TOWARDS AUTOMATED EXTRACTION OF PRODUCT ATTRIBUTES

Hoa Minh Ly (hmly@suffolk.edu) and Dmitry Zinoviev (dzinoviev@suffolk.edu), Mathematics and Computer Science

WORDS IN RETAIL PRODUCT DESCRIPTIONS: IN 3 5G #2 UNIVERSAL 8" 1-1/2" PFJ POLY 6" PREM VALVE LIGHT SOCKET BIT 2G TEE PLANTER PW BROWN AIR KEY CLEAR TROPICAL PROMO COUPLING 5 POT ASSORTED COVER BLACK FLAT BLADE 1" DRIVE 10" #1 EA HANDLE PLASTIC OIL ROUND SET MINI CEDAR PH 10PK LED 2PK NATURAL PALM FAN 18" FLUSH HINGE WITH ASST HD SCREW BATH TOP MACH 1/4" RND SHEET PINE W/ PEX 6PNL OZ ORCHID ELBOW QT OAK HOOK DOUBLE PK BASE ROLLER BASKET & FOR BRONZE PERENNIAL POINSETTIA HOSE MDF PT EXT SPRAY TAPE SATIN 1/2" PLUG S/O BLK HEX 1G NUT - HB #10 WALLPLT C[illegible] 2" 1/4 P[illegible] W[illegible]OW [illegible] BRUSH 1/1 GOLD 12 GAL LILY STEEL MAPLE 10 LOCK MED 2 PVC TRIM WHT SCR X SCREEN BULB ROOF SQ 36" 3G LF WASHER ANNUAL 3PK SHOWER TOOL RH AM 3/8" PBF NICKEL 6 ALUM CORNER KNOB #6 2" IS WHITE END FLOOR FOLIAGE 1 FIR DOOR VINYL SEED LT PANEL FT 3/4" WALL TREE 4IN CHROME 4 FAUCET SQUARE 4PK 9/16 6IN TILE BOARD 36"X80" RIDGID MIX ROSE GRAY ZINC DR DEEP 15A #3 DUTY BOX 5/8 3/4 SHADE 24" PIPE PC SS [illegible] BLUE PRO LB AND WD [illegible] GARDEN METAL SAW LH PREMIUM CASING FL CH GLASS GALV PAINT 4" GAS 1-1/4" CABLE PACK KW ALUMINUM BN DECO GE BREAKER WOOD RING CARPET METRIC BLIND PATIO ALL 3/8 KIT WIRE [illegible] 8 FILTER 5" SN BAR CAP CABINET BRASS POST SOLID PULL RED NIPPLE 11/16 SHELF HDX ROD

