# Data Mining and Analysis

## Market Basket Analysis: 1

CSCE 676 :: Fall 2019
Texas A&M University
Department of Computer Science & Engineering
Prof. James Caverlee

# Agenda

Today:

History, Basic Definitions, Frequent Itemsets

Wednesday:

Apriori

Friday:

Improvements and extensions

# Resources

MMDS: Mining of Massive Datasets [http://www.mmds.org/mmds/v2.1/ch06-assocrules.pdf]

Tan, Steinbach, Karpatne, Kumar. Introduction to Data Mining [https://www-users.cs.umn.edu/~kumar001/dmbook/slides/chap5-association_analysis.pdf]

Carlos Castillo course on Data Mining [https://github.com/chatox/data-mining-course]

Vagelis Papalexakis course on Data Mining [https://www.cs.ucr.edu/~epapalex/teaching/235_S19/index.html]

# (*Traditional*) Market Basket Analysis

Goal: Mine patterns from transaction records (e.g., items for sale at Walmart)

Understand customers

- Purchasing habits, sensitivity to price, promotions

Understand products

- Co-purchases, fast/slow movers

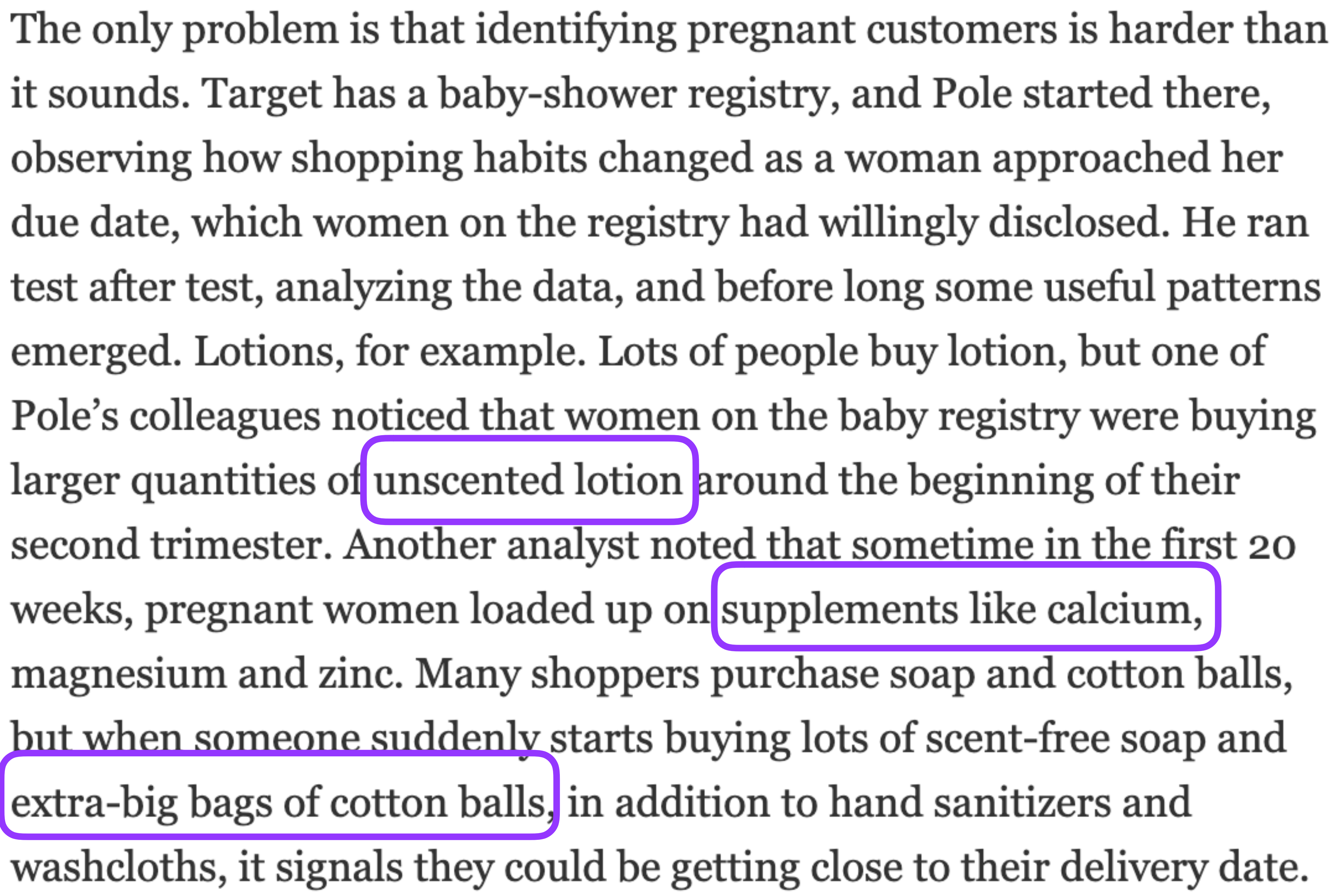
Take action: promotions, store layout, ...

# How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did

Every time you go shopping, you share intimate details about your consumption patterns with retailers. And many of those retailers are studying those details to figure out what you like, what you need, and which coupons are most likely to make you happy. Target, for example, has figured out how to data-mine its way into your womb, to figure out whether you have a baby on the way long before you need to start buying diapers.

Target has got you in its aim

Charles Duhigg outlines in the New York Times how Target tries to hook parents-to-be at that crucial moment before they turn into rampant — and loyal — buyers of all things pastel, plastic, and miniature. He talked to Target statistician Andrew Pole — before https://www.nytimes.com/2012/02/19/magazine/shopping-habits.html



The only problem is that identifying pregnant customers is harder than it sounds. Target has a baby-shower registry, and Pole started there, observing how shopping habits changed as a woman approached her due date, which women on the registry had willingly disclosed. He ran test after test, analyzing the data, and before long some useful patterns emerged. Lotions, for example. Lots of people buy lotion, but one of Pole's colleagues noticed that women on the baby registry were buying larger quantities of unscented lotion around the beginning of their second trimester. Another analyst noted that sometime in the first 20 weeks, pregnant women loaded up on supplements like calcium, magnesium and zinc. Many shoppers purchase soap and cotton balls, but when someone suddenly starts buying lots of scent-free soap and extra-big bags of cotton balls, in addition to hand sanitizers and washcloths, it signals they could be getting close to their delivery date.

# Association Rule Discovery

Supermarket shelf management – Market-basket model:

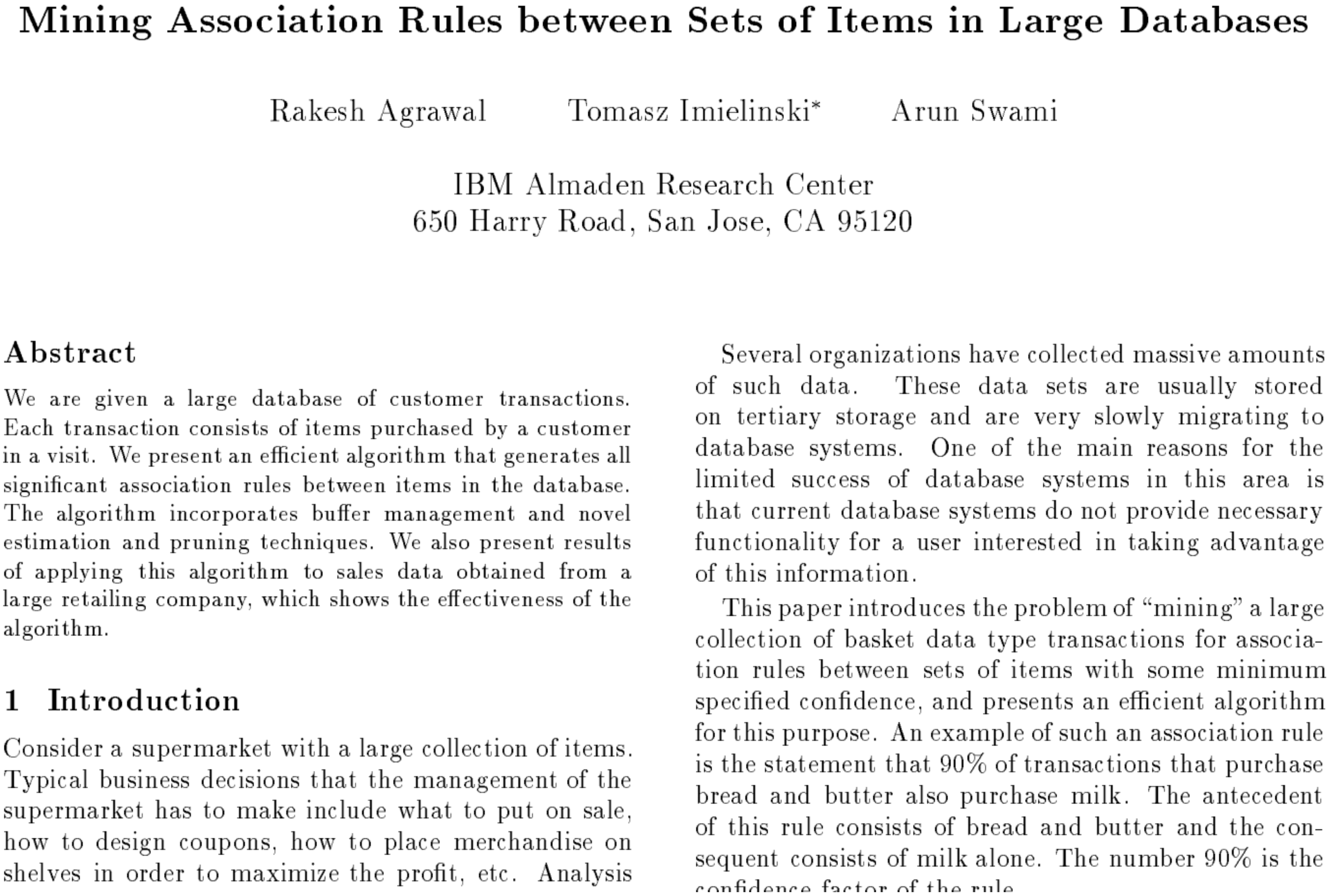
Goal: Identify items that are bought together by sufficiently many customers

