Beyond Boundaries: Capturing Social Segregation on Hypernetworks

Andrea Failla[1,2], Giulio Rossetti[2], and Francesco Cauteruccio[3]

[1] University of Pisa, Department of Computer Science, Pisa 56127, Italy
name.surname@phd.unipi.it
[2] National Research Council, ISTI-CNR, Pisa 56127, Italy
name.surname@isti.cnr.it
[3] University of Salerno, DIEM, Fisciano 84084, Italy
fcauteruccio@unisa.it

Abstract. In recent years, the study of complex social systems has been fueled by the renewed interest in higher-order topologies, thus leading to the emergence of *hypernetwork science*. A critical and interesting phenomenon often characterizing social complex systems is *segregation*, i.e., the extent to which network entities are separated or clustered based on certain semantic attributes or features. This paper introduces a novel approach to studying segregation in hypernetworks. Firstly, we propose a general framework to extend classical segregation measures from dyadic to polyadic network structures. Then, we introduce a novel segregation measure called "Random Walk HyperSegregation" (RWHS), which exploits random walkers to estimate segregation at multiple scales. Through an extensive experimental study involving synthetic and real-world case studies, we illustrate the applicability and effectiveness of our measure. Moreover, we highlight the limits of classical segregation measures when extended to high-order topologies — conversely from RWHS, which effectively captured highly-segregated scenarios.

Keywords: complex system · segregation · hypernetwork science · random walk

1 Introduction

The study of complex systems through the lenses offered by network theory has garnered significant attention, revealing intricate patterns of interactions across diverse fields such as sociology, biology, and information [30]. Traditional network analysis, focusing on pairwise interactions, has been instrumental in understanding the dynamics within these systems [29]. However, as we delve deeper into the complexity of real-world systems, it becomes increasingly clear that many exhibited patterns cannot be adequately captured by pairwise interactions alone [4]. Glancing at different (real) complex systems reveals a whole series of multi-way interactions, such as the ones in biological systems [14] or social ecosystems [15, 19]. The willingness to study these phenomena paved the

way for the so-called hypernetwork science, in which interactions are modeled and analyzed through hypergraphs [1] — conservative generalizations of graphs in which (hyper)edges may connect an arbitrary number of vertices, thereby representing multi-way relationships [4].

A critical yet overlooked aspect of high-order representations lies in the understanding of the phenomenon of *segregation* — i.e., the extent to which system entities are separated or clustered based on certain attributes or features. The concept of segregation has its roots in studies on the residential organization of cities [27, 35], and undoubtedly, its analysis provides valuable insights into the underlying structure and dynamics of the system, with profound implications ranging from social equity and cohesion to system resilience and functionality. Indeed, segregation has been extensively studied in networks exhibiting pairwise interactions, such as social networks [5]. However, to the best of our knowledge, investigations regarding such phenomenon have not been conducted in the realm of hypernetworks, where due to the inherent complexity of dealing with higher-order interactions, defining and studying segregation in this context poses significant conceptual and methodological challenges. This paper aims to define a methodology that can help measure segregation in hypernetwork contexts. To such an extent, our contribution is twofold. Firstly, we formally define a general framework for extending pairwise segregation measures to hypernetworks. We refer to the extensions in this framework as *conservative* because they are transpositions of traditional measures originally envisioned for pairwise topologies. Secondly, we propose a novel measure tailored explicitly for hypernetworks, called *Random Walk HyperSegregation* (RWHS). We propose two variants of RWHS, namely *(i) meet-wise* and *(ii) jump-wise* RWHS, each capturing different aspects of node segregation. Moreover, to validate our approach, we present an extensive experimental study assessing the effectiveness and applicability of both segregation measuring strategies. To such an extent, we apply and discuss the proposed measures on synthetically generated and real-world hypernetwork topologies, showing how segregation can be captured through conservative extensions and RWHS.

The paper is organized as follows: in Section 2, we provide an overview of related literature; Section 3 covers the necessary background on hypernetwork modeling; Sections 4 and 5 introduce the general framework for extending classical segregation measures and define RWHS, respectively; Section 6 focuses on the experimental validation of the proposed measures; Finally, Section 7 concludes the paper and delineates some possible future researches.

2 Related Works

Historically, segregation has been formalized as a phenomenon depicting two or more groups coexisting separately in the same environment. Such a definition has been applied to study heterogeneous contexts, e.g., by analyzing gender and ethnic segregation [9, 26]. Several studies on social dynamics stemmed from the seminal work in [35], which focused on residential segregation. Indeed, segrega-

tion has naturally been embraced by social network analysis research that usually treats such a concept as a complex system emerging property. This section presents a bird's-eye view of related literature on segregation and its quantification, focusing on sociological results and methodological approaches.

