Supplementary Information
Higher-order motif analysis in hypergraphs

Quintino Francesco Lotito, Federico Musciotto, Alberto Montresor, and Federico Battiston

SUPPLEMENTARY NOTE 1. DATASETS

The datasets gathered to perform our motif analysis of real-world higher-order systems come from a variety of domains: sociology (proximity contacts, votes), technology (e-mails), biology (gene/disease, drugs) and co-authorship.

- **Gene/disease**: Nodes correspond to genes and each hyperedge is the set of genes associated to a disease.
- **NDC_classes**: Nodes correspond to classification labels and each hyperedge is the set of all the labels applied to a drug. Each drug is timestamped with the day it was first marketed.
- **NDC_substances**: Nodes correspond to substances and each hyperedge is the set of substances in a drug. Each drug is timestamped with the day it was first marketed.
- **DBLP**: Nodes correspond to authors and each hyperedge is the set of authors on a scientific publication tracked by DBLP. Each paper is timestamped with the year of publication.
- **History**: Nodes correspond to authors and each hyperedge is the set of authors on a scientific publication in the field of history. Each paper is timestamped with the year of publication.
- **Geology**: Nodes correspond to authors and each hyperedge is the set of authors on a scientific publication in the field of geology. Each paper is timestamped with the year of publication.
- **PACS**: Nodes correspond to authors and each hyperedge is the set of authors on a scientific publication in the field of physics. Each paper is timestamped with the year of publication and with its PACS classification. The latter allows to split the set of papers in ten subfields of physics by using the highest hierarchical level of PACS classification.
- **email-EU**: Nodes correspond to users in a European research institution and each hyperedge consists of the sender and all recipients of an email. Each email is timestamped.
- **email-ENRON**: Nodes correspond to Enron employees and each hyperedge consists of the sender and all recipients of an email. Each email is timestamped.
- **Wiki**: Nodes correspond to Wikipedia users and each hyperedge is the set of users that expressed the same vote in a Wikipedia voting event since the inception of Wikipedia until January 2008.
- **Justice**: Nodes correspond to justices of the Supreme Court in the US and each hyperedge is the set of justices that expressed the same vote in a case from 1946 to 2019.
- **Primary school**: Nodes are students of a primary school. Wearable sensors are exploited to construct a network of proximity contacts among the students. Contacts are aggregated in time-windows of 20 seconds. Each hyperedge is a maximal clique in each layer (i.e. each interval) of the temporal network of contacts.
- **High school**: Nodes are students of a high school. Wearable sensors are exploited to construct a network of proximity contacts among the students. Contacts are aggregated in time-windows of 20 seconds. Each hyperedge is a maximal clique in each layer (i.e. each interval) of the temporal network of contacts.
- **Conference**: Nodes are participants of a conference. Wearable sensors are exploited to construct a network of proximity contacts among people. Contacts are aggregated in time-windows of 20 seconds. Each hyperedge is a maximal clique in each layer (i.e. each interval) of the temporal network of contacts.
- **Hospital**: Nodes are people at a hospital. Wearable sensors are exploited to construct a network of proximity contacts among people. Contacts are aggregated in time-windows of 20 seconds. Each hyperedge is a maximal clique in each layer (i.e. each interval) of the temporal network of contacts.

Dataset	N	E	E_2	E_3	E_4	E_5	Domain
Gene/disease	9703	11181	1311	614	443	363	Bio
NDC_classes	1161	1088	297	121	125	94	Bio
NDC_substances	5311	9906	1130	745	535	500	Bio
DBLP	1924991	2466799	693363	667291	419434	205970	Co-auth
History	1014734	895439	160885	47423	19120	8775	Co-auth
Geology	1256385	1203895	275736	227950	159509	99140	Co-auth
PACS0	98478	75985	12168	8373	14068	3064	Co-auth
PACS1	67055	43957	10532	7869	4647	2015	Co-auth
PACS2	47475	27504	5424	3452	2224	1411	Co-auth
PACS3	33479	16977	2099	2105	2590	1169	Co-auth
PACS4	57533	41133	8977	6571	5139	2897	Co-auth
PACS5	14455	5306	1062	1041	719	477	Co-auth
PACS6	71989	41663	5140	3293	4946	1715	Co-auth
PACS7	82964	55227	9664	9325	8294	4915	Co-auth
PACS8	13451	5618	1659	1447	910	485	Co-auth
PACS9	18666	7515	2361	2143	1034	397	Co-auth
email-EU	998	25027	12753	4938	2294	1359	Tech
email-ENRON	143	1512	809	317	138	63	Tech
Wiki	6210	5925	593	427	338	304	Socio
Justice	38	2864	216	456	506	560	Socio
Primary school	242	12704	7748	4600	347	9	Socio
High school	327	7818	5498	2091	222	7	Socio
Conference	403	10541	8268	1861	258	63	Socio
Hospital	75	1825	1108	657	58	2	Socio
Workplace	92	788	742	44	2	0	Socio
Baboons	13	231	78	142	11	0	Socio

