# Association Rules Demo

Data Mining
Prof. Dawn Woodard
School of ORIE
Cornell University

# Reading

Reading is Chap.13 in SPB

# Association Rules Demo

- Install the “arules” R package, which implements association rules.
- Load the “arules” package into R
- Load the “Epub” data, which is in the “arules” package:

  $>$ data(Epub)

# Association Rules Demo

- Get information about the “Epub” data:

  $>$ help(Epub)

  What do these data represent? How many transactions (user sessions) are in the dataset? How many electronic publications have been downloaded?

- Unlike our previous datasets, the Epub object is not a data.frame. It is an object of class “transactions”, which is a class defined in “arules” for datasets of purchase transactions:

  $>$ class(Epub)

- There are special functions defined in “arules” for manipulating and viewing “transactions” objects, such as “summary”, “length”, and “image” functions.

# Association Rules Demo

Get more details about the data by calling summary(Epub)

```
transactions as itemMatrix in sparse format with
 15729 rows (elements/itemsets/transactions) and
 936 columns (items) and a density of 0.001758755

most frequent items:
doc_11d doc_813 doc_4c6 doc_955 doc_698 (Other)
    356     329     288     282     245   24393

element (itemset/transaction) length distribution:
sizes
    1     2     3     4     5     6     7     8     9    10    11    12
11615  2189   854   409   198   121    93    50    42    34    26    12
   14    15    16    17    18    19    20    21    22    23    24    25
   10     6     8     6     5     8     2     2     3     2     3     4
   27    28    30    34    36    38    41    43    52    58
    1     1     1     2     1     2     1     1     1     1

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   1.000   1.000   1.646   2.000  58.000

includes extended item information - examples:
   labels
1 doc_11d
2 doc_13d
3 doc_14c

includes extended transaction information - examples:
      transactionID           TimeStamp
10792  session_4795 2003-01-01 20:59:00
10793  session_4797 2003-01-02 07:46:01
10794  session_479a 2003-01-02 10:50:38
```

# Association Rules Demo

- You can check the # of transactions by calling length(Epub)
- You can get subsets of the transactions using the same format as for vectors:

  > Epub[1:10000]

- Take a look at the first 5 transactions using the “inspect” function:

  > inspect(Epub[1:5])

```
> inspect(Epub[1:5])
  items      transactionID            TimeStamp
1 {doc_154}   session_4795 2003-01-01 20:59:00
2 {doc_3d6}   session_4797 2003-01-02 07:46:01
3 {doc_16f}   session_479a 2003-01-02 10:50:38
4 {doc_11d,
   doc_1a7,
   doc_f4}    session_47b7 2003-01-02 18:55:50
5 {doc_83}    session_47bb 2003-01-02 21:27:44
```

# Association Rules Demo

In what year did the last session in this dataset occur?

