Sequences of sets

Austin R. Benson, Ravi Kumar, and Andrew Tomkins

arb@cs.cornell.edu, ravi.k53@gmail.com, atomkins@gmail.com

Code & data ⟶ https://github.com/arbenson/Sequences-of-Sets

Talk on Thursday in RT17: Methodology I (10am-12n).

Sequential behavior and sets.

What are *sequences of sets*?

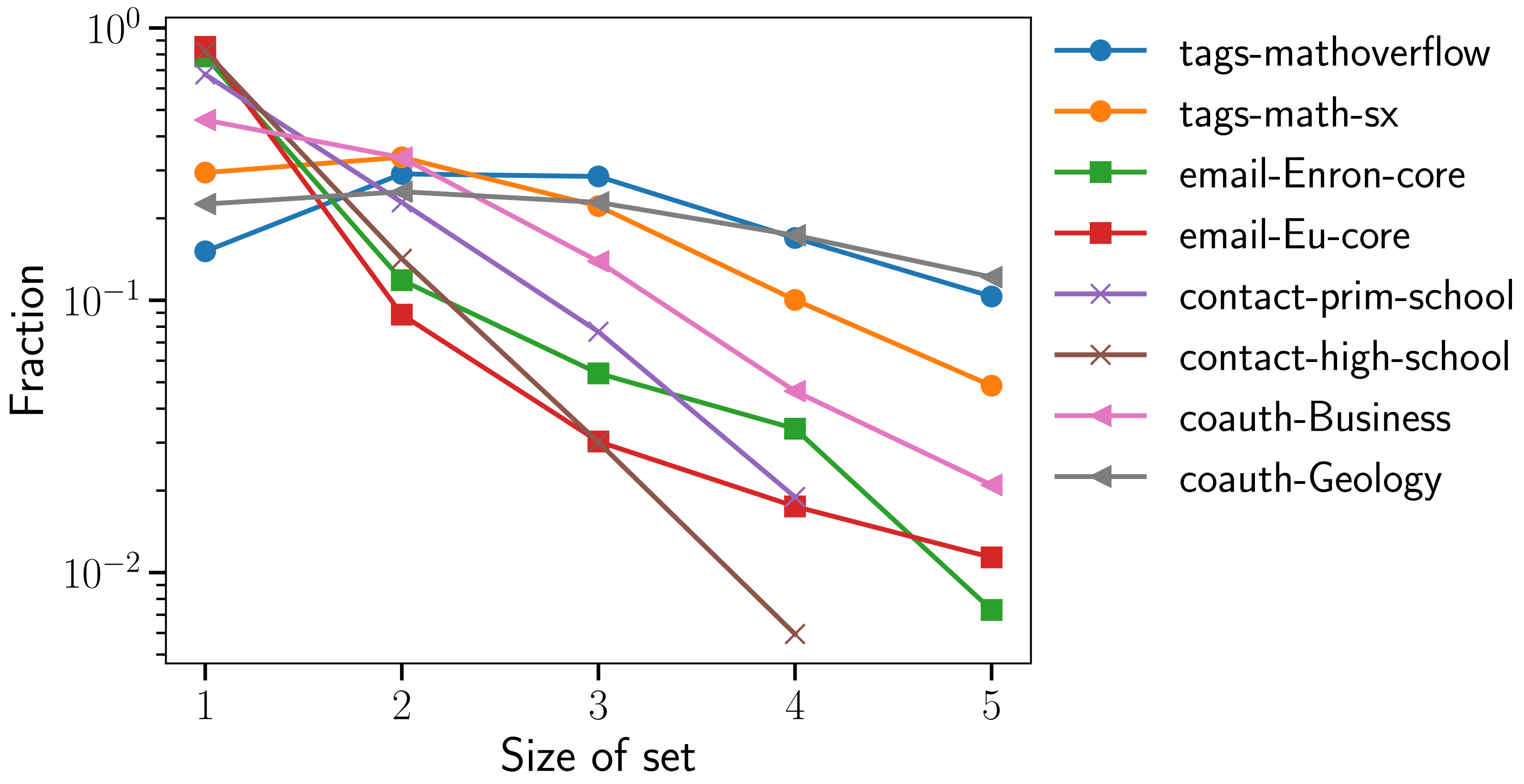
- Sets of reicipients on your emails over time.
- Sets of coauthors on your papers over time.
- Sets of tags that you apply to posts on Stack Overflow over time.
- Sets of friends that you meet for dinner over time.

