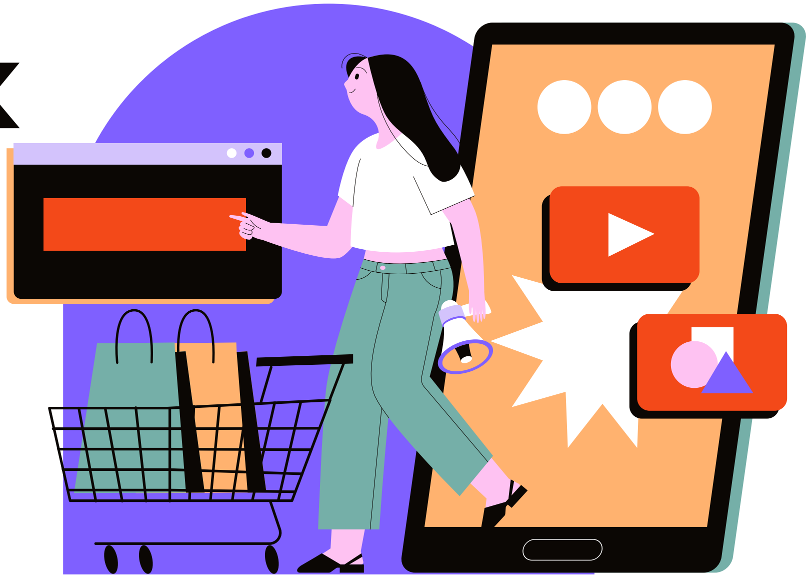
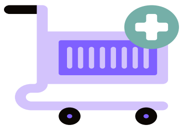
Social Network Analysis Project (SNAP)



Amazon Product Co-Purchasing Data



Original Dataset

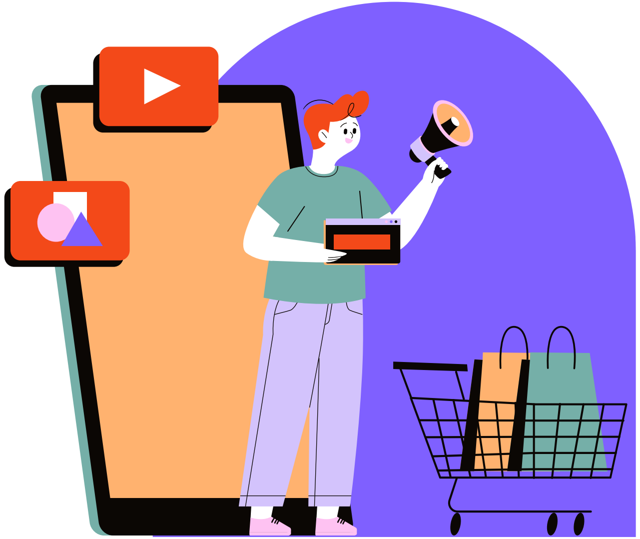
On June 1st 2003, web crawling was conducted on the Amazon e-commerce site in order to create a network based on the *Customers Who Bought This Item Also Bought* feature on the site.

- 403,494 nodes
- 3,387,388 directed edges

Subset

A highly interconnected cluster from the original dataset was selected to improve processing efficiency

- 355 nodes
- 812 directed edges



Problem Statement

Can Amazon optimize their recommended products based on what products are frequently bought with other products in order to encourage more sales?

Questions and Methodology

Questions	Methodology
What products are frequently purchased together?	Community Detection
Are there certain products that result in longer tailed connections/more associated purchases?	Node-level metrics (eigenvector centrality)
What products are the most influential (lead to purchasing the most number of other products)?	Node-level metrics (degree centrality)
Can purchasing patterns be predicted? For example, is there a tendency for item A to be bought with item B if item A is commonly bought with many other products?	ERGM

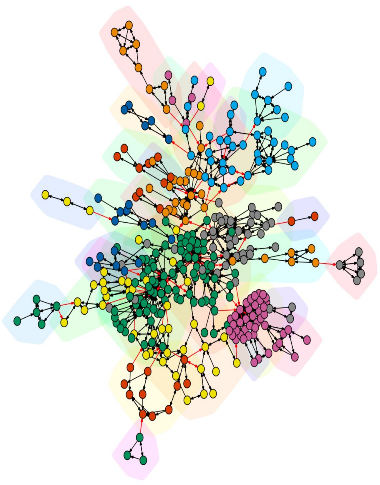
What products are purchased together?

