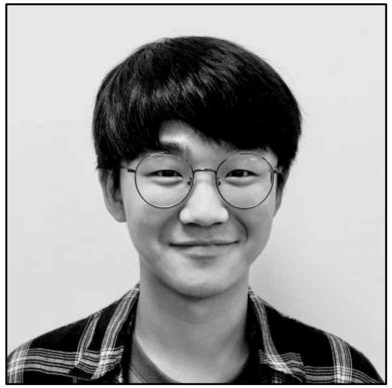
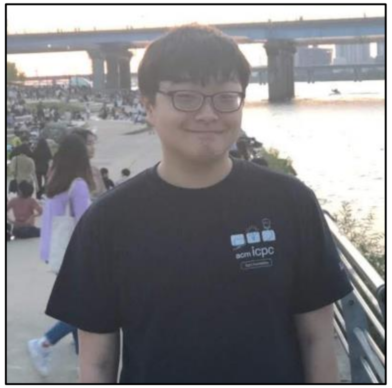
Evolution of Real-world Hypergraphs: Patterns and Models without Oracles



Yunbum Kook

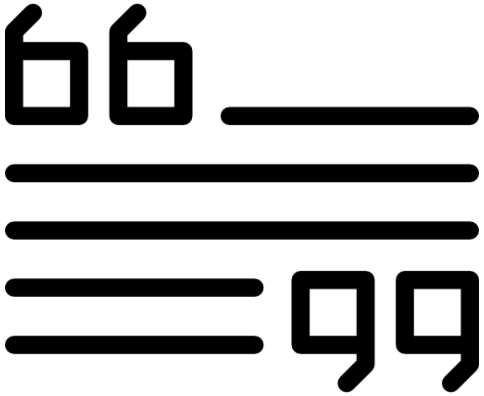


Jihoon Ko

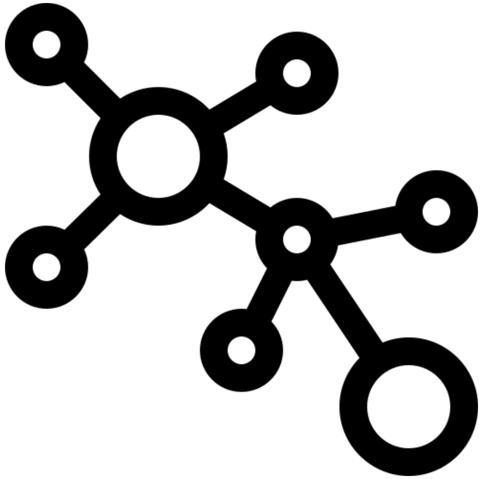
Kijung Shin

Graphs are Everywhere!

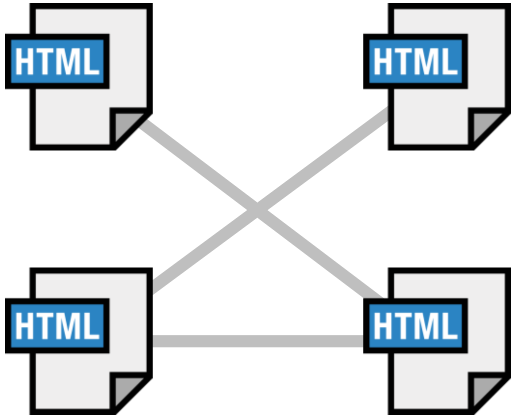
4B+ web pages

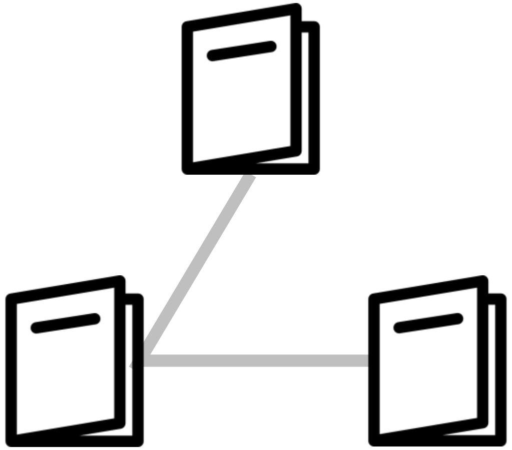


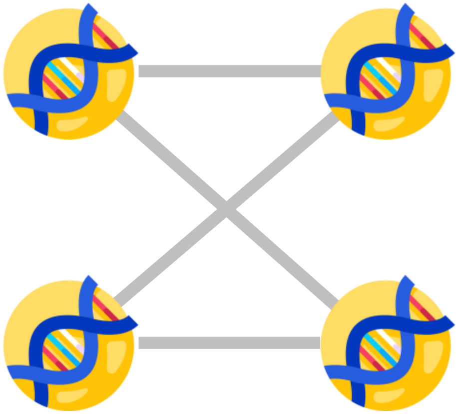
5M papers



6K+ proteins







Icon made by Freepik from www.flaticon.com

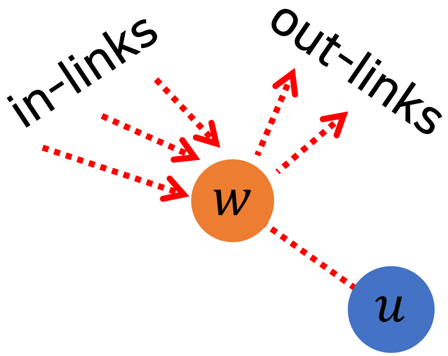
Examples of well-known properties of graphs

- Power-law distributions of spectra and degrees [TM77]
- Densification [LKF07]
- Shrinking diameter over time [LKF07]
- Triadic closure [HTWLF14]
- Temporal Locality [S17]

→ These properties are useful in design and analysis of graph algorithms [CRS12, GS12, KKS20]

Models proposed for the patterns

- Preferential Attachment [BA99]
- Forest Fire [LKF07]



- Kronecker graphs [LCKFG10]

0.8	0.6
0.6	0.2

0.64	0.48	0.48	0.36
0.48	0.16	0.36	0.12
0.48	0.36	0.16	0.12
0.36	0.12	0.12	0.04

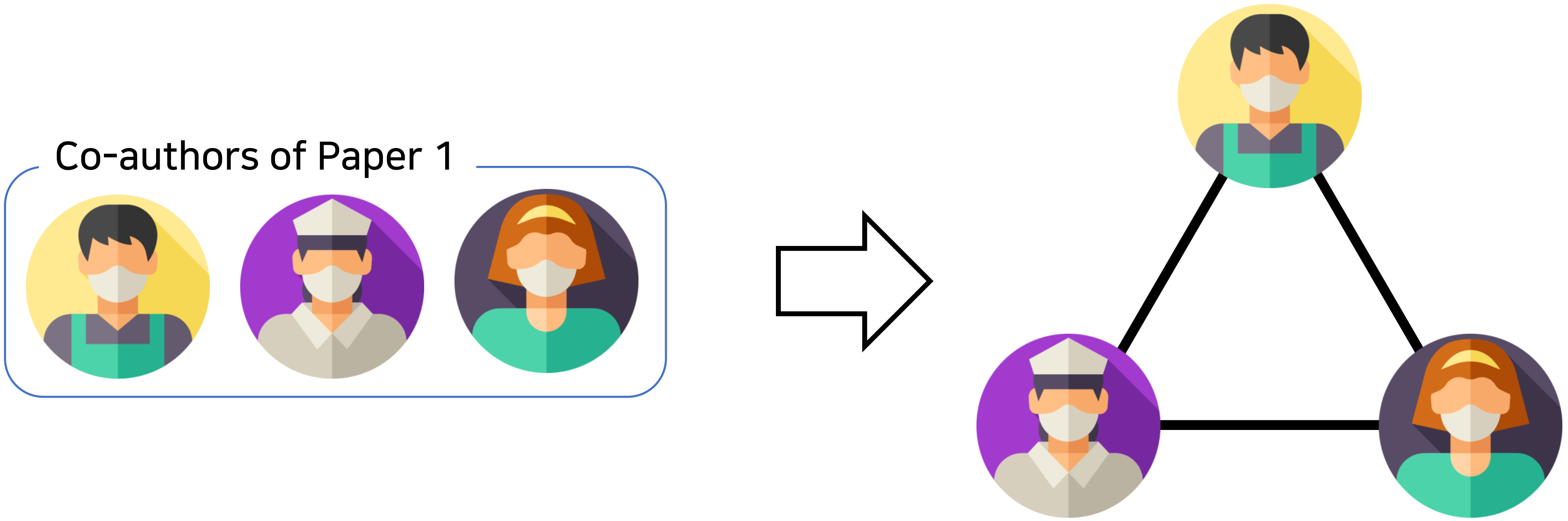
Not all datasets are through "Graph"