One of the most thorough investigations into quantifying segregation within social networks is presented in [5], where a range of segregation metrics — e.g., E-I index and Gupta's Q — were adapted and tested within a social network analysis framework. Analogously, [22] delves into network segregation through a Markov-chain-based model inspired by [35], illustrating how even mild biases against dissimilar nodes can foster segregated network structures. An advanced iteration of the model from [35] is further developed in [20], where the framework is adapted to accommodate a more dynamic social network setting. In particular, nodes are categorized into two groups, and each node can strategically form or dissolve connections based on the group composition of their immediate network vicinity. The findings suggest that employing random strategies for forming these connections results in higher segregation levels than strategies based on discriminatory behavior, indicating that network structure has a negligible impact on segregation outcomes. [3] introduces a data-centric methodology to capture the role of online social networks in intensifying segregation and opinion polarization — particularly focusing on the risk of echo chamber formation and pluralism reduction. Leveraging metrics such as the evenness and exposure indexes [27], the authors outline the segregation discovery problem, drawing inspiration from the classical problem of itemset mining. Noteworthy, [38] illustrates — leveraging data from an online social network of approximately 2 million individuals across 500 towns in Hungary — how sociality and geography play crucial roles in exacerbating wealth and income equalities. Similarly, in [33], a measure of social segregation combining mobile phone data and income register data is proposed, allowing the authors to estimate the association between income differences and communication intensity while considering spatial proximity. The investigation by [13], focused on credit card transactions and online social media data, examines contemporary forms of segregation, highlighting distinct divisions among various socioeconomic and ethnic groups — manifesting in patterns of urban visitation and online information consumption. Segregation was also identified in co-authorship networks [23], where the authors underlined that highly segregated communities tend to be closer to the network periphery, and researchers receive more citations from their community members. [36] highlights that spatial segregation encompasses several aspects, such as spatial exposure and isolation. Therefore, it has a multidimensional nature. To capture such a multifaceted reality, the authors introduce a framework based on graph random walks that gathered insights into different class organizations within cities. Although such an approach shares the same rationale as our RWHS measure, it only focuses on capturing spatial segregation by random walk processes on simple graphs, not higher-order ones, thus neglecting the effects of group interactions. In [18], the authors propose an agent-based network formation model under uncertainty — i.e., assuming agents in the network have only partial information about each

other's groups — and leverage the generalized Freeman's segregation index [17] to quantify segregation. [11] identified high political segregation in a Twitter retweets dataset by leveraging a combination of manually-labeled data and unsupervised clustering algorithms. [32] introduces a high-dimensional approach to measure online polarization. Here, the authors define a high-dimensional network — akin to the classical multiplex network — and elucidate different measures to compute polarization and segregation. The approach is tested on synthetic networks and a Twitter dataset representing discussion regarding COVID-19 vaccines. Concerning segregation measures, the work in [12] proposes a generalization of the classical E-I index to the fuzzy case: in fact, it considers that nodes may belong to a group with a certain membership degree. Finally, [16] introduces a preliminary work on segregation in higher-order networks, applying a network fragmentation to measure segregation — the Borgatti's F measure [6] — to randomly generated hypergraphs with different distributions. A first attempt to approach segregation in high-order networks is performed in [16], where segregation is only estimated on artificially generated hypergraphs.

3 High-order network modeling

In the following, we will model high-order interactions leveraging hypergraphs: this section provides definitions and notations needed to frame our contributions better.

Definition 1 (Hypergraph). *A hypergraph $H = (V, E)$ is a pair consisting of a set $V = \{v_1, \ldots, v_n\}$ of elements called nodes, and a set of sets $E = (e_1, \ldots, e_m)$ called hyperedges.*

A hyperedge represents a relation between a subset of nodes of V, that is, $e_i \subseteq V$ for all $i = 1, \ldots, m$. The order of H is the number of its nodes $n = |V|$, while the size of H is the number of its edges $m = |E|$. We say a node $v_j \in V$ belongs to a hyperedge $e_i \in E$ if $v_j \in e_i$.

To measure segregation, one typically deals with some finite set of attributes associated with each node. The most common example is group membership; network nodes are assigned to a group, and this assignment is known. To maintain this assumption, we introduce the concept of node-attributed hypergraph [15].

Definition 2 (Node-attributed Hypergraph). *Let L be a set of labels, where each label denotes a group. A node-attributed hypergraph $H_L = (V, E, L)$ is a hypergraph H where to each node is assigned a label (a group) $l \in L$.*

We denote the group is assigned to a node $v_j \in V$ as $\gamma(v_j)$, and we also say that node v_j belongs to the group $\gamma(v_j)$. The number of groups is therefore $h = |L|$ — we assume $h \geq 2$. Also, we denote with $E^l \subseteq E$ the set of edges containing at least one node belonging to the group l, i.e., $E^l = \{e_i \in E : \exists v_j \in e_i, \gamma(v_j) = l\}$.

4 Extending Classical Segregation Measures

This section provides a conservative extension of classical segregation measures to high-order topologies. To such an extent, we provide a general formulation encompassing concepts leveraged in measuring segregation, extending them to the hypergraph context. As a result, instances of the proposed general schema provide what is commonly called the system segregation index.

Definition 3 (Segregation measure schema). *We define segregation functions as instances of the tuple*

$$\mathfrak{F} = \langle H_L, f^{ie}(\cdot), \rho(\cdot) \rangle$$

where: (i) $H_L = (V, E, L)$ *is a node-attributed hypergraph; (ii)* $f^{ie} : E \to [0, 1]$ *is a generalized hyperedge type function; (iii)* $\rho : H_L \to [-1, 1]$ *is a generalized segregation measure. Given* H_L*,* $\rho(H_L)$ *measures its segregation, thus providing its segregation index.*

To describe the generalized hyperedge type function $f^{ie}(\cdot)$, it is useful to introduce the concepts of internal and external hyperedges. We borrow these terms from the context of social network analysis, where they refer to intra- and inter-group connections — which the classical segregation measures are based on [5].