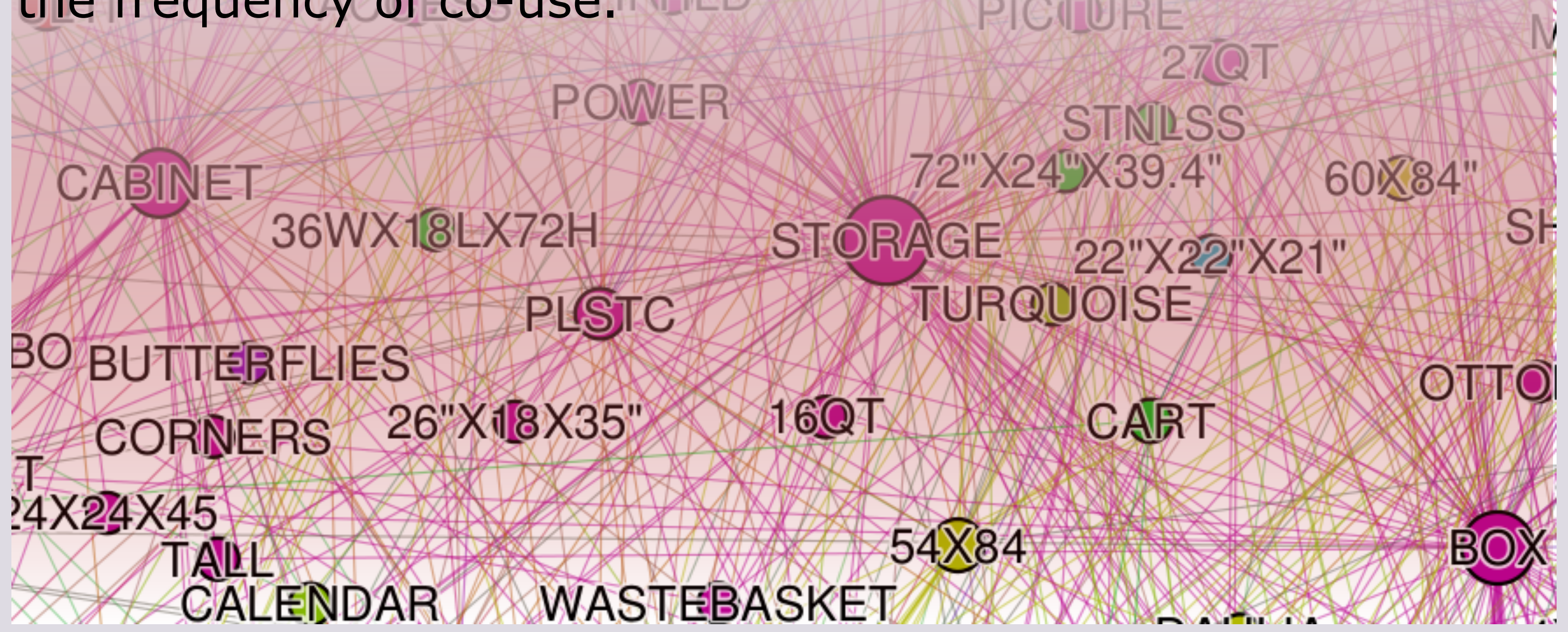
ORANGE PLASTIC **MASONRY TRAY**

attributes product

CAN PRODUCT ATTRIBUTES BE DETECTED AUTOMATICALLY?

CONSTRUCT A NETWORK OF WORDS

A network node represents a word. Two nodes are connected if the words are used together in at least one product description. The strength of the connection is proportional to the frequency of co-use.



CALCULATE NETWORK MEASURES

Being an attribute is a word's role. An automatic role extraction algorithm groups nodes based on their network measures: degree, weight, assortativity, clustering coefficient, betweenness, closeness, and eigenvector centrality.

	Assort.	Clust.	Degree	Eigenv.	Weigh	Betw.	Clos.
ORANGE	.00510	.00136	REMOVED	.01098	REMOVED	.01200	.00608
MASONRY	.00006	.00000		.00026		.00000	.00016
OAK	.00433	.00092		.03869		.03701	.03365
PLASTIC	.00604	.00087		.05961		.07239	.04583
TRAY	.00764	.00213		.00679		.00871	.00224
CLOCK	.00652	.00178		.01228		.01656	.00833

REMOVE CORRELATED MEASURES

Some network measures in a complex network may be strongly correlated or anticorrelated and, therefore, redundant. We removed two redundant measures: Degree and Weight.

	Assort.	Clust.	Degree	Eigenv.	Weight	Betw.	Clos.
Assortativity	1.00	0.11	-0.08	0.07	-0.07	-0.06	0.45
Clustering	0.11	1.00	-0.44	-0.49	-0.34	-0.21	-0.39
Degree	-0.08	-0.44	1.00	0.94	0.93	0.88	0.44
Eigenvector	0.07	-0.49	0.94	1.00	0.85	0.74	0.56
Weight	-0.07	-0.34	0.93	0.85	1.00	0.85	0.35
Betweenness	-0.06	-0.21	0.88	0.74	0.85	1.00	0.25
Closeness	0.45	-0.39	0.44	0.56	0.35	0.25	1.00

