# i. Background:

In online shopping, shopping sequence data is available - buying item category A and then B, C, D, etc. This forms a network in shopping data. Chart 2 shows an item category network in Amazon purchasing data. Buying "video game" is followed by "Travel" book, "Home and Garden" and so on. On the right hand side of Chart 2, "Indie Music[266023]" plays broker role in that removing this splits network into two components. Chart 1 shows network in amazon0302 data. It shows star network with high density at central area. **The purpose of this paper is to segment such a Core Item Group.** This core group is clearly different from peripheral outlier nodes as you can see in Chart 1. The core group is also a broker who is not single node but actually very large group so that we could think it as core group surrounded by smaller peripheral group. Core group has strong tie within the group meaning that customers are more likely to do co-purchasing within the group where peripheral group doesn't. This helps you to understand promo and sales strategy.

# ii. Data Set:

Amazon co-purchasing data (http://snap.stanford.edu/data/index.html#amazon) contains network data and meta data. Network data is simply 2 columns table – "FROM" and "TO" where co-purchasing indicates co-purchased item "TO" given that 1st item purchased as in "FROM". Meta data is attributes list such as category and review. We start to use "amazon0302" with amazon-meta to illustrate procedure.

| Name | Type | Nodes | Edges | Description | Data Size |
|---|---|---|---|---|---|
| amazon0302 | Directed | 262,111 | 1,234,877 | Amazon product co-purchasing network from March 2 2003 | 1,234,877 |
| amazon0312 | Directed | 400,727 | 3,200,440 | Amazon product co-purchasing network from March 12 2003 | 13,420,742 |
| amazon0505 | Directed | 410,236 | 3,356,824 | Amazon product co-purchasing network from May 5 2003 | 14,059,847 |
| amazon0601 | Directed | 403,394 | 3,387,388 | Amazon product co-purchasing network from June 1 2003 | 14,219,665 |
| amazon-meta | Metadata | 548,552 | 1,788,725 | Amazon product metadata: product info and all reviews on around 548,552 products. | 15,010,574 |

**amazon0xxx data attributes**

| | |
|---|---|
| **FROM** | Purchased item |
| **TO** | Co-Purchased item |

**amazon-meta data attributes**

| | |
|---|---|
| **Id** | Id of item product |
| **ASIN** | Amazon Standard Identification Numbers |
| **title** | Title of item product |
| **group** | Highest category |
| **salesrank** | Sales ranking |
| **categories** | 7 levels of category from highest 1 to lower sub category |

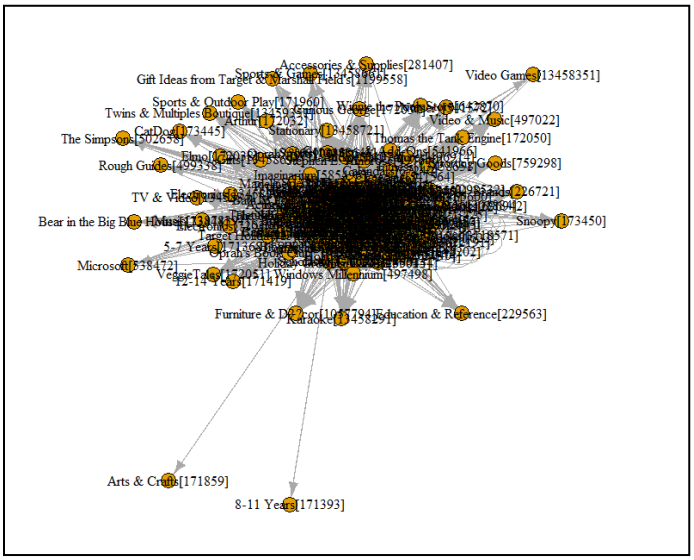
### ii-i. Item Category

Raw data – item product is too detail to compute and to configure broker relationship. Going up to subcategory level 4 and sizing down to manageable level of data is done to create network. Actual data used is level 4 item subcategory. Table 3 shows a sample of the subcategories used.