Given a social network represented by a node-labeled graph, links between users with the same attribute are called internal ties, whereas other links are called external ties. Since these terms have not been defined consistently for node-labeled hypergraphs, we generalize existing formulations within the hyperedge type function f^{ie}. Specifically, $f^{ie} : E \to [0, 1]$ categorizes each hyperedge $e_i \in E$ as internal (1) or external (0) w.r.t. a provided internal/external ties definition instance. For example, we can strictly define a hyperedge e_i internal only if all its nodes belong to the same group. In our investigation, we experiment with three definition instances, namely, *(i)* strict, *(ii)* majority, and *(iii)* linear. The majority definition instance states that a hyperedge is internal if more than half of its nodes belong to the same group. The linear one measures the extent to which a hyperedge is internal. Thus, the hyperedge contributes to each group proportionally w.r.t. its nodes' group memberships.

Leveraging f^{ie}, in conjunction with the generalized segregation measure $\rho(\cdot)$, classical segregation measures defined on graphs can be easily extended to hypergraphs. $\rho(\cdot)$ is generic; that is, its definition can accommodate different measures. In particular, to show the flexibility of the proposed segregation schema, we focus on two classical segregation measures: *(i)* E-I index [24], and *(ii)* Gupta's Q [21]. Our analysis focuses on those two measures due to their simplicity and adaptability; nonetheless, other measures, such as Freeman's segregation index [17], can be easily extended following the same approach.

E-I index. Although originally designed to assess homophily, the E-I index has

been commonly used to measure segregation. It calculates the ratio of the net difference between inter-group and intra-group ties to the overall number of ties, serving as a normalization factor. Let us denote with E_{int} the number of internal hyperedges of H_L, that is $E_{int} = \sum_{e_i \in E} \mathbb{1}[f^{ie}(e_i) \neq 0]$. Analogously, let E_{ext} be the number of external hyperedges, that is defined as $E_{ext} = \sum_{e_i \in E} \mathbb{1}[f^{ie}(e_i) = 0]$ — where $\mathbb{1}[\phi]$ is a binary indicator function that equals to 1 if the condition ϕ is satisfied, 0 otherwise. Then, the extension of the E-I index to hypergraphs becomes:

$$\rho_{\text{E-I}}(H_L) = \frac{E_{ext} - E_{int}}{E_{ext} + E_{int}}$$

$\rho_{\text{E-I}}(H_L)$ ranges in [-1, 1]. A value close to 1 indicates that groups in the hypergraph tend to have more external connections, while a value close to -1 indicates that groups tend to have more internal connections.

Gupta's Q. This index was introduced to analyze the effects of mixing patterns of sexual contacts on the spread of the HIV epidemic and captures the assortativity of a network in terms of integration and segregation. It considers within-group mixing, focusing on the connections of internal hyperedges against the actual connections. Given a group $l \in L$, we define $r_l = \frac{E^l_{int}}{|E^l|}$ as the ratio between the number of internal hyperedges and the total number of hyperedges containing at least one node belonging to the group l. Here, E^l_{int} is the number of internal hyperedges computed over the subset of hyperedges E^l. Then, the extension of Gupta's Q index to hypergraphs becomes:

$$\rho_Q(H_L) = \frac{\sum_{l \in L} r_l - 1}{h - 1}$$

A high value of $\rho_Q(H_L)$ suggests a strong tendency for nodes within the same group to be connected by hyperedges, thus indicating a higher level of segregation. Moreover, the study of single values of r_l could shed light on group dynamics — e.g., if certain groups predominantly form internal hyperedges, it can indicate a higher level of cohesion.

5 Random Walk HyperSegregation

As discussed, the schema introduced in the previous section allows extending classical segregation measures, initially designed for graphs, to hypergraph seamlessly. However, such conservative extensions can be inaccurate. Hypergraphs encompass non-dyadic relationships that the aforementioned segregation measures may not effectively capture. Moreover, segregation can be observed through a multifaceted lens: while measuring it from a global point of view helps identify overarching patterns and uncover structures of interest, analyzing it from a different granularity can offer insights into how individual entities or groups within the network are experiencing the segregation itself. We approach the latter strategy by leveraging random walks as effective proxies for information flow — thus

analyzing walkers' behaviors to infer whether segregation occurs.

Random Walker model. To maintain the formalization simple without loss of generality, we formalize a generic random walk model in terms of nodes — being the same formalization symmetrically applicable to hyperedges.

We denote with ω_i a random walk rooted at node v_i — i.e., a stochastic process consisting of the random variables $\omega_i^1, \omega_i^2, \ldots, \omega_i^k$ such that ω_i^{k+1} identifies a node visited at random within the neighborhood of ω_i^k. This is generally expressed by a probability $P(\omega_i^{k+1} = v_{j+1} | \omega_i^k = v_j)$, also called transition probability: practically speaking, it represents the probability that given the node v_j at the k-th step of the random walk, the next node is v_{j+1}. This probability can be computed in different ways [40]. In our case, it is deliberately general to accommodate different definitions, which will be given in the following. Moreover, we consider random walks of finite length; therefore, given a value $t > 0$, we denote with ω_i^t a random walk rooted at vertex v_i of length t.