- Example: Co-authorship

Q. How to represent it by using a graph?

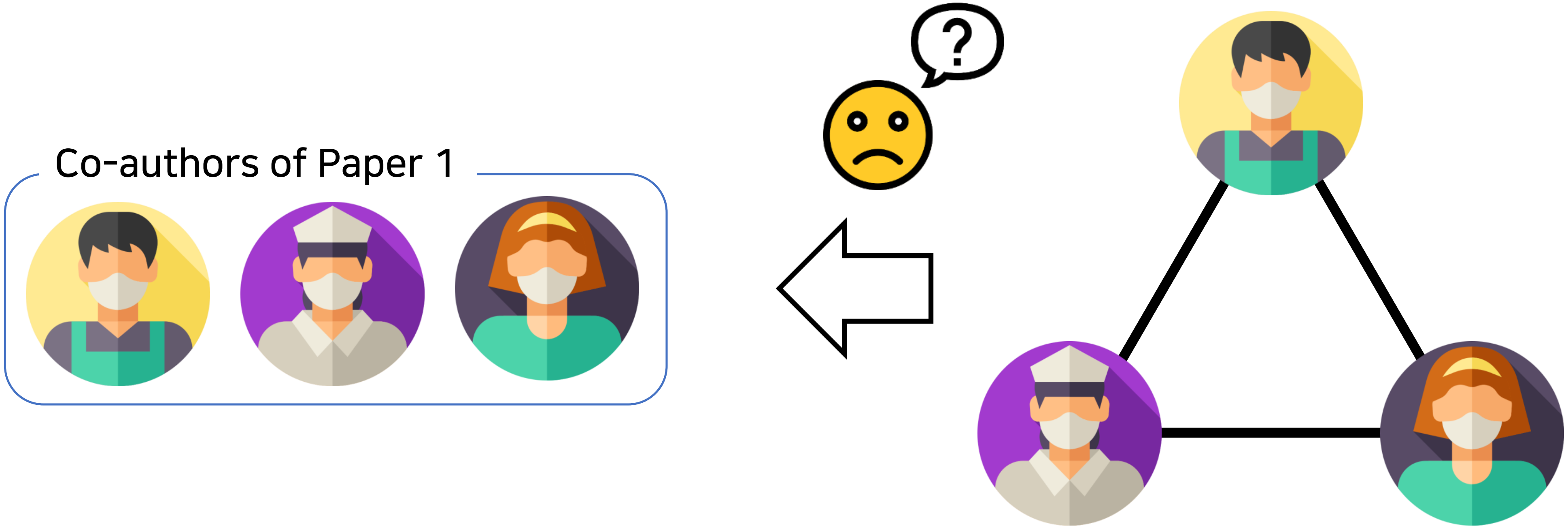
Not all datasets are through "Graph" (cont.)

- Simple reduction to pairwise interaction



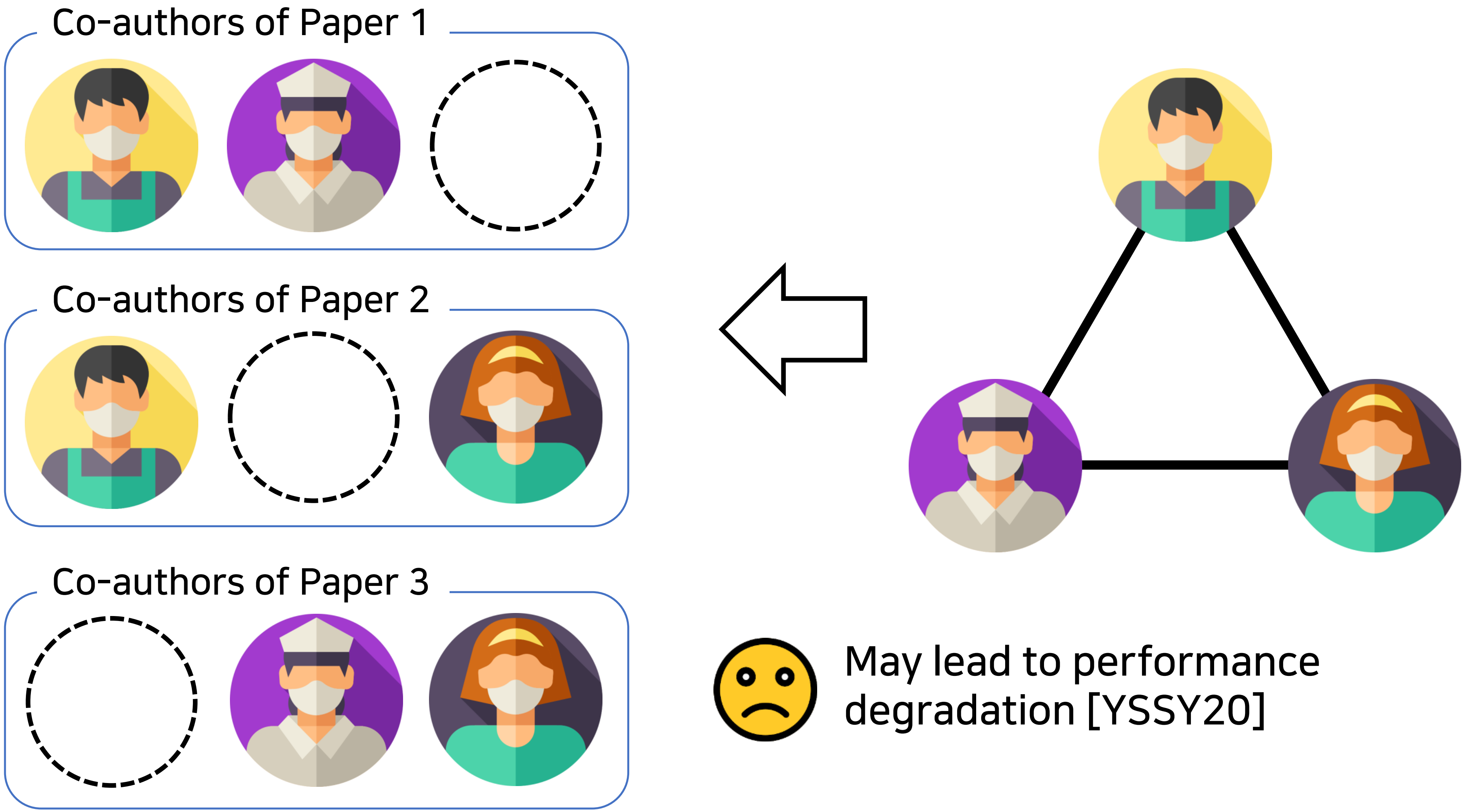
Not all datasets are through "Graph" (cont.)

- Simple reduction to pairwise interaction



Not all datasets are through "Graph" (cont.)

- Simple reduction to pairwise interaction

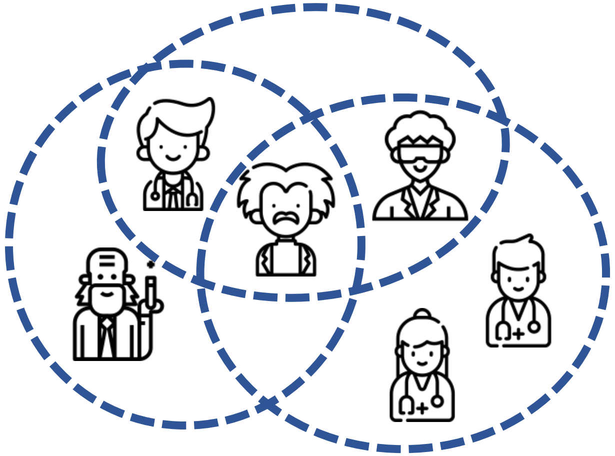


May lead to performance degradation [YSSY20]

Hypergraph: natural extension of graph

- Hypergraphs consist of nodes and hyperedges.
- Each hyperedge is a subset of any number of nodes.

Co-purchases of Items



Collaborations of Researchers

Using high-order information is helpful

- The harder tasks are, the more useful utilizing information on high-order interactions is. [YSSY20]

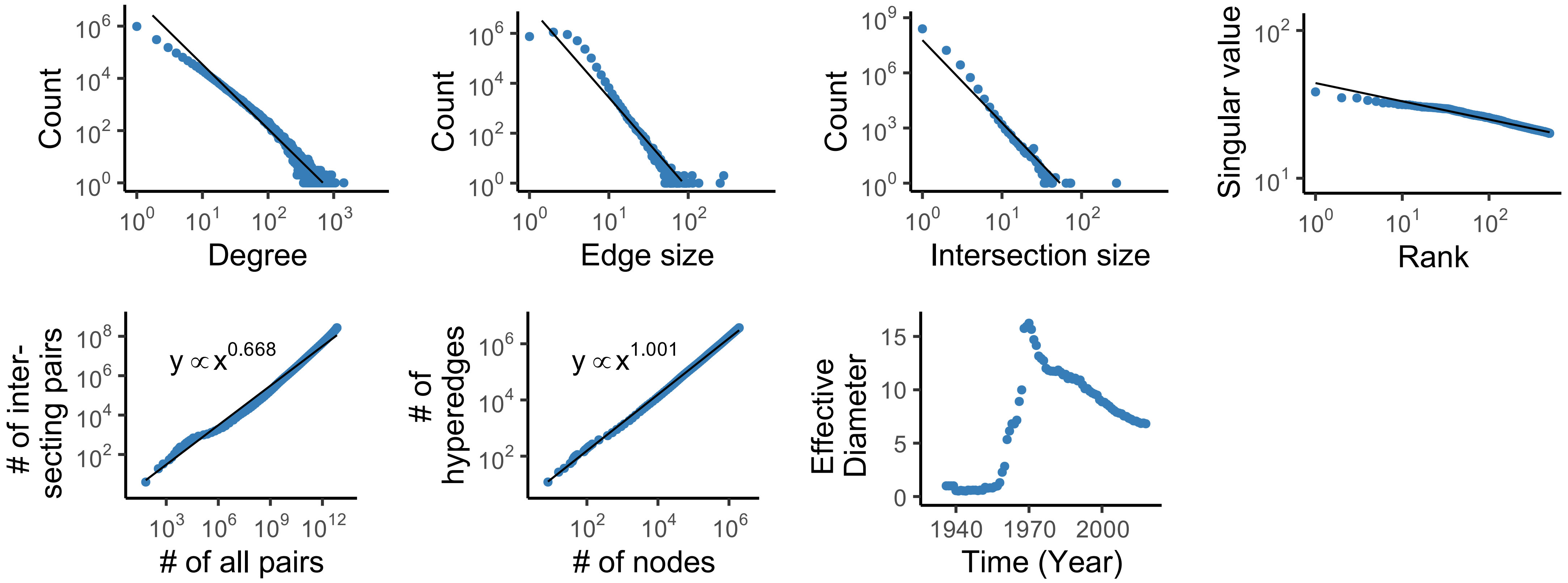
- Hypergraphs are considered in many domains.
  - Computer Vision [HLM09]
  - Recommendation [LL13]
  - Graph Learning [FYZJG19]

Previous study

- Our understanding of real-world hypergraphs is not as thorough as that of graphs.
- Hard to investigate them.
  - Daunting complexity arises from various edge size.
- As an alternative, these are reduced to pairwise interactions and then are investigated.