### ii-ii. Compact Data Size

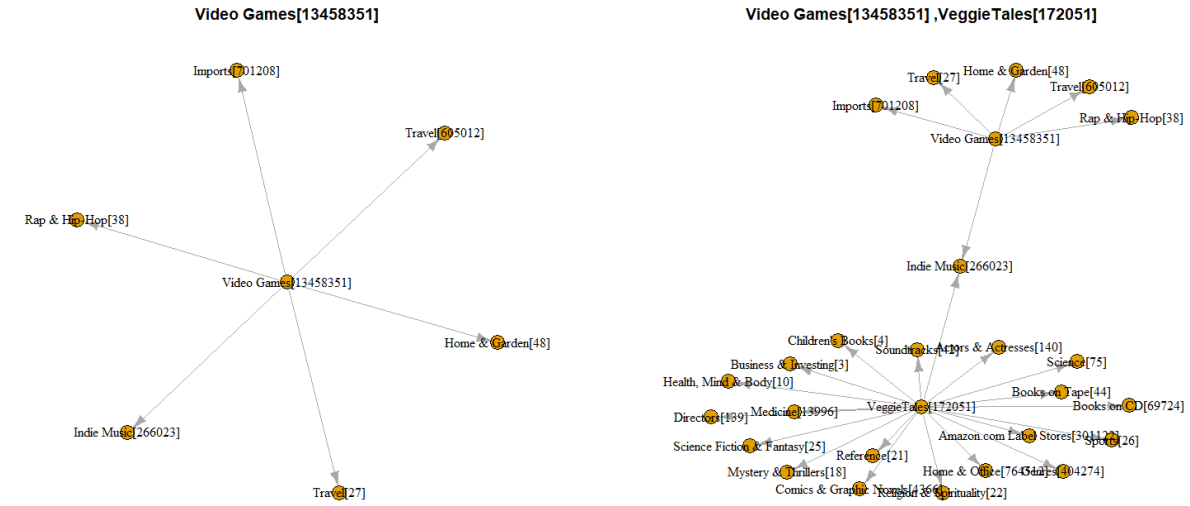
Compact data size is required in data cleansing besides removing duplicated vertices and self-looping. Chart 1 shows raw data network whose center area is very congested or dense.

**Chart 1**




Some network visualization techniques were pretested but due to high degree and density levels, it in fact failed to produce broker group in network. Amazon shopping data needs to compact by reasonable logic. High density level makes it very difficult to configure broker and separate components in network. Reconsider network generating process as below.

**Chart 2**




Online shopping sequence data gives bundling purchase or co-purchasing information. It gives a key to compact size of item category network. “Video Game” purchase leads to “Travel” book purchase, “Indies Music” and so on as shown in Table1 below. This forms a network as in Chart 2 above. Including “VeggieTales” category given that “Video Game” network in presence forms a new network accommodating more other categories as shown in right hand side of Chart 2. Adding one category after another, repeatedly adding some categories, covers up all categories. Pick a category from “FROM”. This means customer purchase an item from an item category then buy an item from another item

category from “TO” as Table 1 shows. That is, “Video Game” is connected to other 6 categories. This means smaller subset of categories “FROM” can form complete set of category after adding some categories in that the formed network is connected through co-purchasing.

**Table1**

| FROM | | | TO | | Transaction Count |
|---|---|---|---|---|---|
| **Video Games** | **Book** | **→** | **All Other Categories** | | **6 (Total)** |
| Video Games | Book | → | Travel | Book | 1 |
| Video Games | Book | → | Imports | Music | 1 |
| Video Games | Book | → | Indie Music | Music | 1 |
| : | : | : | : | : | : |

**Table2**

| FROM | | | TO | | Transaction Count |
|---|---|---|---|---|---|
| **VeggieTales** | **Video** | **→** | **All Other Categories** | | **22 (Total)** |
| VeggieTales | Video | → | Health, Mind & Body | Book | 2 |
| VeggieTales | Video | → | Actors & Actresses | Video | 1 |
| VeggieTales s | Video | → | Indie Music | Music | 1 |
| : | : | : | : | : | : |

Compacting network data seeks for smaller subset data with support of more number of transactions. Table 2 shows that more transactions start from “VeggieTales” category comparing to 6 transactions in “Video Game” in Table 1. It makes reasonable foundation that “VeggieTales” is more important category

Table 3 shows process of adding some categories to complete all categories and lists up top category by transaction count until all 155 categories covered. The largest number or 252,128 of transactions go through “Nonfiction[53]” and we define it as the biggest major category. Starting with “Nonfiction[53]” category, create its network and constitutes of 137 category network as Table 3 shows. Add next biggest category Children's Books[4] and repeat this process one by one to complete network until all 155 categories covered as shown in Table 3.

