Informing Amazon Recommendations from Co-purchasing Networks

Joe Zhang, Samuel Hamilton, Shirley Liu, Aria Wang

Background

- Traditional advertising relies on advertising high-value products to a general population.

- This paradigm is less effective with online retailers, who can supply a larger variety of products. Top 20% of the products from online resources only amount to half of sales.

- Data-driven targeted advertising algorithms offer a solution to connecting consumers to niche products while keeping costs low.

Goals

- Promotions can drive volume while giving the consumer a sense of value. Therefore, we aim to implement a system that identifies products for personalized bundles, and recommendations.

- Sub-aims:
    - Identify product groups most amenable to bundling.
    - Develop an algorithm to recommend products most likely to drive sales.

Data source and description

- **Dataset name**: Amazon product co-purchasing network

- **Dataset source**: Leskovec, J. "Large Network Dataset Collection." Stanford University. http://snap.stanford.edu/data/.

- **Collection date**: March 2, 2003

- Data information: Network was collected by crawling Amazon website. It is based on the "Customers Who Bought This Item Also Bought" feature of the Amazon website. Each node represents a product. If a product i is frequently co-purchased with product j, the graph contains a directed edge from i to j.

Data overview

Network data

```
FromNodeId    ToNodeId
0       1
0       2
0       3
0       4
0       5
1       0
1       2
1       4
1       5
1       15
2       0
2       11
2       12
2       13
2       14
3       63
3       64
3       65
3       66
3       67
4       7
4       16
4       17
```
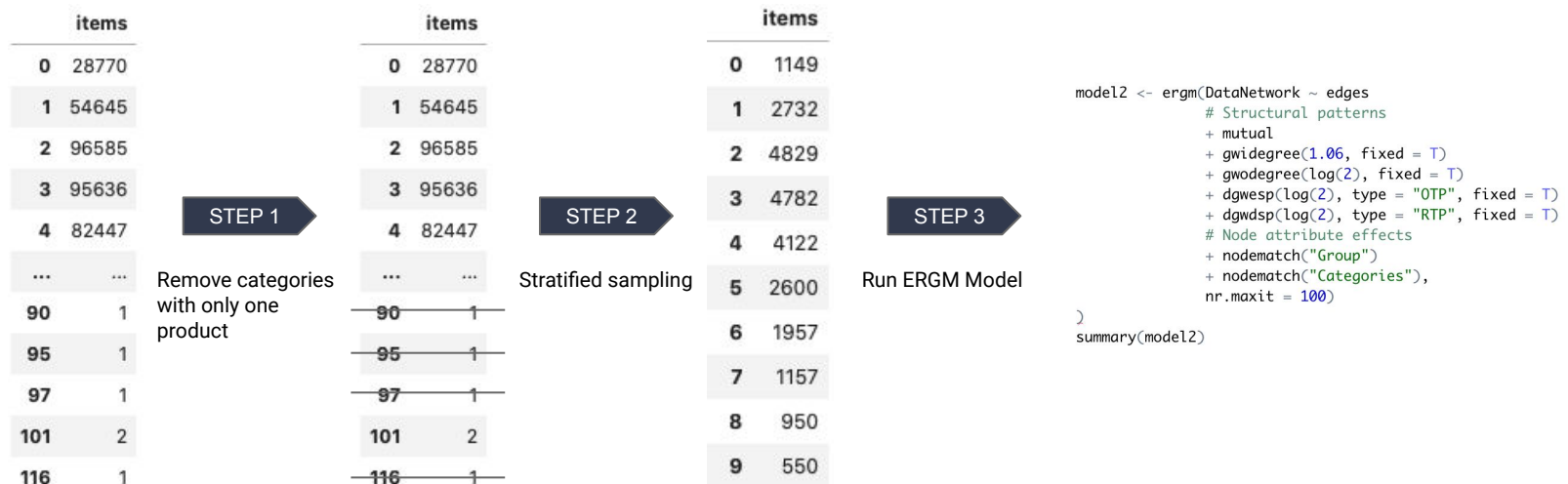
Product metadata

```
Id:    1
ASIN: 0827229534
  title: Patterns of Preaching: A Sermon Sampler
  group: Book
  salesrank: 396585
  similar: 5  0804215715  156101074X  0687023955  0687074231  082721619X
  categories: 2
    |Books[283155]|Subjects[1000]|Religion & Spirituality[22]|Christianity[12290]|Clergy[12360]|Preaching[12368]
    |Books[283155]|Subjects[1000]|Religion & Spirituality[22]|Christianity[12290]|Clergy[12360]|Sermons[12370]
  reviews: total: 2  downloaded: 2  avg rating: 5
    2000-7-28  cutomer: A2JW670Y8U6HHK  rating: 5  votes:  10  helpful:   9
    2003-12-14  cutomer: A2VE83MZF98ITY  rating: 5  votes:   6  helpful:   5

Id:    2
ASIN: 0738700797
  title: Candlemas: Feast of Flames
  group: Book
  salesrank: 168596
  similar: 5  0738700827  1567184960  1567182836  0738700525  0738700940
  categories: 2
    |Books[283155]|Subjects[1000]|Religion & Spirituality[22]|Earth-Based Religions[12472]|Wicca[12484]
    |Books[283155]|Subjects[1000]|Religion & Spirituality[22]|Earth-Based Religions[12472]|Witchcraft[12486]
  reviews: total: 12  downloaded: 12  avg rating: 4.5
    2001-12-16  cutomer: A11NCO6YTE4BTJ  rating: 5  votes:   5  helpful:   4
    2002-1-7  cutomer:  A9CQ3PLRNIR83  rating: 4  votes:   5  helpful:   5
    2002-1-24  cutomer: A13SG9ACZ905IM  rating: 5  votes:   8  helpful:   8
    2002-1-28  cutomer: A1BDAI6VEYMAZA  rating: 5  votes:   4  helpful:   4
```