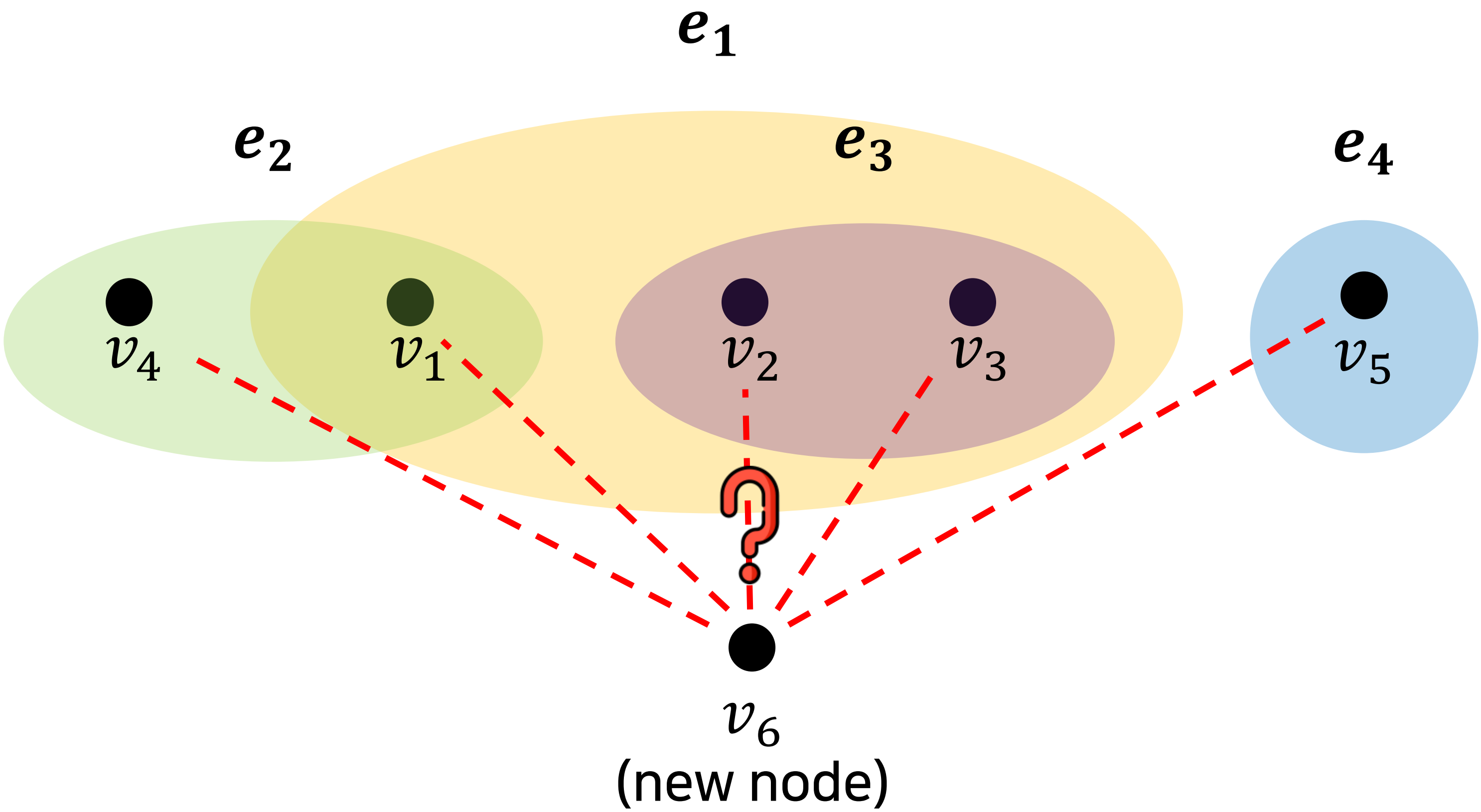
Our Questions

- **Q1.** What kind of macroscopic structural and dynamical patterns can we observe in real-world hypergraphs?



Our Questions

- **Q2.** What can be underlying local dynamics on individuals, which ultimately lead to the observed patterns, beyond apparently random evolution?

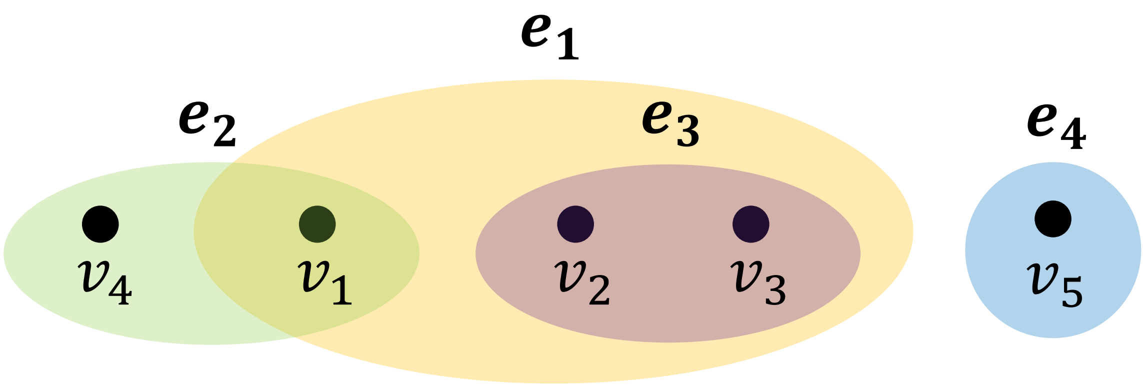


Outline

- Preliminaries
- Macroscopic patterns
  - Static patterns
  - Temporal patterns
- HyperFF: random hypergraph model
- Summary

Preliminaries

- Hypergraph $G = (V, E)$
  - a set V of nodes
  - a set $E \subseteq 2^V$ of hyperedges



Example

Size of e_1: 3
Degree of v_1: 2

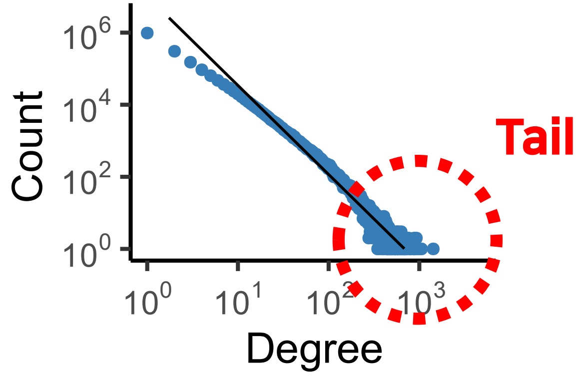
Neighborhood of v_1:
$\{v_2, v_3, v_4\}$

- Incidence Matrix $I \in \{0,1\}^{|V| \times |E|}$
  - Indicates the membership of the nodes V in the hyperedge E
  - Each (i, j)-th entry I_{ij} of I is 1 iff j-th hyperedge in E contains the i-th node in V.

	e_1	e_2	e_3	e_4
v_1	1	1	0	0
v_2	1	0	1	0
v_3	1	0	1	0
v_4	0	1	0	0
v_5	0	0	0	1

Preliminaries

- Heavy-tailed distribution
  - Tails of heavy-tailed distributions decay slower than exponential distributions
  - Typical example: Power-law distribution
    - $P(X) \propto 1/|X|^{\alpha}$ for some constant $\alpha > 0$
    - Straight line on the log-log plot of the probability distribution



- Goodness of Fit
  - It is difficult to argue that an empirical dataset genuinely follows a target distribution
  - Utilize the log likelihood-ratio test to compare two or more candidate distributions

Datasets

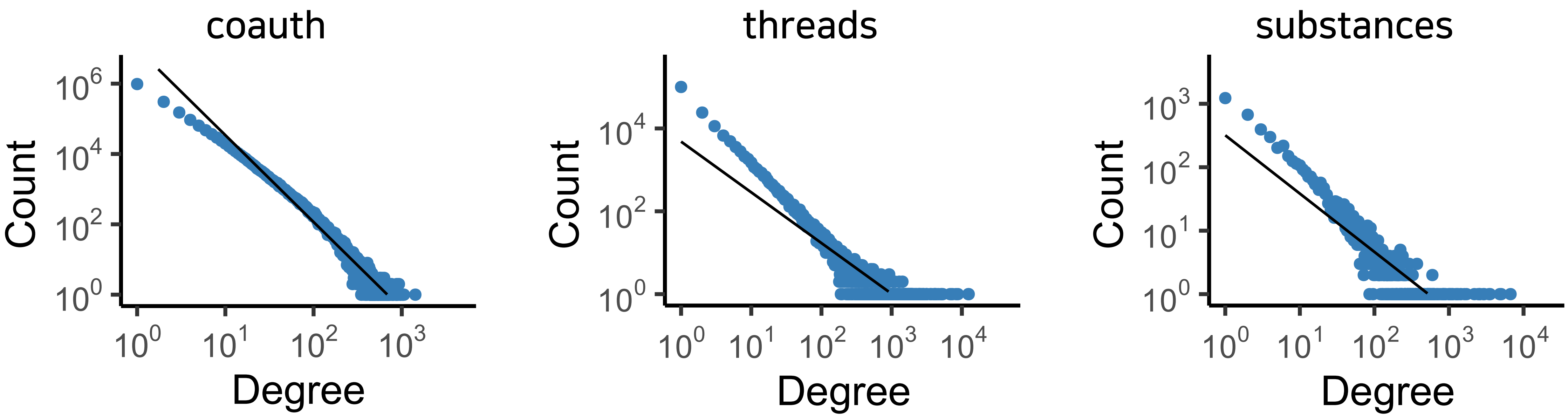
Dataset name	Number of nodes	Number of hyperedges	Brief Summary
contact	327	172,035	Social Interaction
email	1,005	235,263	Email
tags	3,029	271,233	Q&A
substances	5,556	112,919	Drug
threads	176,445	719,792	Q&A
coauth	1,924,991	3,700,067	Co-authorship

