1: Diagram illustrating the decision procedure for classifying the relationship between two variables (hyperedge size and node degree) into one of four categories: non-monotonic, monotonic (but not linear), linear, or no relationship. Decisions are based on successive ANOVA comparisons of nested models: unrestricted GAM, monotonic GAM, and OLS.

If the null is not rejected in the first test, the relationship is assumed monotonic, and a second ANOVA test compares the monotonic GAM to a linear OLS model. Rejection of linearity indicates a monotonic (but non-linear) pattern. If the test fails to reject the linear model, a final F-test is performed to check whether the linear model explains any significant variation in the data (i.e., whether $R^2 > 0$). If this is not the case, the relationship is classified as "no relationship." Thus, the decision tree ensures a structured, conservative, and data-driven assignment of relationship type.

2.4 Data and Computation Overview

Our analysis draws upon 36 empirical hypergraphs spanning diverse domains—including physical contact networks, online user interactions, political affiliations, and biomedical associations. This breadth ensures that observed statistical relationships are robust across domains and not artifacts of a single data type. Detailed descriptions of all datasets, their assigned segments, and statistical summaries of node degrees and hyperedge sizes are presented in Appendix A.2.

To process and analyze these datasets, we developed a flexible computational pipeline tailored to varied file formats and scales. Hypergraphs were represented using sparse matrix structures to ensure memory efficiency and speed. Different sparse formats were tested and selected based on input type and construction cost, with

final conversion to CSR (Compressed Sparse Row) format to enable fast indexing and vectorized computations. Our implementation makes use of Python libraries such as `scipy`, `numpy`, and, where applicable, `numba` for optimization. All statistical analyses, including correlation computations, model fitting, and figure generation, were performed using the R programming language [92]. The full technical details of data ingestion, matrix construction, and performance considerations are documented in Appendix A.3.

All reproducible code used in this study, including both Python and R scripts for data processing, statistical analysis, and figure generation, is available in the public repository: `https://github.com/AleksanderWWW/hypergraph-properties`.

2.5 Limitations and Scope of Inference

While our analyses are systematic and transparent, the study has limitations. We outline these limitations below to clarify the scope of our conclusions and inform future work.

Absence of causal or dynamical validation. The present study is limited to structural, cross-sectional analysis and does not provide causal or dynamical validation of the observed correlations. As stated in the Conclusions, although degree–hyperedge size correlations may influence diffusion dynamics, this remains a testable hypothesis. Empirical validation is complicated by confounding structural features that co-vary with degree–size correlation in real datasets. Addressing this limitation will require generative hypergraph models that vary the correlation in a controlled manner while holding other properties constant.

Dependence on static datasets. Our analysis is restricted to static hypergraphs, including datasets that represent snapshots of potentially evolving systems and are therefore treated as effectively static. As a result, temporal dependencies and structural evolution are not captured.

Assumptions of independence in statistical testing and model estimation. Observations in the bipartite representation are not strictly independent, as multiple degree–hyperedge pairs may share the same node or the same hyperedge. Such structural overlap induces correlations that violate classical independence assumptions underlying regression and model comparison procedures. In the present analysis, we did not apply clustering, resampling, or random-effects adjustments to account for these dependencies.

Beyond structural overlap, domain-specific mechanisms may introduce additional correlations. For example, recurring organizational interactions, such as repeated mailing lists with identical recipient sets, can generate clusters of nearly identical degree–size pairs. These correlated clusters may contribute to local fluctuations in flexible models, potentially creating the appearance of greater functional complexity than is actually present. Addressing these dependencies could be done by explicitly modelling shared nodes or hyperedges, or by applying principled de-duplication or hierarchical modelling strategies.

Multiple testing. The analysis involves multiple hypothesis tests conducted across 36 empirical hypergraphs. Within each dataset, however, hypothesis testing follows a hierarchical and conditional procedure, with at most three nested tests performed and, in most cases, fewer being decisive. In particular, datasets classified as non-monotonic are determined by a single test, while those classified as monotonic rely on two tests. Accordingly, although multiple testing influences cumulative Type I error, this bias is highly unlikely to influence the main conclusions.

Domain imbalance and generalization. Although our study spans multiple semantic segments, including email, physical contact, and online interaction datasets, domain representation is uneven. Consequently, aggregate patterns may partially reflect properties of more frequently sampled domains, and shouldn't be overgeneralized. That said, statistically detectable associations between node degree and hyperedge size are observed across all analyzed segments, indicating that the phenomenon is not confined to a single domain.

Interpretability in large hypergraphs. In large datasets, even minor deviations from linearity may become statistically significant, potentially complicating interpretation. To mitigate this risk, we adopted a highly stringent significance threshold of $\alpha = 10^{-5}$ and provided complete scatterplots, fitted GAM curves, and ΔR^2 measures of incremental explained variance for all datasets, allowing readers to assess whether detected non-linearities reflect substantive structural patterns or merely statistically detectable fluctuations.

3 Results

This section is divided into three parts. In the first subsection 3.1, we justify the optimal choice of hypergraph data preprocessing. In the second subsection 3.2, we delve into the selection of one of three correlation measures: Pearson, Spearman, or Kendall. In the third subsection 3.3, we characterize the observed relationships. We

	Pearson	Spearman	Kendall
node-centric	0.5380	0.6062	0.4360
edge-centric	0.6096	0.4354	0.4360
bipartite representation	0.6656	0.6782	0.6605

Table 1: Eta-squared (η^2) values measuring the proportion of variance in correlations values explained by the hypergraph segment. Rows correspond to type of hypergraph data pre-processing, and columns indicate the correlation type.

classify them into one of four categories: non-monotonic, monotonic (but non-linear), linear, or no apparent relationship.

3.1 Methodological assessment I: comparison of preprocessing strategies

The goal of this subsection is to determine which hypergraph data preprocessing strategy is most suitable for capturing the relationship between hyperedge size and node degree. The criterion for this comparison is the alignment between the computed correlation values and the categorical segmentation of hypergraphs. This alignment is quantified using the Eta-squared statistic (η^2).

3.1.1 Comparing preprocessing strategies via η^2

To compute the correlation between hyperedge size and node degree, two design choices must be made: the hypergraph preprocessing method (node-, edge-centric, and bipartite) and the type of correlation coefficient (Pearson, Spearman, and Kendall). These choices yield a total of nine (3×3) possible combinations. For each such combination, a correlation value is computed for every hypergraph in the dataset. These correlation values are then evaluated for their segment-level alignment using η^2. Table 1 presents the η^2 values for each combination of hypergraph data preprocessing method (rows) and correlation measure (columns). Higher values of η^2 indicate stronger alignment between the correlation values.

The most striking observation is that the choice of data preprocessing method exerts the greatest influence on η^2. Across all three correlation measures, the bipartite representation consistently yields the highest η^2 scores, clearly outperforming both the node-centric and edge-centric approaches. Specifically, bipartite representation achieves $\eta^2 = 0.6656$ with Pearson, 0.6782 with Spearman, and 0.6605 with Kendall, all substantially higher than their respective scores under alternative preprocessing methods. This robustness suggests that the bipartite structure more faithfully preserves segment-level variability relevant to the correlation between hyperedge size and node degree.

In contrast, the choice of correlation measure appears to matter less, especially within the bipartite setting, where all three correlations perform comparably. This implies that once an appropriate data structure is chosen, the precise choice of correlation coefficient has limited impact.

3.1.2 Explaining η^2 values via within-segment variability in correlation estimates

To better understand the η^2 values reported in Table 1, we visualize the distribution of correlation coefficients within 12 hypergraph segments in Figure 5. The figure displays the variability of correlation values for six selected combinations of data preprocessing method and correlation coefficient, i.e., all pairings of Pearson and Spearman correlations[3] with the three preprocessing strategies: node-centric, edge-centric, and bipartite representation.

The figure illustrates how well each combination discriminates among the 12 hypergraph segments. Lower variability in correlation values across hypergraphs within the same segment implies stronger between-group effects, resulting in higher η^2. A pattern emerges: the bipartite representation consistently shows low within-segment variability across all 12 categories, regardless of whether Pearson or Spearman is used. This explains its dominant performance in Table 1. For segments such as `diseases and genes`, `drugs`, `email`, `part-whole`, `user-review`, and `person-place`, both correlations under the bipartite representation show minimal within-segment variability, making them clearly superior to other combinations. For `participant-conference`,

[3]We omit Kendall's τ for clarity, focusing on the two more commonly used and better-performing measures.

user-answer and tag-question segments, all six combinations show tight clustering of correlation values. In the largest physical contact segment, both bipartite and edge-centric preprocessing yield relatively consistent correlation values, whereas node-centric shows a much wider spread. Only within the user-thread segment, the node-centric view shows lower variability than a bipartite representation. These patterns reinforce the advantage of bipartite preprocessing for producing stable, segment-discriminative correlation estimates across a diverse range of hypergraph types.

3.1.3 Segment-level patterns and consistency of Pearson correlations across preprocessing strategies

A deeper analysis comparing three data preprocessing strategies is provided in Figure 6 in Appendix A.5. It compares the three preprocessing strategies by examining pairwise relationships between their Pearson correlation estimates. The bipartite representation shows moderate but consistent agreement with both node- and edge-centric approaches ($R^2 \approx 0.60$), while node- and edge-centric correlations are less aligned ($R^2 = 0.34$), reflecting structural differences in how each method aggregates information. The theoretical basis for these empirical relationships is discussed in Section 4.1. Distributional analysis confirms that bipartite correlations are the most stable and tightly clustered.

To further understand the behaviour and consistency of correlation estimates across different preprocessing steps, we examine Pearson correlation values with their 95%-CI for all hypergraphs, computed under three data preprocessing methods: node-centric, edge-centric, and bipartite representation in Figure 2. The hypergraphs are sorted by decreasing Pearson correlation under bipartite representation. This ordering allows us to visually identify clusters of hypergraphs that exhibit similar correlation structure, both in magnitude and in sign.

Several coherent segment-level patterns emerge. For instance, the physical contact segment, comprising hospital-lyon ($r = 0.337$), contact-high-school ($r = 0.180$), contact-primary-school ($r = 0.089$), InVS13 ($r = -0.030$), InVS15 ($r = 0.020$), Malawi-village ($r = 0.034$), and Science-Gallery ($r = 0.086$), appears mostly in the upper half of the ranking, exhibiting generally positive Pearson correlations under bipartite representation. Similarly, the Drugs segment: NDC-classes ($r = 0.191$) and NDC-substances ($r = 0.091$), and the user-thread group: threads-ask-ubuntu ($r = 0.104$), threads-math-sx ($r = 0.099$), twitter ($r = 0.035$), also cluster together with consistently positive correlation values.

Conversely, several hypergraphs appear at the lower end of the ranking with negative correlation values. These include disgenenet ($r = -0.166$) and diseasome ($r = -0.067$) from the Diseases and genes segment; geometry ($r = -0.129$) and algebra ($r = -0.0968$) from user-answer; and 3 hypergraphs from the user-review segment, including music-blues-reviews ($r = -0.132$), restaurant-reviews ($r = -0.079$), and vegas-bars-reviews ($r = 0.037$), which show weak to moderately negative or near-zero correlations. These trends illustrate how semantic categories often align with shared correlation patterns, hinting at underlying structural regularities that will be explored further in the following analyses.

3.2 Methodological assessment II: comparison of correlations coefficients

This subsection aims to provide practical guidelines for selecting the most appropriate correlation measure, once the bipartite representation has been adopted as the preprocessing method, as recommended in Section 3.1. As shown earlier in Table 1, the η^2 scores do not provide a decisive basis for choosing among Pearson, Spearman, and Kendall. In the bipartite setting, all three perform comparably. To resolve this ambiguity, we proceed with a more detailed investigation focused on comparing Pearson and Spearman coefficients[4]. Our analysis is based on the alignment of their signs with the global trends identified by monotonic GAM models.

Additionally, quantitative comparison of Pearson and Spearman correlations – detailed in Appendix A.5.2 – shows a strong alignment overall, with $R^2 = 79\%$, showing rather modest quantitative differences. However, identified there qualitative discrepancies in sign between Pearson and Spearman are especially consequential and motivate the subsequent part of paper.

3.2.1 Assessing Alignment Between GAM Monotonicity and Correlation Coefficients' Signs

We propose an additional criterion that is not captured by η^2: the degree to which the sign of a correlation coefficient reflects the global trend in the data. This trend is determined using monotonic GAMs, fitted separately

[4]Since Spearman and Kendall are almost perfectly correlated in our dataset, with over 99% agreement, we focus only on Spearman as a representative of the nonparametric class.

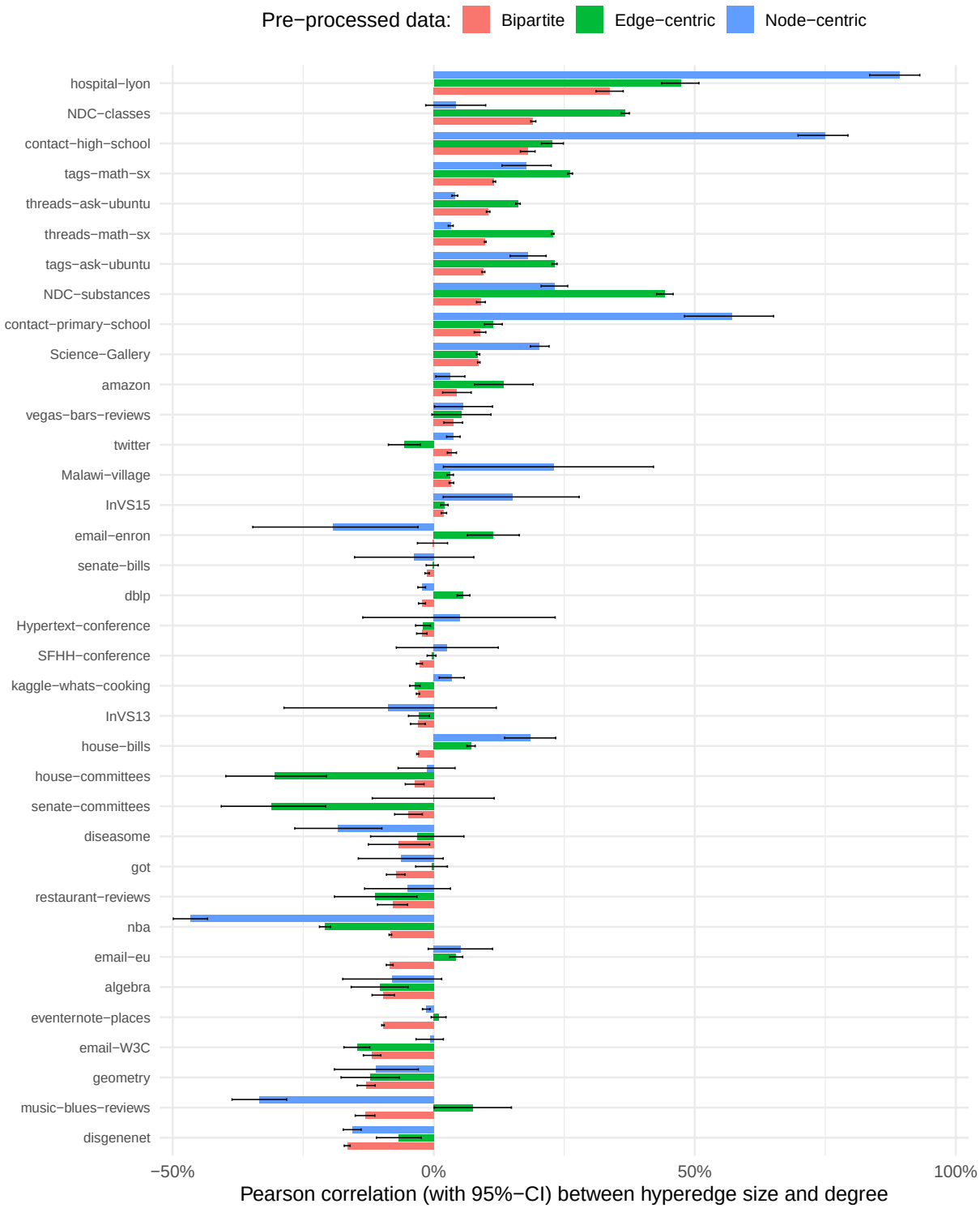


Figure 2: Pearson correlations (with 95% confidence intervals) between hyperedge size and node degree for three pre-processing schemes: bipartite representation, edge-centric, node-centric. Confidence intervals are computed using the Fisher z transformation with standard error $1/\sqrt{N-3}$. Hypergraphs are ordered in decreasing magnitude of the bipartite correlation to make similarities within semantic groups more apparent.

under increasing and decreasing shape constraints, see Section 2.3. Although Pearson measures linear dependence and Spearman/Kendall monotonic relationships, we observe in following subsection 3.3.2 that approximately half of the empirical hypergraphs exhibit complex, non-monotonic relationships, often with a clear dominant trend. For such cases, it is desirable that the sign of the selected correlation coefficient be aligned with the direction of this dominant trend.

To operationalize this, we classify each empirical hypergraph in two ways. First, based on the correlation coefficient (Pearson, Spearman), we assign it to one of three categories: (i) significantly negative, (ii) non-significant, or (iii) significantly positive. Second, based on the monotonic GAM fit, we classify it into: (i) decreasing trend, (ii) no significant trend, or (iii) increasing trend. The alignment between these two classifications is evaluated using contingency tables in Table 2.