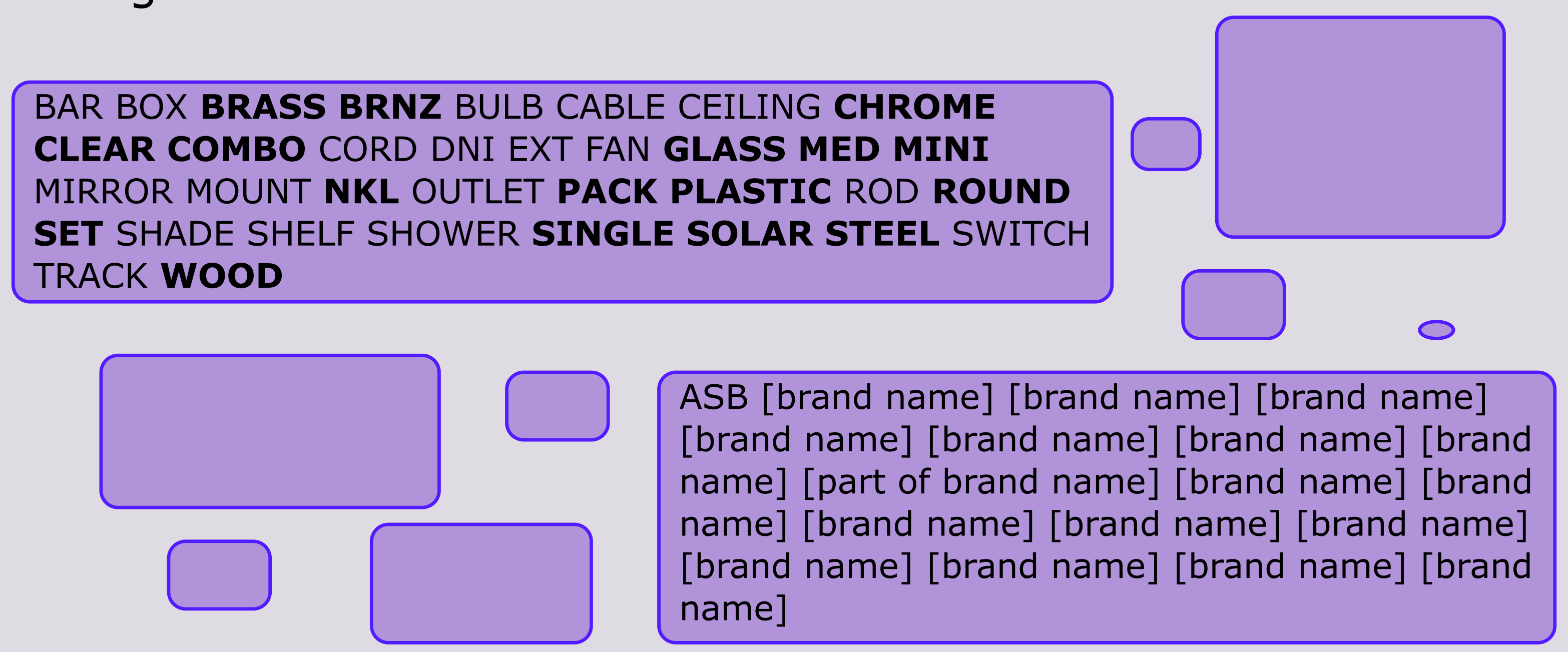
BUILD A VECTOR SPACE

The collection of all surviving network measures represents the "coordinates" of a node in a multi-dimensional vector space.

...	
ORANGE:	{.00510, .00136, .01098, .01200, .01200, .00608}
MASONRY:	{.00006, .00000, .00026, .00000, .00000, .00016}
OAK:	{.00433, .00092, .03869, .03701, .03701, .03365}
PLASTIC:	{.00604, .00087, .05961, .07239, .07239, .04583}
TRAY:	{.00764, .00213, .00679, .00871, .00871, .00224}
CLOCK:	{.00652, .00178, .01228, .01656, .01656, .00833}
...	

PERFORM CLUSTERING

Node vectors can be clustered using the k-mean algorithm. We chose the number of clusters (37) to maximize the likelyhood of having an attribute-rich cluster.



BAR BOX **BRASS BRNZ** BULB CABLE CEILING **CHROME CLEAR COMBO** CORD DNI EXT FAN **GLASS MED MINI** MIRROR MOUNT **NKL** OUTLET **PACK PLASTIC** ROD **ROUND SET** SHADE SHELF SHOWER **SINGLE SOLAR STEEL** SWITCH TRACK **WOOD**

ASB [brand name] [brand name] [brand name] [brand name] [brand name] [brand name] [brand name] [part of brand name] [brand name] [brand name] [brand name] [brand name] [brand name] [brand name] [brand name] [brand name] [brand name]

LOCATE ATTRIBUTE-RICH CLUSTERS

The attribute words have been previously hand-picked.

Cluster ID	Size (S)	# of atts (A)	p=A/S	Z-score
19	283	143	.51	4.84
10	148	72	.49	3.03
33	565	241	.43	2.95
18	139	67	.48	2.82
17	46	25	.54	2.49
12	564	233	.41	2.29
14	584	240	.41	2.22
...				
32	316	88	.28	-3.25
11	67	8	.12	-4.20

TAKE-HOME MESSAGE: ATTRIBUTES DIFFER FROM OTHER WORDS, BUT AT THE MOMENT CANNOT BE IDENTIFIED AUTOMATICALLY