Supplementary Material

1 ILLUSTRATION OF HYPERGRAPH DEFINITIONS

In Figure S1, we illustrate some of the hypergraph concepts introduced in Section 2.1.

2 LABELLED VERSION OF FIGURE 2B

In Figure 2B in the main text, we did not label individual curves. In Figure S2 we split the curves from Figure 2B into 3 different panels and provide curve labels. We use a different naming convention here than we did for 3-patterns in the rest of the manuscript. We do so because the other (less complicated) notation could uniquely describe 1- 2- and 3-patterns, but cannot uniquely map all 4-patterns. The naming convention is described in the caption of Figure S2.

3 PROOF OF THEOREM 2.4

PROOF. Without loss of generality, let us focus on the pure pattern consisting only of k-node hyperedges. Let us denote the prevalence of this pattern $P(X_k)$. By Eq. (1), the analytical formula for $P(X_k)$ is,

$$P(X_k) = p_k^{\binom{m}{k}} \prod_{i=k+1}^{m} (1-p_i)^{\binom{m}{i}}. \tag{S1}$$

According to Lemma 2.11, for any $\epsilon > 0$ and large enough N, we can choose a p such that all factors in this product are arbitrarily large. All other patterns will either have factors of $(1-p_k)$ in the analytical expression, or factors of p_l in the expression, where $l \geq k+1$. By Lemma 2.11, N can be chosen large enough to make any such factors arbitrarily close to 0 if $0 < p_k < 1$. This proves the theorem.

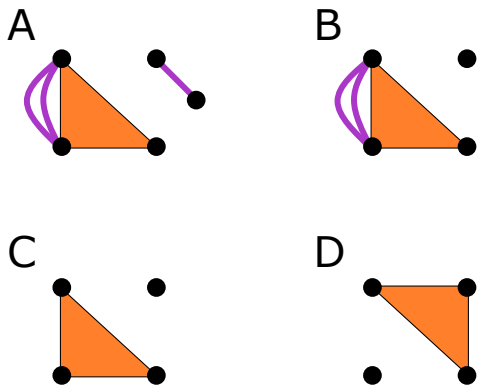

Figure S1. Illustration of some concepts introduced in Section 2.1. **A** 5-node hypergraph. **B** Induced subhypergraph of (A) on the 4 left-most nodes. **(C)** Maximal induced subhypergraph of (A) on 4 left-most nodes. **D** A 4-pattern. (C) happens to be an instance of this 4-pattern in the hypergraph in (A). Labelling nodes $\{0,1,2,3\}$ starting with the label 0 in the top-left corner and increasing labels by 1 in the clockwise direction, (C) and (D) are also examples of two different labelled 4-patterns.

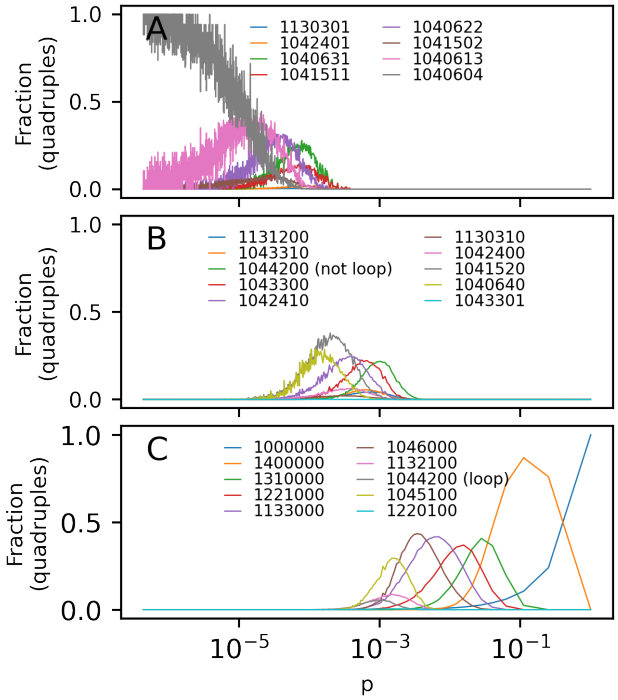

Figure S2. Labelled version of Figure 2B split into 3 panels to make plot colors easier to distinguish. The naming convention used for 4-patterns is different than that used for 3-patterns in the rest of the manuscript. The pattern $1ABCDEF$ has A 3-node hyperedges filled and B not filled, C 2-node hyperedges filled and D not filled, E 1-node hyperedges filled and F not filled. The pattern 1000000 is the 4-pattern consisting of a single 4-node hyperedge. There are 2 possible 4-patterns with the name 1044200 (consisting of 4 2-node hyperedges and all other possible hyperedges missing): one where the hyperedges form a loop and another where they do not. The analytical solutions are not plotted in this figure.