Sign of Pearson	**GAM direction** Dec.	Non-sign.	Inc.	**Sum**
Negative	16	0	0	16
Non-sign.	3	3	1	7
Positive	0	0	13	13
Sum	19	3	14	36

(a) Pearson correlation.

Sign of Spearman	**GAM direction** Dec.	Non-sign.	Inc.	**Sum**
Negative	8	0	0	8
Non-sign.	8	2	0	10
Positive	3	1	14	18
Sum	19	3	14	36

(b) Spearman or Kendall correlations.

Table 2: Contingency tables comparing the sign of correlations with the direction of monotonicity inferred from monotonic GAMs. Statistical significance evaluated at $\alpha = 10^{-5}$.

In both panels of Table 2, all 36 empirical hypergraphs are classified according to the monotonicity direction inferred from monotonic GAM fits (columns). According to this classification, the majority, 19 out of 36 hypergraphs (approximately 53%), exhibit a statistically significant *decreasing* relationship between hyperedge size and node degree in bipartite representation. A slightly smaller group, 14 hypergraphs (39%), shows a statistically significant *increasing* relationship. Only 3 hypergraphs (about 8%) are found to have no statistically significant monotonic trend at a stringent threshold of $\alpha = 0.00001$. Next, we evaluate how these GAM-derived trend directions align with the sign of Pearson, and Spearman.

Alignment Between GAM Monotonicity and Pearson Sign The left panel of Table 2 compares the GAM monotonicity classification with the sign of Pearson correlations. The most common group consists of hypergraphs with negative correlations, 16 out of 36 cases (approximately 44%), all of them belong to 19 hypergraphs that monotonic GAM classifies as decreasing. The second most frequent group comprises hypergraphs with significantly positive Pearson correlations—13 out of 36 (36%), again, all of them belong to 14 increasing trend cases identified by the GAM. The least represented group consists of non-significant Pearson correlations, occurring in 7 hypergraphs, while GAM identifies only 3 hypergraphs with no significant monotonicity-all of these 3 hypergraphs belong to 7 hypergraphs indicated by Pearson correlation. In total, the alignment between Pearson sign and monotonic GAM direction is 32 out of 36 cases (about 89%).

The 4 misaligned cases are all classified as non-significant by Pearson (second row), while GAM assigns 3 of them (`house-committees`, `senate-committees`, `diseasome` as visible in Table 7) to the decreasing category and 1 (`email-enron`) to the increasing category. The Pearson coefficients for these four hypergraphs are all close to zero: -0.0669 (`diseasome`), -0.0483 (`senate-committees`), -0.0365 (`house-committees`), and -0.0024 (`email-enron`), see Table 7. While these values are not significant at $\alpha = 0.00001$, three out of four are significant at a more conventional threshold of $\alpha = 0.05$, with p-values of 0.0260, 0.8720, 0.0000713, and 0.00038, respectively. If we relax the significance threshold to $\alpha = 0.05$, the total alignment between Pearson and monotonic GAM increases to 35 out of 36 hypergraphs (approximately 97%). Moreover, this change would also reclassify three other hypergraphs, `amazon`, `InVS13`, and `vegas-bars-reviews`, from non-significant sign of Pearson to either increasing or decreasing, again in agreement with the GAM classification. Thus, under a more conventional $\alpha = 0.05$ threshold, the alignment remains consistently high at around 97%.

Alignment Between GAM Monotonicity and Spearman and Kendall Signs Turning to the right panel of Table 2, Spearman and Kendall coefficients exhibit notably lower alignment with the monotonic GAM

direction, with only 24 out of 36 hypergraphs (67%) classified in agreement. While these non-parametric coefficients correctly identify all 14 cases with an increasing trend, they misclassify 11 out of 19 hypergraphs with a decreasing trend—labeling 8 as non-significant and, in 3 cases, incorrectly assigning them a statistically significant positive correlation.

Notably, three hypergraphs—`kaggle-whats-cooking`, `house-bills`, and `email-eu`—are assigned an opposite classification. While monotonic GAM identifies these relationships as decreasing, Spearman and Kendall both indicate statistically significant positive correlations. The corresponding Spearman values are 0.014, 0.031, and 0.037, and the Kendall values are 0.009, 0.020, and 0.024, respectively. Although these correlation values are small and only marginally above zero, they are statistically significant at a stringent significance level of $\alpha = 0.00001$.

These examples are particularly striking, as they not only indicate opposite directional trends but do so with very low p-values, which may lead to overconfident and misleading conclusions. Visually, the relationships in these cases appear strongly non-monotonic or globally trending in the opposite direction, further reinforcing the argument that nonparametric coefficients may fail as reliable indicators of global trend direction. The additional two hypergraphs: `dblp` and `SFHH-conference`, are also misaligned in the direction of the correlation sign significant at $\alpha = 0.01$, see Table 7.

A total of five hypergraphs where Spearman and Kendall correlations exhibit a statistically significant sign opposite to the monotonic trend identified by GAM at the $\alpha = 0.01$ level. This underscores a substantial limitation in using nonparametric correlation coefficients as standalone indicators of relationship direction in empirical hypergraph data.

3.2.2 Interpretation Corner: Segment-Level Monotonicity Patterns

Given the empirical nature of the hypergraphs under analysis, we now aim to interpret the observed relationship signs through the lens of what hyperedges and nodes represent in each case. This interpretation goes beyond pure measurement and offers rationale grounded in the semantics of each dataset.

Figure 3 presents the frequency of monotonicity directions (increasing, decreasing, or non-significant) detected by monotonic GAMs, grouped by hypergraph segment. The figure reveals strong homogeneity of GAM monotonicity direction within hypergraph segments: 10 out of 13 categories exhibit perfect consistency in trend direction across all included hypergraphs. Of the remaining three, two segments (`Physical Contact` and `User-review`) contain a single outlier hypergraph (`InVS13` and `vegas-bars-reviews`, respectively), both classified as non-significant.

Positive monotonic trends dominate several hypergraph categories, including `Physical Contact`, `User-Thread`, `Tag-Question`, and `Drugs`. In the **Physical Contact** segment, nodes represent individuals equipped with sensors, and hyperedges correspond to physical group interactions over brief time intervals. An individual participating in larger group interactions (high hyperedge size) is likely to engage with more people overall, thus appearing in more interactions (high node degree). Hence, larger hyperedges naturally imply higher node degrees, leading to a positive correlation.

Several hypergraph segments exhibit a dominant *negative* monotonic trend between hyperedge size and node degree, particularly the `Political`, `Participant-Conference`, and `Person-Place` categories. In the **Political** segment, such as `house-bills` and `house-committees`, nodes represent political actors (e.g., members of Congress), and hyperedges represent either legislative bills or committee memberships. A negative relationship in this context indicates that politicians involved in large coalitions (e.g., large bills with many cosponsors) tend to participate in fewer overall bills. This aligns with political specialization: high-frequency participants may focus on narrow, small-scale initiatives, while those contributing to large, broad coalitions do so more occasionally. Moreover, committee memberships are often limited in number due to institutional constraints, and members of large committees may serve on fewer committees overall, reinforcing the observed inverse relationship.

This interpretive analysis demonstrates that the direction of monotonic relationships observed between hyperedge size and node degree is not random but meaningfully structured across semantic hypergraph segments. The strong within-segment consistency in monotonic direction, along with interpretable domain-based rationales, supports the validity of our classification procedure. More broadly, these results highlight the importance of domain-specific interpretation. While the existence of a systematic relationship between node degree and hyperedge size appears to be a common feature of empirical hypergraphs, its direction and functional form vary substantially across application domains. Consequently, our findings should not be interpreted as supporting a universal positive or negative dependence. Rather, the semantic meaning of nodes and hyperedges appears

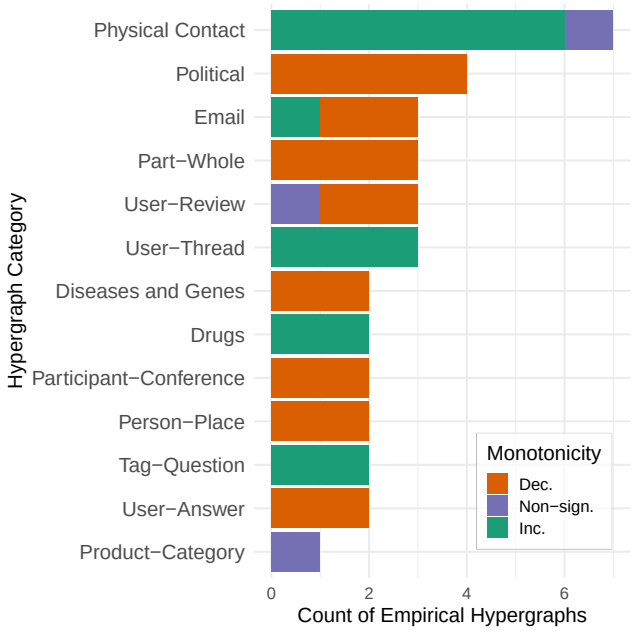


Figure 3: Frequency of monotonic trend by hypergraph category. Each horizontal bar shows the count of hypergraphs in that category with an increasing (Inc.), decreasing (Dec.), or non-significant (Non-sign., if neither an increasing nor a decreasing GAM model fits significantly better than a flat baseline) monotonic GAM fit.

to play a central role in determining whether the relationship is increasing, decreasing, linear, monotonic, or non-monotonic.

3.3 Empirical assessment: Prevalence of degree–hyperedge size relationships

The goal of this subsection is to shift the focus from a quantitative summary to a qualitative understanding of the relationship types observed across empirical hypergraphs. We classify each of the 36 empirical hypergraphs into one of four qualitative categories that characterize the nature of the relationship between hyperedge size and node degree: (1) non-monotonic (including unimodal or multimodal patterns), (2) monotonic, (3) linear, and (4) no relationship. This classification is carried out using the statistical procedure introduced in Subsection 2.3, which combines models' fitting with sequential hypothesis testing.

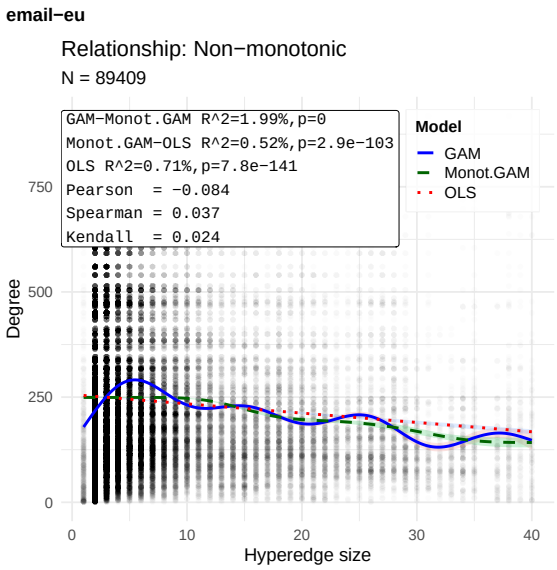
3.3.1 Examples of Identified Relationship Types

In this subsection, we present four representative hypergraphs, each illustrating one of the relationship types identified in our classification scheme. We begin with the most complex case of non-monotonic relationship and proceed through monotonic and linear examples, concluding with a case of no apparent relationship.

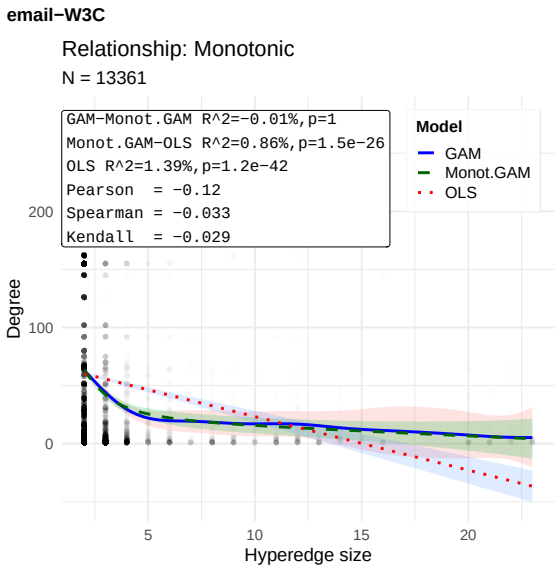
Non-monotonic Relationship Figure 4a displays the scatterplot for `email-eu`, which consists of 89,409 data points in its bipartite representation. The unrestricted GAM (blue solid line) reveals a complex, multimodal structure, characterized by several statistically significant fluctuations, as evidenced by the narrow confidence intervals[5] estimated at a significance level of $\alpha = 0.05$. These complex fluctuations may stem from dependencies among observations, violating the assumption of independence, which is discussed in Section 2.5.

Visual inspection of Figure 4a, along with the monotonic GAM fit (green dashed line), suggests a globally decreasing relationship, consistent with the negative linear regression line and Pearson coefficient of -0.084. Nevertheless, both the Spearman and Kendall coefficients are positive (0.037 and 0.024, respectively), which likely results from the initial upward trend for small hyperedge sizes.

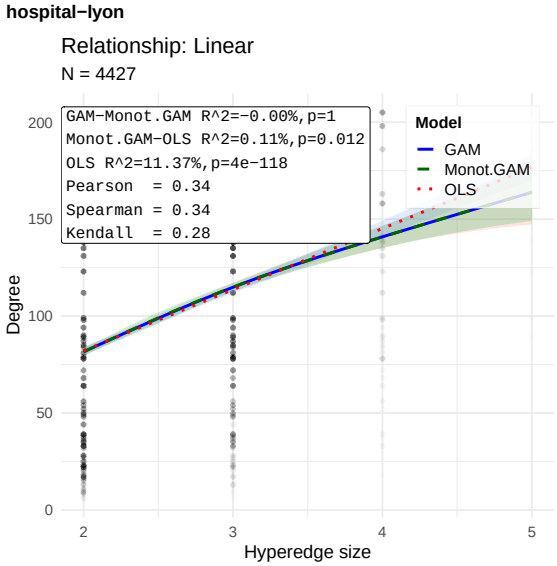
[5]Confidence intervals are computed under the assumption that observations are independent. This assumption is discussed in section 2.5. CIs are derived analytically as Wald-type intervals using model-based standard errors from fitted generalized additive models [48], as implemented in the `mgcv` package [25, 112].



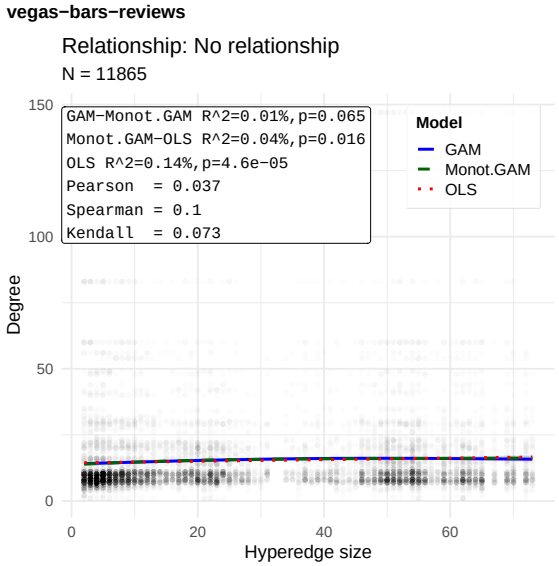
(a) Non-monotonicity between the number of recipients of a given e-mail (hyperedge size) and the total number of emails sent and received (degree). Potentially "clustered" e-mails, such as newsletters or other regularly distributed messages, may induce non-monotonic fluctuations arising from non-i.i.d. error structures.



(b) Decreasing relationship between the number of recipients of a given e-mail (hyperedge size) and the total number of emails sent and received by a given sender or recipient (degree). Illustrates a potential capacity constraint of e-mail users, resulting in a trade-off between sending many e-mails to a small number of recipients (< 5) and sending fewer e-mails to larger groups ($\geqslant 5$).



(c) Proportional relationship between #hospital staff and patients attending a gathering (hyperedge size) and #gatherings attended by an individual (degree). Interpretation: knowing X times more people results in $\approx X$ times more interactions.



(d) No relationship between #reviews for a bar in a given month (hyperedge size) and #reviews submitted by each user (degree). This suggests that months with a higher volume of reviews are not driven by systematically more or less engaged users.

Figure 4: Four types of relationships between hyperedge size and node degree: non-monotonic (top left), monotonic (top right), linear (bottom left), and no relationship (bottom right).

To formalize this "non-monotonic" classification, we focus on the top-left inset of Figure 4a, which displays two ANOVA tests and one F-test as per our methodology. The first ANOVA test compares an unconstrained GAM (blue line) against a monotonic GAM (green line), with the null hypothesis assuming equivalence. The unconstrained GAM fits data better by $\Delta R^2 = 1.99\%$ (p-value$< 10^{-5}$) than the constrained model, revealing a non-monotonic pattern. As the result of this first test is decisive, the procedure terminates at this step, classifying the relationship as non-monotonic. The other two test results are reported for completeness but do not affect the classification.