**(A)** 2003

**(B)** 2008

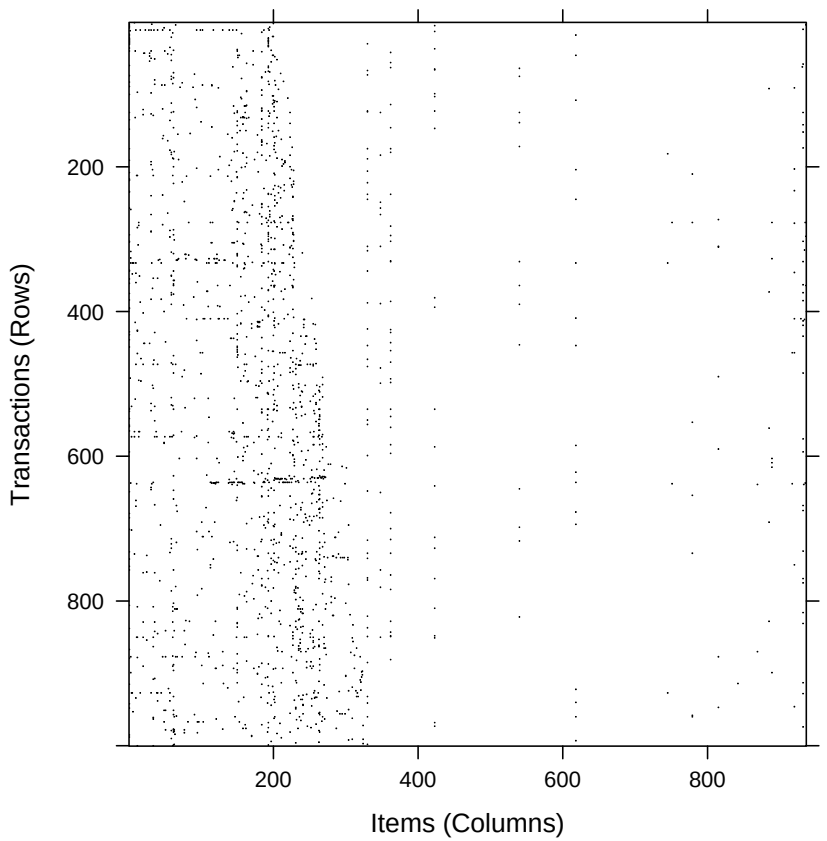
**(C)** 2010

**(D)** 2012

# Association Rules Demo

- You can visualize the transaction data using the “image” function. Do this for the first 1,000 transactions in the dataset:

  > image( Epub[1:1000] )



# Association Rules Demo

- Association rules can be found using the “apriori” function. Look at the help file for this function.
- Focus on the arguments “data” and “parameter”. “parameter” specifies the support threshold for the apriori algorithm, the cutoff value for the confidence, and other quantities.
- First try just specifying “data”, since “parameter” has some default values.

  $>$ rules = apriori( data = Epub )
- Then print out the association rules by calling

  $>$ inspect(rules)
- There’s a problem (what output did you get?); the method did not produce any association rules.

# Association Rules Demo

- What are the default values of “support” and “confidence” for the parameter argument?
- Why are these values causing a problem for this dataset?
- For a document that was downloaded 10 times, the support of the single-item set containing that document is $10/15729 = .0006$. Let’s try using this as the cutoff for the support.
- What value of the confidence cutoff to use? Let’s choose it so that we get a moderate # of association rules that we can inspect manually. A cutoff of .8 produces 48 rules; let’s use that.

# Association Rules Demo

```
> rules = apriori( data = Epub, parameter = list( support = .0006, confidence = .8 ) )
> inspect( rules )
```

```
   lhs                rhs            support confidence      lift
1  {doc_c21,
    doc_cce}       => {doc_c69} 0.0006357683  0.8333333 189.96377
2  {doc_6e8,
    doc_6e9}       => {doc_6e7} 0.0010808062  0.8947368 402.09474
3  {doc_6e7,
    doc_6e9}       => {doc_6e8} 0.0010808062  0.8500000 417.80156
4  {doc_6e7,
    doc_6e8}       => {doc_6e9} 0.0010808062  0.8095238 454.75000
5  {doc_3c4,
    doc_764}       => {doc_574} 0.0006357683  0.8333333 257.00980
6  {doc_574,
    doc_764}       => {doc_3c4} 0.0006357683  0.8333333 267.50000
7  {doc_3c4,
    doc_574}       => {doc_4b4} 0.0008264988  0.8125000 190.74347
8  {doc_3c4,
    doc_764}       => {doc_4b4} 0.0006993452  0.9166667 215.19776
9  {doc_574,
    doc_764}       => {doc_4b4} 0.0006993452  0.9166667 215.19776
10 {doc_3fc,
    doc_800}       => {doc_803} 0.0007629220  0.9230769 191.04049
11 {doc_5be,
    doc_649}       => {doc_5ca} 0.0006357683  0.9090909 170.22727
12 {doc_972,
    doc_a77}       => {doc_8f9} 0.0007629220  0.8000000  65.53750
13 {doc_26b,
    doc_9b}        => {doc_424} 0.0006357683  0.9090909 148.94886
14 {doc_26b,
```