Community Structure

Modality score of 0.8198 and 32 clusters

High modularity score indicates a well-defined community structure.

32 somewhat distinct groups of products that are purchased together

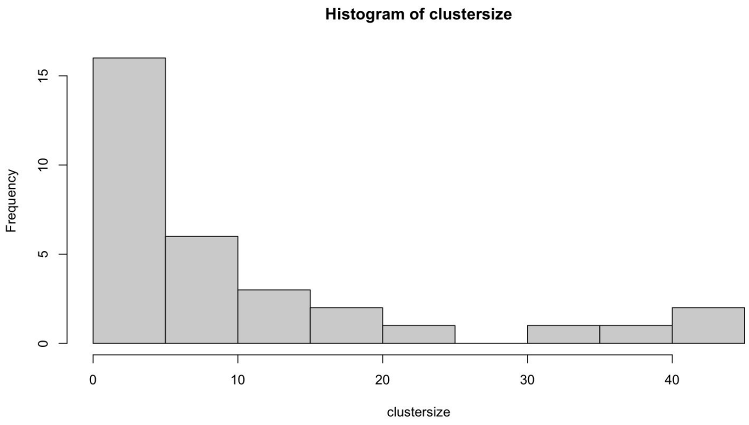


Cluster Sizes

Community sizes range from 1- 45

There are some clusters that only consist of one or two nodes, suggesting that some products are purchased alone or in very small groups.

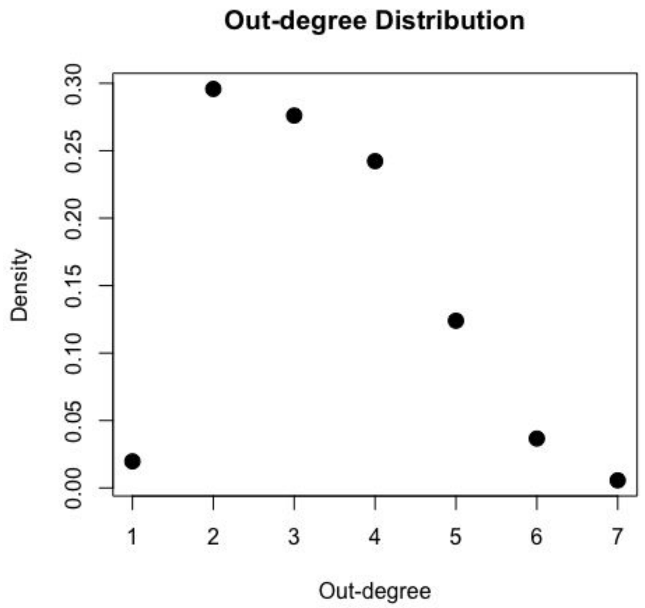
A miscellaneous category should be made or certain categories may need to be grouped together



Outdegree Distribution

Most products are purchased in groups of 3 and no nodes are truly isolated from the graph

May explain why Amazon recommends “Frequently bought together” items in groups of 3.



Product Bundling

- Products within the same cluster can be bundled up to drive more sales of each product through a combined bundle
- Bundles of 3 are ideal
- Further steps can be taken to identify the types of products within each cluster

Which products are the most influential?

