Subsequences and substrings

- Change the episode definition slightly: **consecutive** requirement
- For episode α to appear, require that all event types in α must occur one after the other, with no extra events in between
- If α is the parallel episode AB, then it occurs only if in the sequence we see AB or BA, close enough to each other
- ACB will not count as an occurrence
- With this requirement serial episodes are more or less equivalent to substrings
- The first definition for episodes is **subsequences**

Finding frequently occurring substrings

- Suffix tries: a very efficient data structure

Chapter 5: Complexity of finding frequent patterns

5. Complexity of finding frequent patterns

- How difficult is it to find frequent patterns?
- Examples of some simple theoretical analyses
- Very simple lower bounds
- Border of a theory
- Guess-and-correct algorithm
- Borders and hypergraph transversals

Complexity of finding frequent sets

- data set with n rows, p attributes
- Find all frequent sets for some frequency threshold
- What is the complexity+
- We have to read the whole dataset $\Rightarrow \Omega(n)$ (at least linear in n)
- The result has to be output: in the worst case 2^p frequent sets, each of size from 1 to $p \Rightarrow \Omega(2^p)$
- The levelwise algorithm takes time $O(npC)$, where C is the total number of candidates considered

A very simple lower bound

- Sometimes finding frequent sets takes exponential time in the number of attributes
- But is this just because the output can be large
- Is this the only reason why the problem can be exponential?
- No
- Model of computation: questions of the form "is X frequent?"
- How many such questions have to be asked **to identify the answer**?
- We don't have to output the answer

A very simple lower bound, cont.

- Simple case: p attributes, only one maximal frequent set, with size k
- $\binom{p}{k}$ different possible answers
- Each question "Is X frequent?" provides 1 bit of information
- $$\log \binom{m}{k} \approx k \log(m/k)$$
  questions are needed to identify the single frequent set

A very simple lower bound, cont.

- Simple case: p attributes, many maximal frequent sets, each of size k
- $S = \binom{p}{k}$ different possible maximal frequent sets
- $T = 2^S$ different collections
- Each question "Is X frequent?" provides 1 bit of information
- $$\log 2^S = S = \binom{p}{k}$$
  questions are needed
- If $k = p/2$, then $\binom{p}{k}$ is exponential in p
- Thus identifying the answer can be difficult

Verifying the answer

- Suppose somebody tells us that the frequent sets of a dataset are ABC, CD, and BCE, and all their subsets (attributes $ABCDE$)
- Which questions should we ask to verify that this is indeed true?
- Test that ABC, CD, and BCE indeed are frequent
- If so, the claim is at least partly true
- There might be some other sets that could still be frequent

Verifying the answer, cont.

- ABC, CD, and BCE are frequent
- The claim is that no set other than the subsets of these are frequent
- What is the smallest collection of sets that we should test to verify this?
- Claim: if some other set is frequent, then one of AE, AD, DB, DE is frequent
- Why?

Why?

- If something else than ABC, CD, and BCE and their subsets is frequent, then that set X cannot be a subset of any of those
- The minimal sets X that are not subsets of any of ABC, CD, BCE
- The minimal sets that intersect the complements of ABC, CD, BCE
- The minimal sets that intersect DE, ABE, AD
- These are AE, AD, DB, DE

The border of a collection $\mathcal{F}$ of frequent sets

- A collection $\mathcal{F}$ of frequent sets
- Closed under subsets
- *positive border* $\mathcal{B}d^+(\mathcal{F})$: the sets that are in $\mathcal{F}$, but whose all proper supersets are outside $\mathcal{F}$
- The *negative border* $\mathcal{B}d^-(\mathcal{F})$: sets that are not in $\mathcal{F}$, but whose all proper subsets are in $\mathcal{F}$

Example

- Above we had $\mathcal{F}$ = subsets of ABC, CD, BCE, i.e.,

$$\mathcal{F} = \{\emptyset, A, B, , C, D, E, AB, AC, BC, CD, BE, ABC, BCE\}$$

- $\mathcal{B}d^+(\mathcal{F}) = \{ABC, CD, BCE\}$
- $\mathcal{B}d^-(\mathcal{F}) = \{AE, AD, DB, DE\}$

Another example

- $R = \{A, \ldots, F\}$

  $\{\{A\}, \{B\}, \{C\}, \{F\}, \{A, B\}, \{A, C\}, \{A, F\}, \{C, F\}, \{A, C, F\}\}$.

- The negative border

$$\mathcal{B}d^-(\mathcal{F}) = \{\{D\}, \{E\}, \{B, C\}, \{B, F\}\}$$

- The positive border, in turn, contains the maximal frequent sets, i.e.,

$$\mathcal{B}d^+(\mathcal{F}) = \{\{A, B\}, \{A, C, F\}\}$$

Verification problem

- Verifying that $\mathcal{F}$ is the collection of frequent sets of a database requires
$$|\mathcal{B}d^{+}(()\mathcal{F})| + |\mathcal{B}d^{-}(()\mathcal{F})|$$
queries of the form "Is X frequent?"