**Table3**

| | Category as in “FROM” | Transaction Count | Category Coverage | Core |
|---|---|---|---|---|
| 1 | Nonfiction[53] | 252,128 | 137 | 40 |
| 2 | Children's Books[4] | 210,172 | 140 | 40 |
| 3 | Religion & Spirituality[22] | 194,952 | 143 | 40 |
| 4 | Professional & Technical[173507] | 191,342 | 146 | 40 |
| 5 | Literature & Fiction[17] | 186,123 | 149 | 40 |
| 6 | Home & Office[764512] | 166,629 | 150 | 40 |
| 7 | Travel[605012] | 164,721 | 150 | 40 |
| 8 | Health, Mind & Body[10] | 152,178 | 150 | 40 |
| 9 | Amazon.com Outlet[517808] | 142,977 | 151 | 40 |
| 10 | Indie Music[266023] | 135,493 | 152 | 40 |
| 11 | Business & Investing[3] | 130,993 | 152 | 40 |
| 12 | History[9] | 126,651 | 152 | 40 |
| 13 | Reference[21] | 121,956 | 152 | 40 |

| | | | | |
|---|---|---|---|---|
| 14 | Science[75] | 116,650 | 152 | 40 |
| 15 | Genres[404274] | 116,026 | 153 | 40 |
| 16 | Special Features[408328] | 92,729 | 153 | 40 |
| 17 | Genres[404276] | 91,048 | 153 | 40 |
| 18 | Computers & Internet[5] | 84,491 | 153 | 40 |
| 19 | Home & Garden[48] | 84,139 | 153 | 40 |
| 20 | Entertainment[86] | 83,891 | 153 | 40 |
| 21 | Biographies & Memoirs[2] | 80,876 | 155 | 40 |

It is found that top major 21 item categories from “FROM” can form complete 155 category network all over as shown in Table 3. Using this compacted data to recreate network. Total of transactions of this compacted dataset is 2,926,165 out of 5,241,631 or 56% of total. This compact data set is used to build a network and clarify separate peripheral item categories.

# iii. Analysis:

Since we limit “FROM” categories with top 21 item categories only, its network truncates many vertices from less important categories such as category having only 12 transactions. Chart 3 is the compact data set network.