Supplementary Table 1: Details of the real-world networked datasets considered for our experiments. Each higher-order network is described by its number of nodes, its total number of hyperedges and its number of hyperedges of size 2, 3, 4 and 5.

- **Workplace**: Nodes are employees of a company. Wearable sensors are exploited to construct a network of proximity contacts among people. Contacts are aggregated in time-windows of 20 seconds. Each hyperedge is a maximal clique in each layer (i.e. each interval) of the temporal network of contacts.

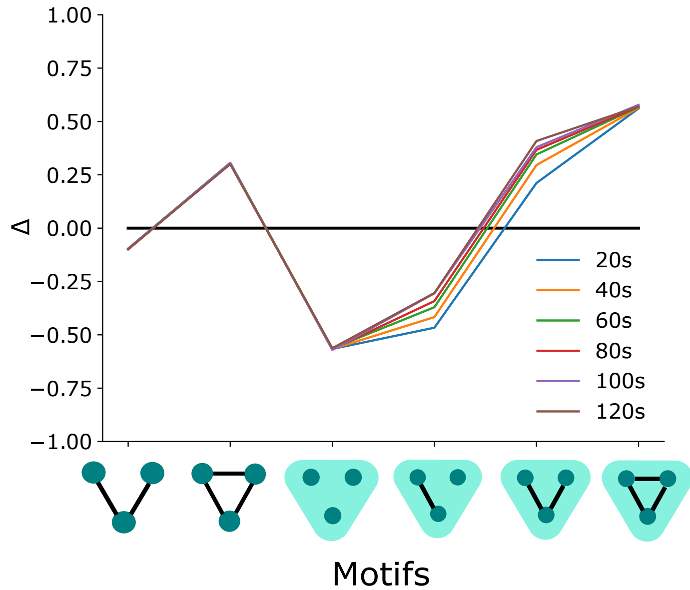
- **Baboons**: Nodes are baboons in an enclosure of a Primate Center in France. Wearable sensors are exploited to construct a network of proximity contacts among baboons. Contacts are aggregated in time-windows of 20 seconds. Each hyperedge is a maximal clique in each layer (i.e. each interval) of the temporal network of contacts.

The summary statistics of the datasets are reported in Supplementary Table 1.

SUPPLEMENTARY NOTE 2. BUILDING HYPEREDGES FROM DYADIC DATA

While in some datasets higher-order structures are naturally encoded as hyperedges (e.g. three authors collaborating on the same paper), in others (e.g. face-to-face interactions) one need to infer higher-order structures from dyadic interactions. The simplest way to perform this task is to set time-windows of size t, build the aggregated network for each time-window, and promote cliques of size n to hyperedges of size n. The idea is that if every node in a set of n nodes is interacting with every other node in the set, than the original interaction was probably a group interaction.

The choice of the size of the time window could be an important factor when evaluating the experimental results. The smallest timescale available in most of the datasets is 20 seconds, which represents our originally selected time window. In Supplementary Figure 1, we report the change in the distribution of the occurrences of the higher-order motifs of order 3 with changes in the considered time window. We considered the dataset High School and time windows of 20, 40, 60, 80, 100 and 120 seconds. At these scales, the choice of the time window does not seem to take a role in the results. Please notice that larger time windows would badly approximate the notion of *group interaction*, since they would aggregate pairwise interactions happened far apart in time.



Supplementary Figure 1: Effect of the choice of the time window for promoting cliques to group interactions on the SP of the dataset High School. Each time window is expressed in seconds. Δ is the relative abundance of each motif with respect to the null model.

SUPPLEMENTARY NOTE 3. CLUSTERING HYPERGRAPHS VIA THEIR 4-MOTIFS PROFILES

Higher-order motifs of order 4 are able to better capture structural information compared to their counterparts of order 3. Intuitively, one can notice a richer hierarchical intra-cluster organization from the dendrograms on the left of the correlation matrices (Fig. 2b and 3b). Moreover, a better separation between the two clusters can be noticed from the lowering of the mean correlation between the SPs of hypergraphs belonging to different clusters (practically, we can notice less red squares and more blue squares that separate the clusters).