Importantly, the reported ΔR^2 values allow us to move beyond binary significance testing and assess the practical importance of each component of the relationship. In the example shown in Figure 4a, the OLS model explains only $R^2 = 0.71\%$ of the variance. Allowing for a non-linear but monotonic relationship contributes an additional $\Delta R^2 = 0.52\%$, increasing the total explained variance to 1.23%. In contrast, allowing for non-monotonicity contributes a further $\Delta R^2 = 1.99\%$, increasing the total explained variance to 3.22%. Thus, the non-monotonic component accounts for the largest share of explained variance in this example. This decomposition demonstrates that the identified non-monotonicity is not only statistically significant but also practically relevant, contributing substantially more explanatory power than either the linear or monotonic components. More generally, the reported ΔR^2 values provide an interpretable measure of how much additional structure is captured by increasingly flexible models and therefore complement the formal hypothesis tests used for relationship classification.

Monotonic Relationship Figure 4b shows an example of a monotonic relationship in the `email-W3C` hypergraph. Visually, the unconstrained GAM (blue line) closely follows the monotonic GAM (green line). The visual similarity and overlapping confidence intervals suggest that the additional flexibility of the unconstrained GAM is unnecessary. This is confirmed by the first ANOVA test, which yields a high p-value, indicating no significant difference of $\Delta R^2 = -0.01\%$ between the models. Following our procedure, we proceed to test whether a simple linear regression (dotted red line) could adequately describe the data. Here, the regression line visibly diverges from the monotonic GAM fit "losing" non-negligible $\Delta R^2 = 0.86\%$ being statistically significant at p-value $= 1.5 \times 10^{-26}$, confirming that a linear model is insufficient. Therefore, the final classification is monotonic. No further testing is required. The monotonic relationship illustrates a potential capacity constraint of e-mail users, resulting in a trade-off between sending many e-mails to a small number of recipients (< 5) and sending fewer e-mails to larger groups ($\geqslant 5$).

Linear Relationship The `hospital-lyon` hypergraph, depicted in Figure 4c, provides a clear example of a linear relationship. This dataset captures group interactions among healthcare workers and patients. Model fits show a consistent positive trend: individuals involved in larger group interactions tend to accumulate more total interactions. This proportional relationship implies that, for instance, knowing X times more people results in $\approx X$ times more interactions. This aligns with intuition in the hospital setting, where participation in larger gatherings, such as rounds or shift changes, implies broader involvement.

Formal testing support the linear interpretation. The first ANOVA test comparing the unconstrained GAM and a monotonic GAM yields $\Delta R^2 = 0.00\%$ and p-value close to 1, suggesting no evidence against monotonicity. A second ANOVA test comparing the monotonic GAM to a linear model produces $\Delta R^2 = 0.11\%$ and a p-value of 0.012. Under our stringent significance threshold of $\alpha = 0.00001$, this result is not sufficient to reject the linear model. Additionally, a comparison between the linear fit and a constant mean model strongly favours the former (with $R^2 = 11.37\%$ and a p-value of 4×10^{-118}), reinforcing the classification as a linear relationship.

No Relationship Figure 4d illustrates a case of no significant relationship in the `vegas-bars-reviews` hypergraph. This dataset captures Yelp users (nodes) who reviewed the same bar in Las Vegas within a one-month period (hyperedges). The number of reviews per bar per month (i.e., hyperedge size) ranges from 2 to 73, while individual users have submitted between 1 and 147 reviews overall, with an average of 9.6. Despite this wide variability, the relationship between the number of reviews a bar receives in a given month and the total number of reviews submitted by its reviewers appears to be flat, with no clear trend.

This is confirmed through a sequence of statistical tests. The first ANOVA test comparing an unconstrained GAM to a monotonic GAM yields a $\Delta R^2 = 0.01\%$ and p-value of 0.065, suggesting limited evidence against monotonicity. The second test comparing the monotonic GAM to a linear model produces a $\Delta R^2 = 0.04\%$ and p-value of 0.016, indicating that the linear model fits the data sufficiently well. However, an F-test comparing the linear model to a constant mean returns a $R^2 = 0.14\%$ and p-value of 4.6×10^{-5}, leading to the rejection

of the linear model under our strict significance threshold of $\alpha = 0.00001$. Thus, despite a slight upward trend suggested by the positive Pearson correlation ($r = 0.037$), the final classification for this configuration is "no relationship." It is worth noting that under a more conventional threshold (e.g., $\alpha = 0.01$), the relationship would be classified as linear.

From an interpretive standpoint, one might expect a weak positive relationship in this setting. Bars that attract many reviewers in a given month may be more popular or prominent, and such venues are likely to be visited and reviewed by more active Yelp users, who tend to submit reviews more frequently overall. However, the noisy nature of user behaviour, varying reviewing habits, and the casual context of online review platforms likely dilute any clear structural trend, resulting in only a mild correlation that fails to reach significance under strict criteria. This suggests that months with a higher volume of reviews are not driven by systematically more or less engaged users; rather, review activity appears to be generated by users with similar levels of engagement across months.

3.3.2 Distribution of Relationship Types

This subsection reports the distribution of the four identified relationship types across the 36 empirical hypergraphs and provides inferential insight into what might be expected in the broader population of hypergraphs.

Table 8 in the Appendix presents, for each hypergraph, the size effects ΔR^2 of competing models and its p-values from the three statistical tests and the relationship type classification. To summarize these detailed results, Table 3 shows the empirical distribution of relationship types across the dataset. Among the 36 hypergraphs, only 3 (8.3%) exhibit no discernible relationship between hyperedge size and node degree. A majority of hypergraphs (18/36, 50%) show a monotonic relationship (including linear), while the remaining 15 (41.7%) demonstrate a non-monotonic relationship. This distribution already provides strong empirical evidence against no relationship in real-world hypergraph data.

Relationship Type	**Count**	**Share (%)**	**Cumulative Share (%)**
No relationship	3	8.3	8.3
Linear	6	16.7	25.0
Monotonic	12	33.3	58.3
Non-monotonic	15	41.7	100.0

Table 3: Distribution of four identified relationship types—non-monotonic, monotonic, linear, and no relationship—across 36 empirical hypergraphs.

Although the use of a highly stringent significance threshold of $\alpha = 10^{-5}$ is non-standard, it is adopted deliberately to reduce the likelihood of false positives and to reinforce the robustness of our classifications in the presence of large datasets under extremely conservative inference. Such a low threshold strongly favours simpler explanations, including linear or even non-existent relationships. Despite this, under $\alpha = 10^{-5}$ only three out of 36 empirical hypergraphs are classified as exhibiting no relationship, while the majority already display statistically detectable dependencies that are predominantly non-monotonic and non-linear. To complement this conservative analysis, we additionally report the same classification using a more conventional threshold of $\alpha = 0.05$. Under this standard choice, the number of hypergraphs classified as exhibiting no relationship decreases from three to zero, and the number classified as linear decreases from six to two, while the number of non-monotonic relationships increases from 15 to 21, corresponding to 58.3% of all empirical hypergraphs. This comparison shows that non-existent relationships between node degree and hyperedge size are very rare, and that the prevalence of non-monotonic and non-linear relationships is not an artifact of liberal significance testing but persists, and indeed becomes more pronounced, under standard inference.

To generalize beyond our finite sample, we conduct Bayesian inference using non-informative uniform priors and derive posterior Beta distributions for selected proportions. Specifically, we compute Bayesian Credible Intervals (BCIs) for the fractions of hypergraphs showing no relationship and those showing non-monotonic relationships. For the "no relationship" category (3 out of 36), the posterior mean is 10.5%, with a 95% BCI of $(3.0\%, 21.9\%)$ and a more conservative 99% BCI of $(1.9\%, 26.6\%)$. These intervals suggest that, in the general hypergraph population, the fraction of cases with no relationship is likely below 25%, reinforcing the idea that such cases are relatively rare.

Conversely, the proportion of non-monotonic relationships is estimated at 42.1% (posterior mean), with a 95% BCI of $(27.1\%, 57.9\%)$ and a 99% BCI of $(22.0\%, 62.7\%)$. This indicates that non-monotonic relationships may not only be common but could even constitute the majority in the broader population. Taken together, these findings highlight the prevalence and diversity of structural dependencies in real-world hypergraphs and call into question modeling assumptions that treat group size and individual connectivity as unrelated.

Given this empirical evidence, it would be misleading to assume that no relationship exists between hyperedge size and node degree in a randomly chosen hypergraph. On the contrary, one should generally expect at least a monotonic relationship, if not a more complex, non-monotonic pattern.

4 Discussion

4.1 Methodological result I: bipartite representation as the preferred preprocessing strategy

Our first and major methodological result is that the bipartite representation is the preferred preprocessing strategy for capturing the relationship between hyperedge size and node degree , as demonstrated in Section 3.1. Below, we discuss why the choice of a preprocessing strategy is both relevant and not *a priori* obvious, summarize the empirical evidence supporting the bipartite representation, examine the theoretical relationships between the three considered approaches, and clarify the interpretational scope and appropriate use cases of each preprocessing strategy.

The choice of preprocessing strategy is not *a priori* obvious. Each of the three strategies considered in this study corresponds to a distinct sampling perspective and addresses a different structural research question. The node-centric and hyperedge-centric approaches are not merely technical alternatives to the bipartite representation, but meaningful summaries in their own right, capturing the behaviour of a typical node or a typical hyperedge, respectively. Although there are conceptual arguments favouring bipartite representations, such as their non-lossy nature and the symmetric treatment of nodes and hyperedges, these arguments alone do not determine which preprocessing strategy is most appropriate for empirical analysis. Consequently, without systematic evidence, it remains unclear which representation should be preferred in practice.

Empirical evidence supporting the bipartite representation. As was demonstrated in Section 3.1.1, a bipartite representation consistently yields the highest Eta-squared (η^2) values — our evaluation criterion — across all 3 correlation measures, i.e. Pearson, Spearman, or Kendall, indicating strong alignment with the semantic segmentation of hypergraphs and low within-segment variability (Section 3.1.2). While the choice of correlation coefficient has some influence on Eta-squared, it is secondary to the choice of representation: in the bipartite framework, all three correlation coefficients yield similar results. These findings emphasize that selecting an appropriate preprocessing method, i.e., the bipartite representation, is more crucial than the specific correlation measure when summarizing the hyperedge size–degree relationship.

Theoretical relationship between preprocessing strategies. The three preprocessing strategies considered in this study are theoretically linked through the structure of the hypergraph incidence matrix. This connection is already apparent empirically. As shown in Section 3.1.3, pairwise comparisons of Pearson correlation coefficients obtained under different preprocessing strategies reveal a clear positive association. In particular, Pearson correlations derived from the bipartite and edge-centric (or node-centric) representations exhibit a coefficient of determination of $R^2 \approx 0.60$, indicating a moderate but systematic alignment between bipartite representation and two other approaches. Here, we provide a theoretical explanation for this empirical relationship by focusing on the Pearson correlation coefficient and showing that its covariance term admits an exact correspondence across preprocessing strategies once appropriate weighting is applied.

It can be shown (see Appendix A.1) that the covariance between node degree k_i and hyperedge size s_j computed under the bipartite preprocessing is equivalent to the covariance between k_i and the node-averaged hyperedge size $\bar{s}_i$ in the node-centric representation, provided that nodes are weighted by their degree, that is $w_i = k_i$. An analogous identity holds between the bipartite and hyperedge-centric representations, where the same bipartite covariance can be expressed as a size-weighted covariance between the hyperedge-averaged degree

$\bar{k}_j$ and the hyperedge size s_j, with weights $w_j = s_j$. Together, these relationships yield the following covariance identities:

$$\mathrm{Cov}_{(i,j):\, b_{ij}=1}\big(k_i, s_j\big) = \mathrm{Cov}^{n}_{i=1\,;\, w_i=k_i}\big(k_i, \bar{s}_i\big) = \mathrm{Cov}^{m}_{j=1\,;\, w_j=s_j}\big(\bar{k}_j, s_j\big). \tag{6}$$

Here $\bar{s}_i = \sum_j b_{ij} s_j / k_i$ and $\bar{k}_j = \sum_i b_{ij} k_i / s_j$, with node and hyperedge weights given by $w_i = k_i$ and $w_j = s_j$, respectively. Equation (6) establishes an exact equivalence between the three preprocessing strategies at the level of covariances.

Despite the above correspondence for covariances, no such general relationship exists for the variances appearing in the denominator of the Pearson correlation coefficient. As a consequence, Pearson correlation coefficients obtained from different preprocessing strategies are not exactly related through weighting. Nevertheless, because the Pearson denominator is always positive, the weighted-covariance equivalence implies that, after applying the appropriate weighting, the signs of the corresponding Pearson coefficients are consistent across approaches.

Interpretational differences and methodological choice. Each preprocessing strategy corresponds to a distinct sampling perspective and therefore answers a different structural question. The node-centric preprocessing characterizes the behaviour of a typical node, while the hyperedge-centric approach focuses on a typical hyperedge. The bipartite preprocessing, by contrast, corresponds to uniformly sampling node–hyperedge incidences and quantifies assortative mixing at the level of individual memberships. This perspective is particularly appropriate when the goal is to characterize degree–size dependence as an emergent structural property of the hypergraph rather than as an average property of nodes or hyperedges. At the same time, when the research question explicitly concerns typical nodes or typical hyperedges, the alternative preprocessing strategies remain meaningful and appropriate. From this viewpoint, the three approaches should be regarded as complementary rather than competing, with the bipartite representation being preferred when no single perspective is clearly privileged.

4.2 Methodological result II: Pearson correlation as the preferred measure of dependence

A second methodological result concerns the choice of a correlation coefficient for summarizing the relationship between hyperedge size and node degree. Although Pearson, Spearman, and Kendall correlations achieve comparable segment-level alignment as measured by η^2 (ranging from 0.66 to 0.68 under bipartite preprocessing; see Table 1), they differ substantially in how reliably they capture the direction of the global relationship. Below, we summarize the empirical evidence supporting Pearson correlation as the preferred default measure, clarify the interpretational limitations of rank-based alternatives, and discuss the theoretical relationship between correlation signs and monotone GAMs, which, although informative, does not fully account for the observed empirical patterns and therefore motivates the empirical analysis.

Empirical evidence supporting Pearson correlation. Pearson correlation exhibits the most consistent agreement with the monotonicity direction inferred from shape-constrained GAMs. Specifically, Pearson aligns with GAM monotonicity in 89% of cases at $\alpha = 10^{-5}$ and in 97% of cases at $\alpha = 0.05$ (Table 2). By contrast, Spearman and Kendall correlations align in only approximately two-thirds of cases and result in sign reversals.

Discrepancies arise in hypergraphs with multimodal, weakly curved, or U-shaped relationships. In such settings, Pearson typically returns values close to zero, reflecting the absence of a clear global trend, whereas rank-based correlations often report stronger associations with potentially misleading signs. These differences indicate that Spearman and Kendall correlations are more sensitive to local structure, while Pearson more robustly captures the dominant global direction of dependence.

Taken together, these results suggest that Pearson correlation is better suited for global characterization of relationship. Accordingly, we recommend Pearson correlation as the default measure of dependence, particularly when the goal is to summarize overall trends. Rank-based correlations remain valuable for detecting monotonic associations; however, in the presence of multimodality , they require careful interpretation.

Theoretical relationship between correlation signs and monotone GAMs. The strong alignment between the sign of the Pearson correlation and the monotonicity direction selected by a monotone GAM, demonstrated in Section 3.2.1, is not incidental and can be partially rationalized on theoretical grounds. Let

$m(x) = \mathbb{E}[Y \mid X = x]$ denote the conditional expectation function. A standard identity yields $\mathrm{Cov}(X, Y) = \mathrm{Cov}(X, m(X))$, see [7]. If $m(\cdot)$ is monotone increasing, then $\mathrm{Cov}(X, m(X)) \geqslant 0$ and hence a non-negative population Pearson numerator. Thus, when the underlying conditional mean is monotone, Pearson sign agreement is expected in the population. An analogous implication can be formulated for rank-based measures such as Spearman's correlation, but under substantially stronger conditions. A non-negative Spearman sign generally requires that the conditional distribution of Y given X is stochastically increasing in X, for example in the sense of first-order stochastic dominance. Consequently, there exist settings in which $m(\cdot)$ is increasing but the stochastic dominance condition is violated, e.g. due to higher variance, in which case Spearman's correlation may be negative even though the Pearson correlation is non-negative. Consistent with this theoretical expectation, for all 18 hypergraphs classified in Section 3.3 as exhibiting monotonic (12) or linear (6) relationships, the monotonicity direction identified by the monotone GAM agrees with the sign of both the Pearson and Spearman correlations, although such agreement for the latter is not guaranteed on theoretical grounds.

When the conditional expectation $m(\cdot)$ is non-monotone, no general sign equivalence holds between Pearson (or Spearman) correlation and the direction selected by a monotone GAM. In this setting, the sign of the Pearson correlation is governed by $\mathrm{Cov}(X, m(X))$ and therefore reflects the net, global trend of $m(\cdot)$. As a consequence, sign alignment remains common when a non-monotone pattern exhibits a clear overall increase or decrease, while disagreements may arise when $\mathrm{Cov}(X, Y)$ is close to zero. In such cases, the Pearson sign is unstable.

This situation is exemplified by the `email-enron` dataset, which is classified as non-monotonic. The increasing model yields a slightly better fit in terms of sum of squared errors, while the Pearson correlation is $r = -0.002$ and statistically insignificant even at $\alpha = 0.05$. In these cases, the sign of the Pearson coefficient may be either positive or negative depending on slight asymmetries in the curve. In this example, the selected monotonic GAM indicates an increasing trend, creating a formal misalignment with the negative Pearson coefficient.