Centrality

Highest eigenvector centrality nodes: 27656, 24718, and 27570

- Both highly purchased and lead to the purchase of other highly purchased items

Highest in-degrees nodes: 26776 with 39, 21722 with 36, and 27565 with 30

- Good add-on products to recommend to customers

Highest betweenness centrality nodes: 83673, 83672, and 43519

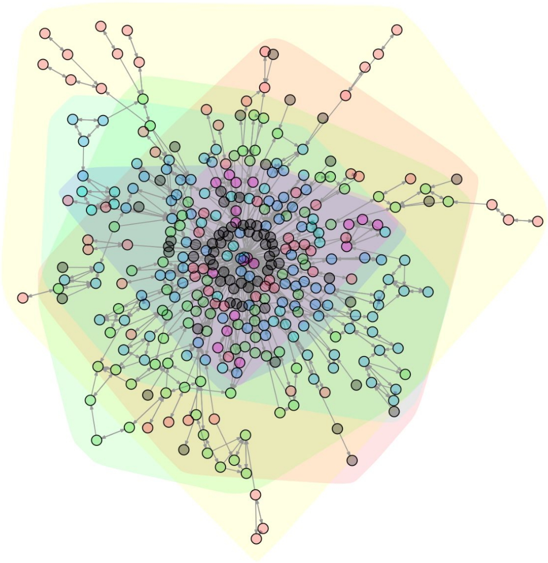
- May guide customers to other different types of products and can increase the variety of products that customers are exposed to

Nodes with high authority scores include 27565, 24718, and 27570. Nodes with high hub scores include 326674, 143365, and 237829.

- High authority scores are products that are highly purchased, and products with elevated hub scores serve as pivotal points, recommending customers to a diverse range of reputable products.

K Core Decomposition

- Higher k-cores are more influential in a network and nodes that are more central have a higher k-core.
- Maximum k-core in our data was 6, suggesting that each node has at most 6 direct co-purchasing ties to other nodes in the subgraph
- Network is relatively interconnected
- Nodes with a degree of 6 include: 16770, 16771, 35897, and 35898



Can purchasing patterns be predicted?

Hypotheses

Hypothesis 1 (edges): Holding everything else constant, the probability that a tie exists between any two products will be very low.

Hypothesis 2 (mutual): Holding everything else constant, if item A is purchased with item B, it will be more likely that item B is purchased with item A.

Hypothesis 3 (gwodeg): Holding everything else constant for all, there is a tendency for item A to be bought with item B if item A is commonly bought with many other products.

Hypothesis 4 (gwideg): Holding everything else constant for all, there is a tendency for item A to be bought with item B if many items are commonly purchased with item B.

ERGM findings

Hypothesis 1 (edges): Supported

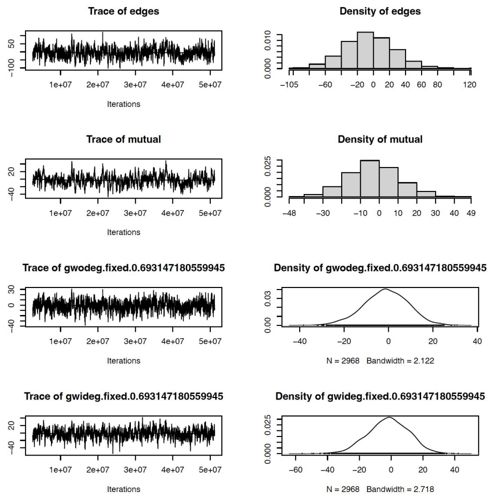
Hypothesis 2 (mutual): Supported

Hypothesis 3 (gwodeg): Contradicted

Hypothesis 4 (gwideg): Supported

```
> summary(model1)
Call:
ergm(formula = item ~ edges + mutual + gwodegree(log(2), fixed = T) +
    gwidegree(log(2), fixed = T))

Monte Carlo Maximum Likelihood Results:

                                  Estimate Std. Error MCMC % z value Pr(>|z|)
edges                              -5.9002     0.1104      0 -53.420   <1e-04 ***
mutual                              6.4791     0.1524      0  42.526   <1e-04 ***
gwodeg.fixed.0.693147180559945      2.6619     0.2892      0   9.203   <1e-04 ***
gwideg.fixed.0.693147180559945     -2.8390     0.1365      0 -20.805   <1e-04 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

     Null Deviance: 174216  on 125670  degrees of freedom
 Residual Deviance:   7702  on 125666  degrees of freedom

AIC: 7710  BIC: 7749  (Smaller is better. MC Std. Err. = 3.086)
```