We are interested in analyzing the *"realizations"* of a random walk on the hypergraph H_L. Fixed a v_i, we denote with $\Omega(\omega_i^t)$ a realization of ω_i^t, namely

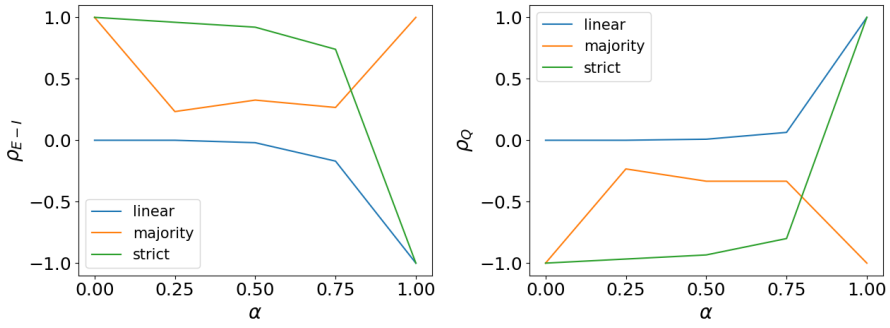
$$\Omega(\omega_i^t) = (v_1, v_2, \ldots, v_t)$$

where $v_j \in V$, for $j = 1, \ldots, t$, and v_j is the node visited at the j-th step of the realization of the random walk ω_i^t. Practically speaking, $\Omega(\omega_i^t)$ is the sequence of node visits obtained by a random walk of length t rooted in v_i. Note that $\Omega(\omega_i^t)$ does not include the random walk source node v_i.

Given the stochastic nature of random walks, it is crucial to acknowledge the potential for high variability between different realizations of the same random walk. Specifically, let $\Omega^1(\omega_i^t)$ and $\Omega^2(\omega_i^t)$ represent two distinct realizations of a random walk rooted at node v_i with length t. Even though both $\Omega^1(\omega_i^t)$ and $\Omega^2(\omega_i^t)$ originate from the same starting node, have the same length and follow the same stochastic rules, the sequences of visited nodes $(v_1^1, v_2^1, \ldots, v_t^1)$ and $(v_1^2, v_2^2, \ldots, v_t^2)$ can be significantly different.

Starting from $\Omega(\omega_i^t)$, we can characterize the nodes that occurred in it, thus gathering insights from them; however, a single realization might not be enough to gather stable analytical insights. To overcome this issue, we focus on collections of realizations. Given a node v_i and the values $t, k > 0$, we define a *collection of realizations* W_i^t as a set of k different realizations of a random walk rooted at node v_i with length t, formally: $W_i^t = \{\Omega^1(\omega_i^t), \Omega^2(\omega_i^t), \ldots, \Omega^k(\omega_i^t)\}$. To avoid burdening the notation, when it is clear from the context, we drop the indication of the random walk, and we write $W_i^{t,k} = \{\Omega^1, \Omega^2, \ldots, \Omega^k\}$. Having described the random walk-based model, we can introduce the Random Walk HyperSegregation (RWHS) measure.

RWHS is a novel segregation measure specifically designed for hypergraphs. A significant distinction w.r.t. extensions of classical segregation measures is that it is defined at the element level, where an element is either a node or a hyperedge. RWHS quantifies the segregation level of each element of the hyper-



(a) (b)

Fig. 1: Extensions of E-I index (a) and Gupta's Q (b) on hypergraphs with controlled homophily. Averages over 50 iterations.

graph *per-se*. Indeed, locality is a desired property since it enables fine-grained analysis of segregation patterns, overcoming known limitations of global measures [31, 34]. Given a node-attributed hypergraph $H_L = (V, E, L)$, RWHS is parametric in *(i)* the node $v_i \in V$ whose level of segregation we want to assess, *(ii)* the value $t > 0$ representing the length of random walks, and *(iii)* the value $k > 0$ representing the number of realizations we want to consider. Having fixed such values, we define two variants of the RWHS, which we call *meet*-wise and *jump*-wise RWHS, respectively.

Definition 4 (*Meet*-wise RWHS). *Based on a collection of realizations* $W_i^{t,k} = \{\Omega^1, \Omega^2, \ldots, \Omega^k\}$*, we denote it as*

$$\phi_m^{t,k}(v_i) = \frac{1}{k} \sum_{r=1}^{k} \frac{|\{v_j \in \Omega^r : \gamma(v_j) = \gamma(v_i)\}|}{t}$$

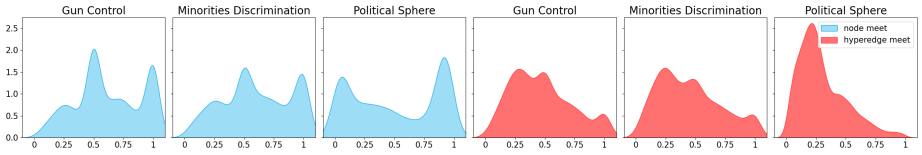
Meet-wise RWHS is the ratio of nodes in the same group as v_i that appear in a realization within $W_i^{t,k}$, averaged over all realizations: it quantifies the exposure of an element to peers belonging to the same group. *Meet-wise RWHS* is bounded in $[0, 1]$: it approaches 1 when elements are enclosed in similarly-valued neighborhoods and 0 when elements are enclosed in differently-valued neighborhoods. This key property allows us to detect not only segregated environments as traditionally defined in the literature (i.e., fragmented areas populated by similar individuals, in which case $\phi_m^{t,k}(v_i) = 1$), but also environments where an element is isolated from its peers, and surrounded only by others belonging to different groups ($\phi_m^{t,k}(v_i) = 0$). In this sense, $\phi_m^{t,k}$ also acts as a measure of *marginalization*.