**Chart 3**

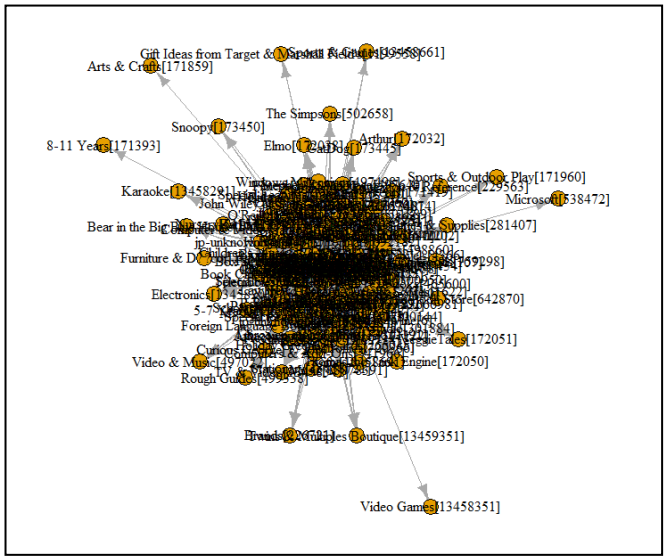



**Chart 4**

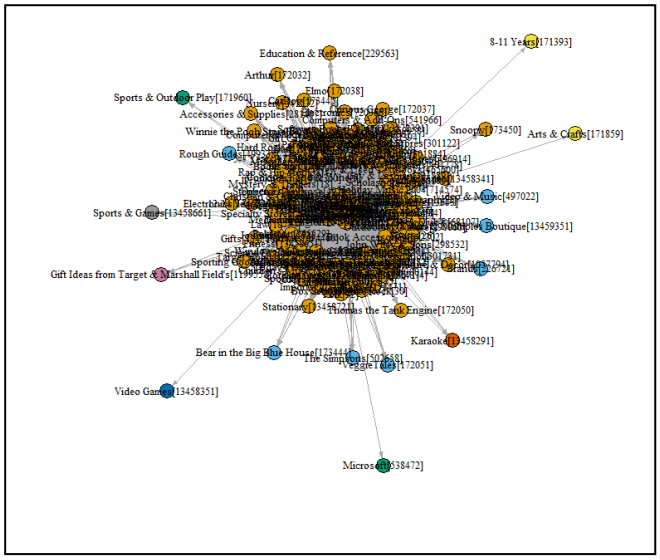
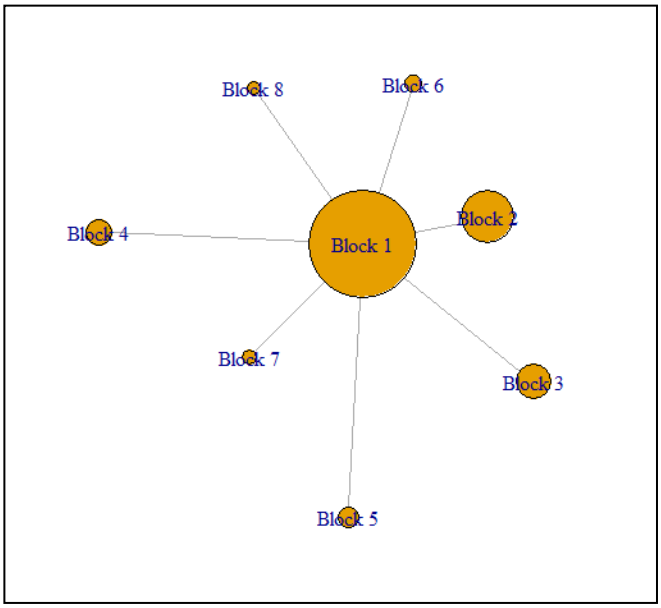



Chart 4 shows CONCOR approach to segment item category and Table 5 is its segmentation list. Block 1 is the largest segmentation constituting of broker group surrounded by other blocks and Block 1 is defined as **Core Item Category**. Block 2-8 is defined as **Peripheral Item Category**. Relationship between block 1 and other blocks is shown in Chart 5. Block 1 is seen as central segmentation. Chart 6 is network after removing block 1 and shows block 1 is also broker segmentation.

## Chart 5




## Chart 6


## Table4

| | ALL | Broker |
|---|---|---|
| size | 155 | 140 |
| Distance | 1.47 | 1.37 |
| Diameter | 3 | 3 |
| degree centralization | 0.35 | 0.31 |
| closeness centralization | 0.19 | 0.18 |
| betweenness centralization | 0.03 | 0.02 |
| degree centralization sd | 88.1 | 76.89 |
| closeness centralization sd | 0.13 | 0.13 |
| betweenness centralization sd | 122.14 | 77.13 |
| Density | 0.54 | 0.63 |
| Average Degree | 167.2 | 175.79 |
| Cohesion | 1 | 1 |
| Compactness | 0.71 | 0.76 |
| Global Clustering Coefficient | 0.84 | 0.87 |
| Mean Core | 136.45 | 145.12 |
| Median Core | 180 | 178 |