ERGM Model Diagnostics



Convergence

Goodness-of-fit for in-degree

	obs	min	mean	max	MC p-value
idegree0	91	99	120.420	140	0.00
idegree1	103	36	54.590	77	0.00
idegree2	63	25	41.915	63	0.01
idegree3	37	19	37.590	60	0.94
idegree4	26	13	32.605	49	0.24
idegree5	9	17	27.240	40	0.00
idegree6	9	9	18.525	31	0.02
idegree7	2	4	11.530	20	0.00
idegree8	0	0	6.085	14	0.01
idegree9	3	0	2.850	8	1.00
idegree10	3	0	1.070	5	0.18
idegree11	0	0	0.425	2	1.00
idegree12	1	0	0.115	2	0.22
idegree13	2	0	0.035	1	0.00
idegree15	1	0	0.005	1	0.01
idegree17	1	0	0.000	0	0.00
idegree23	1	0	0.000	0	0.00
idegree30	1	0	0.000	0	0.00
idegree36	1	0	0.000	0	0.00
idegree39	1	0	0.000	0	0.00

Goodness-of-fit for out-degree

	obs	min	mean	max	MC p-value
odegree0	7	3	10.695	22	0.35
odegree1	105	68	90.385	113	0.06
odegree2	98	103	124.920	152	0.00
odegree3	86	56	77.995	100	0.35
odegree4	44	18	34.195	52	0.10
odegree5	13	2	12.025	22	0.87
odegree6	2	0	3.625	9	0.61
odegree7	0	0	0.920	6	0.91
odegree8	0	0	0.190	2	1.00
odegree9	0	0	0.045	1	1.00
odegree11	0	0	0.005	1	1.00

Goodness-of-fit for edgewise shared partner

	obs	min	mean	max	MC p-value
esp.OTP0	388	621	776.245	838	0
esp.OTP1	256	2	22.905	112	0
esp.OTP2	128	0	1.900	41	0
esp.OTP3	37	0	0.235	7	0
esp.OTP4	3	0	0.000	0	0

Goodness-of-fit for minimum geodesic distance

	obs	min	mean	max	MC p-value
1	812	723	801.285	849	0.71
2	948	1580	1970.190	2280	0.00
3	710	2780	4363.630	5666	0.00
4	563	4710	8549.040	11612	0.00
5	442	7058	13451.595	17212	0.00
6	343	9168	15536.630	18356	0.00
7	272	9803	12868.525	15529	0.00
8	199	4769	8010.635	11483	0.00
9	153	1516	4058.235	8506	0.00
10	112	312	1796.530	6104	0.00
11	111	33	725.960	3946	0.10
12	44	0	276.900	2433	0.26
13	33	0	103.165	1443	0.87
14	20	0	38.635	878	0.63
15	0	0	14.375	482	0.96
16	0	0	5.440	254	1.00
17	0	0	2.005	126	1.00
18	0	0	0.765	76	1.00
19	0	0	0.310	42	1.00
20	0	0	0.155	25	1.00
21	0	0	0.055	10	1.00
22	0	0	0.010	2	1.00
23	0	0	0.005	1	1.00
Inf	120908	45026	53095.925	62107	0.00