How to compute the negative border?

- Given a collection of frequent sets
- Computing the positive border is quite simple: just find the maximal elements
- Computing the negative border is more difficult
- Negative border: the minimal sets that intersect all the complements of the sets in the positive border
- Hypergraph transversal problem
- An interesting combinatorial question

When computing frequent sets

- Candidates = frequent sets + negative border
- Why?

Examples: random data sets

Independent attributes, probability of a 1 is p

p	*min_fr*	$\lvert\mathcal{F}\rvert$	$\lvert\mathcal{B}d^{+}(\mathcal{F})\rvert$	$\lvert\mathcal{B}d^{-}(\mathcal{F})\rvert$
0.2	0.01	469	273	938
0.2	0.005	1291	834	3027
0.5	0.1	1335	1125	4627
0.5	0.05	5782	4432	11531

Experimental results with random data sets.

min_fr	$\lvert\mathcal{F}\rvert$	$\lvert\mathcal{B}d^+(\mathcal{F})\rvert$	$\lvert\mathcal{B}d^-(\mathcal{F})\rvert$
0.08	96	35	201
0.06	270	61	271
0.04	1028	154	426
0.02	6875	328	759

Experimental results with a real data set.

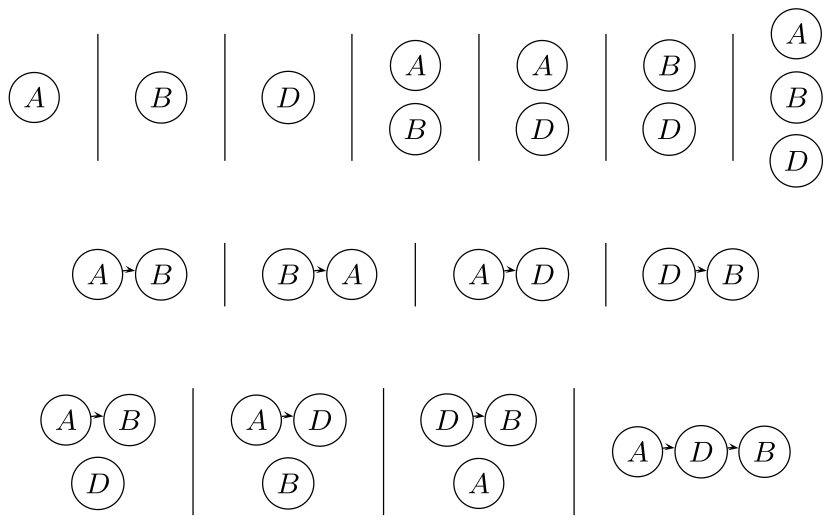
Borders for other types of patterns

- Can be defined in exactly the same way
- Result of finding frequent patterns is a collection of patterns closed under generalizations
- Positive border: most specific patterns in the collection
- Negative border: most general patterns not in the collection

Example for strings

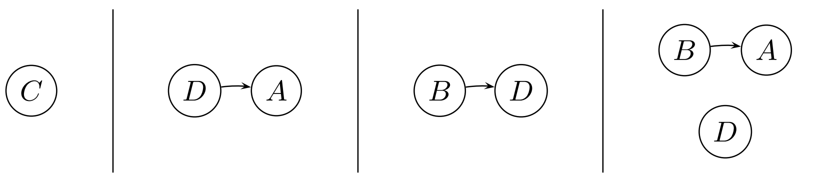
- $\mathcal{P}$: substrings over an alphabet $\sum$
- q: how frequently the substring occurs
- (substrings vs. subsequences $\approx$ sequential episodes)
- $\sum = \{a, b, c\}$
- $\mathcal{F} = \{a, b, c, ab, bc, abc, cb\}$
- positive border $\{abc, cb\}$
- negative border $\{ca, aa, bb, ba, cc, ac\}$(?)

Example for episodes



A collection $\mathcal{F}(\mathrm{s}, \textit{win}, \textit{min\_fr})$ of frequent episodes.

The positive border $\mathcal{B}d^{+}(\mathcal{F}(\mathbf{s}, \mathit{win}, \mathit{min\_fr}))$.



The negative border $\mathcal{B}d^{-}(\mathcal{F}(\mathbf{s}, \mathit{win}, \mathit{min\_fr}))$. (Tends to be tricky to check; is this correct?)

Complexity of the levelwise algorithm

- The levelwise algorithm compute the frequency of the frequent patterns and the patterns in the negative border

The guess-and-correct algorithm

- Levelwise search: safe but sometimes slow
- Especially if there are frequent patterns that are far from the bottom of the specialization relation
- An alternative: start finding $\mathcal{F}$ from an initial guess $\mathcal{S} \subseteq \mathcal{P}$, and then correcting the guess by looking at the database
- If the initial guess is good, few iterations are needed to correct the result