These sequences of sets show complex repetition behavior, sometimes repeating prior sets wholesale, and sometimes creating new sets from partial copies or partial merges of earlier sets.

We provide a stochastic model to capture these patterns.

Datasets that are sequences of sets.

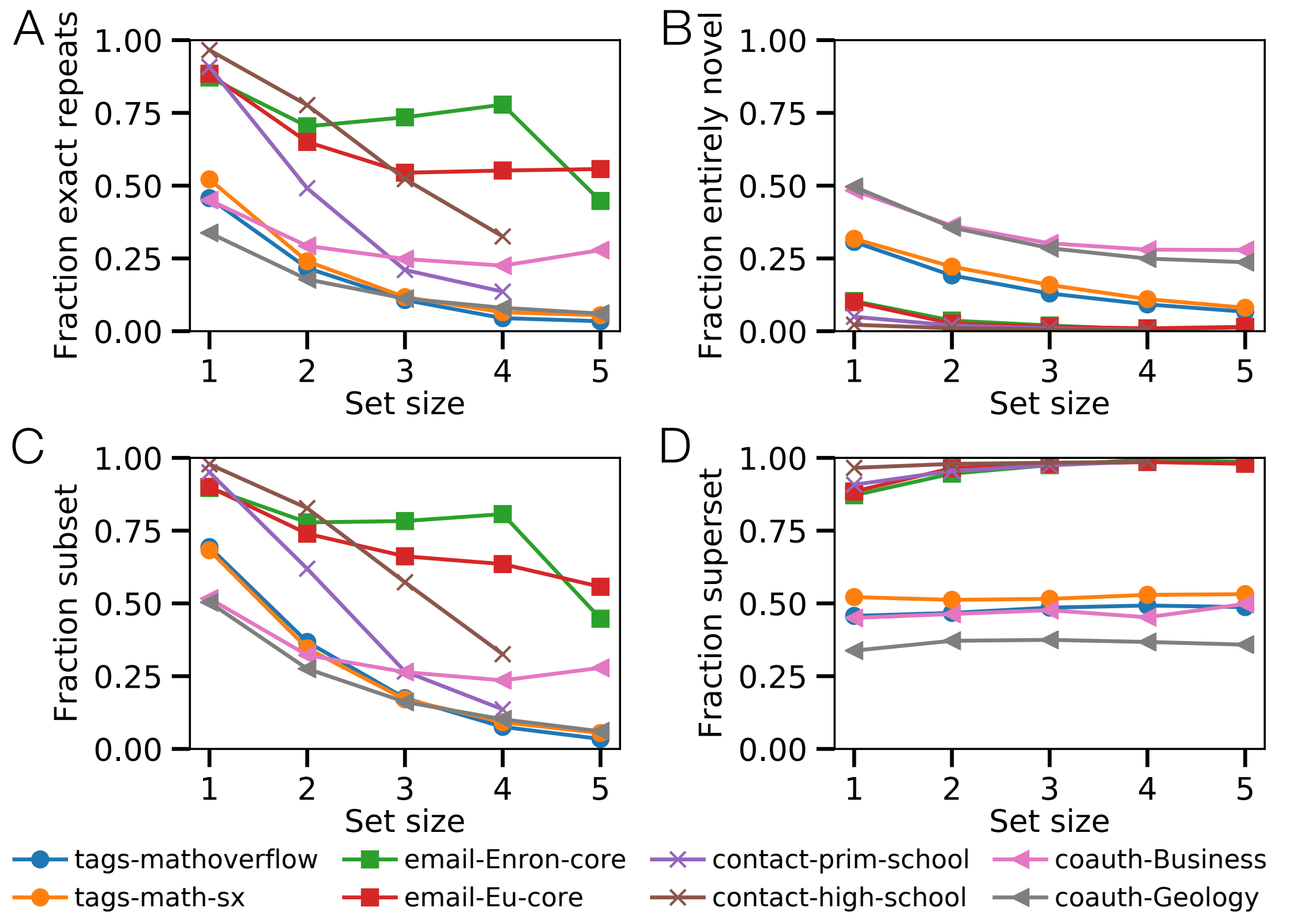
- **tags-math-overflow & tags-math-sx.**
  Sequences of tags applied by users to questions on Stack Exchange
- **email-Enron-core & email-Eu-core**
  Sequences of recipients on emails sent to individuals
- **contact-prim-school & contact-high-school**
  Sequences of groups that an individual interactions with
- **coauth-Business & coauth-Geology**
  Sequences of coauthors of individuals in published academic papers