4 PROOF OF LEMMA 2.14

PROOF. By Lemma 2.11, if $p_k > \frac{1}{2}$, N can be chosen large enough to make the value of p_l arbitrarily close to 1 for $l \leq k-1$. The prevalence of the pure pattern with k-node hyperedges is

$$P(X_k) = p_k^{\binom{m}{k}}\Omega, \tag{S2}$$

where $\Omega = \prod_{i=k+1}^{m}(1-p_i)^{\binom{m}{i}}$. The prevalence of the non-pure patterns containing x_k k-node and x_{k-1} $(k-1)$-node hyperedges is,

$$P(X'_{k-1,k}) = p_{k-1}^{x_{k-1}} p_k^{x_k}(1-p_k)^{\binom{m}{k}-x_k}\Omega. \tag{S3}$$

By Lemma 2.11, for any $\epsilon > 0$ and large enough N, the first factor in this last expression can get arbitrarily close to 1. In this limit, the pure and non-pure patterns therefore cross when the remaining factors in Eqs. (S2) and (S3) are equal. This happens at when $p_k = (1-p_k)$; in other words, $p_k = \frac{1}{2}$. For any $p_k > \frac{1}{2}$, $p_k > (1-p_k)$. Comparing Eqs. (S2) and (S3), the pure pattern dominates in this case.

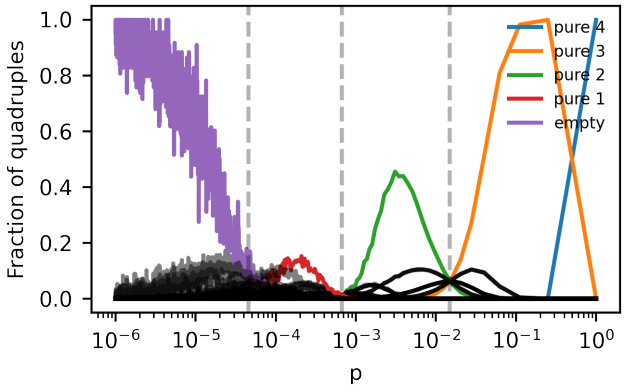

Figure S3. Frequency of labelled m-patterns in the $G^{(m)}(N,p)$ model with pure k-node hyperedge patterns in colors. Each datapoint plots the average prevalence of a labelled m-pattern in 10 simulations of the model for the given p value and $m=4$, $N=100$. Vertical gray dashed lines indicate values pf p where $p_k = 1/2$ for $1 \leq k \leq 4$. As argued in the proof of Theorem 2.18, many prevalence curves cross at these values of p.

5 PROOF OF LEMMA 2.15

PROOF. By Lemma 2.11, if p_{k+1} is non-zero, we can choose N large enough $p_l \to 1$ as $N \to \infty$ for $l \leq k-1$. The prevalence of the pure pattern with k-node hyperedges is

$$P(X_k) = p_k^{\binom{m}{k}}(1-p_{k+1})^{\binom{m}{k+1}}\Omega', \tag{S4}$$

where $\Omega' = \prod_{i=k+2}^{m}(1-p_i)^{\binom{m}{i}}$. The prevalence of the non-pure patterns containing x_k k-node and x_{k+1} $(k+1)$-node hyperedges is,

$$P(X'_{k-1,k}) = p_k^{x_k} p_{k+1}^{x_{k+1}}(1-p_{k+1})^{\binom{m}{k+1}-x_{k+1}}\Omega'. \tag{S5}$$

By Lemma 2.11, for any $\epsilon > 0$ and large enough N, the first factor in both Eqs (S4) and (S5) can get arbitrarily close to 1. Hence, the pure and non-pure patterns above cross when the remaining factors are equal in Eqs (S4) and (S5). This happens when $p_{k+1} = \frac{1}{2}$. For lower values of p_{k+1}, $(1-p_{k+1}) > p_{k+1}$. This proves the lemma.

6 ILLUSTRATION FROM PROOF OF THEOREM 2.18

In the proof of Theorem 2.18, we apply Lemma 2.17. This Lemma states that many prevalence curves cross at values of p where $p_k = 1/2$. In Figure S3, we confirm this by simulations of the $G^{(m)}(N,p)$ model.

7 PROOF OF THEOREM 2.19

PROOF. Let us refer to the patterns as X_A and X_B. The prevalence of the patterns can be written down explicitly,

$$P(X_A) = \gamma_A p_{k-1}^{x_{k-1}^{(A)}} p_k^{x_k} (1-p_k)^{\binom{m}{k}-x_k} \Omega, \tag{S6}$$

$$P(X_B) = \gamma_B p_{k-1}^{x_{k-1}^{(B)}} p_k^{x_k} (1-p_k)^{\binom{m}{k}-x_k} \Omega, \tag{S7}$$

where $\Omega = \prod_{i=k+1}^{m} (1-p_i)^{\binom{m}{i}}$ and $x_j^{(L)}$ is the number of j-node hyperedges in X_L. By Lemma 2.11, N can be chosen large enough to make the p_{k-1} factor arbitrarily close to 1 for both $P(X_A)$ and $P(X_B)$. Hence, for increasing N,

$$\frac{P(X_A)}{P(X_B)} \to \frac{\gamma_A}{\gamma_B}. \tag{S8}$$

We conclude that the pattern with the smallest combinatorial factor is bound to be less prevalent than the other pattern.