More formally, we can consider the scatter plot in Supplementary Figure 2. Each point $p = (x, y)$ is described by $x = corr_3(d_1, d_2)$ and $y = corr_4(d_1, d_2)$ where $corr_n(d_1, d_2)$ is the correlation between the SPs of the higher-order motifs of order n of the datasets d_1 and d_2. Every possible pair of datasets is considered. In red we plot points in which the two datasets belong to the same big cluster (e.g. both datasets in Bio/Co-auth), in blue we plot points in which the two datasets belong to different clusters (i.e. one in Bio/Co-auth and the other in Socio/Tech). We can observe a significant drop in the correlations of the SPs of datasets from different clusters when considering motifs of order 4.

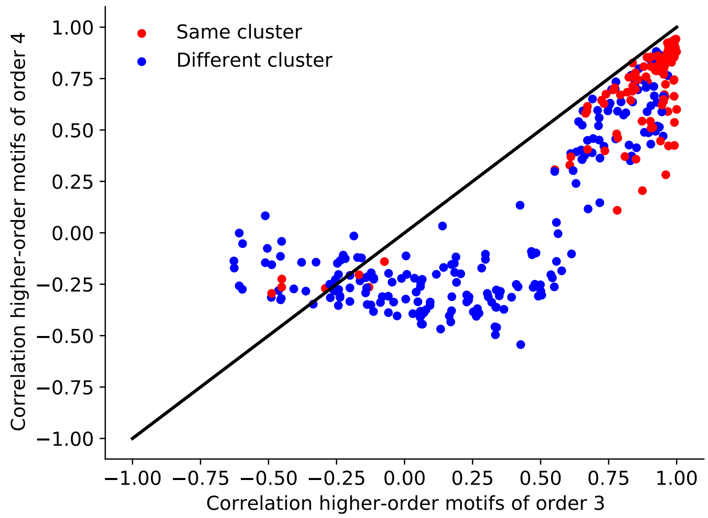
Taken together, these results confirm that scaling up the analysis of higher-order motifs to larger orders inform our tool with more information, capturing more nuanced and fine-grained wiring patterns.

SUPPLEMENTARY NOTE 4. DISAGGREGATED 3-MOTIFS

In Supplementary Figure 3 we compute and show all the SPs of the real-world hypergraphs considered in our experiments, subdivided by their domain. In the main paper, we group and average them to extract a fingerprint of a domain. We can notice that hypergraphs from the same domain tend to have very correlated SPs, while hypergraphs from different domains do not display such correlation.