Acknowledgements.
ARB supported in part by a Simons Investigator Award and NSF NSF TRIPODS Award #1740822.

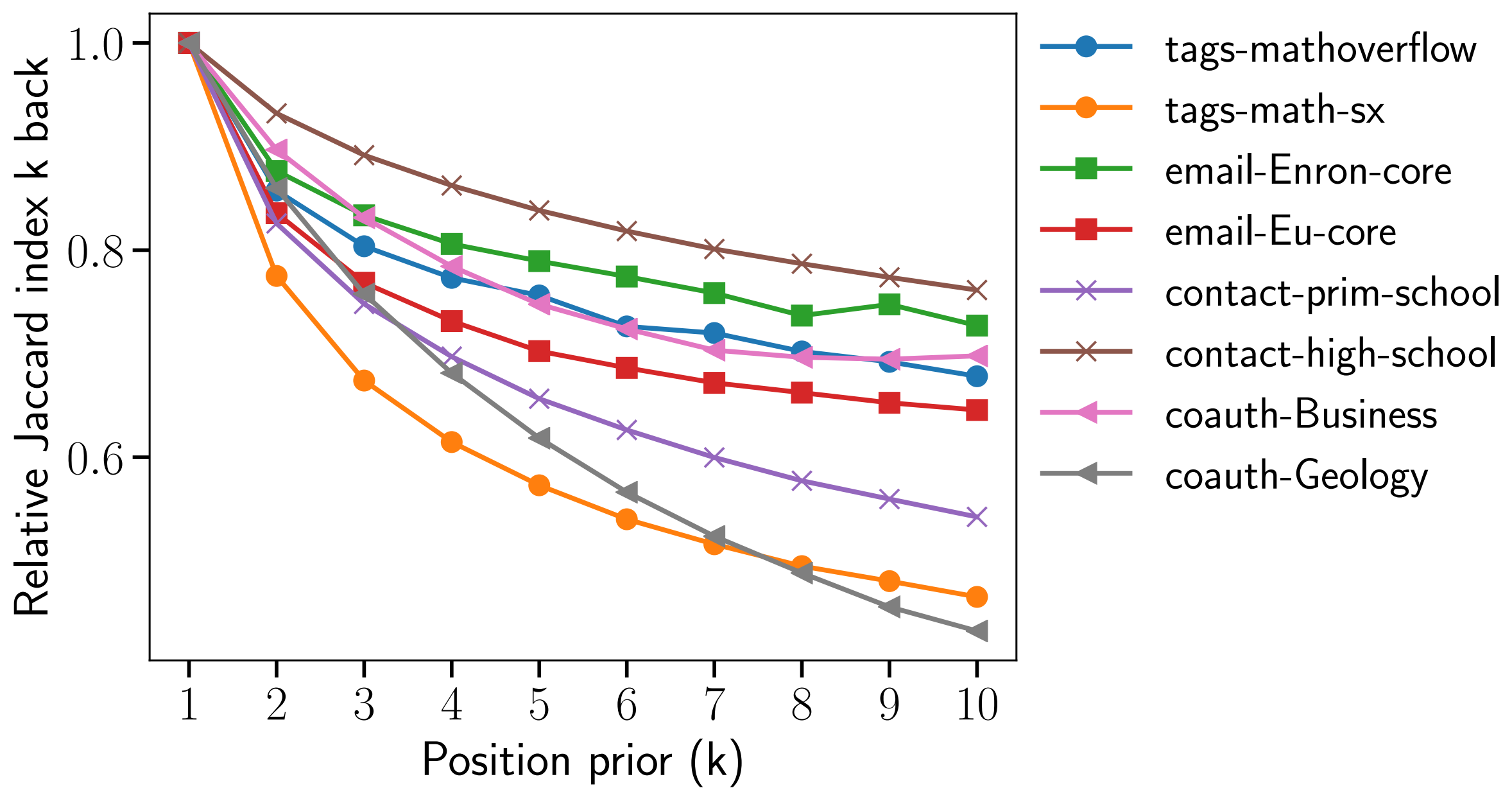
Data analysis and common trends.