Approach: Process the sales data collected with barcode scanners to find dependencies among items

A classic rule:

> If someone buys diapers and milk, then he/she is likely to buy beer

SIGMOD 1993



# Mining Association Rules between Sets of Items in Large Databases

Rakesh Agrawal Tomasz Imielinski* Arun Swami

IBM Almaden Research Center
650 Harry Road, San Jose, CA 95120

## Abstract

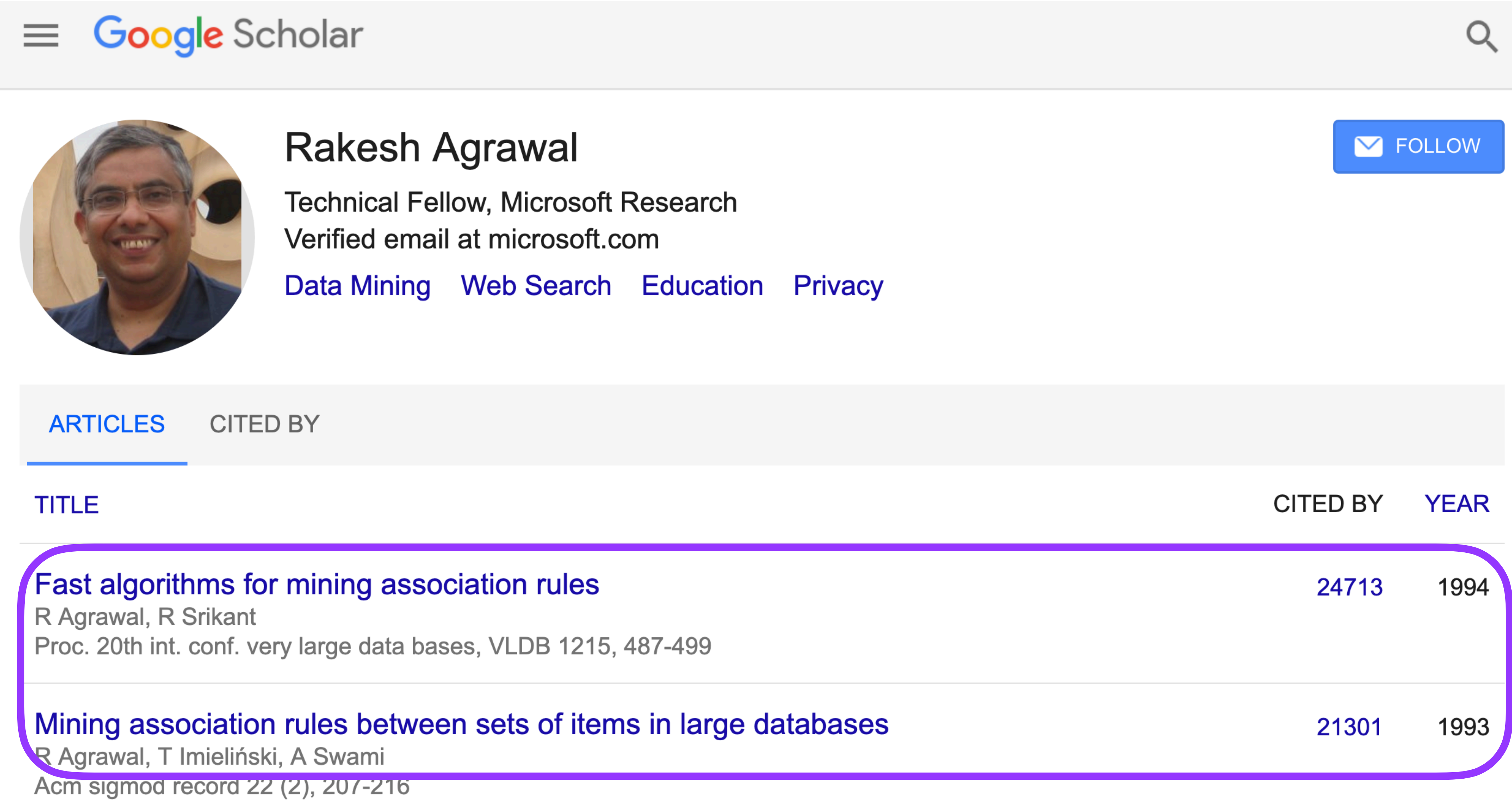
We are given a large database of customer transactions. Each transaction consists of items purchased by a customer in a visit. We present an efficient algorithm that generates all significant association rules between items in the database. The algorithm incorporates buffer management and novel estimation and pruning techniques. We also present results of applying this algorithm to sales data obtained from a large retailing company, which shows the effectiveness of the algorithm.

## 1 Introduction

Consider a supermarket with a large collection of items. Typical business decisions that the management of the supermarket has to make include what to put on sale, how to design coupons, how to place merchandise on shelves in order to maximize the profit, etc. Analysis

Several organizations have collected massive amounts of such data. These data sets are usually stored on tertiary storage and are very slowly migrating to database systems. One of the main reasons for the limited success of database systems in this area is that current database systems do not provide necessary functionality for a user interested in taking advantage of this information.

This paper introduces the problem of "mining" a large collection of basket data type transactions for association rules between sets of items with some minimum specified confidence, and presents an efficient algorithm for this purpose. An example of such an association rule is the statement that 90% of transactions that purchase bread and butter also purchase milk. The antecedent of this rule consists of bread and butter and the consequent consists of milk alone. The number 90% is the confidence factor of the rule