Outline

- Preliminary
- Macroscopic patterns
  - Static patterns
  - Temporal patterns
- HyperFF: random hypergraph model
- Summary

Static patterns

- Pattern 1. Heavy-tailed degree distribution

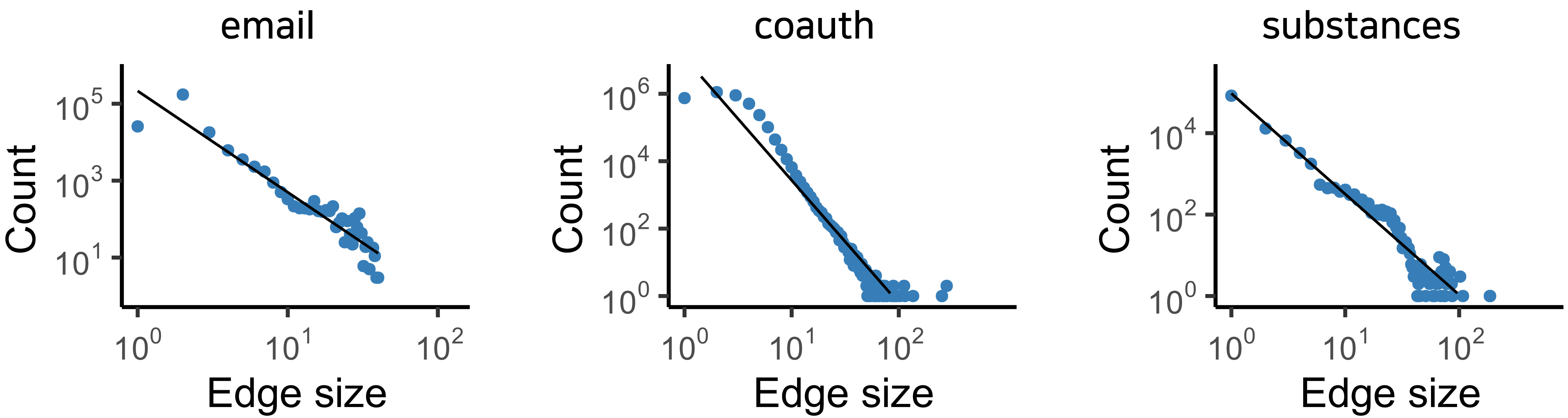


Notation

Degree of a node v is the number of hyperedges containing v

Static patterns

- Pattern 2. Heavy-tailed hyperedge size distribution

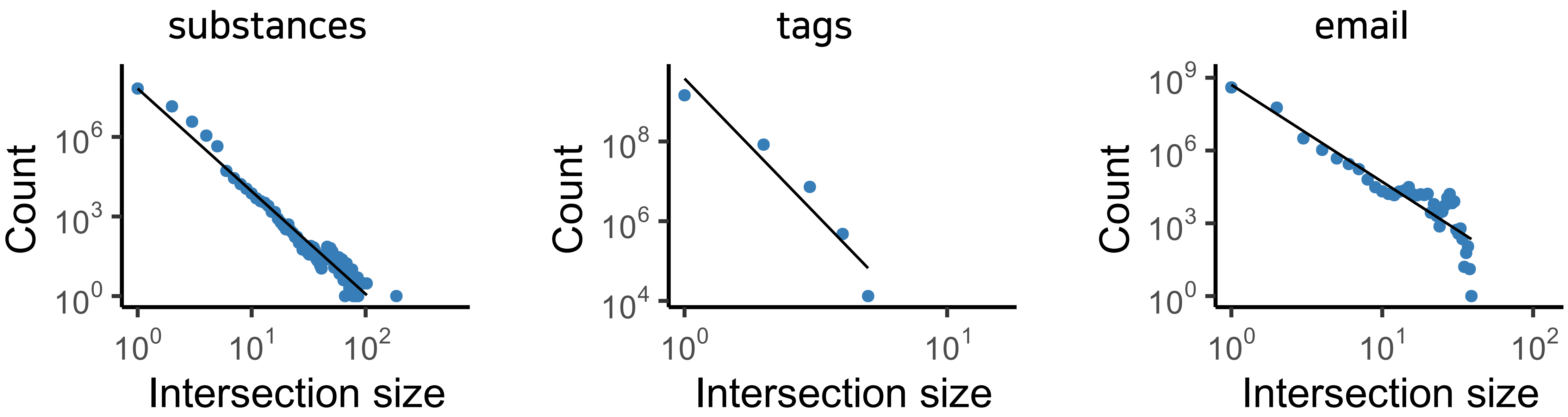


Notation

Size of a hyperedge e is the number of nodes in the hyperedge e

Static patterns

- Pattern 3. Heavy-tailed intersection size

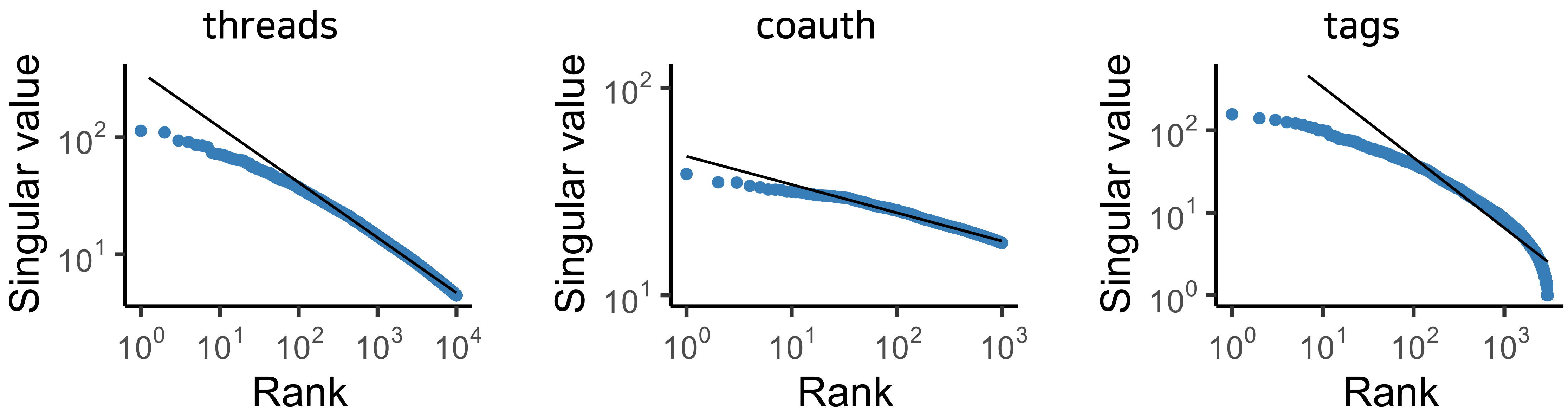


Notation

The intersection size of two hyperedges is the number of nodes commonly contained in two hyperedges