Repeats are common.



(A) Fraction of sets that are exact repeats of a previous set. (B) Fraction of sets that are made up of completely novel items that have not appeared earlier in the sequence. (C) Fraction of sets that are (not necessarily proper) subsets of a previous set in the sequence. (D) Fraction of sets that are (not necessarily proper) supersets of a previous set in the sequence.

There is recency bias.



Average Jaccard index of a set in a sequence with the set appearing k steps prior in the sequence, normalized to $k = 1$.

In all datasets, similarity is higher when k is small.

Similarity is roughly monotonically decreasing in k.

There are subset correlations.

	size-2 subset counts		size-3 subset counts	
Dataset	data	null model	data	null model
email-Enron-core	5.82	4.34 ± 0.043	4.23	2.67 ± 0.038
email-Eu-core	4.46	3.11 ± 0.008	3.23	2.08 ± 0.007
contact-prim-school	2.36	1.87 ± 0.003	1.35	1.09 ± 0.002
contact-high-school	4.49	3.26 ± 0.007	2.09	1.35 ± 0.004
tags-mathoverflow	1.49	1.41 ± 0.002	1.18	1.15 ± 0.002
tags-math-sx	1.49	1.31 ± 0.001	1.21	1.12 ± 0.001
coauth-Business	1.50	1.30 ± 0.001	1.40	1.24 ± 0.001
coauth-Geology	1.29	1.15 ± 0.000	1.15	1.07 ± 0.000

For each sequence in each dataset, we count the number of times each size-2 and size-3 subset appears.

We perform the same counts under a null model where elements are randomly placed into sets (reported mean +/- one s.d., 100 trials).