## Table5

| block | Item Category | block | Item Category | block | Item Category |
|---|---|---|---|---|---|
| 1 | Business & Investing[3] | 1 | Books on Tape[44] | 1 | New & Used Textbooks[465600] |
| 1 | Computers & Internet[5] | 1 | Books, Music & More[559958] | 1 | New Age[36] |
| 1 | Entertainment[86] | 1 | Boxed Sets[13622191] | 1 | O'Reilly[69860] |
| 1 | Home & Office[764512] | 1 | Calendars[67240] | 1 | Opera & Vocal[84] |
| 1 | Professional & Technical[173507] | 1 | Camera & Photo[502394] | 1 | Oprah's Book Club®[68107] |
| 1 | Amazon.com Outlet[517808] | 1 | Classical[85] | 1 | Oprah®[68100] |
| 1 | Genres[404276] | 1 | Comics & Graphic Novels[4366] | 1 | Outdoors & Nature[290060] |
| 1 | Special Features[408328] | 1 | Computer & Video Game Books[696942] | 1 | Pokémon[265529] |
| 1 | Biographies & Memoirs[2] | 1 | Cooking, Food & Wine[6] | 1 | Pop[37] |
| 1 | Children's Books[4] | 1 | Curious George[172037] | 1 | Romance[23] |
| 1 | Health, Mind & Body[10] | 1 | Engineering[13643] | 1 | Scholastic[571658] |
| 1 | History[9] | 1 | Feeding[541564] | 1 | Science Fiction & Fantasy[25] |
| 1 | Home & Garden[48] | 1 | Foreign Language Books[3118571] | 1 | Sheet Music & Scores[1622] |
| 1 | Literature & Fiction[17] | 1 | Gay & Lesbian[301889] | 1 | Software Books[727676] |
| 1 | Nonfiction[53] | 1 | Gifts[13458691] | 1 | Special Features[498864] |
| 1 | Reference[21] | 1 | Hard Rock & Metal[67207] | 1 | Sports[26] |
| 1 | Religion & Spirituality[22] | 1 | Holiday Greeting Cards[265045] | 1 | Star Wars[281446] |
| 1 | Science[75] | 1 | Horror[49] | 1 | Stationary[13458721] |
| 1 | Travel[605012] | 1 | HOWdesign Studio[301884] | 1 | Stephen E. Ambrose[497640] |
| 1 | Genres[404274] | 1 | Imaginarium[585496] | 1 | Teens[28] |
| 1 | Indie Music[266023] | 1 | John Wiley & Sons[298532] | 1 | Teletubbies[265530] |
| 1 | Gift Categories[13984131] | 1 | Journals[642562] | 1 | The Simpsons[502700] |
| 1 | Nursery[541572] | 1 | jp-unknown1[1061348] | 1 | Travel[27] |
| 1 | Parenting & Families[20] | 1 | jp-unknown2[1061350] | 1 | Windows Millennium[497498] |
| 1 | Accessories[265040] | 1 | jp-unknown3[1061352] | 1 | Electronics[13458201] |
| 1 | Arthur[172032] | 1 | Large Print[300950] | 1 | Actors & Actresses[404278] |
| 1 | Arts & Photography[1] | 1 | Law[10777] | 1 | Alternative Rock[30] |
| 1 | Barbie[171471] | 1 | Libros en español[301731] | 1 | Blues[31] |
| 1 | Bargain Books[45] | 1 | Madeline[13729211] | 1 | Broadway & Vocalists[265640] |
| 1 | Bath & Potty[541586] | 1 | McGraw-Hill[300144] | 1 | Christian & Gospel[173429] |
| 1 | Book Accessories[642202] | 1 | Medicine[13996] | 1 | Classic Rock[67204] |
| 1 | Book Clubs[292203] | 1 | Miscellaneous[35] | 1 | Country[16] |
| 1 | Books on CD[69724] | 1 | Mystery & Thrillers[18] | 1 | Dance & DJ[7] |
| block | Item Category | block | Item Category | | |

| | | | |
|---|---|---|---|
| 1 | Directors[403502] | 1 | 5-7 Years[171368] |
| 1 | Education & Reference[229563] | 1 | Music[13878391] |
| 1 | Electronics[720366] | 1 | Actors & Actresses[140] |
| 1 | Fitness[13458651] | 1 | CatDog[173445] |
| 1 | Folk[32] | 1 | Directors[139] |
| 1 | Formats[692182] | 1 | Specialty Stores[498860] |
| 1 | Formats[694466] | 1 | Sporting Goods[759298] |
| 1 | International[33] | 1 | Thomas the Tank Engine[172050] |
| 1 | Jazz[34] | 1 | TV & Video[13458341] |
| 1 | Latin Music[289122] | 2 | Bear in the Big Blue House[173444] |
| 1 | R&B[39] | 2 | Twins & Multiples Boutique[13459351] |
| 1 | Rap & Hip-Hop[38] | 2 | Brands[226721] |
| 1 | Rock[40] | 2 | Rough Guides[499338] |
| 1 | Soundtracks[42] | 2 | The Simpsons[502658] |
| 1 | Specialty Stores[498862] | 2 | Video & Music[497022] |
| 1 | Sports[13458641] | 2 | VeggieTales[172051] |
| 1 | Target Holiday[13666981] | 3 | Microsoft[538472] |
| 1 | Accessories & Supplies[281407] | 3 | Sports & Outdoor Play[171960] |
| 1 | Amazon.com Label Stores[301122] | 4 | 8-11 Years[171393] |
| 1 | Box Sets[291920] | 4 | Arts & Crafts[171859] |
| 1 | Children's Music[173425] | 5 | Video Games[13458351] |
| 1 | Computers & Add-Ons[541966] | 6 | Karaoke[13458291] |
| 1 | DVD-Audio[574074] | 7 | Gift Ideas from Target & Marshall Field's[1199558] |
| 1 | Elmo[172038] | 8 | Sports & Games[13458661] |
| 1 | Furniture & Décor[1057794] | | |
| 1 | Holiday Music 2004[714574] | | |
| 1 | Imports[701208] | | |
| 1 | Snoopy[173450] | | |
| 1 | Special Features[496914] | | |
| 1 | Today's Deals in Music[287454] | | |
| 1 | Winnie the Pooh Store[642870] | | |
| 1 | 12-14 Years[171419] | | |