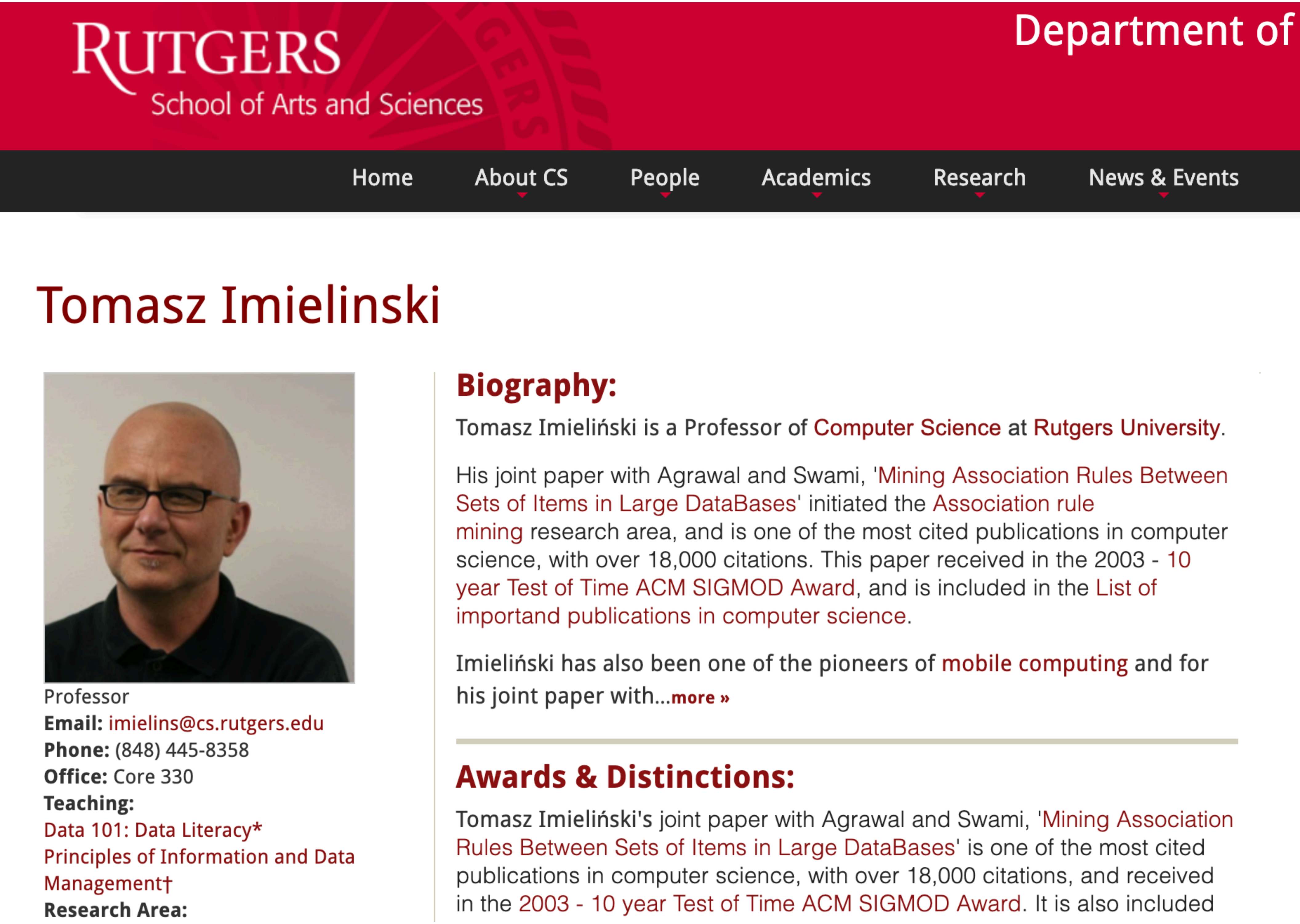


Rakesh Agrawal

Technical Fellow, Microsoft Research
Verified email at microsoft.com

Data Mining Web Search Education Privacy

| TITLE | CITED BY | YEAR |
|---|---|---|
| Fast algorithms for mining association rules<br>R Agrawal, R Srikant<br>Proc. 20th int. conf. very large data bases, VLDB 1215, 487-499 | 24713 | 1994 |
| Mining association rules between sets of items in large databases<br>R Agrawal, T Imieliński, A Swami<br>Acm sigmod record 22 (2), 207-216 | 21301 | 1993 |



# Tomasz Imielinski

Professor
**Email:** imielins@cs.rutgers.edu
**Phone:** (848) 445-8358
**Office:** Core 330
**Teaching:**
Data 101: Data Literacy*
Principles of Information and Data Management†
**Research Area:**

## Biography:

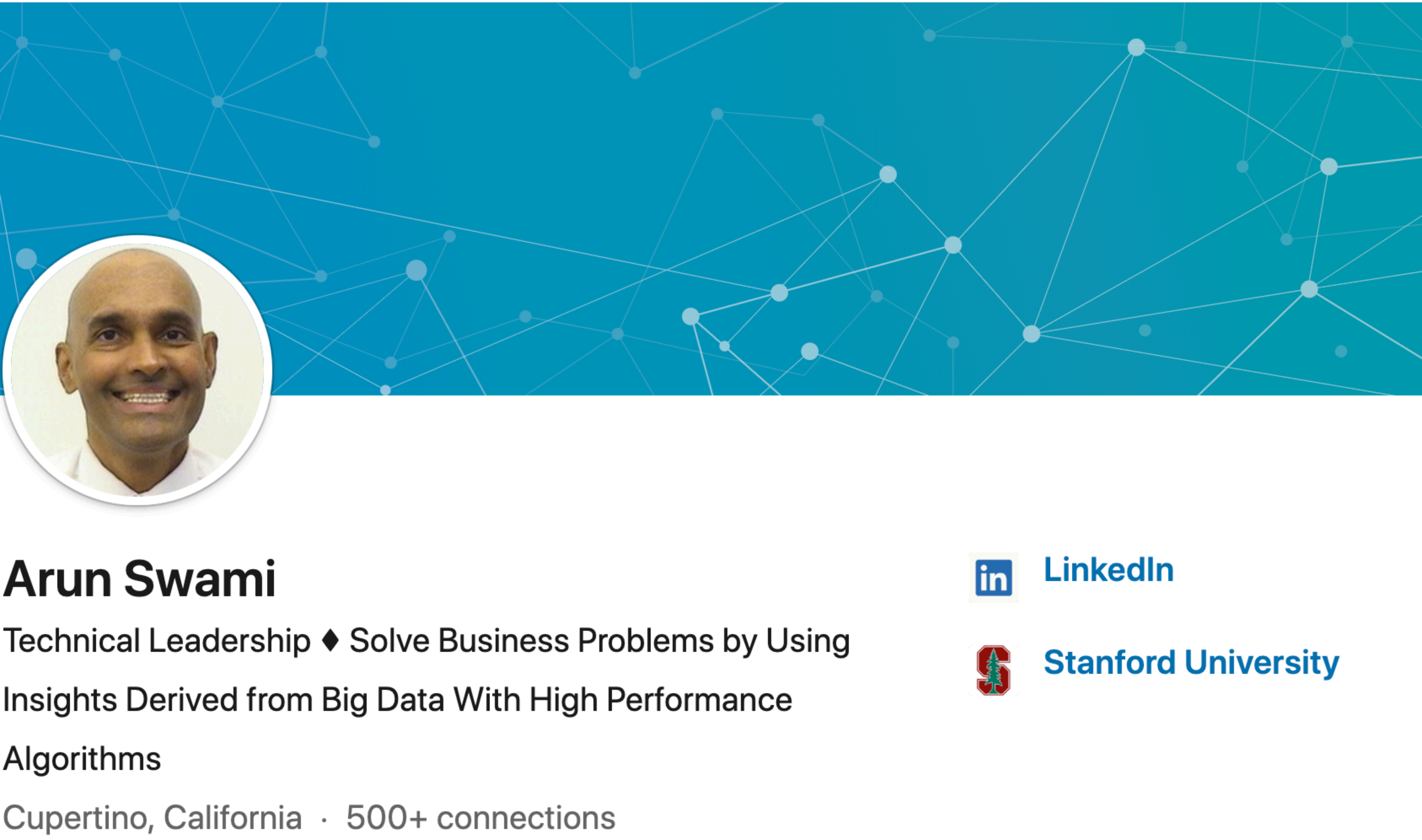
Tomasz Imieliński is a Professor of Computer Science at Rutgers University.

His joint paper with Agrawal and Swami, 'Mining Association Rules Between Sets of Items in Large DataBases' initiated the Association rule mining research area, and is one of the most cited publications in computer science, with over 18,000 citations. This paper received in the 2003 - 10 year Test of Time ACM SIGMOD Award, and is included in the List of importand publications in computer science.

Imieliński has also been one of the pioneers of mobile computing and for his joint paper with...**more »**

## Awards & Distinctions:

Tomasz Imieliński's joint paper with Agrawal and Swami, 'Mining Association Rules Between Sets of Items in Large DataBases' is one of the most cited publications in computer science, with over 18,000 citations, and received in the 2003 - 10 year Test of Time ACM SIGMOD Award. It is also included