However, the above example of `email-enron` is the only instance of disagreement with Pearson among the 15 hypergraphs classified as non-monotonic. By contrast, five non-monotonic hypergraphs exhibit similar disagreement between Spearman sign and monotone GAM monotonicity, namely `dblp`, `email-eu`, `kaggle-whats-cooking`, `house-bills`, and `SFHH-conference`[6]. The larger number of discrepancies for Spearman is not predicted by existing theoretical considerations and therefore necessitates empirical investigation.

While strong theoretical arguments support sign alignment between Pearson correlation, and to a lesser extent Spearman correlation, and monotone GAMs in the case of genuinely monotonic relationships, no analogous guarantee exists for non-monotonic conditional mean structures. Such non-monotonic relationships constitute a substantial fraction of the empirical cases considered here, accounting for 41.7% of identified relationships at $\alpha = 10^{-5}$ and 58.3% at $\alpha = 0.05$, as shown in Section 3.3.2. Given this prevalence of non-monotonic relationships, neither Pearson- nor Spearman-based theoretical implications alone are sufficient to characterize expected sign alignment with monotone GAMs. Consequently, the empirical assessment of sign alignment reported in Section 3.2 is both necessary and informative for evaluating the practical consistency of these measures in the presence of non-monotonic relationships.

Robustness Check: Logarithmic Feature Transformation An important design decision concerns whether to apply natural logarithm transformations of either the hyperedge size or node degree

Our empirical evaluation shows that logarithmic transformation does not alter the conclusions of our study. Specifically, the Pearson correlations between outcomes obtained from logarithmic and non-logarithmic setups exceed 96% across all cases. Therefore, for clarity and consistency, we focus our main results on the non-logarithmic setup.

4.3 Empirical result: prevalence of edge size–degree relationships

A main empirical result concerns the typical form of the relationship between hyperedge size and node degree. Across 36 empirical hypergraphs, dependencies are the rule rather than the exception. Below, we summarize the empirical evidence and discuss the robustness of this conclusion.

Empirical evidence for prevalence and complexity. Only 3 out of 36 hypergraphs (8.3%) are classified as exhibiting no relationship between hyperedge size and node degree. The remaining 91.7% display , split between monotonic relationships (18/36, including 6 linear cases) and non-monotonic relationships (15/36). When the

[6]This number differs from that reported in Section 3.2 because here we exclude hypergraphs classified as exhibiting no relationship and focus solely on sign agreement, irrespective of statistical significance.

threshold is relaxed to the conventional level of $\alpha = 0.05$, the prevalence of structural dependence becomes even more pronounced: the number of "no relationship" cases drops to zero, linear cases decrease from six to two, and non-monotonic relationships increase to 21 out of 36 (58.3%). Taken together, these results show that independence or simple linear dependence between hyperedge size and node degree is atypical , while non-linear and often non-monotonic relationships are common.

Robustness with respect to variable assignment. The prevalence of complex relationships is robust to the choice of which variable is treated as the predictor. Repeating the entire classification procedure after reversing the roles of hyperedge size and node degree leaves the number of "no relationship" cases unchanged at three, while substantially increasing the number of hypergraphs classified as non-monotonic, from 15 to 22. At the same time, the number of linear classifications drops sharply, from six to one. More than half of the hypergraphs retain identical classifications across both configurations, and most discrepancies correspond to shifts toward more complex categories rather than simplification. This pattern suggests that our original results, obtained with hyperedge size as the predictor, if anything understate the prevalence of non-monotonic structure and therefore provide a conservative view of the underlying dependencies.

4.4 Theoretical implications for social dynamics

The primary motivation for this study is to identify and quantify the relationship between two fundamental structural properties of hypergraphs: hyperedge size and node degree. This relationship is of particular interest due to its potential influence on dynamical processes that unfold on such higher-order structures. This section reviews selected processes modelled on hypergraphs and proposes hypotheses on how structural correlations might impact these dynamics. By doing so, we aim to highlight the broader relevance of our empirical findings to modelling and understanding complex systems through a higher-order lens.

Spreading Dynamics of Social Contagions and Epidemics Traditional models of contagion dynamics, such as SIS or SIR , have been extensively studied in the context of pairwise networks [5, 50, 61]. More recently, researchers have extended these models to higher-order structures, particularly hypergraphs, which allow for the direct modelling of group interactions that cannot be reduced to dyads [94, 16, 103, 29]. However, this line of research has not yet addressed the potential role of correlation between hyperedge size and node degree.

We hypothesize that a positive correlation — where highly active nodes populate exceptionally large groups — acts as a structural accelerator for simple contagions by maximizing broadcast reach and lowering epidemic thresholds. However, it simultaneously hinders complex contagions, which require dense, overlapping environments to provide sufficient social reinforcement. For complex contagions, a negative correlation (hubs concentrating in small groups) is hypothesized to be far more efficient at driving rapid adoption.

Social Influence Diffusion Process The core problem in this context is known as Social Influence Maximization (SIM): identifying a small seed set of influential individuals in a network whose activation leads to maximal spread of influence [68]. SIM and its variants have been studied on graphs where nodes represent individuals and edges represent pairwise interactions. One prominent formulation is the Target Set Selection (TSS) problem [59], which seeks the smallest set of initially activated nodes that can eventually activate the entire network under a diffusion model, typically the linear threshold (LT) model [46, 98]. In reality, however, social interactions are often group-based rather than pairwise. Hence, hypergraphs offer a natural representation. Extensions of SIM and TSS to hypergraphs are therefore crucial for modelling higher-order social influence.

An open question concerns how the correlation between node degrees and hyperedge sizes impacts the required size of the seed set. We hypothesize that this correlation directly impacts how influence is distributed. If high-degree nodes systematically dominate massive hyperedges, influence is highly centralized. In such systems, a very small, targeted seed set can efficiently trigger global cascades. Conversely, a negative correlation implies a more balanced, decentralized participation structure, meaning optimal influence maximization would require a larger, more distributed seed set.

Cooperation in the Public Goods Game The public goods game is a model in evolutionary game theory that extends the prisoner's dilemma to group interactions [99, 87]. In its simplest form, each of the G players in a group decides whether to contribute a token (cooperate) or not (defect). The total contributions are then multiplied by a synergy factor $r > 1$ and equally distributed among all group members, regardless of their

strategy. If N_c players cooperate, then a cooperator receives $\pi_C = rN_c/G - t$ (paying cost t), and a defector receives $\pi_D = rN_c/G$. Thus, the game captures the essential social dilemma: while defection is individually rational, collective cooperation yields the highest group payoff. Multiplayer games such as this are typically studied on classical graphs by randomly selecting an edge, i.e., a pair of neighbouring players (i, j).

In evolutionary game theory, large groups typically exacerbate the free-rider problem, destabilizing cooperation. We hypothesize that if highly active individuals are concentrated in small hyperedges (a negative correlation), they can form tight-knit, stable cooperative clusters that resist defection. If their activity is instead diluted across large hyperedges (a positive correlation), the structural support for cooperation is expected to collapse.

Summary Our focus is strictly on developing rigorous methodology to identify empirical relationship between node degree and hyperedge size that constrains potential dynamics, rather than simulating the emergent behavioural processes themselves. While the dynamic implications are left for future research, establishing this structural baseline allows us to formulate concrete, testable hypotheses.

4.5 Implications for Null Models and Generative Hypergraph Models

The identified degree–hyperedge size relationships may also have important implications for the development of randomized reference models and generative hypergraph models. A common strategy in network science is to construct null models that preserve selected structural constraints while randomizing all remaining aspects of the system, allowing researchers to determine whether an observed pattern represents an independent structural feature or merely emerges from more basic properties. This approach has been widely applied to bipartite and higher-order systems using entropy-based and configuration-model frameworks [64, 86, 95, 97, 96, 90].

From this perspective, the relationship between node degree and hyperedge size can itself be viewed as a structural observable whose significance should be assessed relative to suitable randomized reference models. Existing hypergraph generators and null models often preserve degree sequences and hyperedge-size distributions while typically implicitly assuming no systematic dependence between them. Our findings suggest that this assumption may overlook an important aspect of higher-order structure. Future work could therefore develop null models that explicitly preserve, control, or manipulate degree–hyperedge size relationships, for instance, in a similar way that [115] propose a method to control degree correlations in generated graphs. Such models would enable controlled experiments isolating the impact of this dependency on contagion dynamics, influence diffusion, cooperation, and other higher-order processes, in much the same way that graph configuration models are used to study the effects of degree distributions and assortativity in classical networks.

5 Conclusions and Further Research

The structural relationship between hyperedge size and node degree is a fundamental yet understudied property of hypergraphs. Understanding whether and how these two features co-vary is essential not only for characterizing real-world hypergraph data but also for designing more realistic generative models and interpreting dynamical processes. This section summarizes our key findings and offers guidance for future research.

Conclusions This study introduces a generalizable statistical framework to systematically investigate the empirical relationship between hyperedge size and node degree across 36 real-world hypergraphs, using linear models and flexible Generalized Additive Models (GAMs). We classified each hypergraph according to the complexity and direction of its underlying trend: linear, monotonic, non-monotonic, or absent, using a sequence of statistical tests. Our results reveal that such relationships are not only widespread but also often complex: nearly 42% of hypergraphs exhibit non-monotonic patterns, and only a small minority (3 out of 36) show no significant dependency. These findings directly challenge common assumptions of structural independence in generative models.

Importantly, while the existence of a degree–size relationship appears to be a pervasive feature of empirical hypergraphs, its direction and functional form are strongly domain-dependent, despite substantial variation across categories. Thus, our results should not be interpreted as supporting a universal positive or negative dependence, but rather as evidence that non-trivial degree–size relationships are common across diverse application domains. These patterns were interpretable based on the domain-specific meaning of nodes and hyperedges. For

example, a positive relationship is expected in contact networks where individuals participating in larger groups naturally accumulate more contacts, while negative correlations in political networks align with institutional constraints or specialization effects.

An important practical takeaway for data analysts and hypergraph modellers is the critical role of data preprocessing. Among the three examined strategies: node-centric, edge-centric, and bipartite representation, we recommend the bipartite projection as the default. It consistently exhibited the lowest within-segment variability and highest η^2 effect sizes.

Moreover, we found that the sign of classical correlation coefficients, especially Pearson, aligns well with the direction of the dominant trend estimated by GAMs. We also showed that Spearman and Kendall coefficients can misrepresent global trends in the presence of multimodal or U-shaped relationships.

Taken together, this work highlights that the relationship between hyperedge size and node degree is not a marginal feature but a structurally and semantically meaningful property of empirical hypergraphs. It should therefore be measured carefully, interpreted in context, and considered when designing models, generating synthetic data, or studying dynamics on hypergraph-structured systems.

Further Research Directions A direct application of this work is to inform the next generation of generative models for synthetic hypergraphs, which typically neglect relationships between hyperedge size and node degree. One approach is to leverage algorithms designed for degree–degree correlations in bipartite graphs, such as the method proposed by Xulvi-Brunet and Sokolov [115, 57], and adapt them for hypergraph construction via bipartite representations. Since Pearson correlations between hyperedge size and node degree translate directly to degree–degree correlations in bipartite graphs, steering such dependencies during bipartite construction enables generation of synthetic hypergraphs with empirically realistic structure.

However, as our findings show, linear measures like Pearson often fail to capture the full complexity of these relationships, many of which are non-linear and non-monotonic. A promising direction is to model this complexity explicitly by introducing dependencies into the data-generation process. For instance, by duplicating hyperedges according to simulated weights drawn from power-law distributions, one could mimic the empirical fluctuations and multimodal structures observed in real hypergraphs. These patterns often reflect structural dependencies and repeated groupings (e.g., recurring meetings or standard mailing lists), and embedding such mechanisms into generative models could substantially increase their realism.

Another critical avenue for future research is to examine how the correlation between hyperedge size and node degree affect dynamical processes on hypergraphs. From a theoretical perspective, this constitutes an original conceptual step: we use empirically observed structural correlations to motivate specific, falsifiable hypotheses. To make this relationship testable, we hypothesized specific mechanisms linking the node-degree and hyperedge-size correlation to dynamical outcomes across three canonical models of collective behaviour in Discussion subsection 4.4. While our results suggest that such correlations could influence diffusion dynamics, this remains a testable hypothesis for future modelling.

Finally, the statistical procedure developed in this study, implemented in R, provides an end-to-end framework for assessing the relationship type. Such a tool has potential applications in many disciplines that require classification of dependencies. Future research could extend the tool in several directions: accounting for correlated or clustered observations , richer feature engineering, such as logarithmic, Box–Cox, or user-defined , as well as packaging the method into libraries in R, Python, or Julia.

By combining rigorous empirical analysis, domain-informed interpretation, and practical tooling, this study lays the foundation for future theoretical, algorithmic, and applied research into the structural dependencies that underlie real-world hypergraph data.

Acknowledgements

We thank the reviewers for their careful reading and constructive feedback, which has substantially improved the clarity, transparency, and overall quality of the manuscript. We are particularly grateful for the exceptionally thorough and insightful comments, which were not only thoughtful but also highly actionable. These suggestions provided clear and concrete guidance that directly informed our revisions, leading to significant improvements in the presentation of our objectives, the transparency of our methodology, and the interpretability of our results. We believe the manuscript is considerably stronger as a result of this process. We also thank the Co-Editor Prof. Ulrik Brandes for overseeing the review process.

Mateusz Zawisza gratefully acknowledges Jordan Barrett (Department of Mathematics and Statistics, Dalhousie University, Halifax, NS, Canada) for sharing initial versions of code and datasets, as well as for valuable discussions. He also thanks the participants of the Modelling and Mining Complex Networks as Hypergraphs Workshop in 2024 at Toronto Metropolitan University for their questions and feedback, as well as seminar participants at Toronto Metropolitan University, Department of Mathematics, in particular, François Théberge (Tutte Institute for Mathematics and Computing).

CRediT authorship contribution statement

Bogumił Kamiński: Conceptualization, Methodology, Project administration, Supervision, Writing: review & editing. **Paweł Prałat**: Conceptualization, Methodology, Project administration, Supervision, Writing: review & editing. **Aleksander Wojnarowicz**: Data curation, Formal analysis, Investigation, Software, Validation, Roles/Writing: original draft. **Mateusz Zawisza**: Conceptualization, Formal analysis, Methodology, Investigation, Software, Validation, Visualization, Writing: original draft, Writing: review & editing.

Declaration of Interest

The authors declare no competing interests.

Funding

This research study was conducted in the framework of project BPI/PST/2024/1/00129/U/00001 financed by Polish National Agency for Academic Exchange under the Strategic Partnerships programme.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used ChatGPT (OpenAI) in order to assist with language editing, code documentation, and the formulation of section summaries. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

References

[1] Sameer Agarwal, Kristin Branson, and Serge Belongie. Higher order learning with graphs. In *Proceedings of the 23rd international conference on Machine learning*, pages 17–24, 2006.

[2] Sinan G Aksoy, Cliff Joslyn, Carlos Ortiz Marrero, Brenda Praggastis, and Emilie Purvine. Hypernetwork science via high-order hypergraph walks. *EPJ Data Science*, 9(1):16, 2020.

[3] Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74(1):47–97, 2002.

[4] Ilya Amburg, Nate Veldt, and Austin Benson. Clustering in graphs and hypergraphs with categorical edge labels. In *Proceedings of the web conference 2020*, pages 706–717, 2020.

[5] Roy M Anderson and Robert M May. *Infectious diseases of humans: dynamics and control*. Oxford university press, 1991.

[6] Miroslav Andjelković, Bosiljka Tadić, Slobodan Maletić, and Milan Rajković. Hierarchical sequencing of online social graphs. *Physica A: Statistical Mechanics and its Applications*, 436:582–595, 2015.

[7] Joshua D Angrist and Jörn-Steffen Pischke. *Mostly harmless econometrics: An empiricist's companion*. Princeton university press, 2009.

[8] Alessia Antelmi, Daniele De Vinco, and Carmine Spagnuolo. Hypergraphrepository: a community-driven and interactive hypernetwork data collection. In *International Workshop on Algorithms and Models for the Web-Graph*, pages 159–173. Springer, 2024.

[9] Alessia Antelmi, Daniele De Vinco, and Carmine Spagnuolo. Hypergraphrepository: A community-driven and interactive hypernetwork data collection. In Megan Dewar, Bogumił Kamiński, Daniel Kaszyński, Łukasz Kraiński, Paweł Prałat, François Théberge, and Małgorzata Wrzosek, editors, *Modelling and Mining Networks*, pages 159–173, Cham, 2024. Springer Nature Switzerland.

[10] J. Barrett, P. Prałat, A. Smith, and F. Theberge. Counting simplicial pairs in hypergraphs. *Journal of Complex Networks*, 2025. Accepted for publication, 39 pages.

[11] M. Barthelemy. A class of models for random hypergraphs. *Physical review. E*, 106 6-1:064310, 2022.

[12] Federico Battiston, Giulia Cencetti, Iacopo Iacopini, Vito Latora, Maxime Lucas, Alice Patania, Jean-Gabriel Young, and Giovanni Petri. Networks beyond pairwise interactions: structure and dynamics. *Physics Reports*, 874:1–92, 2020.

[13] Austin R Benson, Rediet Abebe, Michael T Schaub, Ali Jadbabaie, and Jon Kleinberg. Simplicial closure and higher-order link prediction. *Proceedings of the National Academy of Sciences*, 115(48):E11221–E11230, 2018.