Table 6 shows statistics for 4 Amazon data – All Category and Core Item Category. Core Item Category has shorter distance and lower degree centralization which means single node is less likely to be central

actor. Standard Deviation is also lower meaning within core item category more homogeneous and so density level is higher. Average degree is also higher with higher cluster coefficient. It gives “small world” nature in Core Item Category network.

**Table6**

| | amazon0302 | | amazon0312 | | amazon0505 | | amazon0601 | |
|---|---|---|---|---|---|---|---|---|
| | **ALL** | **CORE** | **ALL** | **CORE** | **ALL** | **CORE** | **ALL** | **CORE** |
| **size** | 155 | 140 | 161 | 149 | 160 | 153 | 161 | 146 |
| **Distance** | 1.47 | 1.37 | 1.43 | 1.37 | 1.42 | 1.38 | 1.42 | 1.34 |
| **Diameter** | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| **degree centralization** | 0.35 | 0.31 | 0.37 | 0.33 | 0.36 | 0.33 | 0.35 | 0.3 |
| **closeness centralization** | 0.19 | 0.18 | 0.21 | 0.2 | 0.22 | 0.21 | 0.2 | 0.18 |
| **betweenness centralization** | 0.03 | 0.02 | 0.02 | 0.02 | 0.03 | 0.02 | 0.02 | 0.02 |
| **degree centralization sd** | 88.1 | 76.89 | 89.66 | 80.96 | 89.17 | 83.69 | 90.01 | 78.22 |
| **closeness centralization sd** | 0.13 | 0.13 | 0.13 | 0.13 | 0.13 | 0.13 | 0.14 | 0.13 |
| **betweenness centralization sd** | 122.14 | 77.13 | 100.99 | 77.2 | 110.22 | 83.38 | 106.84 | 72.46 |
| **Density** | 0.54 | 0.63 | 0.57 | 0.63 | 0.58 | 0.62 | 0.57 | 0.66 |
| **Average Degree** | 167.2 | 175.79 | 182.82 | 186.05 | 184.8 | 187.74 | 183.9 | 191.01 |
| **Cohesion** | 1 | 1 | 3 | 3 | 2 | 2 | 0 | 2 |
| **Compactness** | 0.71 | 0.76 | 0.72 | 0.75 | 0.73 | 0.75 | 0.72 | 0.77 |
| **Global Clustering Coefficient** | 0.84 | 0.87 | 0.83 | 0.85 | 0.83 | 0.84 | 0.83 | 0.86 |
| **Mean Core** | 136.45 | 145.12 | 145.73 | 149.24 | 147.04 | 149.8 | 145.1 | 152.88 |
| **Median Core** | 180 | 178 | 186 | 182 | 186 | 184 | 184 | 182 |

Table 7 lists up Core Item Category over 4 Amazon data. Top 11 Core Item Category is very stable over 4 Amazon data. Again Core Item Category connects many other item categories each other in relatively smaller world. However, there exist outlier item category - Peripheral Item Category shown in Table 8. Core Item Category network is more chance to connect each other within while it is less chance to connect Peripheral Item Category. Unlike to Core Item Category, Peripheral Item Category varies over 4 Amazon data.