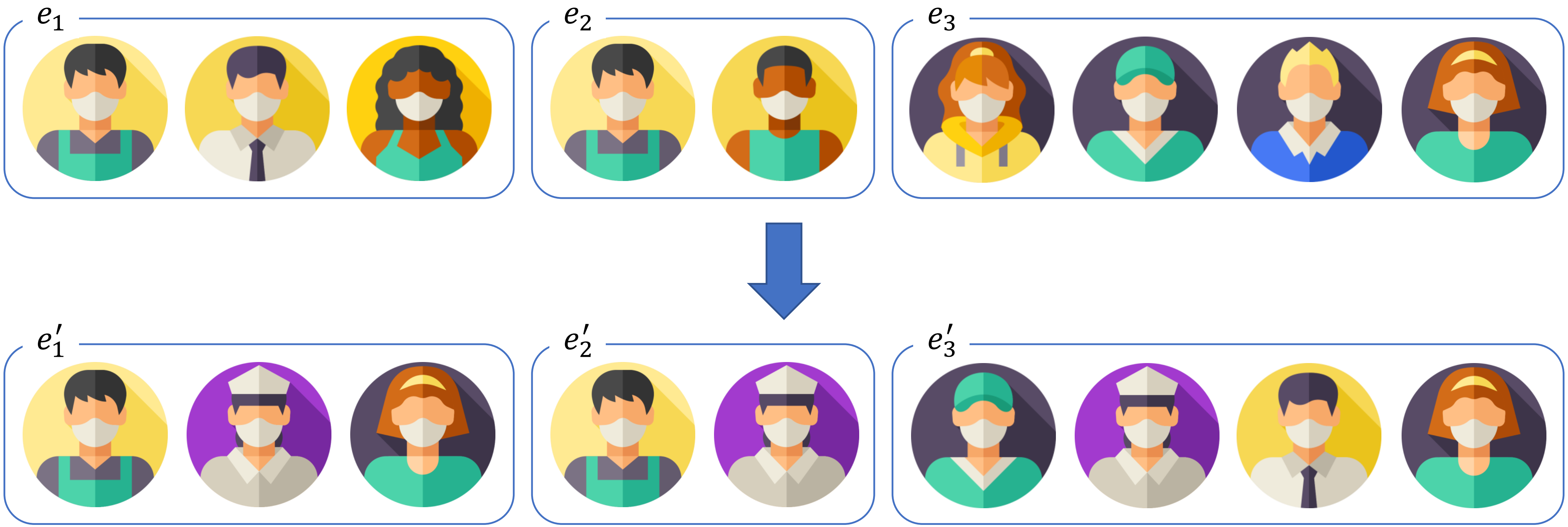
Static patterns

- Pattern 4. Skewed singular values of incidence matrix



Comparison with Null model

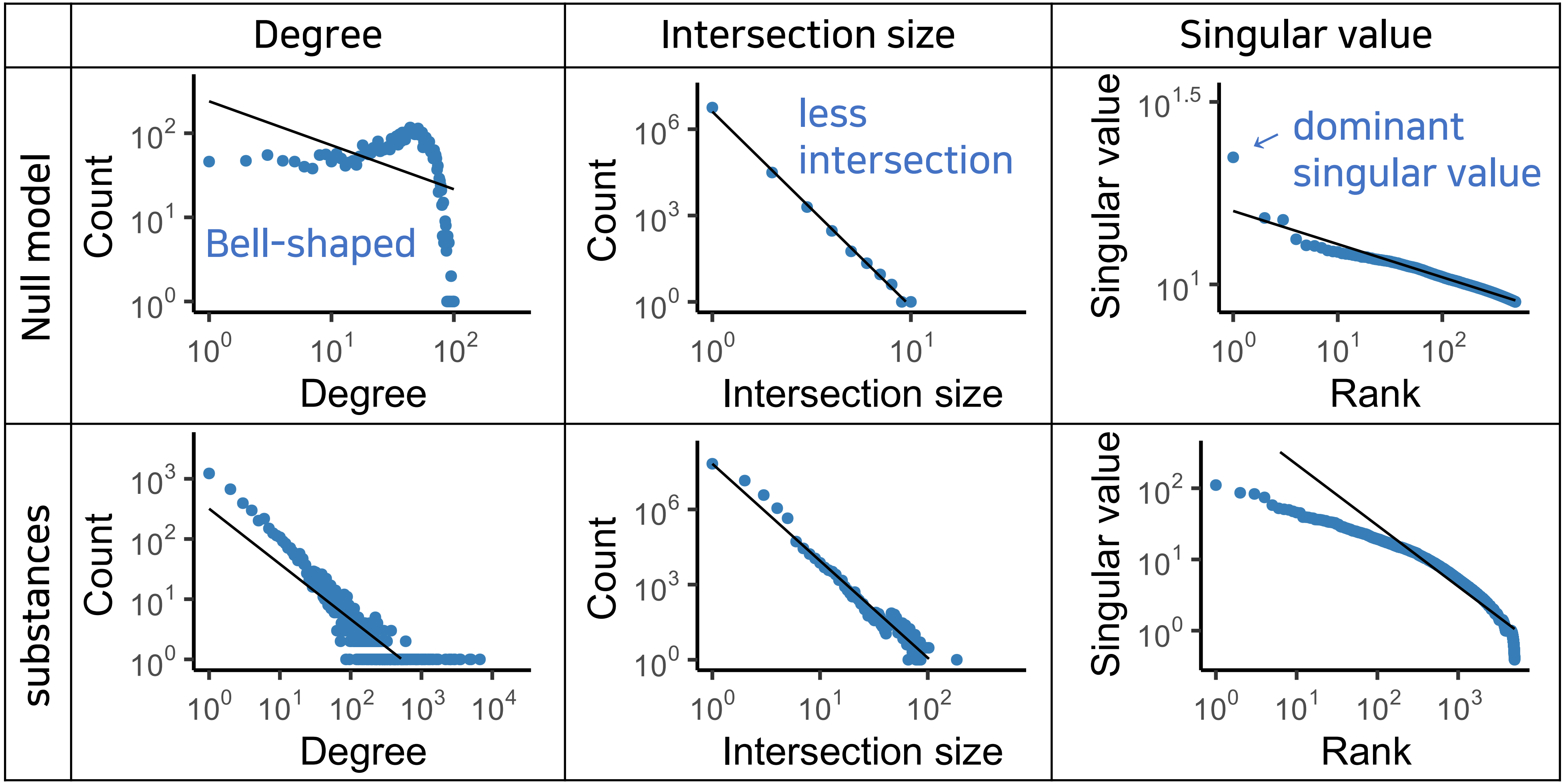
A real sequence of a set of hyperedges



A sequence generated by null model

Comparison with Null model

- Null model generated from *substance* dataset

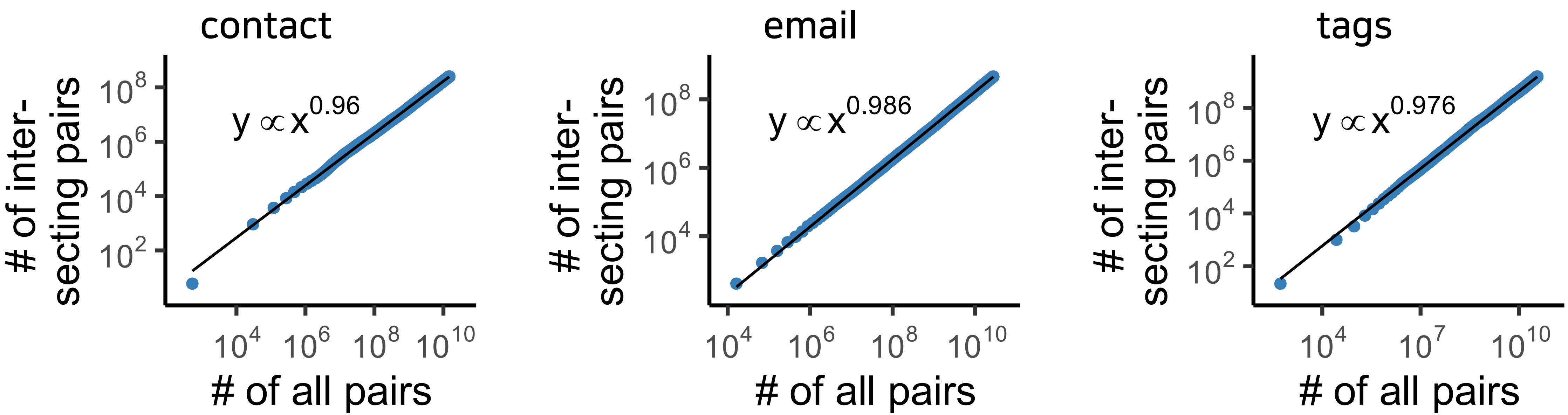


Outline

- Preliminary
- Macroscopic patterns
  - Static patterns
  - Temporal patterns
- HyperFF: random hypergraph model
- Summary

Temporal Patterns

- Pattern 1. Diminishing overlaps

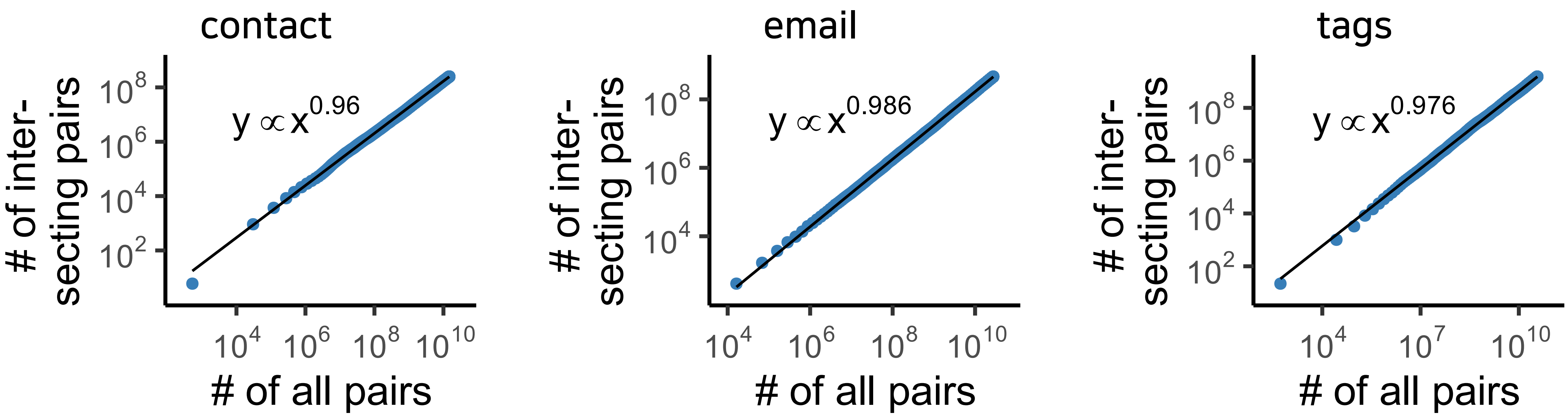


Notation

For a hypergraph $G_t = (V_t, E_t)$ at time t, the number of intersecting pairs at time t is computed as $\left|\{\{e_i, e_j\} \middle| e_i, e_j \in E_t, e_i \cap e_j \neq \phi\}\right|$, and density of interactions (DoI) is the ratio of the number of intersecting pairs to the number of all possible pairs