Aim 1 - Identify product groups most amenable to bundling

- **Methods**:

  - Collect a stratified random sample of the data for computational feasibility.
  - Conduct an ERGM to identify the qualities of nodes which drive co-purchasing.

	items
0	28770
1	54645
2	96585
3	95636
4	82447
...	...
90	1
95	1
97	1
101	2
116	1

STEP 1 Remove categories with only one product

	items
0	28770
1	54645
2	96585
3	95636
4	82447
...	...
90	1
95	1
97	1
101	2
116	1

STEP 2 Stratified sampling

	items
0	1149
1	2732
2	4829
3	4782
4	4122
5	2600
6	1957
7	1157
8	950
9	550

STEP 3 Run ERGM Model

```
model2 <- ergm(DataNetwork ~ edges
        # Structural patterns
        + mutual
        + gwidegree(1.06, fixed = T)
        + gwodegree(log(2), fixed = T)
        + dgwesp(log(2), type = "OTP", fixed = T)
        + dgwdsp(log(2), type = "RTP", fixed = T)
        # Node attribute effects
        + nodematch("Group")
        + nodematch("Categories"),
        nr.maxit = 100)
)
summary(model2)
```

Aim 1 - Identify product groups most amenable to bundling

Monte Carlo MLE results:

	Estimate	Standard error	MCMC %	Z value	Probability
edges	-15.01333	0.24800	0	-60.537	<1e-04***
mutual	12.83730	0.41725	8	30.766	<1e-04***
gwideg.fixed	1.71338	0.20363	0	8.414	<1e-04***
gwodeg.fixed	5.74803	0.21441	0	26.809	<1e-04***
gwesp.OTP.fixed	3.18092	0.35574	3	8.942	<1e-04***
gwdsp.RTP.fixed	8.30550	0.58314	3	14.243	<1e-04***
nodematch.Group	0.06144	0.07483	1	0.821	0.412
nodematch.Categories	0.11775	0.10740	1	1.096	0.273

- Co-purchasing network demonstrate in/out-degree bias
- Groups and categories are not the best predictors of co-purchasing

# Aim 1 - Identify product groups most amenable to bundling

Goodness-of-fit for model statistics:

	obs	min	mean	max	p-value
edges	3076.000	3014	3142	3266	0.22
mutual	849	849	931	1003	0.01
gwideg.fixed	2826	2755	2865	2966	0.47
gwodeg.fixed	2909	2850	2961	3062	0.17
gwesp.OTP.fixed	297	228	251	297	0.01
gwdsp.RTP.fixed	143	143	163	192	0.01
nodematch.Group	1792	1715	1797	1884	0.89
nodematch.Categories	372	344	381	417	0.61

- The observed edges, indegree, outdegree, nodematch.group and nodematch.categories are not significantly different from the simulated value while the observed mutual, OTP and RTP are.
- The model is generally a good one.
- **Conclusion**: Recommend highly co-purchased products, not similar products.

Aim 2: Identify products to promote

- **Goal**: After a product makes a purchase, recommend products based on that purchase and the products that with the highest centrality themselves.
- **Results**:
    - Network is highly fractured
    - Over a subset of 3584 nodes, there were 1364 components.
    - Can develop a product shop based off of these small components
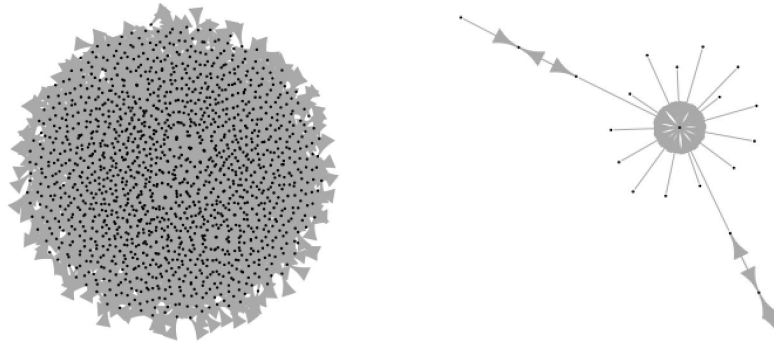
Conclusions

- Network qualities, such as in-degree bias, have a much larger impact on consumer purchasing than the type of product itself.

- Components within the network can be used to produce customized shops after a product purchase.

Thank you!