**Table7 Core Item Category**

| amazon0302 | amazon0312 | amazon0505 | amazon0601 |
|---|---|---|---|
| Nonfiction[53] | Nonfiction[53] | Nonfiction[53] | Nonfiction[53] |
| Children's Books[4] | Children's Books[4] | Children's Books[4] | Children's Books[4] |
| Religion & Spirituality[22] | Religion & Spirituality[22] | Religion & Spirituality[22] | Religion & Spirituality[22] |
| Professional & Technical[173507] | Professional & Technical[173507] | Professional & Technical[173507] | Professional & Technical[173507] |
| Literature & Fiction[17] | Literature & Fiction[17] | Literature & Fiction[17] | Literature & Fiction[17] |
| Home & Office[764512] | Home & Office[764512] | Home & Office[764512] | Home & Office[764512] |

| | | | |
|---|---|---|---|
| Travel[605012] | Travel[605012] | Travel[605012] | Travel[605012] |
| Health, Mind & Body[10] | Health, Mind & Body[10] | Health, Mind & Body[10] | Health, Mind & Body[10] |
| Amazon.com Outlet[517808] | Amazon.com Outlet[517808] | Amazon.com Outlet[517808] | Amazon.com Outlet[517808] |
| Indie Music[266023] | Indie Music[266023] | Indie Music[266023] | Indie Music[266023] |
| Business & Investing[3] | Business & Investing[3] | Business & Investing[3] | Business & Investing[3] |
| History[9] | History[9] | History[9] | |
| Reference[21] | Reference[21] | | |
| Science[75] | Science[75] | | |
| Genres[404274] | Genres[404274] | | |
| Special Features[408328] | | | |
| Genres[404276] | | | |
| Computers & Internet[5] | Computers & Internet[5] | | |
| Home & Garden[48] | Home & Garden[48] | | |
| Entertainment[86] | Entertainment[86] | | |
| Biographies & Memoirs[2] | | | |

Core Item Category Network is easy to connect to all other item categories through co-purchasing except

**Table8 Peripheral Item Category**

| amazon0302 | amazon0312 | amazon0505 | amazon0601 |
|---|---|---|---|
| Bear in the Big Blue House[173444] | Bear in the Big Blue House[173444] | | |
| Twins & Multiples Boutique[13459351] | | | |
| Brands[226721] | | | |
| Rough Guides[499338] | | | |
| The Simpsons[502658] | | | |
| Video & Music[497022] | Video & Music[497022] | | Video & Music[497022] |
| VeggieTales[172051] | | | VeggieTales[172051] |
| Microsoft[538472] | Microsoft[538472] | | |
| Sports & Outdoor Play[171960] | | | Sports & Outdoor Play[171960] |
| 8-11 Years[171393] | | 8-11 Years[171393] | 8-11 Years[171393] |
| Arts & Crafts[171859] | | Arts & Crafts[171859] | Arts & Crafts[171859] |
| Video Games[13458351] | | | |
| Karaoke[13458291] | | | Karaoke[13458291] |

| | | | |
|---|---|---|---|
| Gift Ideas from Target & Marshall Field's[1199558] | Gift Ideas from Target & Marshall Field's[1199558] | | |
| Sports & Games[13458661] | Sports & Games[13458661] | | |
| | Curious George[172037] | | |
| | Sporting Goods[759298] | | |
| | Mr. & Mrs. Potato Head[173448] | | Mr. & Mrs. Potato Head[173448] |
| | TV & Video[13458341] | TV & Video[13458341] | TV & Video[13458341] |
| | Exercise & Fitness[14053631] | Exercise & Fitness[14053631] | Exercise & Fitness[14053631] |
| | Sport[3678571] | Sport[3678571] | Sport[3678571] |
| | Sports Equipment[3395101] | Sports Equipment[3395101] | Sports Equipment[3395101] |
| | | Batman[13964181] | Education & Reference[229563] |
| | | | The Simpsons[502658] |
| | | | Grownups[171439] |

# iv. Final Remark:

Core Item Category discussed here is an alternative analytic metric in retail business. In grocery business, for example, lower level of category is not manageable in its size ranging from 800 to 3,000 categories. Several core item category modeling were invented in different business scenario such as top 10 categories and top 30% sales categories. Business analyst uses such a segmentation simply for conventional but they may or may not be appropriate segmentation. There is no guarantee that top 30% sales categories draw entire business picture nor show how important the segmentation plays a role in retail analytics.

Efficient category core segmentation is minimal subset from all categories, represents entire business picture and has a clear objective. Minimal size of segmentation is usually manageable. 50 categories is rather usable than 800 categories. Top 10 categories completely ignore rest of categories in that analyst even doesn't look at rest of categories at all. How top 30% sales categories contraction makes impact on entire business is unknown.

Core Item Category is guaranteed to connect majority of frequent co-purchasing items and provide outlier items information in network structure form. Business must pay greatest attention to Core Item Category and special attention to Peripheral Item Category. For example most effective (spreading effect) campaign is to start with Core Item Category.