Definition 5 (*Jump*-wise RWHS). *Based on a collection of realizations* $W_i^{t,k} = \{\Omega^1, \Omega^2, \ldots, \Omega^k\}$*, we denote it as*

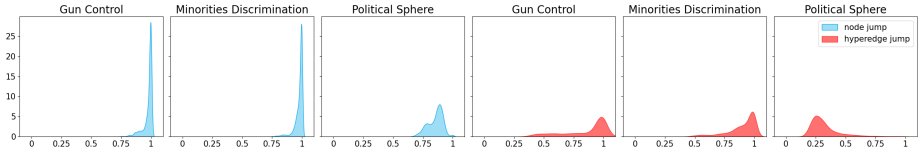
$$\phi_j^{t,k}(v_i) = \frac{1}{k}\sum_{r=1}^{k}\frac{|\{(v_q \in \Omega^r, v_{q+1} \in \Omega^r) : \gamma(v_q) = \gamma(v_{q+1}))\}|}{t-1}$$

Jump-wise RWHS is defined as the ratio of pairs of nodes that belong to the same group and are sequentially adjacent in a realization within $W_i^{t,k}$, averaged over all realizations. Practically, for each realization, it counts the pairs of subsequent steps whose nodes belong to the same group and normalizes it for the total number of pairs, then — to control for random effects — it averages such value across all realizations in the collection. Like its meet-wise counterpart, this measure is also bounded in [0,1]. The key difference is that $\phi_j^{t,k}(v_i)$ is independent from the group $\gamma(v_i)$. Indeed, whether a walker starting at v_i encounters only elements belonging to $\gamma(v_i)$ or only elements belonging to another group is irrelevant. Thus, for walks where all pairs share the same value, $\phi_j^{t,k}(v_i) = 1$ (regardless of which this value is); for walks where each step leads to a differently-labeled element, $\phi_j^{t,k}(v_i) = 0$. We argue that jump-wise RWHS better captures the *fragmentation* aspect of segregation, while the meet-wise RWHS is more akin to homophily estimation.

Some remarks. A peculiarity of working with hypergraphs is that measures can be adapted to be computed not only on nodes but also on hyperedges [1]. In our scenario, the random walk-based model, as well as the RWHS variants based on it, can be seamlessly applied on hyperedges. However, the question of how to label hyperedges arises. A straightforward way to do so is to assign to a hyperedge the most frequent attribute value of its nodes [15, 39]. To clarify this transposition, consider a discussion hypergraph, where nodes are social media users, and hyperedges denote discussion threads such that users interacting under a thread are enclosed in the same hyperedge. Each node is labeled with political leaning on a binary spectrum, e.g., either `republican` or `democrat`. In this case, the meet-wise RWHS on nodes measures how much a node v_i is surrounded (or likely to be influenced) by peers sharing the same political leaning. We can assign each hyperedge a label corresponding to its participants' most frequent political leaning to compute the measure of hyperedges. This way, we implicitly identify `republican`- and `democrat`-dominated threads. In this scenario, the meet-wise RWHS on hyperedges measures how nodes in the same context/thread will likely come across similarly-dominated contexts/threads. In a way, this transposition acts as a meso-scale segregation measure. Another remark concerns how random walks are computed. As mentioned earlier, the transition probability governing them can be computed in several ways [40], especially when dealing with the hypergraph scenario. For instance, in [10], the authors formalize a version of random hypergraph walks via edge-dependent vertex weights; that is, each node contributes to the walk based on a collection of weights that depend on the hyperedges it participates. Conversely, in our work, we leverage random walks as introduced in [7]. Regarding walks on nodes, the transition probability linearly correlates with hyperedge size, and within-edge jumps are more likely than cross-edge ones, coherently with information-spreading properties. Instead, when we



(a)



(b)

Fig. 2: **Reddit Politics.** Meet-wise (a) and jump-wise (b) RWHS score distribution. The X-axis represents RWHS, while the Y-axis represents the density estimate.

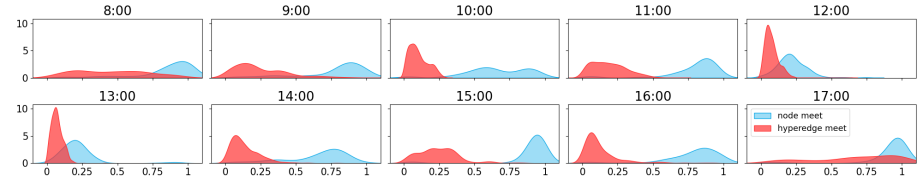
deal with random walks on hyperedges, the transition probability depends on the number of common nodes between hyperedges.

6 Experiments

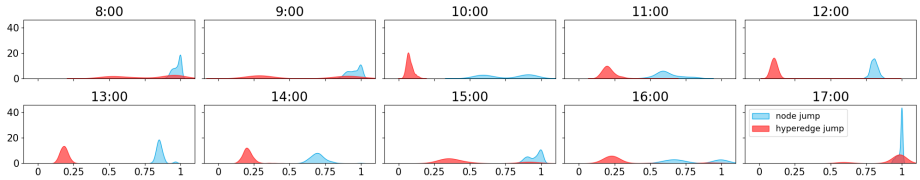
In the following, we apply the proposed segregation measures on synthetic and real-world hypernetworks. Our aim is twofold: (i) validating the extension of pairwise segregation measures to the high-order scenario, underlying its non-triviality by discussing outputs on synthetic hypernetworks with planted node mixing; (ii) validating the RWHS measures in real-world hypernetworks settings as described by face-to-face and online social interactions.