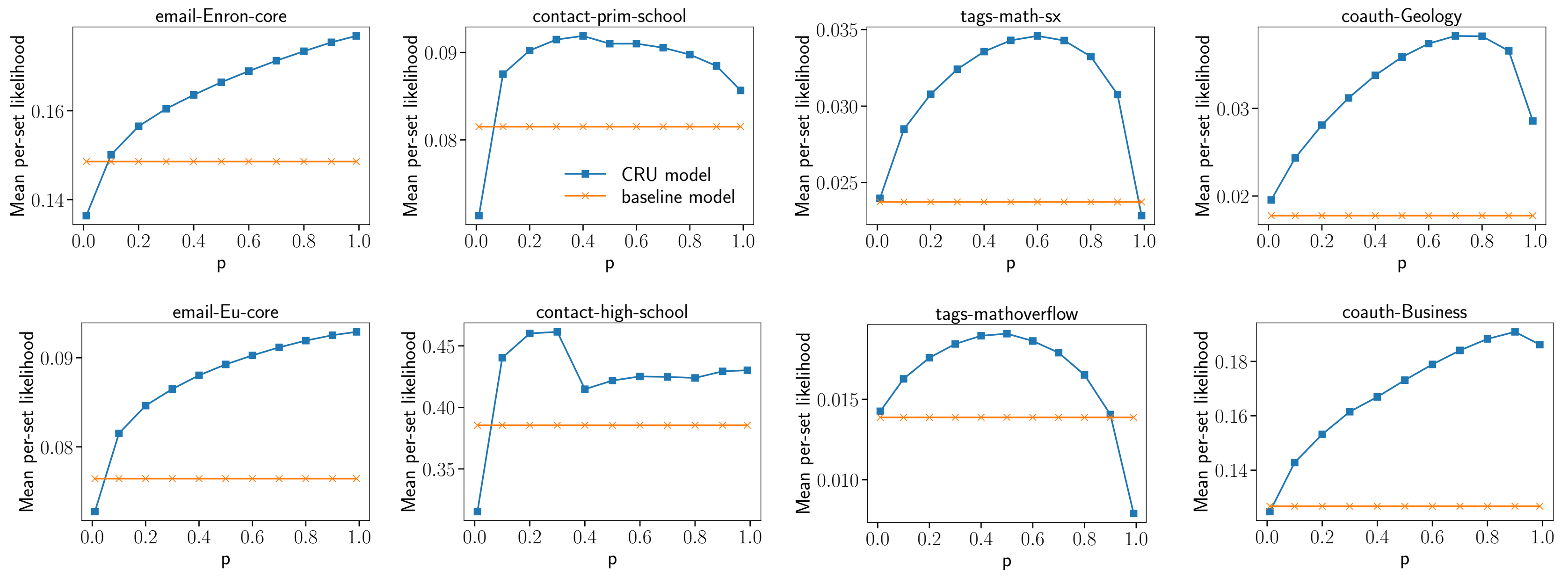
The real data has larger counts.

Correlated Repeated Unions (CRU) model for sequences of sets.

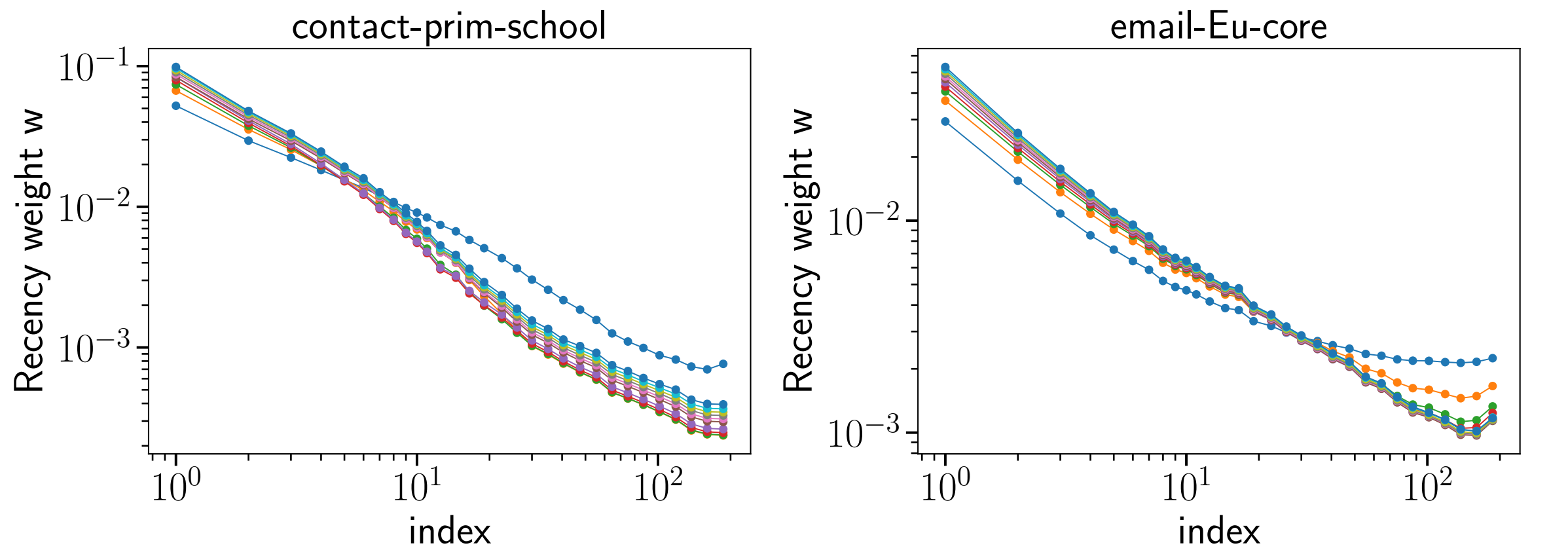
We provide a model capturing repeat elements in sequences of sets.
Observe sequence of sets S_1, …, S_k.
Want to predict repeat elements in S_{k+1}. Start with empty set N.