# Arun Swami

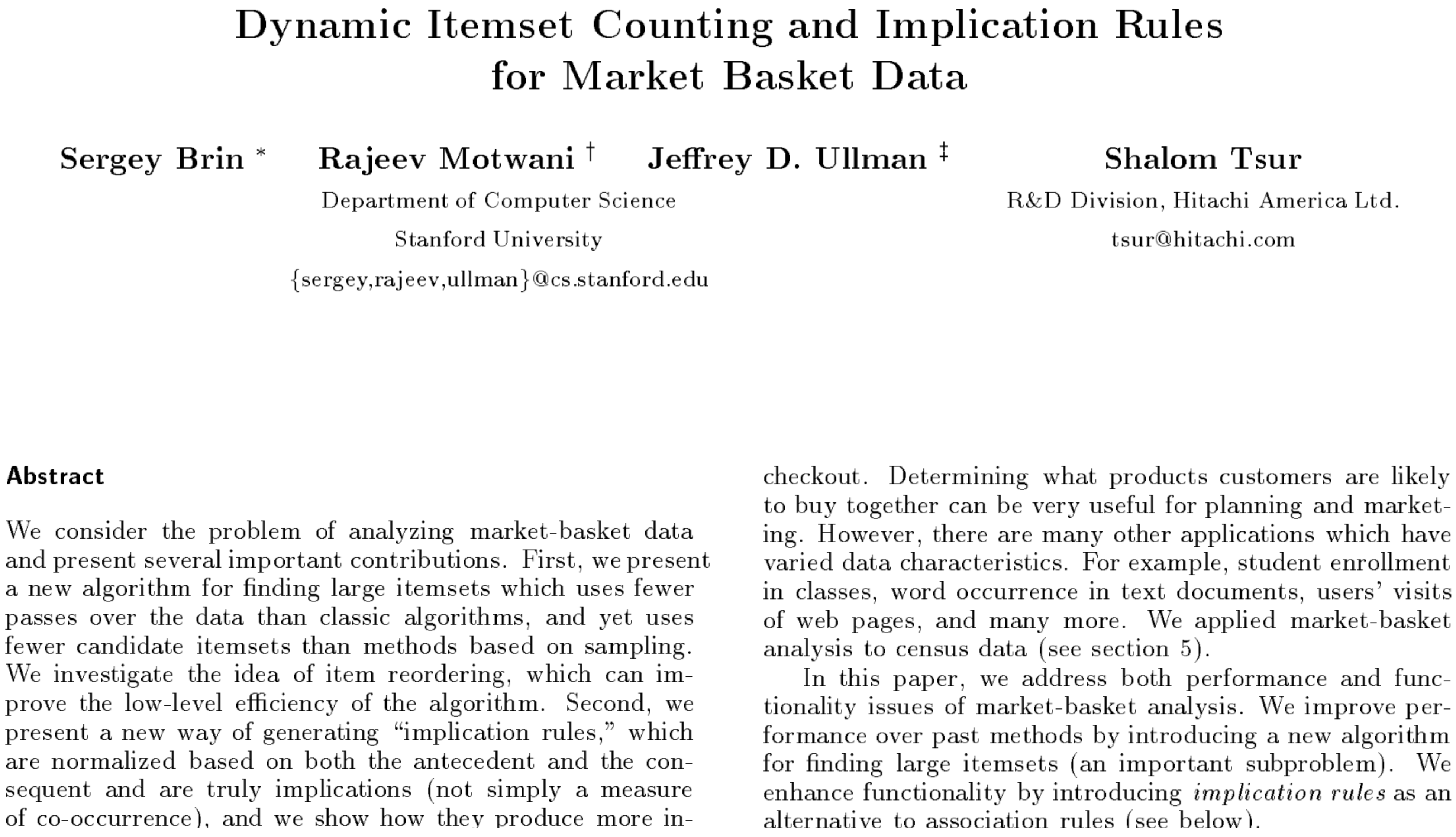
Technical Leadership ♦ Solve Business Problems by Using Insights Derived from Big Data With High Performance Algorithms

Cupertino, California · 500+ connections

LinkedIn

Stanford University

# SIGMOD 1997



## Dynamic Itemset Counting and Implication Rules for Market Basket Data

**Sergey Brin** [*] **Rajeev Motwani** [†] **Jeffrey D. Ullman** [‡]
Department of Computer Science
Stanford University
{sergey,rajeev,ullman}@cs.stanford.edu

**Shalom Tsur**
R&D Division, Hitachi America Ltd.
tsur@hitachi.com

### Abstract

We consider the problem of analyzing market-basket data and present several important contributions. First, we present a new algorithm for finding large itemsets which uses fewer passes over the data than classic algorithms, and yet uses fewer candidate itemsets than methods based on sampling. We investigate the idea of item reordering, which can improve the low-level efficiency of the algorithm. Second, we present a new way of generating "implication rules," which are normalized based on both the antecedent and the consequent and are truly implications (not simply a measure of co-occurrence), and we show how they produce more in-

checkout. Determining what products customers are likely to buy together can be very useful for planning and marketing. However, there are many other applications which have varied data characteristics. For example, student enrollment in classes, word occurrence in text documents, users' visits of web pages, and many more. We applied market-basket analysis to census data (see section 5).

In this paper, we address both performance and functionality issues of market-basket analysis. We improve performance over past methods by introducing a new algorithm for finding large itemsets (an important subproblem). We enhance functionality by introducing *implication rules* as an alternative to association rules (see below).

# The Market-Basket Model

Given: a set of baskets containing items

Typically, #items in a basket is small; #baskets is very large

A basket contains a small subset of items

(basket = transaction)

Goal: Find association rules of the form {x, y} —> z

Input:

| *TID* | *Items* |
|---|---|
| **1** | **Bread, Coke, Milk** |
| **2** | **Beer, Bread** |
| **3** | **Beer, Coke, Diaper, Milk** |
| **4** | **Beer, Bread, Diaper, Milk** |
| **5** | **Coke, Diaper, Milk** |

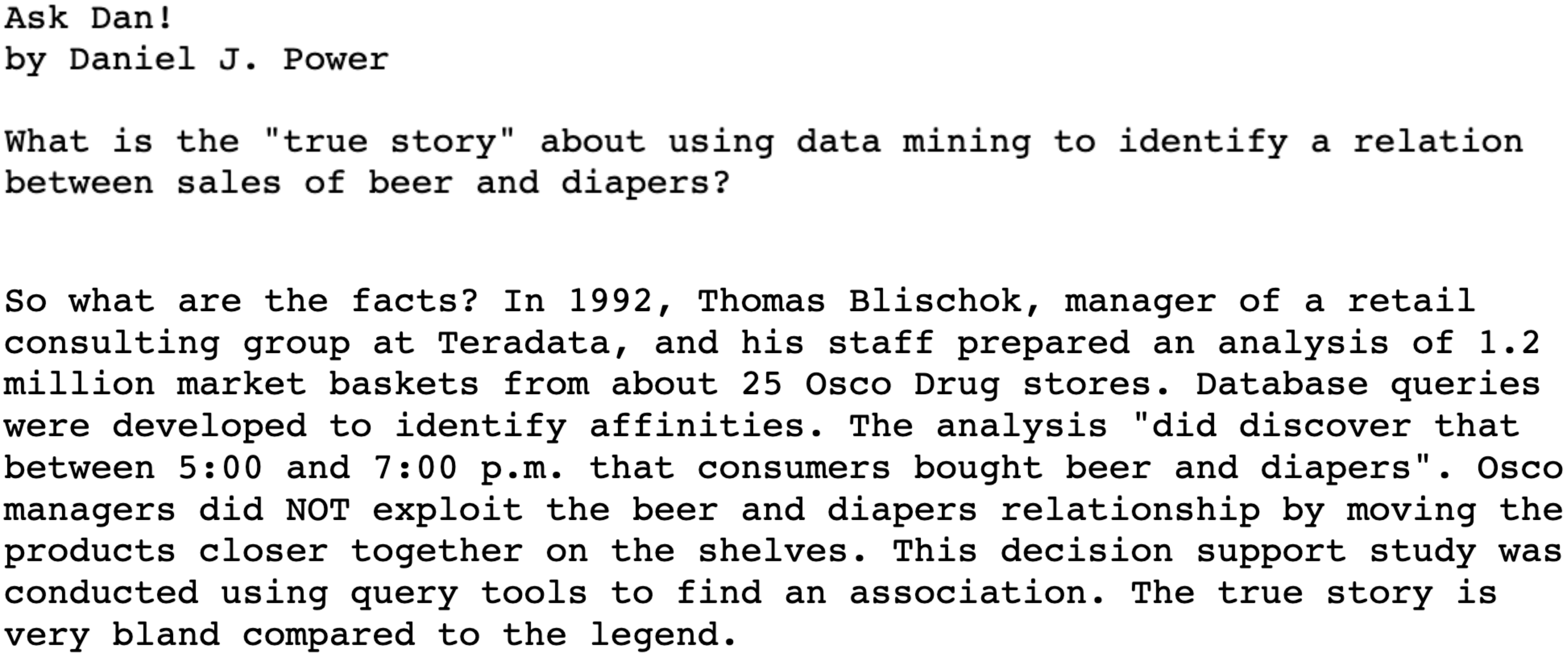
Output:

**Rules Discovered:**

{Milk} --> {Coke}

{Diaper, Milk} --> {Beer}

# Diapers and Beer



Ask Dan!
by Daniel J. Power

What is the "true story" about using data mining to identify a relation between sales of beer and diapers?

So what are the facts? In 1992, Thomas Blischok, manager of a retail consulting group at Teradata, and his staff prepared an analysis of 1.2 million market baskets from about 25 Osco Drug stores. Database queries were developed to identify affinities. The analysis "did discover that between 5:00 and 7:00 p.m. that consumers bought beer and diapers". Osco managers did NOT exploit the beer and diapers relationship by moving the products closer together on the shelves. This decision support study was conducted using query tools to find an association. The true story is very bland compared to the legend.

http://www.dssresources.com/newsletters/66.php

# General Concept: Items and Baskets

Baskets = patients; Items = drugs & side effects

Baskets = sentences; Items = documents

Documents that appear together could indicate plagiarism

Baskets = Instagram accounts; Items = liked photos

Baskets = movies; Items = actors

# Association Rules

If-then rules about the contents of baskets

$\{i_1, i_2, \ldots, i_k\} \rightarrow j$ means:

> “if a basket contains all of $i_1, \ldots, i_k$ then it is likely to contain $j$”

In practice there are many rules, want to find significant/interesting ones!