Conservative extensions and their limitations. To show that generalizing existing segregation measures to hypergraphs is non-trivial, we applied the modified E-I index and Gupta's Q on synthetic hypernetworks with tunable assortativity. To such an extent, we extended a well-known graph generator [25] to account for high-order interactions.

In the following, we first briefly describe the original graph generator, discuss our modified implementation, and ultimately use it to evaluate segregation measures. The original generator is a node-attributed version of the Barabasi-Albert model, where connection probability depends on (i) preferential attachment and (ii) on a parameter α that controls the likelihood of same-class nodes being connected. As outlined by the authors, $\alpha = 1$ results in a perfectly assortative network (i.e., nodes are connected only to same-class nodes); conversely, $\alpha = 0$ results in a perfectly disassortative network (i.e., nodes are connected only to nodes from the other class); lastly, $\alpha = 0.5$ results in a classical Barabasi-Albert model. We choose to extend this model because it seamlessly controls the edge



(a)



(b)

Fig. 3: **Primary School.** Meet-wise (a) and jump-wise (b) RWHS score distribution. The X-axis represents RWHS, while the Y-axis represents the density estimate.

formation mechanism as a linear combination of degrees and in-group preference — which we use as a proxy for segregation. We modify this implementation as follows: (i) each node is assigned z hyperedges; (ii) the probability of attaching to a node depends on its number of neighbors, which are the nodes that share at least one hyperedge with it; (ii) hyperedge size is a random integer $s \in [2, 10]$. We generate synthetic hypergraphs for our experiments with $N = 1000$ nodes, $z = 3$, equal group size, and varying values for α. Results are averaged over 50 iterations.

E-I index. Figure 1a depicts the output of the extended E-I index on hypernetworks with known underlying segregation, highlighting how choosing different definition instances, introduced in Section 4 impacts the measure's output. The *strict* one correctly captures both full segregation and full heterogeneity, but the transition between these extremes is not smooth, remaining constant until $\alpha = 0.5$ and then plummets rapidly to -1. *Linear* cannot capture heterogeneity and instead moves from 0 (no segregation nor heterogeneity) to -1. Finally, *majority* strongly deviates from what one would expect a segregation measure to capture; indeed, it scores high values (~ 1) at both extremes and relieves heterogeneity when tie formation is independent of homophily.

Gupta's Q's extensions (Figure 1b) provides similar results. *Strict* gives coherent results at the extremes, but overestimates heterogeneity; *Linear* correctly identifies complete segregation, but never captures heterogeneity; *Majority* returns unsatisfactory results in all scenarios except for heterogeneity.

Therefore, the performed controlled experiments underline that one-to-one high-order extensions of pairwise measures are subject to severe limitations, not behaving as they were originally intended. Such a negative result is tied to the

inconsistency of semantics while moving from dyadic to polyadic interactions: conservative extensions suffer from not taking such semantic nuances into account, thus leading to inconsistencies or unexpected behaviors.

Hypernetwork based segregation measure. We test RWHS on high-order social interactions data, focusing on (i) online debates between pro/anti-Trump Redditors during the first two years and half of Donald Trump's presidency [28] and on (ii) children face-to-face interactions. Reddit data describe interactions and user ideology labels (`protrump`, `neutral`, `antitrump`) across three topics: *Gun control*, *Minorities Discrimination* and *Political Sphere*[4]. In these scenarios, each hyperedge identifies the set of users that commented on a given post at the same level of the discussion tree — thus modeling individual contexts of a broader discussion. The Primary School dataset [37] (henceforth, PS) from the SocioPatterns project[5] comprises face-to-face interaction data collected via RFID sensors over two days. The RFID sensors capture interactions with a 20-second resolution; we first preprocess the data by aggregating it into a series of static networks so that each network contains all interactions captured over an hour. Each node is enriched with information on their class (10 unique values in total). Subsequently, we infer the system's high-order structure by leveraging the method introduced in [8]. Specifically, if at time t there are $N * (N + 1)/2$ edges between N nodes such that they are involved in a fully connected clique, such links are promoted to form a hyperedge of size N. We limit our analysis to the first day.

Reddit. We compute RWHS scores on nodes and hyperedges. Scores are computed over a collection of realizations of $k = 50$ walks of length $t = 6$ for each element. As shown by Figure 2a, all hypernetworks are characterized by highly- and non-segregated areas. In particular, meet-wise RWHS on nodes show taller peaks towards high values, outlining a tendency towards segregation. Conversely, on hyperedges meet-wise shows lower values, meaning that users participate in a heterogeneous collection of discussions, i.e., a `protrump` user participates in both `pro-` and `antitrump`-dominated discussions. This is coherent with the results of the analysis in [15]. Jump-wise shows extremely right-skewed distributions for nodes and hyperedges — i.e., users are likelier to be located in label-wise homogeneous areas. In all scenarios, curves related to the Political Sphere differ from the other datasets, highlighting strongly heterogeneous areas. We hypothesize that since the discussions taking place in Political Sphere pertain to a wide range of topics, users have higher chances to connect with peers from other groups.
To assess whether the observed segregation is statistically significant, we compare RWHS scores of real data to those obtained on instances of an extension of the Chung-Lu configuration model [2] preserving both hyperedge size and node degree distributions — also tailored to preserve label distribution of the corresponding real hypernetwork. Comparing RWHS distributions via a two-sample

[4] Details on data collection and label inference in [28]
[5] `www.sociopatterns.org`

Kolmogorov-Smirnov test for goodness of fit — which tests the null hypothesis that the distribution underlying two samples is the same — we rejected the null hypothesis with $p < 0.01$: all empirical distributions significantly differ from the corresponding null ones. Moreover, all empirical hypernetworks have higher mean and median than the corresponding generated ones, suggesting a general tendency for higher segregation than expected at random, a characteristic perfectly captured by RWHS.