[14] Austin R Benson, David F Gleich, and Jure Leskovec. Higher-order organization of complex networks. *Science*, 353(6295):163–166, 2016.

[15] Claude Berge. *Hypergraphs: combinatorics of finite sets*, volume 45. Elsevier, 1984.

[16] Ágnes Bodó, Gyula Y Katona, and Péter L Simon. Sis epidemic propagation on hypergraphs. *Bulletin of mathematical biology*, 78(4):713–735, 2016.

[17] Phillip Bonacich. Power and centrality: A family of measures. *American journal of sociology*, 92(5):1170–1182, 1987.

[18] Phillip Bonacich. Simultaneous group and individual centralities. *Social networks*, 13(2):155–168, 1991.

[19] Phillip Bonacich, Annie Cody Holdren, and Michael Johnston. Hyper-edges and multidimensional centrality. *Social networks*, 26(3):189–203, 2004.

[20] Stephen P Borgatti and Martin G Everett. Network analysis of 2-mode data. *Social networks*, 19(3):243–269, 1997.

[21] Ronald L Breiger. The duality of persons and groups. *Social forces*, 53(2):181–190, 1974.

[22] Alain Bretto. *Hypergraph Theory: An Introduction*. Springer, 2013.

[23] Ciro Cattuto, Wouter Van den Broeck, Alain Barrat, Vittoria Colizza, Jean-François Pinton, and Alessandro Vespignani. Dynamics of person-to-person interactions from distributed rfid sensor networks. *PloS one*, 5(7):e11596, 2010.

[24] Damon Centola and Michael Macy. Complex contagions and the weakness of long ties. *American journal of Sociology*, 113(3):702–734, 2007.

[25] John M. Chambers and Trevor J. Hastie. *Statistical Models in S*. Wadsworth & Brooks/Cole, Pacific Grove, CA, 1992.

[26] Philip S Chodrow. Configuration models of random hypergraphs. *Journal of Complex Networks*, 8(3):cnaa018, 2020.

[27] Philip S Chodrow, Nate Veldt, and Austin R Benson. Generative hypergraph clustering: From blockmodels to modularity. *Science Advances*, 7(28):eabh1303, 2021.

[28] Christophe Croux and Catherine Dehon. Influence functions of the spearman and kendall correlation measures. *Statistical methods & applications*, 19:497–515, 2010.

[29] Guilherme Ferraz de Arruda, Giovanni Petri, and Yamir Moreno. Social contagion models on hypergraphs. *Physical Review Research*, 2(2):023032, 2020.

[30] Manh Tuan Do, Se-eun Yoon, Bryan Hooi, and Kijung Shin. Structural patterns and generative models of real-world hypergraphs. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 176–186, 2020.

[31] Patrick Doreian. On the evolution of group and network structure. *Social Networks*, 2(3):235–252, 1979.

[32] Matt Dowle and Arun Srinivasan. data.table: Extension of data.frame. *R package version 1.12.8*, 2019.

[33] David Easley and Jon Kleinberg. *Networks, Crowds, and Markets: Reasoning about a Highly Connected World*. Cambridge University Press, New York, 2010.

[34] F. Essam, H. El, and S. R. H. Ali. A comparison of the pearson, spearman rank and kendall tau correlation coefficients using quantitative variables. *Asian Journal of Probability and Statistics*, 2022.

[35] Ernesto Estrada and Juan A Rodríguez-Velázquez. Subgraph centrality and clustering in complex hyper-networks. *Physica A: Statistical Mechanics and its Applications*, 364:581–594, 2006.

[36] Katherine Faust. Centrality in affiliation networks. *Social networks*, 19(2):157–191, 1997.

[37] David A. Freedman. *Statistical Models: Theory and Practice*. Cambridge University Press, 2009.

[38] John E. Freund and Benjamin M. Perles. *Statistics: A First Course*. Pearson, 9th edition, 2014.

[39] Andrew Gelman. Analysis of variance—why it is more important than ever. *The Annals of Statistics*, 33(1):1–53, 2005.

[40] Mathieu Génois and Alain Barrat. Can co-location be used as a proxy for face-to-face contacts? *EPJ Data Science*, 7(1):1–18, 2018.

[41] Mathieu Génois and Alain Barrat. Can co-location be used as a proxy for face-to-face contacts? *EPJ Data Science*, 7(1):1–18, 2018.

[42] Mathieu Génois, Christian L Vestergaard, Julie Fournet, André Panisson, Isabelle Bonmarin, and Alain Barrat. Data on face-to-face contacts in an office building suggest a low-cost vaccination strategy based on community linkers. *Network Science*, 3(3):326–347, 2015.

[43] Gourab Ghoshal and M E J Newman. Random hypergraphs and their applications. *Physical Review E*, 79(6):061109, 2009.

[44] Kwang-Il Goh, Michael E Cusick, David Valle, Barton Childs, Marc Vidal, and Albert-László Barabási. The human disease network. *Proceedings of the National Academy of Sciences*, 104(21):8685–8690, 2007.

[45] Peter Gould and Anthony Gatrell. A structural analysis of a game: the liverpool v manchester united cup final of 1977. *Social Networks*, 2(3):253–273, 1979.

[46] Mark Granovetter. Threshold models of collective behavior. *American journal of sociology*, 83(6):1420–1443, 1978.

[47] Jacopo Grilli, György Barabás, Matthew J Michalska-Smith, and Stefano Allesina. Higher-order interactions stabilize dynamics in competitive network models. *Nature*, 548(7666):210–213, 2017.

[48] Trevor Hastie and Robert Tibshirani. *Generalized Additive Models*, volume 43 of *Monographs on Statistics and Applied Probability*. Chapman and Hall/CRC, 1990.

[49] Xie He, Philip S. Chodrow, and Peter J. Mucha. Hypergraph link prediction via hyperedge copying. *ArXiv*, abs/2502.02386, 2025.

[50] Herbert W Hethcote. The mathematics of infectious diseases. *SIAM review*, 42(4):599–653, 2000.

[51] Lorenzo Isella, Juliette Stehlé, Alain Barrat, Ciro Cattuto, Jean-François Pinton, and Wouter Van den Broeck. What's in a crowd? analysis of face-to-face behavioral networks. *Journal of theoretical biology*, 271(1):166–180, 2011.

[52] Lorenzo Isella, Juliette Stehlé, Alain Barrat, Ciro Cattuto, Jean-François Pinton, and Wouter Van den Broeck. What's in a crowd? analysis of face-to-face behavioral networks. *Journal of theoretical biology*, 271(1):166–180, 2011.

[53] Bogumił Kamiński, Paweł Misiorek, Paweł Prałat, and François Théberge. Modularity based community detection in hypergraphs. *Journal of Complex Networks*, 12(5):cnae041, 2024.

[54] Bogumił Kamiński, Valérie Poulin, Paweł Prałat, Przemysław Szufel, and François Théberge. Clustering via hypergraph modularity. *PloS one*, 14(11):e0224307, 2019.

[55] Bogumił Kamiński, Paweł Prałat, and François Théberge. Community detection algorithm using hypergraph modularity. In *International Conference on Complex Networks and Their Applications*, pages 152–163. Springer, 2020.

[56] Bogumił Kamiński, Paweł Prałat, and François Théberge. Artificial benchmark for community detection (abcd)—fast random graph model with community structure. *Network Science*, 9(2):153–178, 2021.

[57] Bogumil Kaminski, Pawel Prałat, and François Théberge. *Mining complex networks*. Chapman and Hall/CRC, 2021.

[58] Bogumił Kamiński, Paweł Prałat, and François Théberge. Hypergraph artificial benchmark for community detection (h–abcd). *Journal of Complex Networks*, 11(4):cnad028, 2023.

[59] David Kempe, Jon Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 137–146, 2003.

[60] Maurice G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1-2):81–93, 1938.

[61] William Ogilvy Kermack and Anderson G McKendrick. A contribution to the mathematical theory of epidemics. *Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character*, 115(772):700–721, 1927.

[62] Jung-Ho Kim and K. Goh. Higher-order components dictate higher-order contagion dynamics in hypergraphs. *Physical review letters*, 132 8:087401, 2022.

[63] Nicholas W Landry and Juan G Restrepo. The effect of heterogeneity on hypergraph contagion models. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 30(10), 2020.

[64] Matthieu Latapy, Clémence Magnien, and Nathalie Del Vecchio. Basic notions for the analysis of large two-mode networks. *Social networks*, 30(1):31–48, 2008.

[65] Geon Lee, Fanchen Bu, Tina Eliassi-Rad, and Kijung Shin. A survey on hypergraph mining: Patterns, tools, and generators. *ACM Computing Surveys*, 57(8):1–36, 2025.

[66] Jure Leskovec, Jon Kleinberg, and Christos Faloutsos. Graph evolution: Densification and shrinking diameters. *ACM Transactions on Knowledge Discovery from Data*, 1(1), 2007.

[67] N. Levshina. Relationships between two quantitative variables: Correlation analysis with elements of linear regression modelling. *Quantitative Methods in Linguistics*, 2015.

[68] Yuchen Li, Ju Fan, Yanhao Wang, and Kian-Lee Tan. Influence maximization on social graphs: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 30(10):1852–1872, 2018.

[69] B. Luo. eventernote-places. Zenodo, May 2024.

[70] Stan Matwin, Aristides Milios, Paweł Prałat, Amilcar Soares, and François Théberge. *Generative Methods for Social Media Analysis*. Springer, 2023.

[71] J Miller McPherson. Hypernetwork sampling: Duality and differentiation among voluntary organizations. *Social Networks*, 3(4):225–249, 1982.

[72] Ron Milo, Shai Shen-Orr, Shalev Itzkovitz, Nadav Kashtan, Dmitri Chklovskii, and Uri Alon. Network motifs: Simple building blocks of complex networks. *Science*, 298(5594):824–827, 2002.

[73] Staša Milojević. Principles of scientific research team formation and evolution. *Proceedings of the National Academy of Sciences*, 111(11):3984–3989, 2014.

[74] Danielle Navarro. *Learning statistics with R: A tutorial for psychology students and other beginners. (Version 0.6)*. University of New South Wales, Sydney, Australia, 2015. R package version 0.5.1.

[75] Saket Navlakha and Carl Kingsford. The power of protein interaction networks for associating genes with diseases. *Bioinformatics*, 26(8):1057–1063, 2010.

[76] M. E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45(2):167–256, 2003.

[77] M. E. J. Newman, D. J. Watts, and S. H. Strogatz. Random graph models of social networks. *Proceedings of the National Academy of Sciences*, 99(Suppl 1):2566–2572, 2002.

[78] Mark Newman. *Networks*. Oxford university press, 2018.

[79] Mark EJ Newman. Assortative mixing in networks. *Physical review letters*, 89(20):208701, 2002.

[80] Mark EJ Newman. Modularity and community structure in networks. *Proceedings of the national academy of sciences*, 103(23):8577–8582, 2006.

[81] Mark EJ Newman and Michelle Girvan. Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113, 2004.

[82] Laura Ozella, Daniela Paolotti, Guilherme Lichand, Jorge P Rodríguez, Simon Haenni, John Phuka, Onicio B Leal-Neto, and Ciro Cattuto. Using wearable proximity sensors to characterize social contact patterns in a village of rural malawi. *EPJ Data Science*, 10(1):46, 2021.

[83] Luca Pappalardo, Paolo Cintia, Alessio Rossi, Emanuele Massucco, Paolo Ferragina, Dino Pedreschi, and Fosca Giannotti. A public data set of spatio-temporal match events in soccer competitions. *Scientific data*, 6(1):236, 2019.

[84] Alice Patania, Giovanni Petri, and Franco Vaccarino. The shape of collaborations. *EPJ Data Science*, 6(1):18, 2017.

[85] Karl Pearson. Mathematical contributions to the theory of evolution. iii. regression, heredity, and panmixia. *Philosophical Transactions of the Royal Society of London. Series A*, 187:253–318, 1896.

[86] Jorge Peña and Yannick Rochat. Bipartite graphs as models of population structures in evolutionary multiplayer games. *PLOS ONE*, 7(9):1–13, 2012.

[87] Matjaž Perc, Jesús Gómez-Gardenes, Attila Szolnoki, Luis M Floría, and Yamir Moreno. Evolutionary dynamics of group interactions on structured populations: a review. *Journal of the royal society interface*, 10(80):20120997, 2013.

[88] Janet Piñero, Juan Manuel Ramírez-Anguita, Josep Saüch-Pitarch, Francesco Ronzano, Emilio Centeno, Ferran Sanz, and Laura I Furlong. The disgenet knowledge platform for disease genomics: 2019 update. *Nucleic acids research*, 48(D1):D845–D855, 2020.

[89] Alexis Pister and Marc Barthelemy. Stochastic block hypergraph model. *Physical Review E*, 110(3):034312, 2024.

[90] Giulia Preti, Adriano Fazzone, Giovanni Petri, and Gianmarco De Francisci Morales. Higher-order null models as a lens for social systems. *Physical Review X*, 14(3):031032, 2024.

[91] N. Pya and S.N. Wood. Shape constrained additive models. *Statistics and Computing*, 25(3):543–559, 2015.

[92] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2024.

[93] John T. E. Richardson. Eta squared and partial eta squared as measures of effect size in educational research. *Educational Research Review*, 6(2):135–147, 2011.

[94] Martin Ritchie, Luc Berthouze, Thomas House, and Istvan Z Kiss. Higher-order structure and epidemic dynamics in clustered networks. *Journal of Theoretical Biology*, 348:21–32, 2014.

[95] Fabio Saracco, Riccardo Di Clemente, Andrea Gabrielli, and Tiziano Squartini. Randomizing bipartite networks: the case of the world trade web. *Scientific reports*, 5(1):10595, 2015.

[96] Fabio Saracco, Giovanni Petri, Renaud Lambiotte, and Tiziano Squartini. Entropy-based models to randomise real-world hypergraphs. *Communications Physics*, 8(1):284, 2025.

[97] Fabio Saracco, Mika J Straka, Riccardo Di Clemente, Andrea Gabrielli, Guido Caldarelli, and Tiziano Squartini. Inferring monopartite projections of bipartite networks: an entropy-based approach. *New Journal of Physics*, 19(5):053022, 2017.

[98] Thomas C Schelling. *Micromotives and macrobehavior*. WW Norton & Company, 2006.

[99] Karl Sigmund. *The calculus of selfishness*. Princeton University Press, 2010.

[100] Charles Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904.

[101] Juliette Stehlé, Nicolas Voirin, Alain Barrat, Ciro Cattuto, Vittoria Colizza, Lorenzo Isella, Corinne Régis, Jean-François Pinton, Nagham Khanafer, Wouter Van den Broeck, et al. Simulation of an seir infectious disease model on the dynamic contact network of conference attendees. *BMC medicine*, 9(1):87, 2011.

[102] Hanlin Sun and G. Bianconi. Higher-order percolation processes on multiplex hypergraphs. *Physical review. E*, 104 3-1:034306, 2021.

[103] Qi Suo, Jin-Li Guo, and Ai-Zhong Shen. Information spreading dynamics in hypernetworks. *Physica A: Statistical Mechanics and its Applications*, 495:475–487, 2018.

[104] Johan Ugander, Lars Backstrom, Cameron Marlow, and Jon Kleinberg. Structural diversity in social contagion. *Proceedings of the National Academy of Sciences*, 109(16):5962–5966, 2012.

[105] Edwin van den Heuvel and Zhuozhao Zhan. Myths about linear and monotonic associations: Pearson'sr, spearman's ρ, and kendall's τ. *The American Statistician*, 76(1):44–52, 2022.

[106] Philippe Vanhems, Alain Barrat, Ciro Cattuto, Jean-François Pinton, Nagham Khanafer, Corinne Régis, Byeul-a Kim, Brigitte Comte, and Nicolas Voirin. Estimating potential infection transmission routes in hospital wards using wearable proximity sensors. *PloS one*, 8(9):e73970, 2013.

[107] Stanley Wasserman and Katherine Faust. *Social network analysis: Methods and applications*. Cambridge university press, 1994.

[108] Duncan J Watts and Peter Sheridan Dodds. Influentials, networks, and public opinion formation. *Journal of consumer research*, 34(4):441–458, 2007.

[109] Duncan J Watts and Steven H Strogatz. Collective dynamics of 'small-world'networks. *nature*, 393(6684):440–442, 1998.

[110] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016.

[111] Hadley Wickham, Romain Francois, Lionel Henry, and Kirill Müller. dplyr: A grammar of data manipulation. *R package version 1.0.0*, 2020.

[112] S. N. Wood. *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC, 2 edition, 2017.

[113] W. Xu, C. Chang, Y. S. Hung, and S. K. Kwan. Order statistics correlation coefficient as a novel association measurement with applications to biosignal analysis. *IEEE Transactions on Biomedical Engineering*, 2007.

[114] Weichao Xu, Yunhe Hou, YS Hung, and Yuexian Zou. A comparative analysis of spearman's rho and kendall's tau in normal and contaminated normal models. *Signal Processing*, 93(1):261–276, 2013.

[115] Ramón Xulvi-Brunet and Igor M Sokolov. Changing correlations in networks: assortativity and dissortativity. *Acta Physica Polonica B*, 36(5):1431–1455, 2005.

[116] Hao Yin, Austin R. Benson, Jure Leskovec, and David F. Gleich. Local higher-order graph clustering. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM Press, 2017.