# Interesting Rules?

As a first step, let’s look for rules that have lots of evidence in the dataset (support) —> frequent itemsets

Then let’s find rules based on those itemsets that have high confidence

# The Basics

# Itemsets

itemset = collection of one or more items

Example = {Milk, Bread, Diapers}

k-itemset = an itemset that contains k items

The example above is a 3-itemset

# Big Problem!

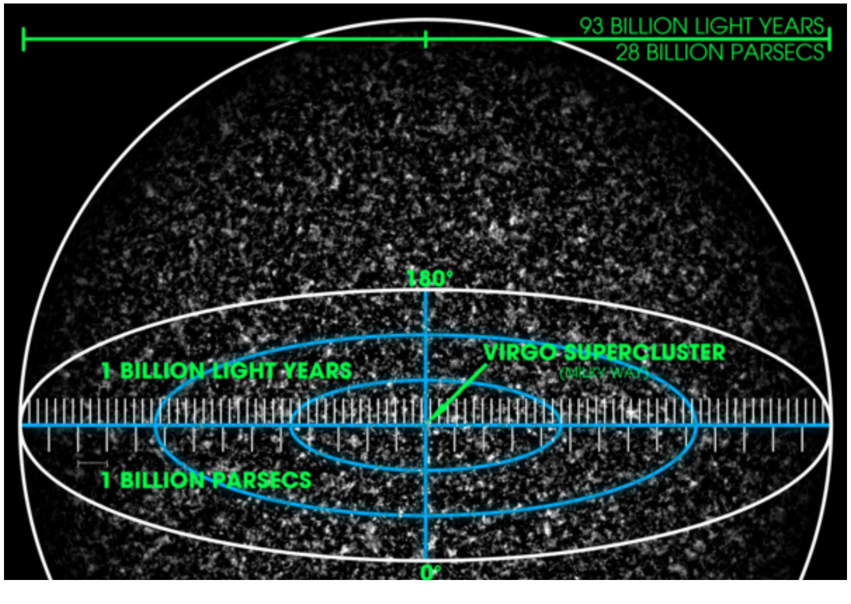
If there are n items

How many possible itemsets are there?

$2^n$

1000 items —> $2^{1000}$ —> $10^{300}$

$10^{80}$ hydrogen atoms in the observable universe!



# Support

Support of item set I

sup(I) = the fraction of baskets that contain I as a subset

> (sometimes support is reported as the number of baskets containing I, not a fraction)

# Frequent itemsets

Sets of items that appear together "frequently" in baskets

Given a support threshold minsup, the sets of items that appear in at least minsup baskets are frequent item sets

# Support

sup(I) = the fraction of baskets that contain I as a subset

| tid | Set of items |
|---|---|
| 1 | Bread, Jam, Juice |
| 2 | Tofu, Juice, Tomatoes |
| 3 | Bread, Strawberries, Tofu, Juice |
| 4 | Tofu, Juice, Tomatoes |
| 5 | Strawberries, Juice, Tomatoes |

sup(bread) = ?

sup(bread, juice) = ?

sup(strawberries, tomatoes) = ?

# Support

The smaller minsup is, the larger the number of frequent itemsets

Support monotonicity property:
if $J \subseteq I$, $sup(J) \geq sup(I)$

(if I is contained in a transaction, J is contained in the same transaction!)

# Frequent Itemsets

Items = {bread, jam, juice, tofu, tomatoes, strawberries}

Support threshold minsup = 2 baskets (or 2/5 of all)

| tid | Set of items |
|---|---|
| 1 | Bread, Jam, Juice |
| 2 | Tofu, Juice, Tomatoes |
| 3 | Bread, Strawberries, Tofu, Juice |
| 4 | Tofu, Juice, Tomatoes |
| 5 | Strawberries, Juice, Tomatoes |

Frequent item sets = ?

# Closed itemset

An itemset is closed if all itemsets containing it are less frequent

| tid | Set of items |
|---|---|
| 1 | Bread, Jam, Juice |
| 2 | Tofu, Juice, Tomatoes |
| 3 | Bread, Strawberries, Tofu, Juice |
| 4 | Tofu, Juice, Tomatoes |
| 5 | Strawberries, Juice, Tomatoes |

Find a closed itemset in this set if transactions

# Closed itemset

| tid | Set of items |
|---|---|
| 1 | Bread, Jam, Juice |
| 2 | Tofu, Juice, Tomatoes |
| 3 | Bread, Strawberries, Tofu, Juice |
| 4 | Tofu, Juice, Tomatoes |
| 5 | Strawberries, Juice, Tomatoes |

Example closed itemset: {Bread, Juice}

sup({Bread, Juice}) = 2

sup({Bread, Juice, Jam}) = 1

sup({Bread, Juice, Strawberries}) = 1

sup({Bread, Juice, Tofu}) = 1

We’ll revisit this in a bit:

Downward closure property: every subset of a frequent itemset is also frequent

# Maximal itemset

An itemset is maximal if

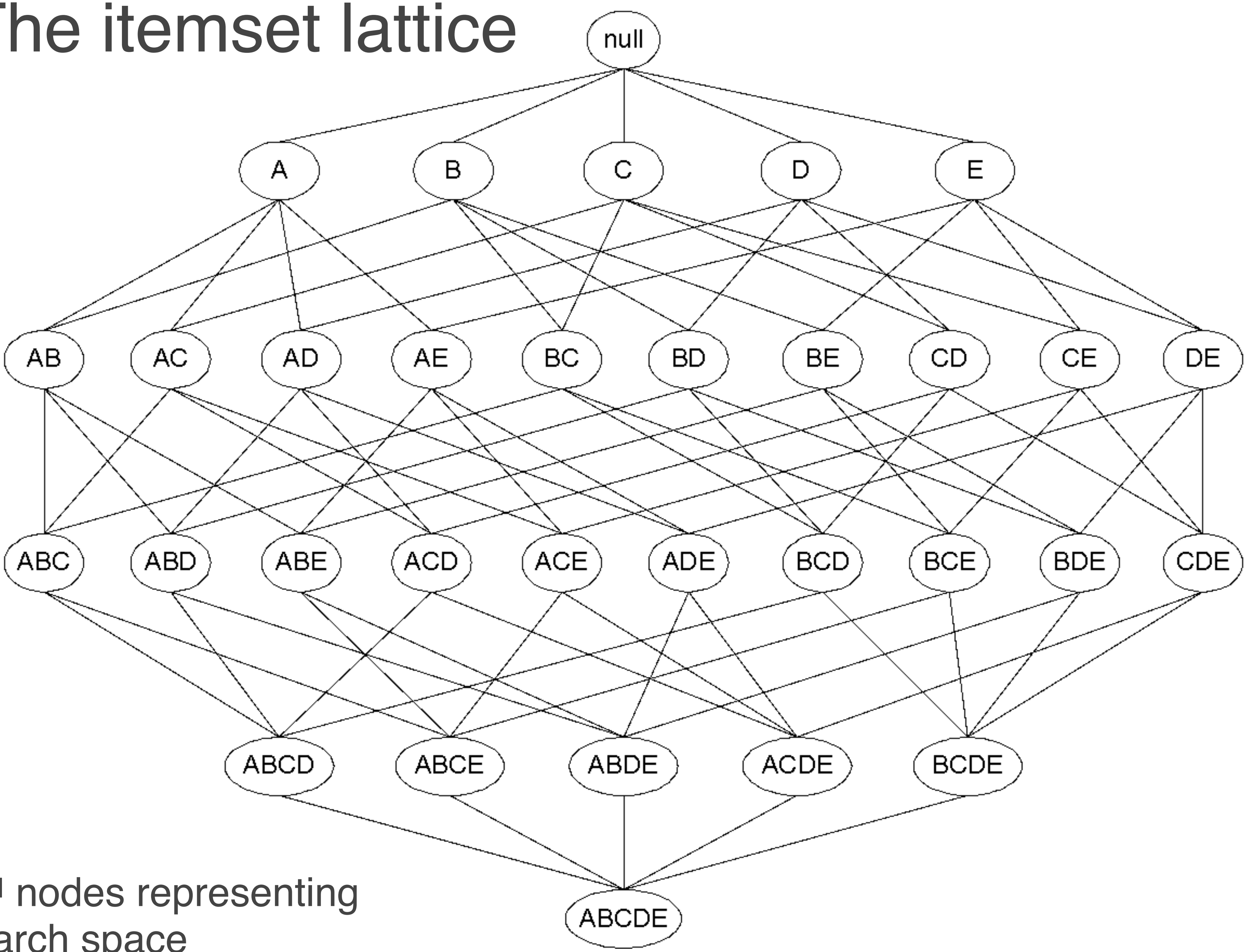
- it is closed and

- it has support ≥ minsup

| tid | Set of items |
|---|---|
| 1 | Bread, Jam, Juice |
| 2 | Tofu, Juice, Tomatoes |
| 3 | Bread, Strawberries, Tofu, Juice |
| 4 | Tofu, Juice, Tomatoes |
| 5 | Strawberries, Juice, Tomatoes |