Temporal Patterns

- Pattern 1. Diminishing overlaps



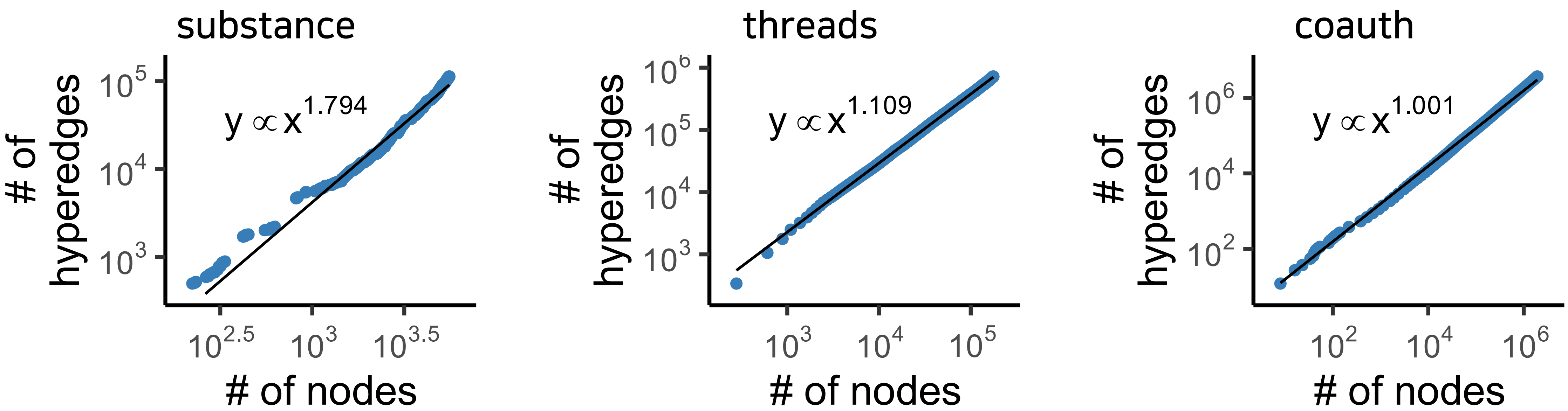
Notation

$y(t)$ = # of intersecting pairs at time t
$x(t)$ = # of all possible pairs at time t
s = the slope of fitted lines on the log-log plot of $y(t)$ over $x(t)$

$$\Rightarrow \quad \mathrm{DoI}(t) \propto x(t)^{-(1-s)}$$

Temporal Patterns

- Pattern 2. Densification



Notation

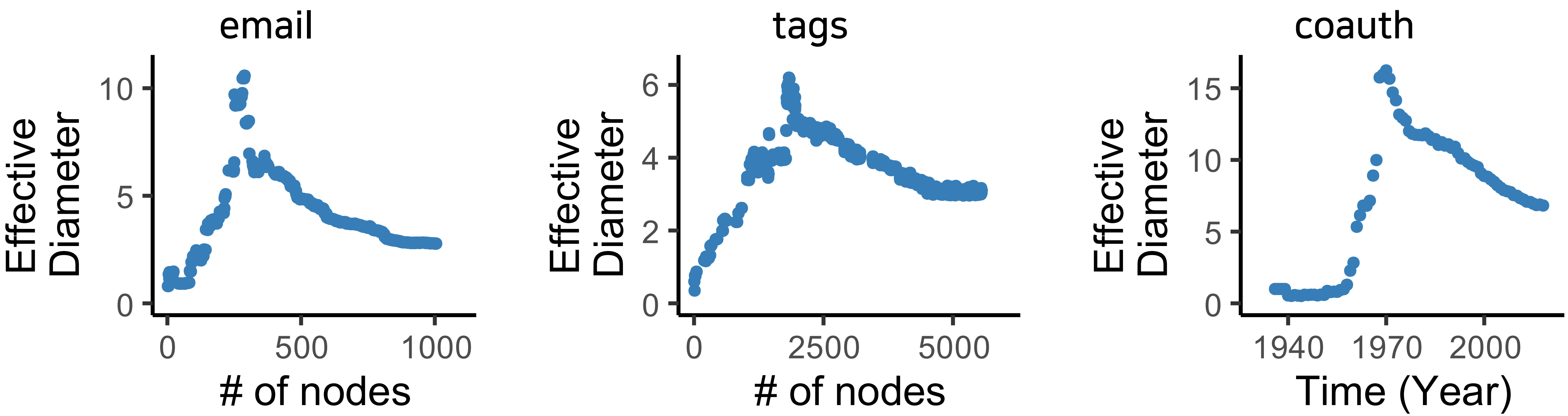
$\mathrm{E(t)}$ = # of hyperedges at time t
$\mathrm{V(t)}$ = # of nodes at time t
s = the slope of fitted lines on the log-log plot of $\mathrm{E(t)}$ over $\mathrm{V(t)}$

$\Rightarrow$

$\mathrm{E(t)} \propto \mathrm{V(t)}^s$
(or equivalently)
Average degree $\propto \mathrm{V(t)}^{s-1}$

Temporal Patterns

- Pattern 3. Shrinking diameter



Notation

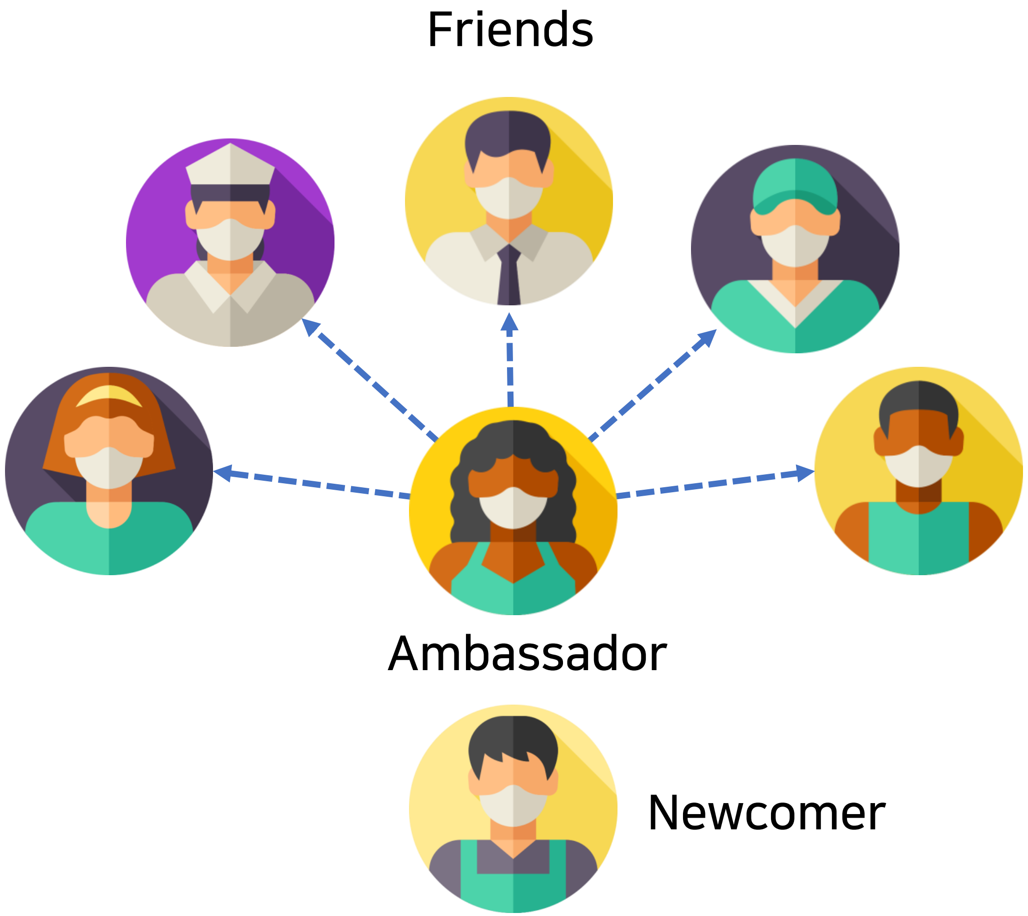
Effective diameter is the smallest d such that the paths of length at most d connect 90% of all reachable pairs of nodes

Outline

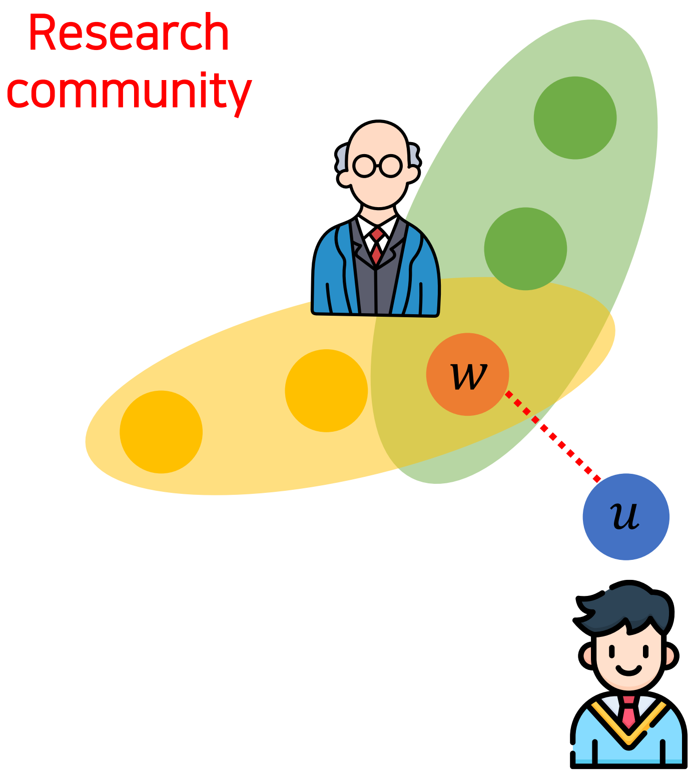
- Preliminary
- Macroscopic patterns
    - Static patterns
    - Temporal patterns
- HyperFF: random hypergraph model
- Summary

Proposed model: Hypergraph Forest Fire

- How do we meet friends at a party?
- How do we identify references when writing papers?