[117] Yuanzhao Zhang, M. Lucas, and F. Battiston. Higher-order interactions shape collective dynamics differently in hypergraphs and simplicial complexes. *Nature Communications*, 14, 2022.

A Appendix

A.1 Proof of the covariance equivalence between preprocessing strategies

Let $L = \{(i,j) : b_{ij} = 1\}$ denote the set of incidences and let

$$M = |L| = \sum_{i,j} b_{ij} \tag{7}$$

be the total number of incidences. We define the incidence-based means

$$\mu_k^{\mathrm{bi}} = \frac{1}{M} \sum_{(i,j)\in L} k_i, \qquad \mu_s^{\mathrm{bi}} = \frac{1}{M} \sum_{(i,j)\in L} s_j. \tag{8}$$

The bipartite covariance between node degree and hyperedge size is then given by

$$\mathrm{Cov}_{\mathrm{bi}}(k,s) = \frac{1}{M} \sum_{(i,j)\in L} \big(k_i - \mu_k^{\mathrm{bi}}\big)\big(s_j - \mu_s^{\mathrm{bi}}\big). \tag{9}$$

Grouping the sum by vertices yields

$$\sum_{(i,j)\in L} \big(k_i - \mu_k^{\mathrm{bi}}\big)\big(s_j - \mu_s^{\mathrm{bi}}\big) = \sum_{i=1}^{n} \big(k_i - \mu_k^{\mathrm{bi}}\big) \sum_{j:\, b_{ij}=1} \big(s_j - \mu_s^{\mathrm{bi}}\big). \tag{10}$$

Using the identities

$$\sum_{j:\, b_{ij}=1} s_j = k_i \bar{s}_i, \qquad \sum_{j:\, b_{ij}=1} 1 = k_i, \tag{11}$$

we obtain

$$\sum_{j:\, b_{ij}=1} \big(s_j - \mu_s^{\mathrm{bi}}\big) = k_i \big(\bar{s}_i - \mu_s^{\mathrm{bi}}\big). \tag{12}$$

Substituting this expression back into the covariance yields

$$\mathrm{Cov}_{\mathrm{bi}}\big(k_i, s_j\big) = \frac{1}{M} \sum_{i=1}^{n} k_i \big(k_i - \mu_k^{\mathrm{bi}}\big)\big(\bar{s}_i - \mu_s^{\mathrm{bi}}\big), \tag{13}$$

which is precisely the vertex-level covariance between k_i and $\bar{s}_i$ weighted by node degree $w_i = k_i$.

An analogous result for the hyperedge-centric representation can be obtained by grouping the incidence sum by hyperedges rather than vertices; since the steps are entirely symmetric, we omit the derivation for brevity.

A.2 Empirical Hypergraph Datasets: Domains and Descriptive Statistics

The dataset analyzed in this study consists of 36 empirical hypergraphs drawn from a wide variety of domains, encompassing both physical and digital social networks, biological networks and political structures. This breadth includes physical contact networks (e.g., `hospital-lyon`, `contact-primary-school`), online user interactions (e.g., `threads-math-sx`, `tags-ask-ubuntu`), institutional affiliations (e.g., `house-committees`, `senate-bills`), and domain-specific scientific data such as drug composition (`NDC-substances`) or disease-gene associations (`diseasome`, `disgenenet`). All 36 hypergraphs are organized into semantically coherent and internally homogeneous segments, such as User-Answer, Physical Contact, Email, and others. The complete list of datasets, along with their assigned segments and detailed interpretations of both nodes and hyperedges, is provided in Table 4. This interpretability is critical, as it allows us to meaningfully analyze and interpret the relationship between node degree and hyperedge size. The node and hyperedge semantics ensure that our calculated correlations are not just statistical artifacts but reflect domain-relevant structural patterns. All datasets are publicly available through their respective sources, and we additionally host them in our GitHub repository: `https://github.com/AleksanderWWW/hypergraph-properties`.

Hypergraph name	**Segment**	**Node interpretation**	**Hyperedge interpretation**
algebra [8]	User-Answer	users of mathoverflow.net	users who answered a particular type of question about algebra within a month
amazon [8]	Product-Category	products reviewed by users on Amazon	groups of similar items
contact-high-school [8]	Physical Contact	people at a high school	interactions at a resolution of 20 seconds
contact-primary-school [8]	Physical Contact	people at a primary school	interactions at a resolution of 20 seconds
dblp [8]	Part-Whole	DBLP paper authors	documents published between January and May 2017
diseasome [44]	Diseases and Gene	diseases	genes associated with diseases
disgenenet [88]	Diseases and Gene	genes associated with diseases	diseases
email-enron [8]	Email	email addresses at Enron	sender and all recipients of the email
email-eu [13] [116] [66]	Email	email addresses at a European research institution	sender and all receivers grouped by timestamp
email-W3C [8]	Email	email addresses on W3C mailing lists	set of email addresses on the same email
geometry [9]	User-Answer	users of mathoverflow.net	sets of users who answered a certain question category about geometry
got [8]	Person-Place	GoT characters	GoT scenes linking characters appearing in the same scene
hospital-lyon [106]	Physical Contact	patients and health-care workers in a hospital ward in Lyon, France	group interactions
music-blues-reviews [8]	User-Review	Amazon users	users who reviewed a blues music product within a month
nba [9]	Part-Whole	NBA players	players involved in a match up to 2012
NDC-classes [9]	Drugs	class labels applied to drugs	drugs
NDC-substances [9]	Drugs	substances making up a drug	drugs
restaurant-reviews [8]	User-Review	Yelp users	users who reviewed restaurants in Madison, WI within a month
tags-ask-ubuntu [13]	Tag-Question	tags	sets of tags applied to questions on askubuntu.com
tags-math-sx [13]	Tag-Question	tags	sets of tags applied to questions on math.stackexchange.com
threads-ask-ubuntu [8]	User-Thread	users on askubuntu.com	users participating in a thread lasting $\leqslant$ 24 hours
threads-math-sx [8]	User-Thread	users on math.stackexchange.com	users participating in a thread lasting $\leqslant$ 24 hours
twitter [8]	User-Thread		
vegas-bars-reviews [8]	User-Thread	Yelp users	users who reviewed the same bar in Las Vegas within a month
evernote-places [69]	Person-Place	artists or artist groups	places where idol/voice actor events took place
house-bills (House) [27]	Political	political affiliation	bill cosponsorship in the US House of Representatives

house-committees (House) [27]	Political	political affiliation	committee membership in the US House of Representatives
Hypertext-conference [51]	Participant-Conference	conference attendees	face-to-face interactions over 2.5 days
InVS13 [42], InVS15 [40], science-gallery [52]	Physical Contact	participants with sensors	snapshots of groups present at specific times
kaggle-whats-cooking [4]	Part-Whole	ingredients	dishes comprising those ingredients
Malawi-village [82]	Physical Contact	individuals living in a village	interactions in a rural Malawi village
house-bills (Senate) [27]	Political	political affiliation	bill cosponsorship in the US Senate
house-committees (Senate) [27]	Political	political affiliation	committee membership in the US Senate
SFHH-conference [41, 101, 23]	Participant-Conference	conference attendees	face-to-face contacts every 20 seconds

Table 4: List of hypergraph datasets with their sources, assigned semantic segments, and interpretations of nodes and hyperedges.

Beyond their domain diversity, the hypergraphs analyzed in this study exhibit substantial variation in structural characteristics, particularly in the size and distribution of node degrees and hyperedge sizes. This variability is essential for assessing the correlation and relationship between these two quantities. In contrast, classical graphs with binary edges lack such variability, as edge size is fixed at 2. Consequently, in standard graphs, the notion of a relationship between edge size and node degree is either undefined or trivially zero. Tables 5 and 6 provide detailed summary statistics of node degrees and hyperedge sizes, including the number of observations (`n`), average value (`avg`), standard deviation (`sd`), skewness (`skew`), and observed range (`range`).

name	**n**	**avg**	**sd**	**range**	**skew**
algebra	423	19.53	34.01	1–375	5.03
amazon	4989	1.02	0.18	1–4	8.98
contact-high-school	327	55.63	27.06	2–148	0.48
contact-primary-school	242	126.98	55.15	28–261	0.31
dblp	71116	1.24	0.80	1–25	7.33
diseasome	516	2.15	2.15	1–22	3.84
disgenenet	12368	9.09	16.87	1–377	6.67
email-enron	143	32.33	24.26	2–118	1.22
email-eu	1005	88.96	116.35	1–918	2.52
email-W3C	5601	2.39	11.43	1–282	17.23
geometry	580	21.53	36.26	1–260	3.72
got	577	20.99	59.79	1–632	5.83
hospital-lyon	75	59.03	48.99	6–205	1.22
music-blues-reviews	1106	9.49	10.72	1–127	3.25
nba	2191	293.95	308.26	1–1476	1.12
NDC-classes	1161	134.53	402.96	1–5357	7.88
NDC-substances	5311	10.08	35.11	1–579	8.85
restaurant-reviews	565	8.14	7.22	1–59	3.51
tags-ask-ubuntu	3029	164.84	606.11	1–12931	10.31
tags-math-sx	1629	364.10	1039.61	1–13950	6.80
threads-ask-ubuntu	125602	2.76	20.78	1–2332	51.55
threads-math-sx	176445	9.13	92.98	1–12511	59.98
twitter	22964	2.21	4.61	1–266	18.03
vegas-bars-reviews	1234	9.62	7.37	1–147	7.85
eventernote-places	71890	9.92	25.02	1–421	5.70
house-bills	1494	835.79	815.06	1–6220	2.10
house-committees	1290	9.18	7.09	1–44	1.16
Hypertext-conference	113	345.56	304.16	2–1446	1.52
InVS13	92	210.65	193.14	5–1089	2.13

name	n	avg	sd	range	skew
InVS15	217	691.01	488.80	1–3192	1.50
kaggle-whats-cooking	6714	63.78	388.31	1–18048	22.99
Malawi-village	86	2338.01	1780.06	12–7636	0.63
Science-Gallery	10972	65.41	56.06	1–486	1.66
senate-bills	294	789.62	640.09	1–3514	1.31
senate-committees	282	19.18	14.85	1–63	0.85
SFHH-conference	403	289.42	311.67	2–1960	2.64

Table 5: Node degree distribution

name	n	avg	sd	skew	range
algebra	1268	6.52	6.58	6.32	2–107
amazon	1176	4.35	2.27	-0.71	1–6
contact-high-school	7818	2.33	0.53	1.38	2–5
contact-primary-school	12704	2.42	0.55	0.88	2–5
dblp	25624	3.45	2.12	5.24	1–69
diseasome	481	2.31	1.50	1.95	1–11
disgenenet	2069	54.36	169.31	7.38	1–2453
email-enron	1514	3.05	2.29	5.92	1–37
email-eu	25148	3.56	3.40	4.51	1–40
email-W3C	6000	2.23	0.99	10.11	2–23
geometry	1193	10.47	15.65	4.11	2–230
got	4165	2.91	2.35	2.33	0–24
hospital-lyon	1824	2.43	0.56	0.92	2–5
music-blues-reviews	694	15.13	14.71	1.81	2–83
nba	31686	20.33	1.89	0.18	14–28
NDC-classes	49724	3.14	2.10	2.66	1–24
NDC-substances	9906	5.40	5.78	1.49	1–25
restaurant-reviews	601	7.66	7.28	1.90	2–43
tags-ask-ubuntu	147222	3.39	1.03	0.04	1–5
tags-math-sx	170476	3.48	0.97	0.02	1–5
threads-ask-ubuntu	192947	1.80	0.80	1.29	1–14
threads-math-sx	719792	2.24	1.04	1.48	1–21
twitter	4065	12.51	16.90	3.40	1–207
vegas-bars-reviews	1194	9.94	13.82	2.65	2–73
eventernote-places	19033	37.48	185.35	12.84	0–6420
house-bills	60987	20.47	33.83	4.27	2–399
house-committees	341	34.73	21.39	-0.03	1–81
Hypertext-conference	19036	2.05	0.24	5.51	2–6
InVS13	9644	2.01	0.10	10.72	2–4
InVS15	73822	2.03	0.18	5.50	2–4
kaggle-whats-cooking	39774	10.77	4.43	0.86	1–65
Malawi-village	99942	2.01	0.11	9.18	2–4
Science-Gallery	338765	2.12	0.35	3.13	2–5
senate-bills	29157	7.96	10.27	3.23	2–99
senate-committees	315	17.17	6.79	-0.53	4–31
SFHH-conference	54305	2.15	0.50	5.34	2–9

Table 6: Hyperedge size distribution

This structural heterogeneity, combined with interpretability and semantic clarity, makes our dataset partic-

ularly suitable for a robust investigation of correlations between hyperedge size and node degree. The richness of the dataset ensures that the findings are not limited to a single domain or structure, while the semantic interpretability allows us to validate the significance of results in real-world terms. Altogether, this provides a strong foundation for generalizable and meaningful analysis.

A.3 Computational Implementation and Complexity

In this section, we outline the implementation details behind the analysis hypergraph properties employed in this paper. The key components involve:

- data ingestion from various source formats,
- construction of efficient data structures for hypergraph representation,
- tools used to optimize computations on hypergraphs.

The data used for the process was obtained from different sources and was therefore stored in diverse file formats. Those included JSON (JavaScript Object Notation), HGF (Hypergraph format), XGI (CompleX Group Interactions) and plain text (.txt). A different strategy was necessary for each. Additionally, for all but for JSON the code for line-by-line reading had to be crafted (Python's built-in `json` library handled JSON files without the need of custom reading and parsing implementations).

Based on the loaded file contents, an instance of a sparse matrix was created. Choosing this type of data structure allowed for efficient storage of large hypergraphs (dense matrices would quickly drain memory resources and cause crashes in the processing pipeline), while remaining on par with the representation used in literature. The latter significantly simplified the translation from theory into software implementation.

For the JSON files, it was possible to use the `scipy.sparse.coo_array` object and construct the entire sparse matrix in one function call. In the other cases, an incremental line-by-line approach was needed. For synthetic hypergraphs generated by the h–ABCD synthetic benchmark [58], `scipy.sparse.lil_array` was the most efficient type for construction, whereas for empirical hypergraphs, `scipy.sparse.dok_array` performed best. Based on this observation, we recommend further research into a potential relationship between a hypergraph type, and the optimal sparse matrix type for incremental construction. In all cases, the constructed object was converted in the `scipy.sparse.csr_array` type, as it is best suited for fast data indexing, crucial to the calculations performed in the following parts of the process.

The achieved space complexity was $O(nnz) + O(n)$, where nnz is the number of non-zero elements and n is the number of rows in the matrix. This is more efficient for hypergraphs, where the nnz will typically be substantially smaller than the dense matrices' $O(n \cdot m)$ complexity (m being the number of columns).

Upon receiving the object representing a hypergraph, the downstream tasks in the pipeline utilized `numpy`'s array and `scipy`'s sparse array methods optimized for fast vector computations, to calculate correlations and descriptive statistics of the data. Initially, `numba` was employed in hopes of taking advantage of the JIT (Just-in-time) compilation. This approach, however, yielded no significant improvements in the processing speed, while increasing the complexity of the implementation details.

Predominant operations in the process of correlation computation were summing over rows, summing over columns and indexing non-zero elements of the CSR matrix. The first two operations are of $O(nnz)$ time complexity, and the indexing of $O(1)$. Those characteristics allowed for efficient data processing even for large hypergraph files.

All statistical analyses, including correlation computations, model fitting, and figure generation, were performed using the R programming language [92]. Correlation measures such as Pearson's r, Spearman's ρ, and Kendall's τ were calculated using base R functions, while statistical modelling was conducted using Generalized Additive Models (GAMs) and shape-constrained additive models (SCAMs). Specifically, unrestricted GAMs were fitted using the `mgcv` package [112], and monotonic (increasing and decreasing) GAMs were implemented with the `scam` package [91], which extends `mgcv` to support monotonicity constraints. For each dataset, GAMs were estimated using penalized regression splines. Unrestricted GAMs employed thin plate regression spline bases for the smooth term. Smoothing parameters were selected automatically using restricted maximum likelihood (REML). Monotonic GAMs were fitted using shape-constrained spline bases enforcing either monotone increase or monotone decrease, with smoothing parameters again estimated via REML. The final monotonic model was selected by comparing the monotone increasing and monotone decreasing fits and retaining the specification that better matched the observed data. All models assumed Gaussian errors, and effective degrees of freedom were determined through penalization rather than fixed a priori.

All statistical analyses, including correlation computations, model fitting, and figure generation, were performed using the R programming language [92]. Correlation measures such as Pearson's r, Spearman's ρ, and Kendall's τ were calculated using base R functions, while statistical modeling was conducted using Generalized Additive Models (GAMs) and shape-constrained additive models (SCAMs). Specifically, unrestricted GAMs were fitted using the `mgcv` package [112], and monotonic (increasing or decreasing) GAMs were implemented with the `scam` package [91], which extends `mgcv` to support monotonicity constraints.

Model comparisons were carried out using ANOVA F-tests from base R functions [25]. For data wrangling and summarization, we employed the `dplyr` [111] and `data.table` [32] packages. Figures were produced using `ggplot2` [110], with `ggrepel` for improved label placement and `ggpubr` for consistent theming. Supplementary LaTeX-ready tables were generated using `xtable`, and eta-squared (η^2) statistics were calculated with the `lsr` package [74]. Altogether, the R ecosystem provided a flexible and reproducible framework for executing the statistical pipeline described in this study.