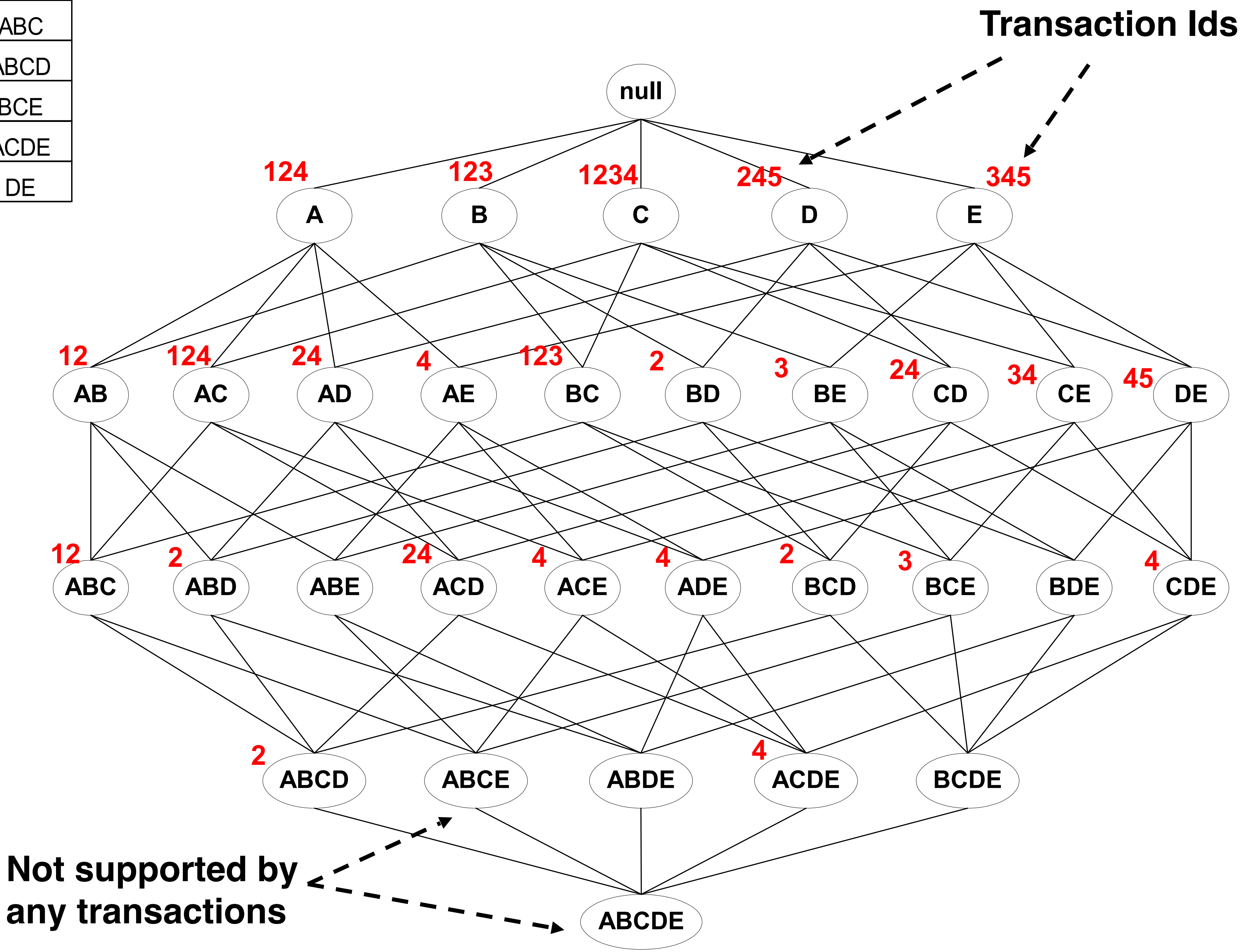
Find three maximal itemsets with minsup = 0.4

# The itemset lattice



$2^{|U|}$ nodes representing search space

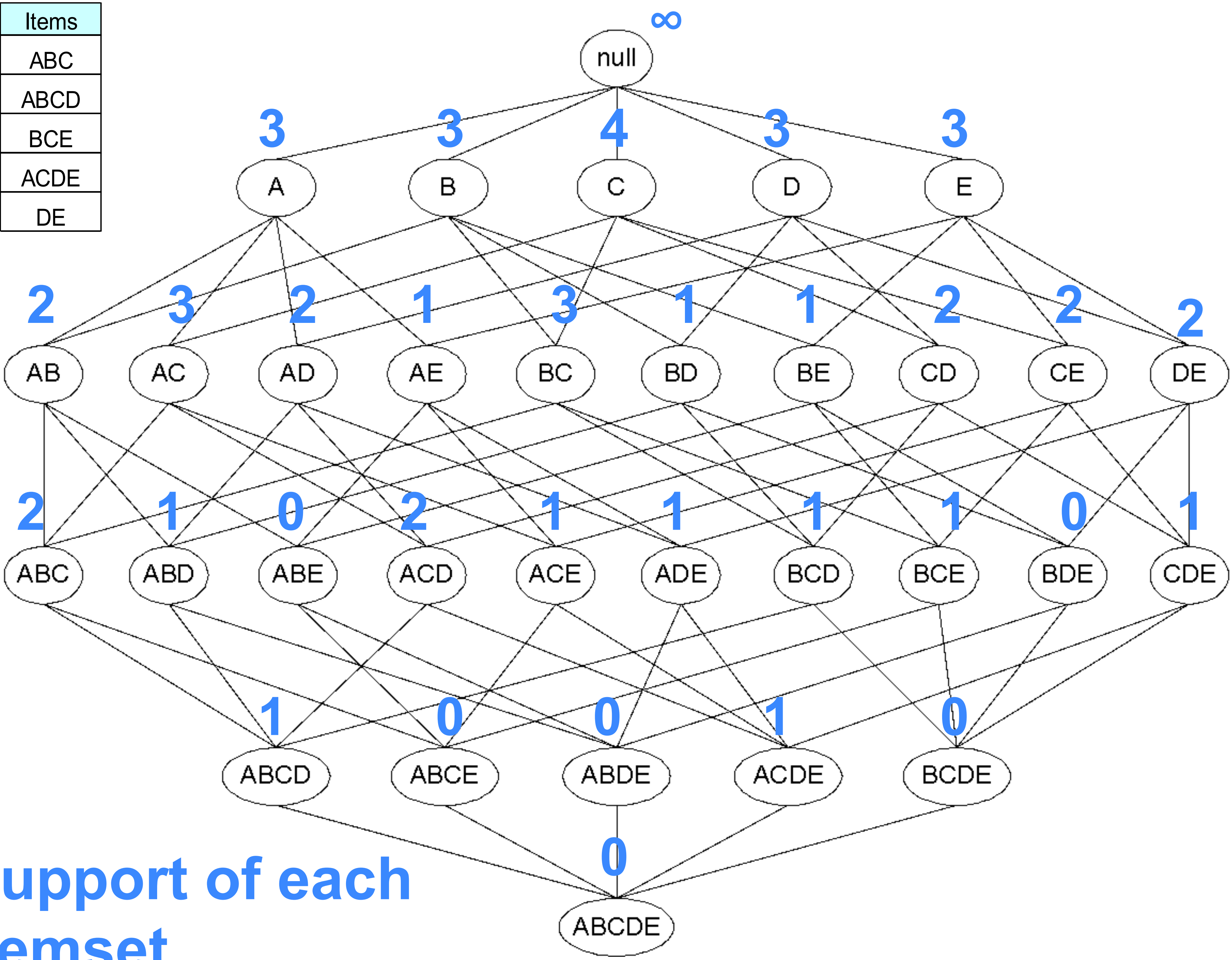
| TID | Items |
|---|---|
| 1 | ABC |
| 2 | ABCD |
| 3 | BCE |
| 4 | ACDE |
| 5 | DE |