Primary School. We perform a temporal analysis of spatial segregation on the PS dataset. Individuals in PS spend most of their school day in their classroom [37]; thus, we expect high levels of segregation. Indeed, the meet-wise scores, Figure 3, are generally high on nodes, outlining the prevalence of same-class group interactions. Conversely, on hyperedges scores tend towards -1, implying that reaching students from other classes is still possible. Symmetrically, jump-wise scores are generally high on nodes — emphasizing that they are surrounded by closed groups (i.e., classes) — while being low on hyperedges (e.g., inter-class interactions still noticeable). Moreover, all RWHS scores capture strong homogeneity when students arrive at school (8:00) and when they leave classes (17:00) — conversely from lunchtime (12:00-13:00) when the highest heterogeneity is captured. These results are coherent with what was previously observed in [37], showing that RWHS correctly captures the intra- and inter-class dynamics we expect to observe, thus providing a suitable and effective measure to capture segregation.

7 Conclusion

While segregation has been extensively studied in classical network setups, it remains an overlooked aspect in higher-order social modeling. We proposed a general schema that extends classical segregation approaches to node-attributed hypernetworks to fill this gap. We tested it by reinterpreting the well-known E-I index and Gupta's Q measures. Moreover, we also defined a new multi-scale segregation index, RWHS, that relies on random walks as effective proxies for information flow. With an exhaustive series of experiments, we illustrated that conservative extensions of classic measures to high-order settings cannot adequately capture segregation. Conversely, RWHS allows us to observe homogeneous and heterogeneous multi-scale mixing patterns. Such a result underlines the limitations of classical approaches that, when applied to high-order topologies, fail to exploit the richer topological semantics they offer. This paper should not be considered as an endpoint for research on this topic but rather as a starting one. As future research, we plan to extend the segregation study to temporal hypernetworks where nodes, attributes, and interactions co-evolve over time — idealistically leveraging time-respecting walks [15] to embed temporal constraints within random walk processes.

Acknowledgments. This work is supported by (i) the European Union – Horizon 2020 Program under the scheme "INFRAIA-01-2018-2019 – Integrating Activities for Advanced Communities", Grant Agreement n.871042, "SoBigData++: European Integrated Infrastructure for Social Mining and Big Data Analytics" (`http:`

//www.sobigdata.eu); (ii) SoBigData.it which receives funding from the European Union – NextGenerationEU – National Recovery and Resilience Plan (Piano Nazionale di Ripresa e Resilienza, PNRR) – Project: "SoBigData.it – Strengthening the Italian RI for Social Mining and Big Data Analytics" – Prot. IR0000013 – Avviso n. 3264 del 28/12/2021; (iii) EU NextGenerationEU programme under the funding schemes PNRR-PE-AI FAIR (Future Artificial Intelligence Research).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. S. G. Aksoy, C. Joslyn, C. O. Marrero, B. Praggastis, and E. Purvine. Hypernetwork science via high-order hypergraph walks. *EPJ Data Science*, 9(1):16, 2020.
2. S. G. Aksoy, T. G. Kolda, and A. Pinar. Measuring and modeling bipartite graphs with community structure. *Journal of Complex Networks*, 5(4):581–603, 2017.
3. A. Baroni and S. Ruggieri. Segregation discovery in a social network of companies. *Journal of Intelligent Information Systems*, 51:71–96, 2018.
4. F. Battiston, G. Cencetti, I. Iacopini, V. Latora, M. Lucas, A. Patania, J.-G. Young, and G. Petri. Networks beyond pairwise interactions: Structure and dynamics. *Physics Reports*, 874:1–92, 2020.
5. M. Bojanowski and R. Corten. Measuring segregation in social networks. *Social networks*, 39:14–32, 2014.
6. S. P. Borgatti. Identifying sets of key players in a social network. *Computational & Mathematical Organization Theory*, 12:21–34, 2006.
7. T. Carletti, F. Battiston, G. Cencetti, and D. Fanelli. Random walks on hypergraphs. *Physical review E*, 101(2):022308, 2020.
8. G. Cencetti, F. Battiston, B. Lepri, and M. Karsai. Temporal properties of higher-order interactions in social networks. *Scientific reports*, 11(1):1–10, 2021.
9. M. Charles and D. B. Grusky. Models for describing the underlying structure of sex segregation. *American journal of Sociology*, 100(4):931–971, 1995.
10. U. Chitra and B. Raphael. Random walks on hypergraphs with edge-dependent vertex weights. In *Proc. of the 36th International Conference on Machine Learning (ICML 2019)*, volume 97, pages 1172–1181. PMLR, 2019.
11. M. Conover, J. Ratkiewicz, M. Francisco, B. Gonçalves, F. Menczer, and A. Flammini. Political polarization on twitter. In *Proc. of the International AAAI Conference on Web and Social Media (ICWSM)*, volume 5, pages 89–96. AAAI, 2011.
12. R. L. de Andrade and L. C. Rêgo. A proposal for the ei index for fuzzy groups. *Soft Computing*, 27(4):2125–2137, 2023.
13. X. Dong, A. J. Morales, E. Jahani, E. Moro, B. Lepri, B. Bozkaya, C. Sarraute, Y. Bar-Yam, and A. Pentland. Segregated interactions in urban and online. *EPJ Data Science*, 9(20):1–22, 2020.
14. G. A. Dotson, C. Chen, S. Lindsly, A. Cicalo, S. Dilworth, C. Ryan, S. Jeyarajan, W. Meixner, C. Stansbury, J. Pickard, N. Beckloff, A. Surana, M. Wicha, L. A. Muir, and I. Rajapakse. Deciphering multi-way interactions in the human genome. *Nature Communications*, 13(1):5498, 2022.
15. A. Failla, S. Citraro, and G. Rossetti. Attributed stream hypergraphs: temporal modeling of node-attributed high-order interactions. *Applied Network Science*, 8(1):1–19, 2023.