All reproducible code used in this study, including both Python and R scripts for data processing, statistical analysis, and figure generation, is available in the public repository: `https://github.com/AleksanderWWW/hypergraph-properties`.

A.4 Correlation Measures for Hyperedge Size and Node Degree

To characterize the statistical association between hyperedge size and node degree, we employ three standard correlation measures: Pearson's r, Spearman's ρ, and Kendall's τ. Each captures a different notion of dependence and responds differently to nonlinearity, outliers, and the shape of the relationship. In this subsection, we briefly define each measure, discuss their strengths and limitations, and outline when their use is most appropriate.

Pearson Correlation Pearson's correlation coefficient r quantifies the strength and direction of a linear relationship between two continuous variables: x and y [85]. It is defined as:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{14}$$

Confidence intervals are computed using the Fisher z transformation with standard error $1/\sqrt{N-3}$. Pearson's r assumes that both variables are linearly related. It is sensitive to outliers and may be misleading in the presence of nonlinear or monotonic but non-linear relationships [38]. However, recent work by van den Heuvel and Zhan (2022) challenges the conventional wisdom distinguishing Pearson's r for linear relationships and Spearman's ρ or Kendall's τ for nonlinear monotonic associations. They argue that "Pearson's correlation coefficient should not be ruled out a priori for measuring nonlinear monotonic associations," and further demonstrate via counterexamples that Pearson's r can be preferred over Spearman's ρ and Kendall's τ in testing dependency even when the association is monotonic but nonlinear [105]. Therefore, Pearson's r tends to be more robust and interpretable in contexts where the global trend described by a single summary of direction and strength is desired.

Spearman Correlation Spearman's rank correlation coefficient ρ is a non-parametric measure that assesses the strength of a monotonic relationship between two variables [100]. It is defined as the Pearson correlation between the ranks of the variables:

$$\rho = \frac{\sum_{i=1}^{n}(R(x_i) - \bar{R}_x)(R(y_i) - \bar{R}_y)}{\sqrt{\sum_{i=1}^{n}(R(x_i) - \bar{R}_x)^2}\sqrt{\sum_{i=1}^{n}(R(y_i) - \bar{R}_y)^2}} \tag{15}$$

where $R(x_i)$ and $R(y_i)$ denote the ranks of x_i and y_i, respectively. In the presence of ties, *midranks* (average ranks) are used, which corresponds to the standard definition of Spearman's ρ. Spearman's rank correlation was computed using the implementation `scipy.stats.spearmanr`. As a robustness check, we examine the sensitivity of the Spearman correlation to alternative tie-handling procedures. In particular, we compare the default average ranking (used throughout the paper whenever Spearman correlation is reported) with the `min`, `max`, and `dense` ranking methods.

Both the `min` and `max` ranking schemes yield Spearman correlations that are virtually indistinguishable from the default average ranking. For the `min` method, the mean difference equals 0.00198, with $R^2 = 0.997$

between the two sets of estimates, and a paired two-sided t-test indicates no statistically significant difference ($p = 0.2489$). Similarly, for the `max` method, the mean difference equals -0.000699, with $R^2 = 0.997$, and the paired two-sided t-test is not significant ($p = 0.702$).

The `dense` ranking produces estimates that are statistically different from the default Spearman correlation (paired two-sided t-test $p = 0.0407$), but the magnitude of the discrepancy remains small: the mean difference equals 0.017 and the association between the two sets of estimates remains strong ($R^2 = 0.925$).

Given the small effect size and the high concordance across ranking schemes, we conclude that the choice of tie-handling method has negligible practical impact on the results. We therefore report only the default average-rank Spearman correlation throughout the paper.

Kendall Correlation Kendall's tau (τ) is another non-parametric measure of monotonic association, based on the number of concordant and discordant pairs in the data [60]. Given the abundance of ties in node degrees and hyperedge sizes in our data, we focus on Kendall's τ_b, which explicitly accounts for ties in both variables. Kendall's τ_b is defined as

$$\tau_b = \frac{C - D}{\sqrt{(C + D + T_x)(C + D + T_y)}}, \tag{16}$$

where C is the number of concordant pairs, D the number of discordant pairs, and T_x and T_y denote the numbers of tied pairs in x and y, respectively. A pair of observations (x_i, y_i) and (x_j, y_j) is said to be *concordant* if the ranks of both elements agree in direction: that is, either $x_i > x_j$ and $y_i > y_j$, or $x_i < x_j$ and $y_i < y_j$. Conversely, the pair is *discordant* if the ranks disagree: one variable increases while the other decreases (e.g., $x_i > x_j$ but $y_i < y_j$). Tied pairs are not counted as concordant or discordant but are incorporated into the normalization of τ_b, making this variant particularly suitable for data with many repeated values.

As a robustness check, we also computed Kendall's τ_c, which differs only in the normalization and is designed for settings with a limited number of distinct values. The results based on τ_c are qualitatively consistent with those for τ_b: the two measures are almost perfectly correlated (Pearson r = 0.995) and do not differ significantly on average (paired two-sample t-test: mean diff = 0.00329, t = 1.88, two-tailed p = 0.068).

Comparison and Application In the context of our hypergraph analysis, Pearson's r offers a direct assessment of global trends between hyperedge size and node degree and is meaningful when such trends are present. Spearman's ρ and Kendall's τ, on the other hand, are more appropriate when the relationship is suspected to be nonlinear but monotonic, especially common in empirical network data. While Spearman tends to be more sensitive, Kendall is more statistically robust and better suited to small or highly discrete datasets.

Non-parametric correlation coefficients such as Spearman's ρ and Kendall's τ are designed to capture rank-based, monotonic associations and are known for their robustness to outliers and non-normal distributions [105, 34]. These coefficients rely on relative ordering rather than the actual magnitudes of data, making them less sensitive to the shape and sign of complex, nonlinear relationships when compared to Pearson's r, which directly assesses covariance between variable magnitudes [67, 113].

The statistical literature supports three relevant findings regarding the use of Pearson's r, Spearman's ρ, and Kendall's τ for measuring association. First, although non-parametric measures like ρ and τ are valued for their robustness, they exhibit higher variance and bias than Pearson's r even under non-normal, contaminated, or curved distributions [114]. Second, Pearson's r remains the most statistically efficient estimator even when the underlying relationship is approximately linear or near-normal, conditions that are frequently approximated in large empirical datasets [28]. Third, while Spearman's and Kendall's measures are designed to detect monotonicity, they may fail to reflect the dominant global trend direction, especially when the relationship is weakly monotonic or contains local non-monotonic variations [105]. In this paper, we further investigate this third point by directly comparing the signs of Pearson's r, Spearman's ρ, and Kendall's τ to the global direction of association estimated by monotonic Generalized Additive Models (GAMs) [48, 112]. This allows us to evaluate how well each correlation measure captures the overarching trend in the data, even when local fluctuations or curvature are present.

A.5 Supplementary Analyses Referenced in Main Text

A.5.1 Methodological result I: bipartite representation as the preferred preprocessing strategy

Explaining η^2 values via within-segment variability in correlation estimates Figure 5 presents the variability of correlation values for six pairings of Pearson and Spearman correlations with the three preprocessing strategies: node-centric, edge-centric, and bipartite representation. It is referenced in subsection 3.1.2.

Pairwise Comparison of Preprocessing Strategies A deeper analysis of the interrelation between the three data preprocessing strategies is provided in Figure 6. Figure (a) displays a scatterplot comparing Pearson correlations under the bipartite representation versus edge-centric processing. The reported R^2 of 0.60 indicates a moderate positive association between these two measurements. This is consistent with Figures 2 and 5, which demonstrate general alignment between these two correlation estimates, although with notable exceptions. Four hypergraphs with the largest discrepancies between bipartite and edge-centric Pearson correlations are labelled in the figure 6(a). Among them, `music-blues-reviews` stands out, with a positive edge-centric Pearson correlation of 0.0753, while its bipartite Pearson is substantially negative at -0.132. Interestingly, its node-centric Pearson is even more negative at -0.335, aligning more closely with the bipartite estimate. Similar discrepancies are observed in `house-committees` ($r_{\text{edge}} = -0.305$, $r_{\text{bipartite}} = -0.037$) and `senate-committees` ($r_{\text{edge}} = -0.310$, $r_{\text{bipartite}} = -0.048$), both from the `political` segment. Again, their node-centric correlations (-0.014 and -0.001, respectively) are closer to bipartite values. The regression line fitted between edge-centric and bipartite Pearson correlations is $r_{\text{bipartite}} = 0.015 + 0.46 \times r_{\text{edge}}$, indicating no systematic bias, as the intercept is small and statistically non-significant.

Figure 6(b) shows the scatterplot of bipartite versus node-centric Pearson correlations. The relationship between these two is similar, with $R^2 = 0.60$ and an estimated regression line of $r_{\text{bipartite}} = 0.013 + 0.48 \times r_{\text{node}}$, again indicating no significant bias. Nonetheless, some hypergraphs exhibit large discrepancies. For example, `contact-high-school` has a node-centric Pearson of 0.750 but a bipartite value of only 0.180, a difference of 0.570. Similar discrepancies are found in `hospital-lyon` ($r_{\text{node}} = 0.893$, $r_{\text{bipartite}} = 0.337$), `contact-primary-school` (0.572 vs. 0.089), and `nba` (-0.467 vs. -0.083). As previously, the third preprocessing measurements, i.e., edge-centric Pearson correlations for these hypergraphs are closer to bipartite correlations.

Finally, Figure 6(c) compares Pearson correlations between node-centric and edge-centric representations. This pair exhibits the weakest relationship, with $R^2 = 0.34$. The estimated regression line is $r_{\text{node}} = 0.020 + 0.52 \times r_{\text{edge}}$, again with a non-significant intercept. The most prominent outliers are again hypergraphs from the `physical contact` segment: `contact-high-school` ($r_{\text{node}} = 0.750$, $r_{\text{edge}} = 0.228$), `contact-primary-school` (0.572 vs. 0.114), `hospital-lyon` (0.893 vs. 0.473), as well as `music-blues-reviews` (-0.335 vs. 0.075). These illustrate that node- and edge-centric preprocessing steps can yield substantially different correlation estimates even when applied to structurally similar hypergraphs.

Figure 6(d) displays overlaid distributions of Pearson correlations for the three hypergraph preprocessing strategies. Consistent with the regression results discussed earlier, the distributions share similar central tendencies, indicating no systematic location bias. However, the distributions differ notably in their spread. The node-centric correlations exhibit the widest dispersion, reflecting the presence of several hypergraphs with exceptionally especially high, but also low correlation values. In contrast, the bipartite-based Pearson correlations are the most concentrated around its mean, suggesting greater stability and less variability across datasets. These observations further reinforce the finding that the bipartite representation yields more stable correlation estimates.

A.5.2 Methodological result II: Pearson correlation as the preferred measure of dependence

Quantitative Comparison of Pearson and Spearman correlations Figure 7 presents a scatterplot of Pearson vs. Spearman correlation coefficients for 36 empirical hypergraphs. Six hypergraphs for which the two coefficients differ in sign are labelled. Visual inspection reveals a general alignment along the identity (dashed) line, with minor fluctuations. The fitted regression (solid) line supports this, showing a slope close to one and an intercept estimate of 2.43% (SE = 0.988%), which is statistically significant at $\alpha = 0.05$ ($p = 0.0192$). This indicates a small but systematic upward bias in Spearman relative to Pearson. The goodness-of-fit of this regression, $R^2 = 79.0\%$, suggests a relative high alignment between the two metrics.

Half of the hypergraphs show an absolute difference between the two metrics of less than 2.95pp, and 75% have a difference below 6.4pp. Nevertheless, a few outliers show discrepancies exceeding 12pp in absolute

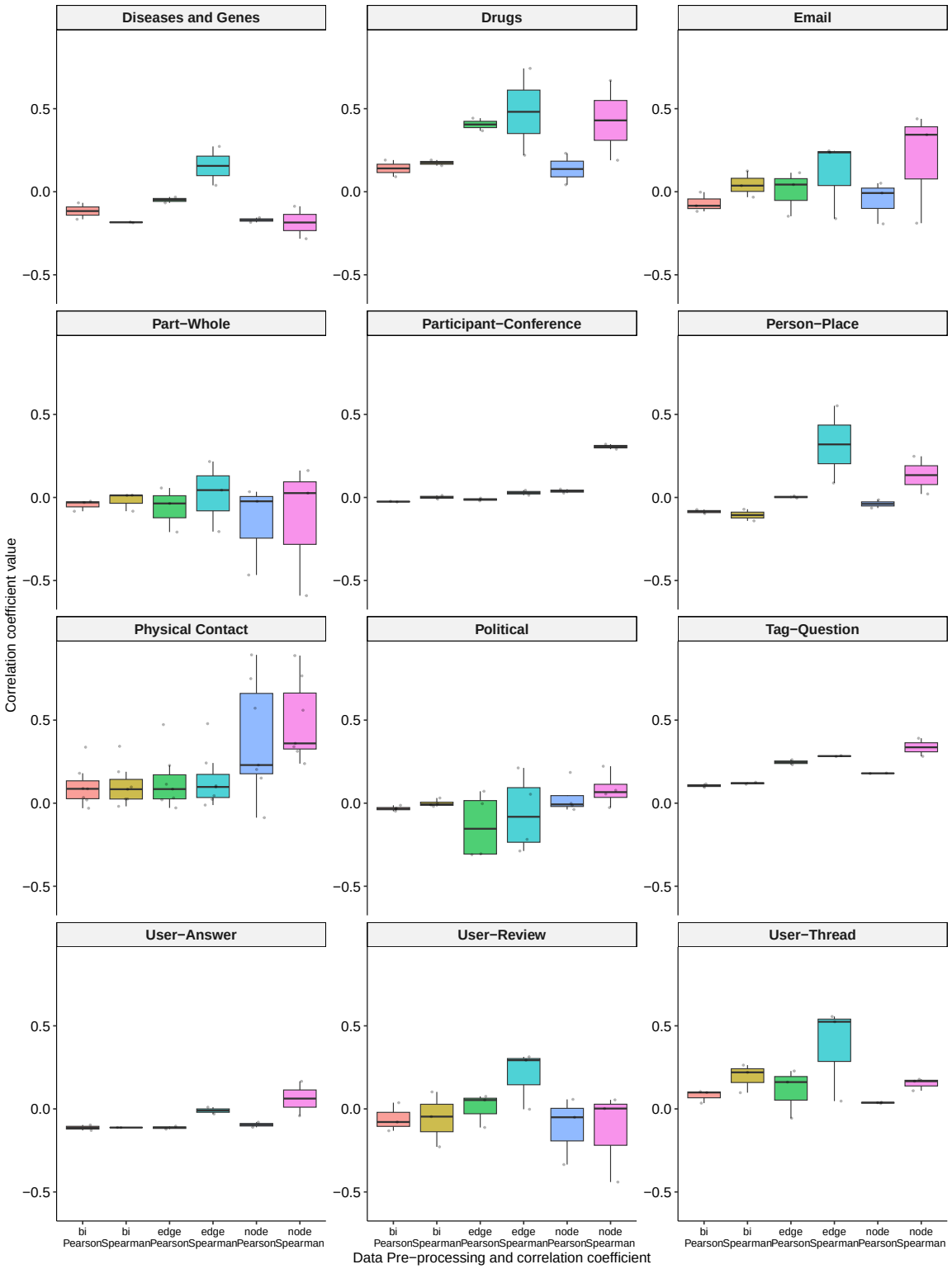


Figure 5: Box plots with points of correlation values across preprocessing methods and correlation types, grouped by hypergraph segments. Labels `bi`, `edge`, and `node` denote bipartite, edge-, and node-centric strategies.

terms: `email-eu` (Spearman vs. Perason difference: 12.1pp), `threads-math-sx` (12.1pp), `diseasome` (-12.3pp), `email-enron` (13.2pp), and `threads-ask-ubuntu` (16pp). Notably, four of these belong to either the `email` or

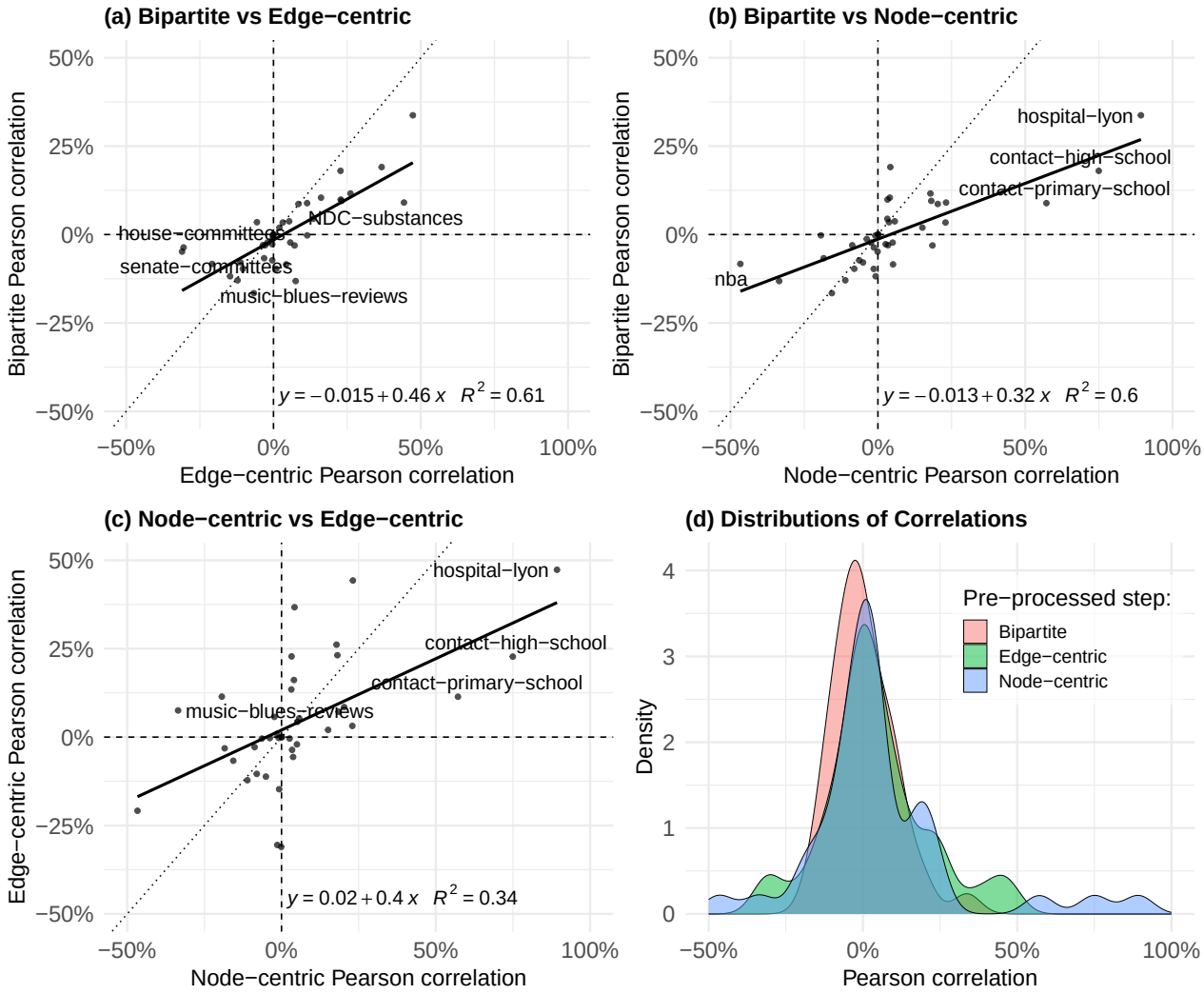


Figure 6: (a–c) Scatterplots of Pearson correlations between pairs of hypergraph preprocessing strategies, with fitted regression line (solid) and identity line (dashed); (d) distribution of Pearson correlations.