Goodness-of-fit for model statistics

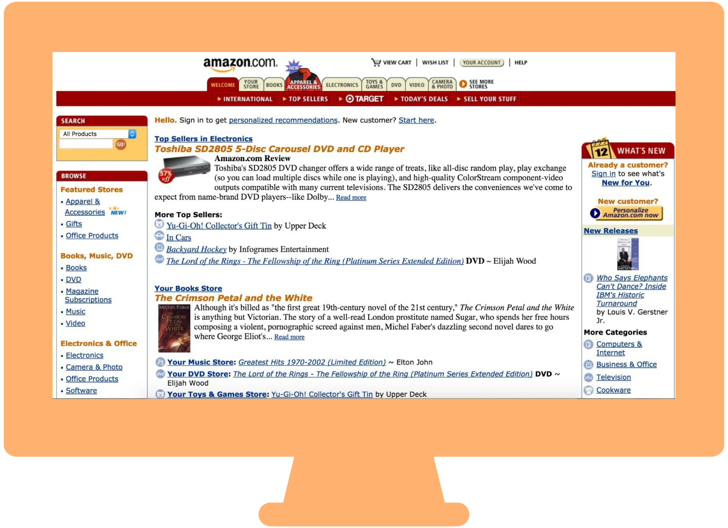
	obs	min	mean	max	MC p-value
edges	812.0000	723.0000	801.2850	849.0000	0.71
mutual	212.0000	184.0000	206.7050	227.0000	0.60
gwodeg.fixed.0.693147180559945	514.1250	484.3438	511.1110	530.7031	0.78
gwideg.fixed.0.693147180559945	380.1064	345.3550	377.6166	403.5508	0.84

Goodness-of-fit Test Results

What can we do with this information and where can we go from here?

Recommendations

- Website and recommendation optimization
- Add-on recommendations
- Explore trends within clusters
- Further network research



Website Optimization

Improved Navigation

- Strategic placement of popular co-purchasing items
- Improved tab system based on categories of products that belong to the larger clusters of our network

Recommendations

- Add more centralized products to the top recommended products list!

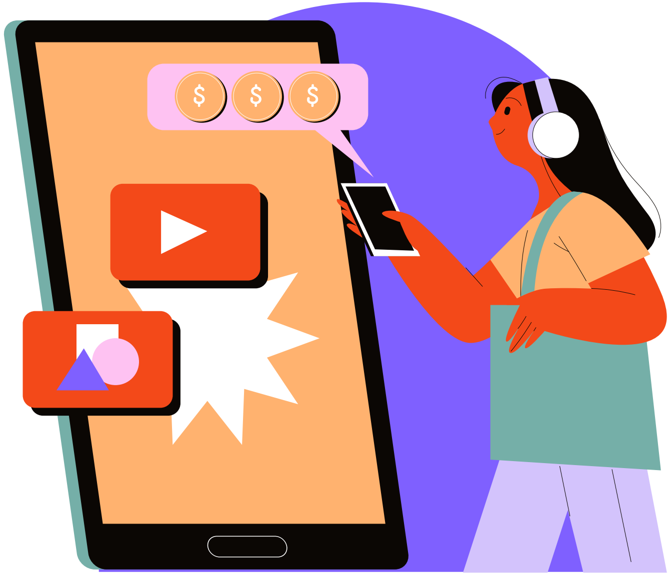
Amazon’s Picks

- Products with high eigenvector centrality are important!

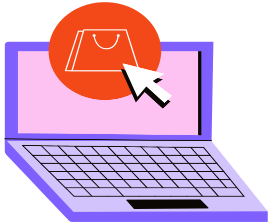
- “Amazon’s Picks” labeling system that prioritize products with high degree centrality

Amazon Prime Day

Prime Day exclusive bundles of commonly bought together products



Supply Chain and Inventory Management



- Low connectivity products should be less emphasised
- Relay this information to their vendors
- Ease the confusion of such a large catalogue of products and create a more streamlined and user friendly website interface

Further Network Analysis

Larger Scale

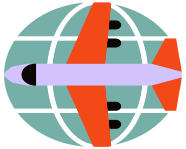
More information can be found within a larger dataset

With Amazon’s computing power this can be done!

Recent Dataset

The dataset we worked on was created in 2003

A more recent dataset would give you more relevant information



Regional Networks

Since 2003, Amazon has gone global!

Taking datasets for certain regions can help create more specific recommendations

Thank you!