1. Sample set S_{k-j} from j steps back with recency weight w_j.
2. Sample T by keeping each item x in S_{k-j} with correlation probability p.
3. N = union(N, T)
4. Repeat 1—3 until N is as big as prescribed.

Learned model parameters.

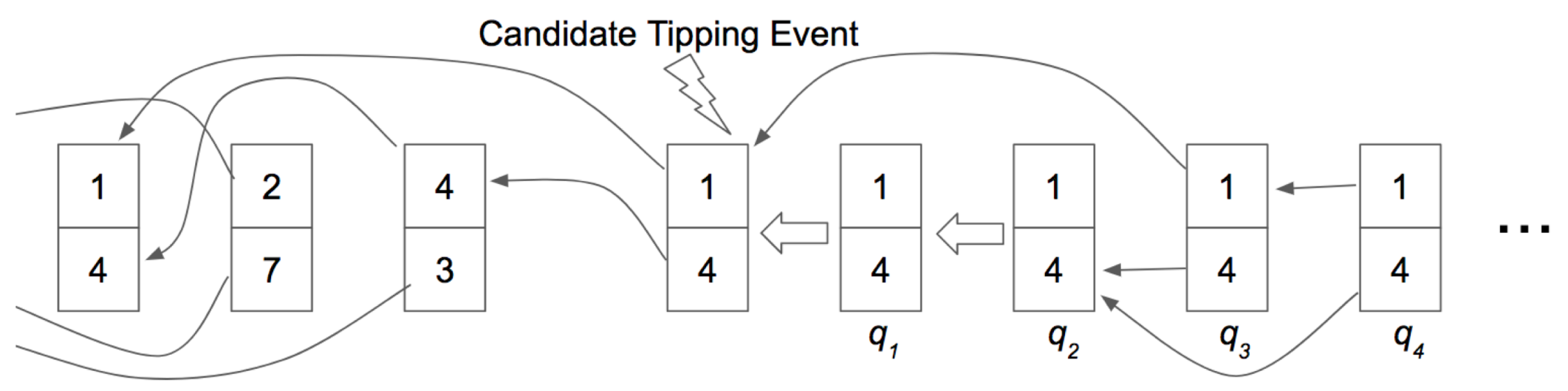


Likelihoods tend to be unimodal in the correlation probability p.
Likelihood functions look similar for datasets in similar domains.



Learned weights tend to be heavy-tailed and (roughly) monotonically decreasing.

Theory of asymptotic tipping behavior.



After a tipping event, a single set forevermore. New sets come from copying individual elements (think arrows) or an entire pair from after the tipping point (block arrows).

Theorem. Let $W = \sum_{j=1}^{\infty} w_{\mathrm{j}}$. If $W < \infty$, then with probability 1, the process *tips* and only a single pair will occur infinitely often.