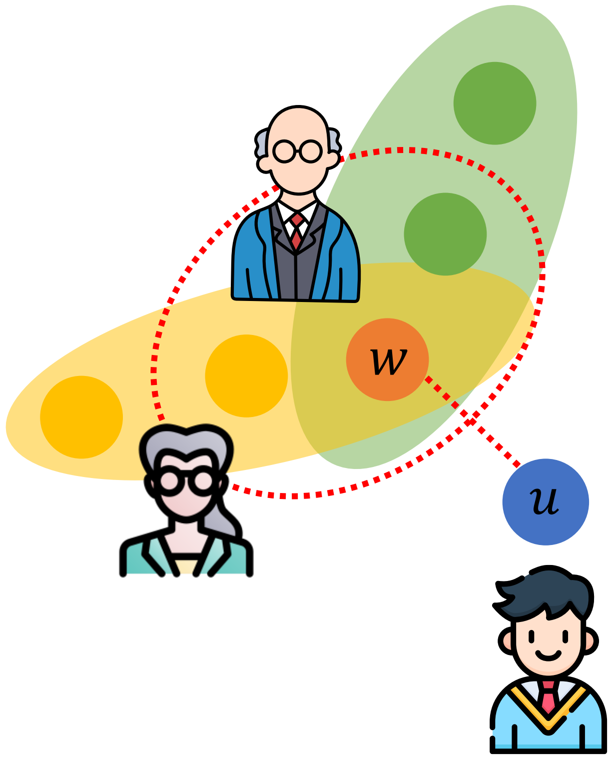


Proposed model: Hypergraph Forest Fire



1) The new node u chooses a random ambassador w from the hypergraph so far and burns the ambassador

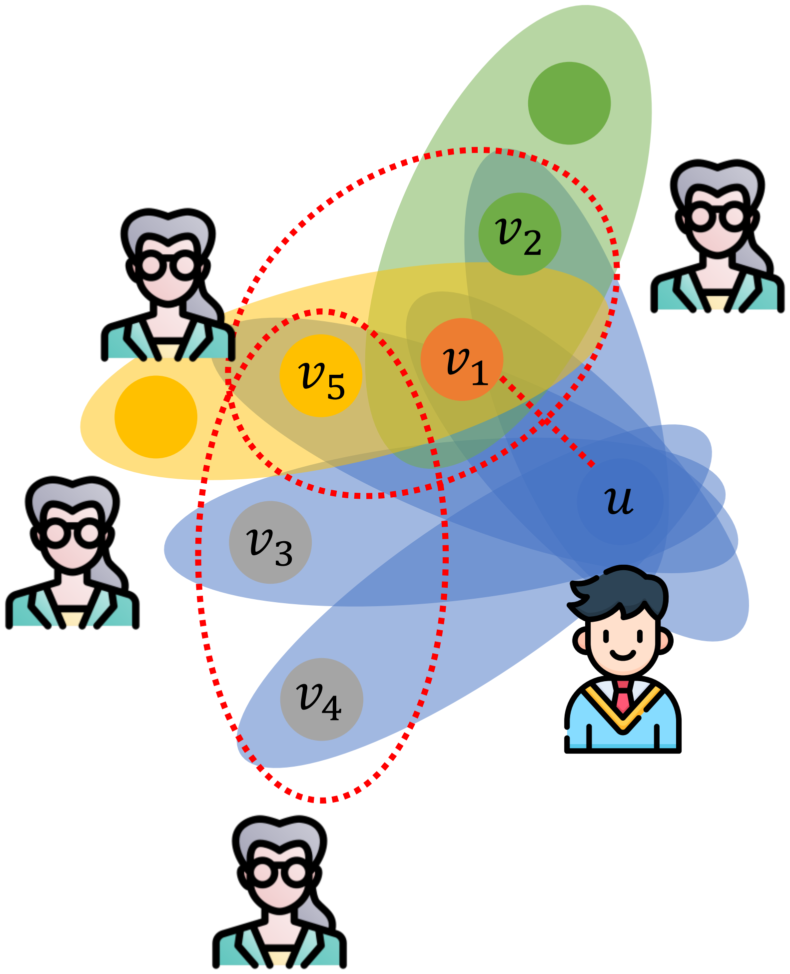
Proposed model: Hypergraph Forest Fire



1) The new node u chooses a random ambassador w from the hypergraph so far and burns the ambassador

2) Burn n neighbors of the ambassador w in descending order of 'tie strength', where n is sampled from the geometric distribution with mean $\frac{p}{1-p}$.
(p: burning probability)

Proposed model: Hypergraph Forest Fire

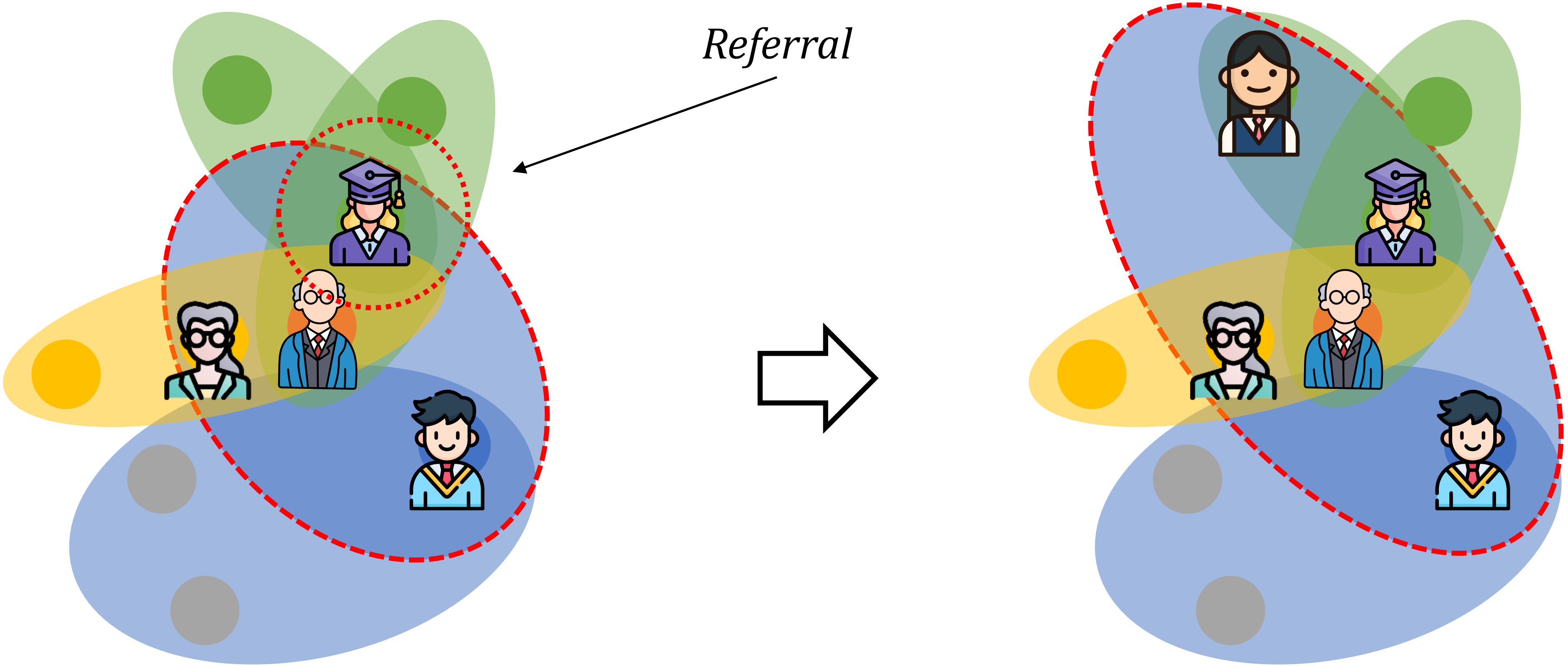


3) Recursively apply (2) to each burned neighbor by viewing a burned neighbor as a new ambassador of the new node u.

4) For each burned node v, form a hyperedge $\{u, v\}$ and increase 'tie strength' of $\{u, v\}$ by 1.

What is a reasonable way to expand each size-2 hyperedge?