`threads` categories. In each of these four cases, Spearman is systematically higher than Pearson by at least 12pp; in two instances (`email-enron`, `email-eu`) this leads to opposite signs. The common factor appears to be an initial upward trend in dense regions of the data, which biases rank-based Spearman estimates, as visible in Figure 4a. While Pearson reflects the global trend, Spearman is heavily influenced by early dense data ranges.

The case of `diseasome` presents a reverse scenario. Here, Spearman (-0.19) is 12.3pp lower than Pearson (-0.067), revealing a monotonic decreasing relationship. Spearman is statistically significant at $\alpha = 0.00001$, whereas Pearson is significant at $\alpha = 0.05$ with $p = 0.026$. This is the one case in which Spearman is better aligned with the monotonic GAM classification than Pearson due to its statistical insignificance, that can be explained by the relatively small dataset size ($N = 1109$).

Assessing Alignment Between GAM Monotonicity and Correlation Coefficients' Signs In Section 3.2, we consider the alignment between the sign of each correlation coefficient and the monotonicity direction inferred from GAM models. This analysis draws extensively on Table 7, which reports Pearson, Spearman, and Kendall correlations (sorted by decreasing Pearson) for all 36 empirical hypergraphs, along with the monotonicity direction of the fitted monotonic GAM in the bipartite representation. This allows us to compare the sign of each coefficient with the trend direction. A summary of this alignment is presented in Table 2.

Hypergraph	Pearson	Spearman	Kendall	GAM Monotonicity	N
hospital-lyon	0.337*** (0.013)	0.343*** (0.013)	0.278*** (0.012)	Inc.	4427
NDC-classes	0.191*** (0.002)	0.157*** (0.002)	0.117*** (0.002)	Inc.	156185
contact-high-school	0.180*** (0.007)	0.189*** (0.007)	0.152*** (0.006)	Inc.	18192
tags-math-sx	0.116*** (0.001)	0.114*** (0.001)	0.087*** (0.002)	Inc.	593121
threads-ask-ubuntu	0.104*** (0.002)	0.263*** (0.002)	0.206*** (0.002)	Inc.	346537
threads-math-sx	0.099*** (0.001)	0.220*** (0.001)	0.166*** (0.002)	Inc.	1610393
tags-ask-ubuntu	0.095*** (0.001)	0.126*** (0.001)	0.101*** (0.002)	Inc.	499298
NDC-substances	0.090*** (0.004)	0.191*** (0.004)	0.129*** (0.003)	Inc.	53528
contact-primary-school	0.089*** (0.006)	0.084*** (0.006)	0.068*** (0.005)	Inc.	30729
Science-Gallery	0.086*** (0.001)	0.097*** (0.001)	0.081*** (0.003)	Inc.	717690
amazon	0.044** (0.014)	0.041** (0.014)	0.040** (0.014)	Non-sign.	5112
vegas-bars-reviews	0.037** (0.009)	0.102*** (0.009)	0.073*** (0.006)	Non-sign.	11865
twitter	0.035*** (0.004)	0.098*** (0.004)	0.073*** (0.003)	Inc.	50850
Malawi-village	0.034*** (0.002)	0.025*** (0.002)	0.022*** (0.003)	Inc.	201069
InVS15	0.020*** (0.003)	0.025*** (0.003)	0.018*** (0.003)	Inc.	149949
email-enron	-0.002 (0.015)	0.125*** (0.014)	0.093*** (0.011)	Inc.	4623
senate-bills	-0.013*** (0.002)	-0.003 (0.002)	-0.004 (0.002)	Dec.	232147
dblp	-0.022*** (0.003)	0.013** (0.003)	0.011** (0.003)	Dec.	88458
Hypertext-conference	-0.023*** (0.005)	-0.010* (0.005)	-0.008* (0.004)	Dec.	39048
SFHH-conference	-0.027*** (0.003)	0.012** (0.003)	0.009** (0.003)	Dec.	116636
kaggle-whats-cooking	-0.030*** (0.002)	0.014*** (0.002)	0.007** (0.002)	Dec.	428249
InVS13	-0.030** (0.007)	-0.019** (0.007)	-0.016** (0.006)	Non-sign.	19380
house-bills	-0.031*** (0.001)	0.031*** (0.001)	0.021*** (0.002)	Dec.	1248666
house-committees	-0.036** (0.009)	-0.011 (0.009)	-0.007 (0.006)	Dec.	11843
senate-committees	-0.048** (0.014)	-0.019 (0.014)	-0.013 (0.009)	Dec.	5408
diseasome	-0.067* (0.030)	-0.186*** (0.029)	-0.138*** (0.023)	Dec.	1109
got	-0.073*** (0.009)	-0.141*** (0.009)	-0.100*** (0.006)	Dec.	12114
restaurant-reviews	-0.079*** (0.015)	-0.046** (0.015)	-0.032** (0.010)	Dec.	4601
nba	-0.083*** (0.001)	-0.083*** (0.001)	-0.059*** (0.002)	Dec.	644051
email-eu	-0.084*** (0.003)	0.037*** (0.003)	0.024*** (0.002)	Dec.	89409
algebra	-0.097*** (0.011)	-0.112*** (0.011)	-0.078*** (0.008)	Dec.	8262
eventernote-places	-0.097*** (0.001)	-0.071*** (0.001)	-0.044*** (0.002)	Dec.	713400
email-W3C	-0.118*** (0.009)	-0.033** (0.009)	-0.029** (0.007)	Dec.	13361
geometry	-0.129*** (0.009)	-0.112*** (0.009)	-0.078*** (0.006)	Dec.	12485
music-blues-reviews	-0.132*** (0.010)	-0.228*** (0.009)	-0.159*** (0.007)	Dec.	10499
disgenenet	-0.166*** (0.003)	-0.182*** (0.003)	-0.125*** (0.002)	Dec.	112471

Table 7: Correlation measures (sorted by decreasing Pearson) with standard error in parentheses for bipartite representation with significance stars (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.00001$) and the monotonicity direction of fitted monotonic GAM. Pearson and Spearman use full N (SE via Fisher z approximation; Spearman uses a large-sample approximation). Kendall uses a tie-corrected normal approximation; for $N > 10^5$, Kendall is computed on a simple random subsample of 10^5 observations without replacement using a fixed random seed. At this sample size, the corresponding asymptotic standard error is approximately 0.002, indicating negligible sampling variability for the reported correlations.

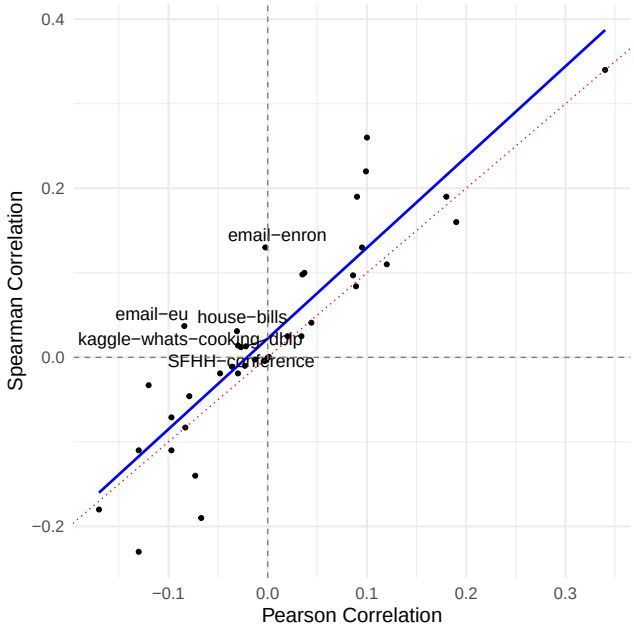


Figure 7: Scatterplot of Pearson vs. Spearman correlation coefficients for 36 hypergraphs. The dashed red line indicates the identity line (45°), and the solid blue line shows the fitted linear regression. The six hypergraphs for which the two correlation coefficients differ in sign are labeled.

A.5.3 Empirical result: prevalence of edge size–degree relationships

Identification of Relationship Types: Results of Statistical Tests Table 8 reports, for each hypergraph, the p-values in the parenthesis from the three nested statistical tests and the resulting classification of the relationship type, following the procedure outlined in Subsection 2.3. The classification is based on a conservative significance threshold of $\alpha = 0.00001$, ensuring robustness against spurious detections. Moreover, we report size effects for comparisons between competing models, as the difference in explained variance measured by ΔR^2 defined as:

$$\Delta R^2 = R^2_{\text{unrestricted}} - R^2_{\text{restricted}}.$$

The overall distribution of relationship types across the 36 empirical hypergraphs is summarized in Table 3 and discussed in Subsection 3.3.2.

Identification of Relationship Types: Visual Inspection Subsection 3.3.1 introduced and discussed four representative examples (out of 36) of hypergraphs, each illustrating one of the four identified relationship types: non-monotonic, monotonic, linear, and no relationship. These examples were visualized in Figures 4 and 4. The remaining 32 empirical hypergraphs are presented in a separate file.

Hypergraph	**Relationship**	ΔR^2 (p-val for H_0 : monotonic)	ΔR^2 (p-val for H_0 : linear)	R^2 (p-val for H_0 : $R^2 = 0$)	**N**
algebra	Monotonic	0.06% (0.11000)	0.34% ($< 10^{-5}$)	0.94% ($< 10^{-5}$)	8262
amazon	No relationship	0.02% (0.08100)	0.04% (0.07700)	0.2% (0.00150)	5112
contact-high-school	Monotonic	-0% (0.19000)	0.16% ($< 10^{-5}$)	3.24% ($< 10^{-5}$)	18192
contact-primary-school	Linear	-0% (0.01600)	0% (0.01600)	0.79% ($< 10^{-5}$)	30729
dblp	Non-monotonic	0.59% ($< 10^{-5}$)	0.14% ($< 10^{-5}$)	0.05% ($< 10^{-5}$)	88458
diseasome	Monotonic	0.22% (0.34000)	0.34% ($< 10^{-5}$)	0.45% (0.02600)	1109
disgenenet	Monotonic	0% (0.89000)	0.45% ($< 10^{-5}$)	2.75% ($< 10^{-5}$)	112471
email-enron	Non-monotonic	1.07% ($< 10^{-5}$)	2.04% ($< 10^{-5}$)	0% (0.87000)	4623
email-eu	Non-monotonic	1.99% ($< 10^{-5}$)	0.52% ($< 10^{-5}$)	0.71% ($< 10^{-5}$)	89409
email-W3C	Monotonic	-0.01% (1.00000)	0.86% ($< 10^{-5}$)	1.39% ($< 10^{-5}$)	13361
eventernote-places	Non-monotonic	0.01% ($< 10^{-5}$)	0.24% ($< 10^{-5}$)	0.95% ($< 10^{-5}$)	713400
geometry	Monotonic	0.11% (0.00078)	0.24% ($< 10^{-5}$)	1.68% ($< 10^{-5}$)	12485
got	Monotonic	0.24% (0.00001)	1.52% ($< 10^{-5}$)	0.53% ($< 10^{-5}$)	12114
hospital-lyon	Linear	-0% (1.00000)	0.11% (0.01200)	11.37% ($< 10^{-5}$)	4427
house-bills	Non-monotonic	0.48% ($< 10^{-5}$)	0.15% ($< 10^{-5}$)	0.1% ($< 10^{-5}$)	1248666
house-committees	Non-monotonic	0.95% ($< 10^{-5}$)	0.99% ($< 10^{-5}$)	0.13% (0.00007)	11843
Hypertext-conference	Linear	-0.01% (0.13000)	0.01% (0.12000)	0.05% ($< 10^{-5}$)	39048
InVS13	No relationship	-0% (0.00990)	0% (0.00990)	0.09% (0.00003)	19380
InVS15	Linear	0% (0.28000)	0% (0.00510)	0.04% ($< 10^{-5}$)	149949
kaggle-whats-cooking	Non-monotonic	0.04% ($< 10^{-5}$)	0.03% ($< 10^{-5}$)	0.09% ($< 10^{-5}$)	428249
Malawi-village	Linear	-0% (0.01000)	0% (0.01000)	0.12% ($< 10^{-5}$)	201069
music-blues-reviews	Non-monotonic	1.39% ($< 10^{-5}$)	0.83% ($< 10^{-5}$)	1.73% ($< 10^{-5}$)	10499
nba	Non-monotonic	0.03% ($< 10^{-5}$)	0.06% ($< 10^{-5}$)	0.69% ($< 10^{-5}$)	644051
NDC-classes	Non-monotonic	11.43% ($< 10^{-5}$)	9.14% ($< 10^{-5}$)	3.64% ($< 10^{-5}$)	156185
NDC-substances	Non-monotonic	0.87% ($< 10^{-5}$)	4.69% ($< 10^{-5}$)	0.82% ($< 10^{-5}$)	53528
restaurant-reviews	Linear	-0% (0.00440)	0% (0.00440)	0.62% ($< 10^{-5}$)	4601
Science-Gallery	Non-monotonic	0.02% ($< 10^{-5}$)	0.18% ($< 10^{-5}$)	0.75% ($< 10^{-5}$)	717690
senate-bills	Non-monotonic	0.02% ($< 10^{-5}$)	0.01% (0.00096)	0.02% ($< 10^{-5}$)	232147
senate-committees	Monotonic	0.2% (0.05600)	1.32% ($< 10^{-5}$)	0.23% (0.00038)	5408
SFHH-conference	Non-monotonic	0.12% ($< 10^{-5}$)	0.22% ($< 10^{-5}$)	0.08% ($< 10^{-5}$)	116636
tags-ask-ubuntu	Monotonic	0% (0.20000)	0.32% ($< 10^{-5}$)	0.91% ($< 10^{-5}$)	499298
tags-math-sx	Monotonic	0% (0.08800)	0.29% ($< 10^{-5}$)	1.35% ($< 10^{-5}$)	593121
threads-ask-ubuntu	Monotonic	0% (0.26000)	0.28% ($< 10^{-5}$)	1.09% ($< 10^{-5}$)	346537
threads-math-sx	Monotonic	-0.03% (1.00000)	0.85% ($< 10^{-5}$)	0.98% ($< 10^{-5}$)	1610393
twitter	Non-monotonic	3.26% ($< 10^{-5}$)	0.89% ($< 10^{-5}$)	0.12% ($< 10^{-5}$)	50850
vegas-bars-reviews	No relationship	0.01% (0.06500)	0.04% (0.01600)	0.14% (0.00005)	11865

Table 8: Difference in explained variance between models measured by ΔR^2 with p-values in parenthesis identifying one of four relationship types, i.e. non-monotonic, monotonic, linear, no relationship, by running following three statistical tests: (1) ANOVA test comparing two models with $H_0 : unrestricted\, GAM = monotonic\, GAM$, (2) ANOVA test comparing two models with H_0 $monotonic\ GAM = OLS$, (3) F-test with $H_0 : R^2 = 0$. Significance level $\alpha = 10^{-5}$.