16. D. Vasques Filho. Cohesion and segregation in higher-order networks. *arXiv e-prints*, pages arXiv–2207, 2022.
17. L. C. Freeman. Segregation in social networks. *Sociological Methods & Research*, 6(4):411–429, 1978.
18. K. Garimella, G. F. Morales, A. Gionis, and M. Mathioudakis. Quantifying controversy on social media. *ACM Transactions on Social Computing*, 1(1):1–27, 2018.
19. Y.-C. Gong, M. Wang, W. Liang, F. Hu, and Z.-K. Zhang. Uhir: An effective information dissemination model of online social hypernetworks based on user and information attributes. *Information Sciences*, 644:119284, 2023.
20. O. B. Gretha, P. M. Cristal, and N. Mauhe. Segregation in social networks: a simple schelling-like model. In *Proc. of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 95–98. IEEE, 2018.
21. S. Gupta, R. M. Anderson, and R. M. May. Networks of sexual contacts: implications for the pattern of spread of hiv. *Aids*, 3(12):807–818, 1989.
22. A. D. Henry, P. Prałat, and C.-Q. Zhang. Emergence of segregation in evolving social networks. *Proceedings of the National Academy of Sciences*, 108(21):8605–8610, 2011.
23. A. M. Jaramillo, H. T.P. Williams, N. Perra, and R. Menezes. The structure of segregation in co-authorship networks and its impact on scientific production. *EPJ Data Science*, 12(1):47, 2023.
24. D. Krackhardt and R. N. Stern. Informal networks and organizational crises: An experimental simulation. *Social psychology quarterly*, pages 123–140, 1988.
25. E. Lee, F. Karimi, C. Wagner, H.-H. Jo, M. Strohmaier, and M. Galesic. Homophily and minority-group size explain perception biases in social networks. *Nature Human Behaviour*, 3(10):1078–1087, 2019.
26. R. Louf and M. Barthelemy. Patterns of residential segregation. *PloS one*, 11(6):e0157476, 2016.
27. D. S. Massey and N. A. Denton. The dimensions of residential segregation. *Social forces*, 67(2):281–315, 1988.
28. V. Morini, L. Pollacci, and G. Rossetti. Toward a standard approach for echo chamber detection: Reddit case study. *Applied Sciences*, 11(12):5390, 2021.
29. M. Newman. *Networks*. 2018. Oxford University Press.
30. M. E.J. Newman. Complex systems: A survey. *arXiv preprint arXiv:1112.1440*, 2011.
31. L. Peel, J.-C. Delvenne, and R. Lambiotte. Multiscale mixing patterns in networks. *Proceedings of the National Academy of Sciences*, 115(16):4057–4062, 2018.
32. S. C. Phillips, J. Uyheng, and K. M. Carley. A high-dimensional approach to measuring online polarization. *Journal of Computational Social Science*, 6(2):1147–1178, 2023.
33. B.-A. Reme, A. Kotsadam, J. Bjelland, P. R. Sundsøy, and J. T. Lind. Quantifying social segregation in large-scale networks. *Scientific Reports*, 12(1):6474, 2022.
34. G. Rossetti, S. Citraro, and L. Milli. Conformity: a path-aware homophily measure for node-attributed networks. *IEEE Intelligent Systems*, 36(1):25–34, 2021.
35. T. C. Schelling. Models of segregation. *The American economic review*, 59(2):488–493, 1969.
36. S. Sousa and V. Nicosia. Quantifying ethnic segregation in cities through random walks. *Nature Communications*, 13(1):5809, 2022.

37. J. Stehlé, N. Voirin, A. Barrat, C. Cattuto, L. Isella, J-F. Pinton, M. Quaggiotto, W. Van den Broeck, C. Régis, B. Lina, and P. Vanhems. High-resolution measurements of face-to-face contact patterns in a primary school. *PloS one*, 6(8):e23176, 2011.
38. G. Tóth, J. Wachs, R. Di Clemente, A. Jakobi, B. Ságvári, J. Kertész, and B. Lengyel. Inequality is rising where social network segregation interacts with urban topology. *Nature communications*, 12(1):1143, 2021.
39. N. Veldt, A. R. Benson, and J. Kleinberg. Combinatorial characterizations and impossibilities for higher-order homophily. *Science Advances*, 9(1):eabq3200, 2023.
40. F. Xia, J. Liu, H. Nie, Y. Fu, L. Wan, and X. Kong. Random walks: A review of algorithms and applications. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 4(2):95–107, 2019.