Transaction Ids

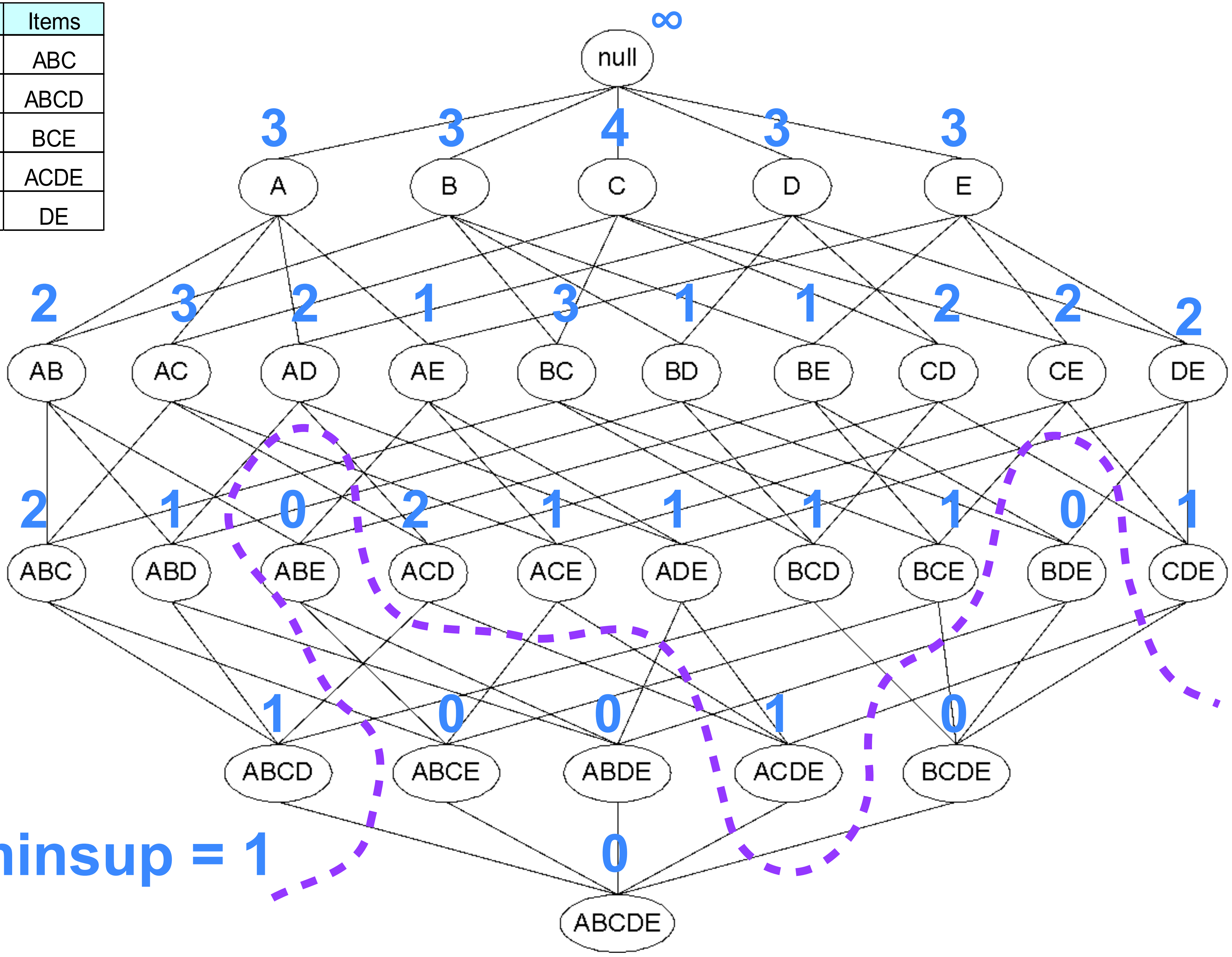
Not supported by any transactions

| TID | Items |
|---|---|
| 1 | ABC |
| 2 | ABCD |
| 3 | BCE |
| 4 | ACDE |
| 5 | DE |



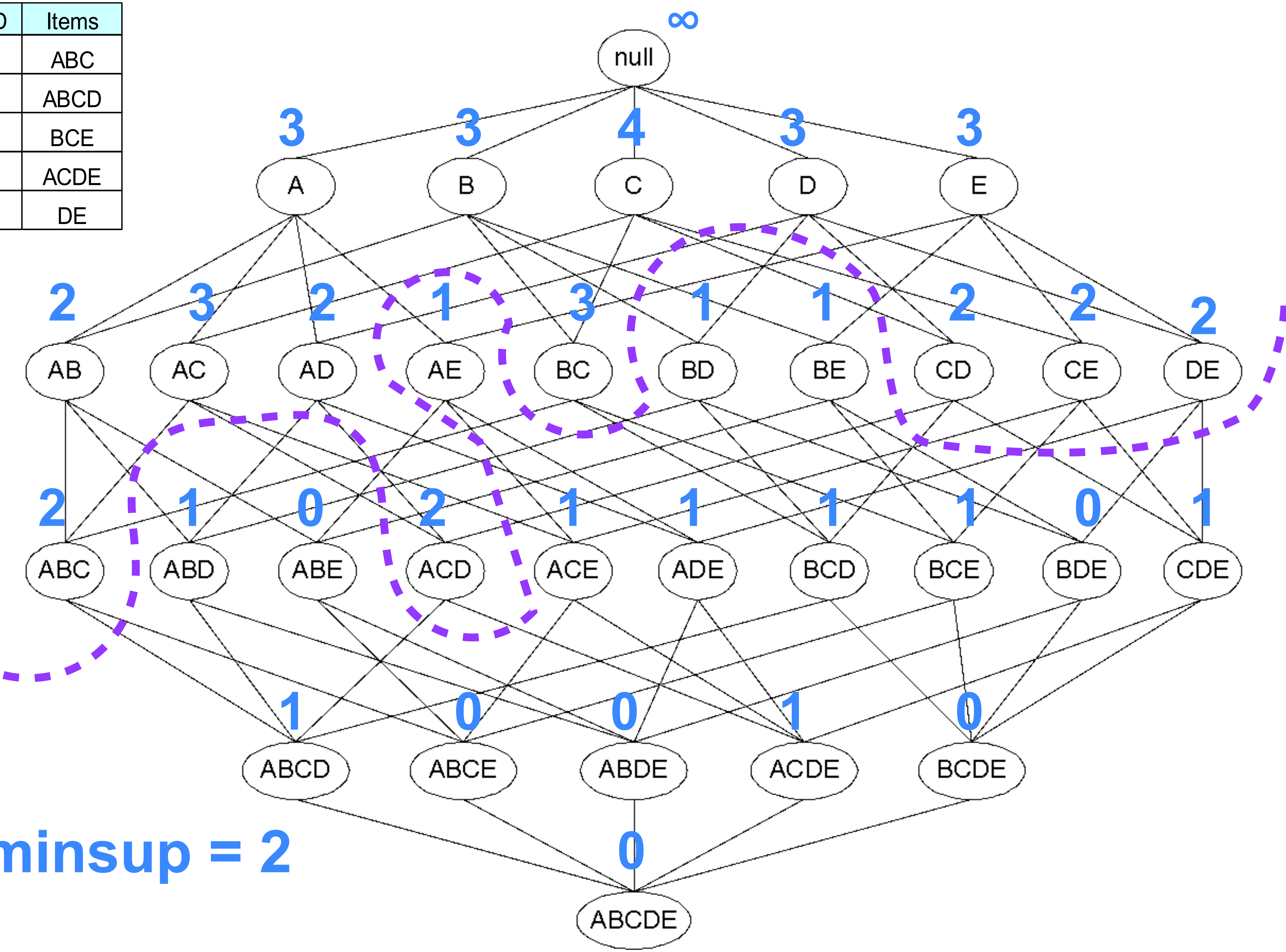
# Support of each itemset

| TID | Items |
|---|---|
| 1 | ABC |
| 2 | ABCD |
| 3 | BCE |
| 4 | ACDE |
| 5 | DE |



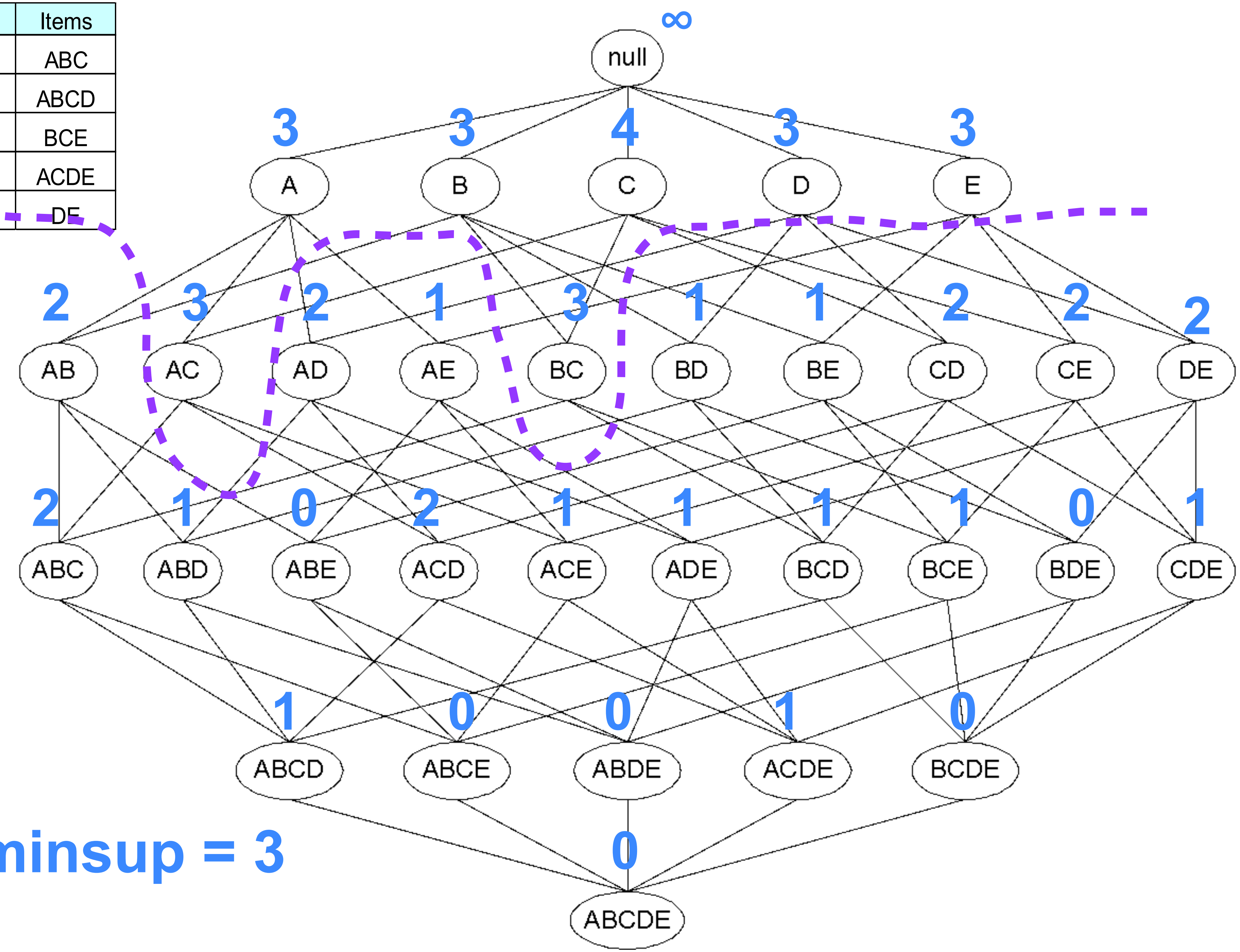
**minsup = 1**

| TID | Items |
|---|---|
| 1 | ABC |
| 2 | ABCD |
| 3 | BCE |
| 4 | ACDE |
| 5 | DE |



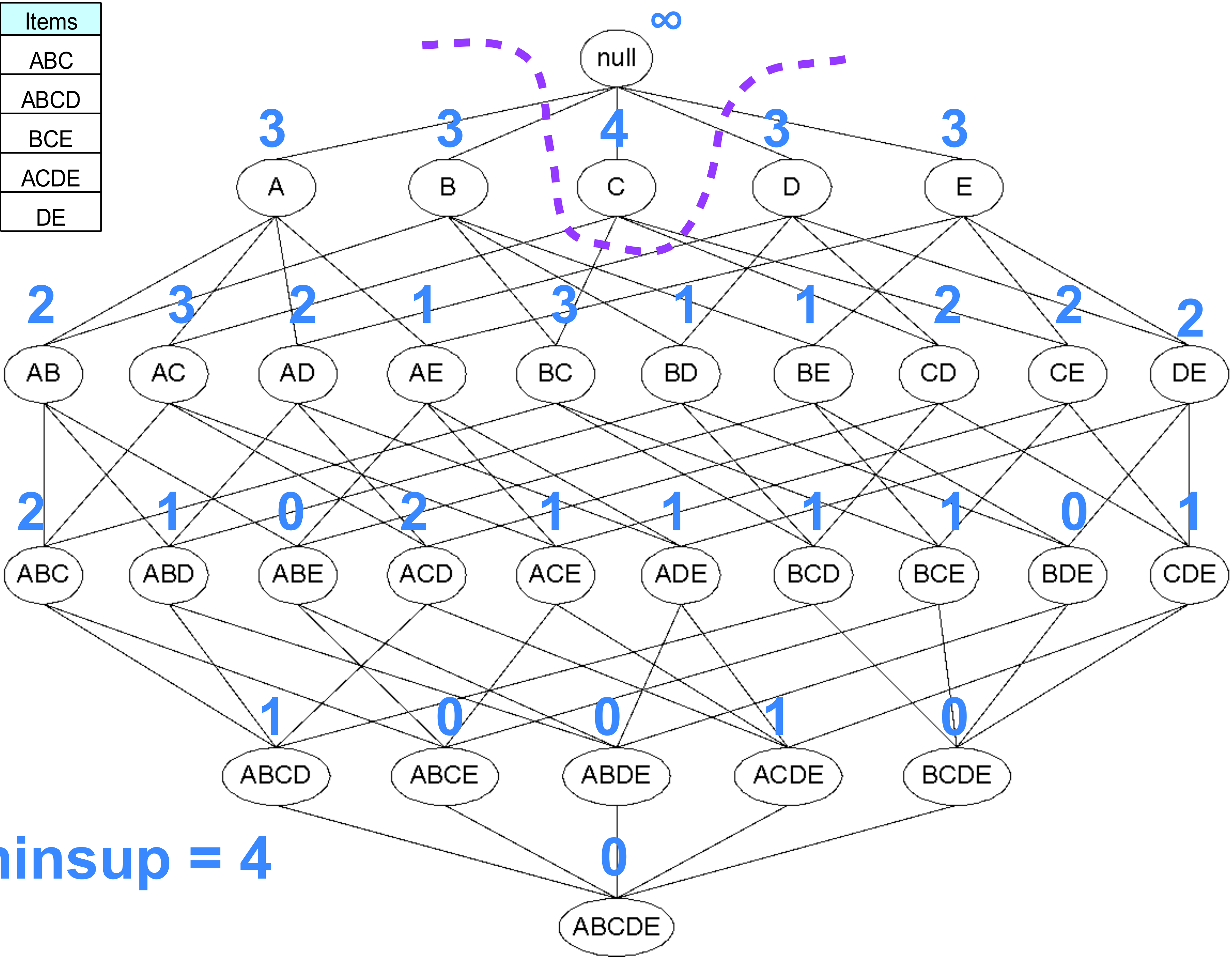
minsup = 2

| TID | Items |
|---|---|
| 1 | ABC |
| 2 | ABCD |
| 3 | BCE |
| 4 | ACDE |
| 5 | DE |



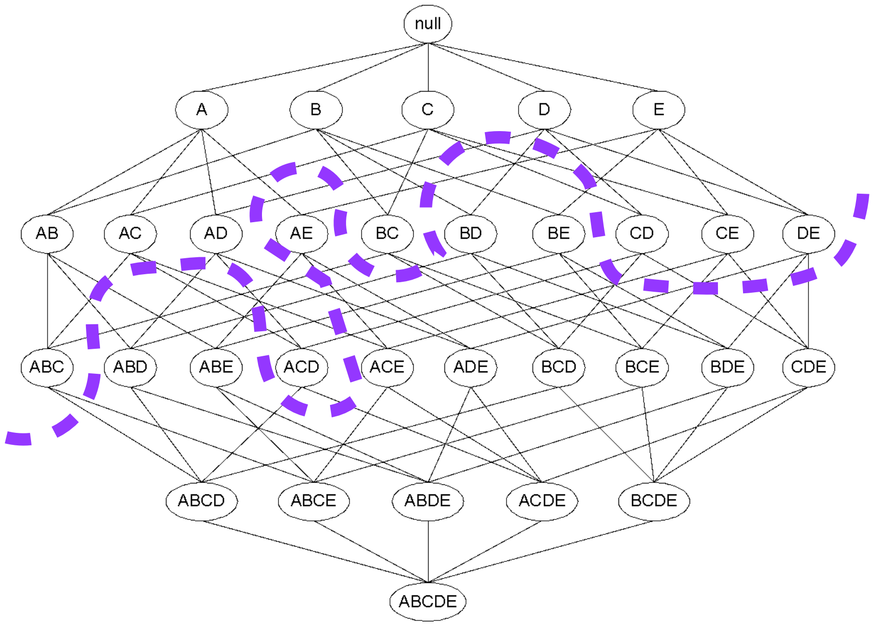
minsup = 3

| TID | Items |
|---|---|
| 1 | ABC |
| 2 | ABCD |
| 3 | BCE |
| 4 | ACDE |
| 5 | DE |



**minsup = 4**

# The border is a graph cut and ...



- All itemsets above the border are frequent
- All itemsets below the border are not frequent
- All maximal frequent itemsets are adjacent to the border
- Any border respects the downward closure property

# Confidence

Given: $\{i_1, i_2, \ldots, i_k\} \rightarrow j$

The confidence of this association rule is the probability of j given $I = \{i_1, \ldots, i_k\}$

$$\mathrm{conf}(I \rightarrow j) = \frac{\mathrm{support}(I \cup j)}{\mathrm{support}(I)}$$

# Confidence

$$\text{conf}(I \rightarrow j) = \frac{\text{support}(I \cup j)}{\text{support}(I)}$$

| tid | Set of items |
|---|---|
| 1 | Bread, Jam, Juice |
| 2 | Tofu, Juice, Tomatoes |
| 3 | Bread, Strawberries, Tofu, Juice |
| 4 | Tofu, Juice, Tomatoes |
| 5 | Strawberries, Juice, Tomatoes |

conf({tofu,juice} —> tomatoes)?