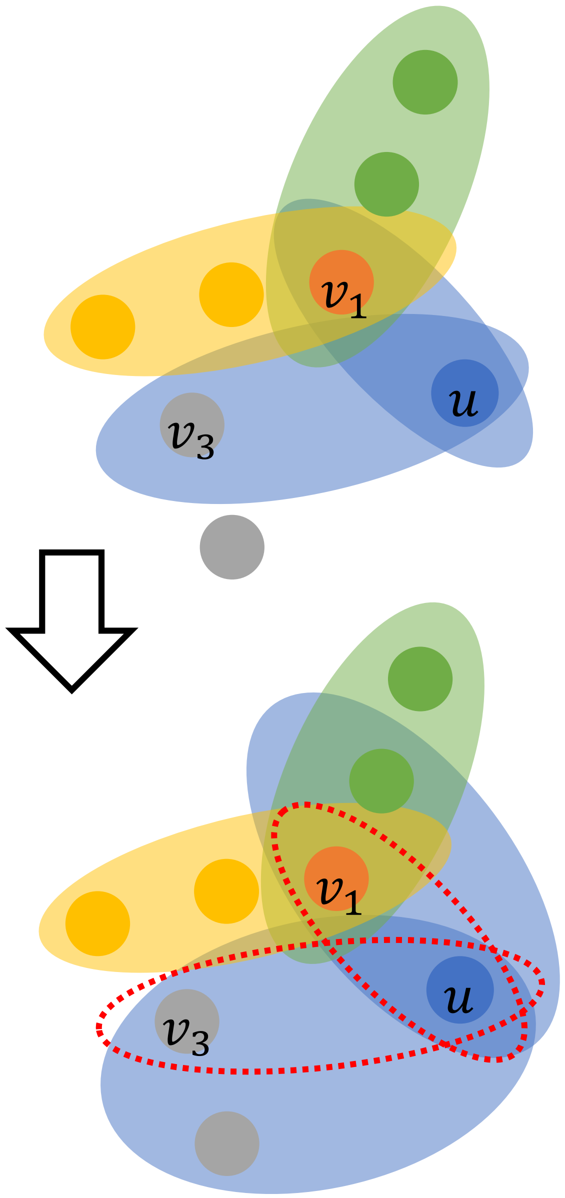
Proposed model: Hypergraph Forest Fire



Research group looking for a new researcher to work with…

Bring the new!

Proposed model: Hypergraph Forest Fire

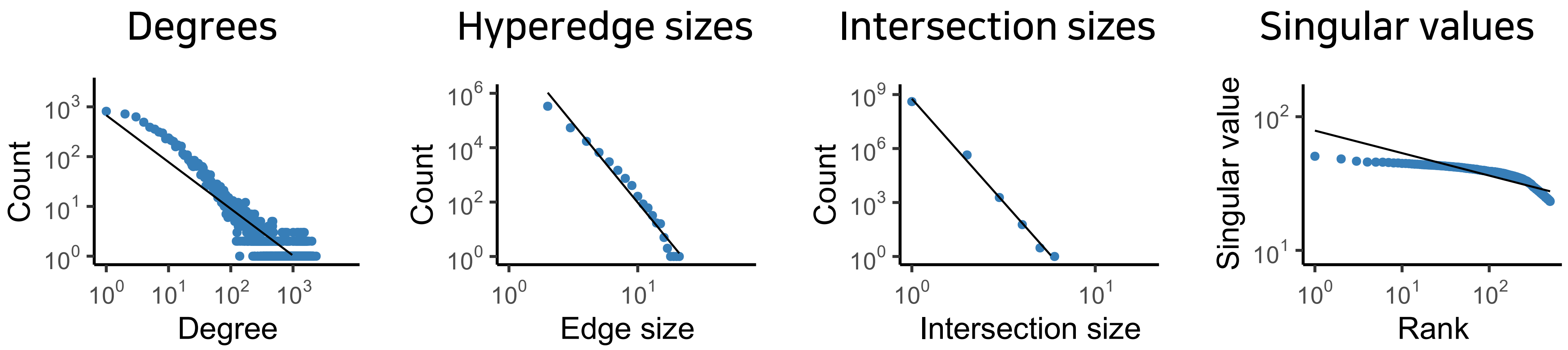


5) For each hyperedge created in (4), reset the burning history and start the burning process at the burned node v in which we use the geometric distribution with mean $\frac{q}{1-q}$, expand the hyperedge until the process ends.

(q: expanding probability)

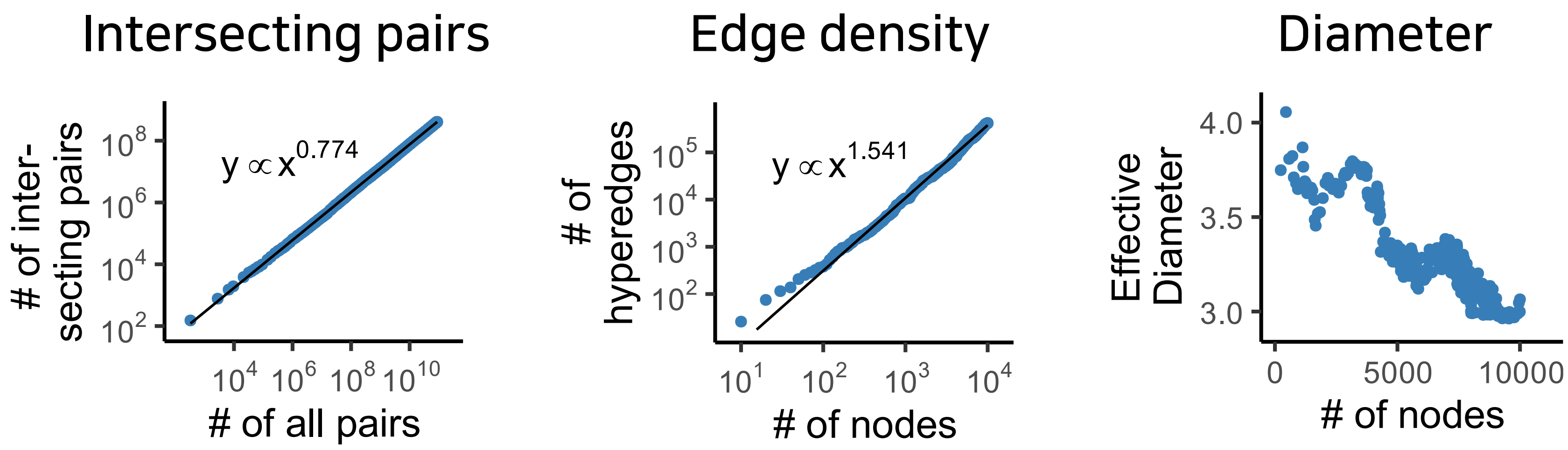
Generated hypergraph using HyperFF

- Empirical Static Patterns with $(p, q) = (0.51, 0.2)$



Generated hypergraph using HyperFF

- Empirical Dynamical Patterns with $(p, q) = (0.51, 0.2)$



Proposed model: Hypergraph Forest Fire

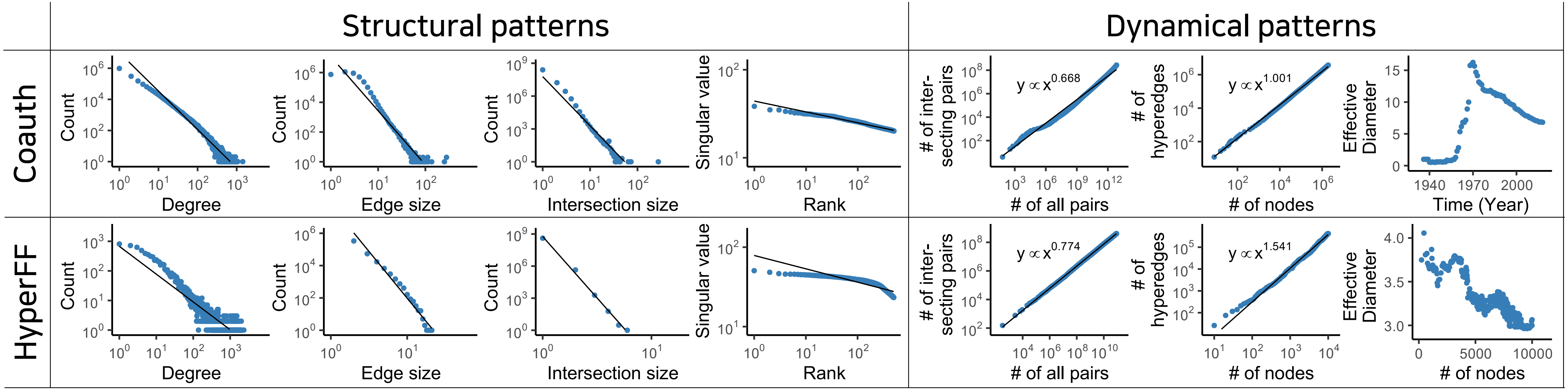
- HyperFF reproduces all examined patterns without relying on exter nal information!
- The mechanisms on individual nodes are simple and intuitive!
  - It has only two scalars as parameter
  - It does not directly impose but eventually lead to the examined patterns

Outline

- Preliminary
- Macroscopic patterns
  - Static patterns
  - Temporal patterns
- HyperFF: random hypergraph model
- Summary

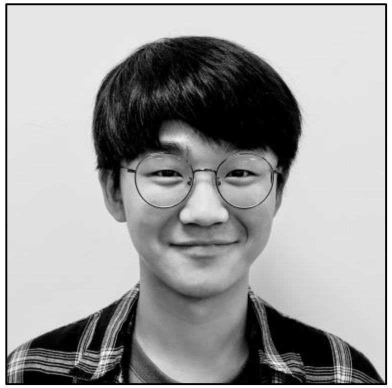
Summary

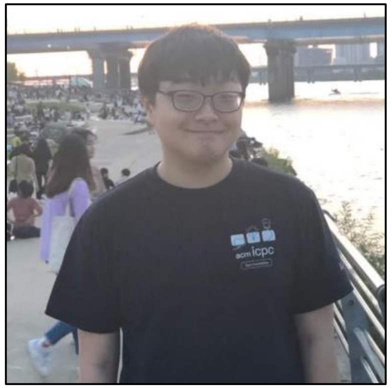
- Structural and dynamical patterns of real-world hypergraphs
- Propose a generative model HyperFF



The code and datasets used in the paper are available at
https://github.com/yunbum-kook/icdm20-hyperff

Evolution of Real-world Hypergraphs: Patterns and Models without Oracles





Yunbum Kook

Jihoon Ko

Kijung Shin