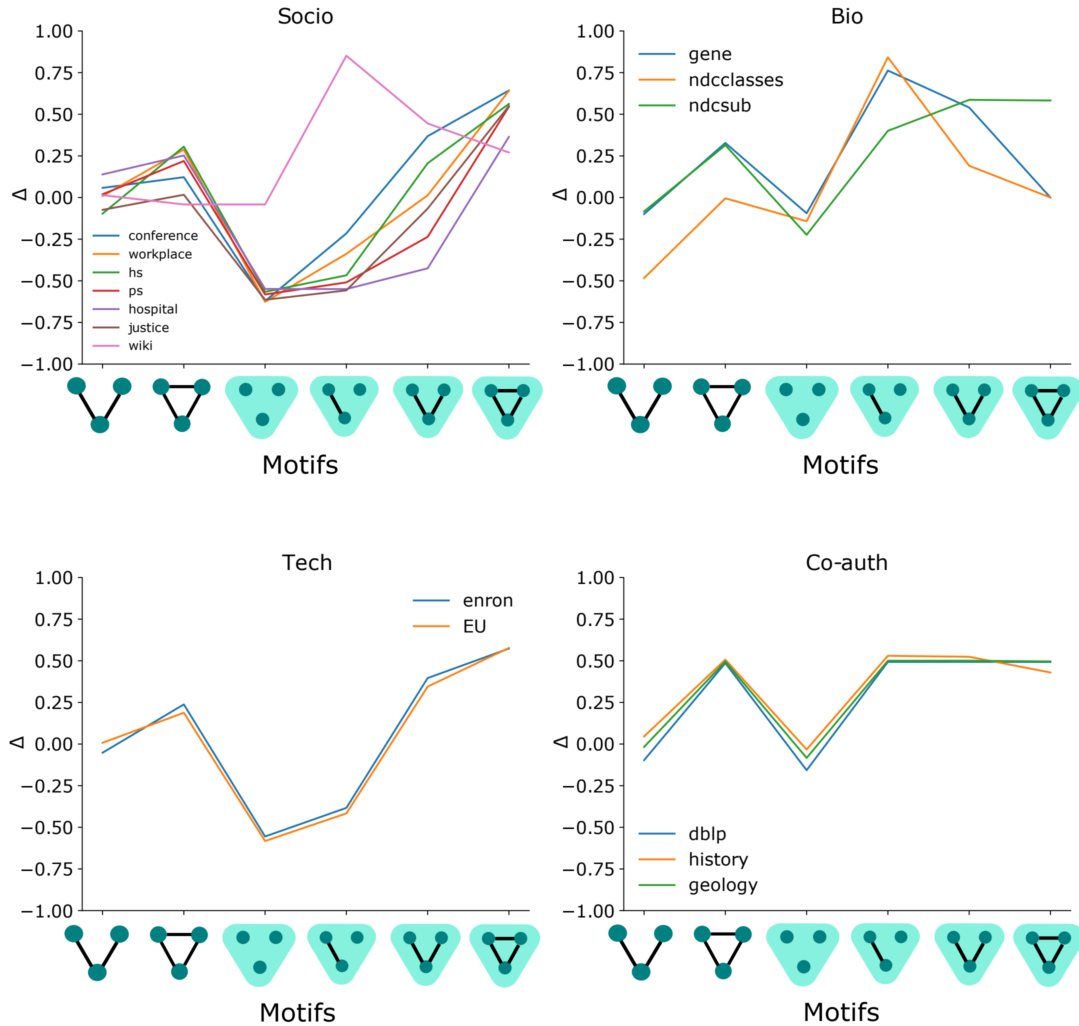
SUPPLEMENTARY NOTE 5. MOST SIGNIFICATIVE 4-MOTIFS

In Supplementary Figure 4 we report the most under-expressed higher-order motifs of order 4 for each domain. Comparing Supplementary Figure 4 with Fig.3c from the main paper (which shows the most over-expressed higher-order motifs of order 4 for each domain) further supports the idea that hypergraphs from the Socio / Tech domain have a preference towards over-expressing structures involving more lower-order nested relations (e.g. dyadic links), while

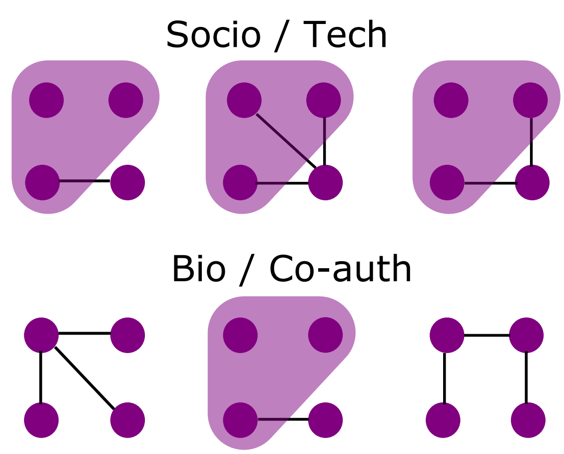


Supplementary Figure 2: Each point $p = (x, y)$ is described by $x = corr_3(d_1, d_2)$ and $y = corr_4(d_1, d_2)$ where $corr_n(d_1, d_2)$ is the correlation between the SPs of the higher-order motifs of order n of the datasets d_1 and d_2. Every possible pair of datasets is considered. In red we plot points in which the two datasets belong to the same big cluster (e.g. both datasets in Bio/Co-auth), in blue we plot points in which the two datasets belong to different clusters (i.e. one in Bio/Co-auth and the other in Socio/Tech).

hypergraphs from the Bio / Co-author domain display a preference towards less nested relations but of higher-order. For example, people interacting in groups are likely to interact also in single pairs, therefore it is unlikely that group interactions in the Socio / Tech domain are not supported by a large number of nested lower-level interactions. In fact, from Supplementary Figure 4, we can notice that patterns involving a group interaction with none or few nested dyadic interactions are penalized. To name another example, representative of the opposite cluster, people tend to write papers in large groups and tend to maintain the same research group over time, with few additions or removals; therefore patterns involving only dyadic relations are penalized.



Supplementary Figure 3: All the SPs of the real-world hypergraphs considered in our experiments, subdivided by their domain. Hypergraphs from the same domain tend to have very correlated SPs, while hypergraphs from different domains do not display such correlation.



Supplementary Figure 4: Most under-expressed higher-order motifs of order 4 from the two clusters.