Constructing and Exploring Composite Items

Gautam Das
University of Texas at Arlington

Joint work with

Senjuti Basu-Roy (UTA), Sihem Amer-Yahia (Yahoo! Research), Ashish Chawla (Yahoo! Inc), and Cong Yu (Google)

What's a composite item?

User searches for an iPhone
Price budget: $350

Central item (iPhone) + Satellite package
Total price: $322.80

Wireless Car Charger $14.95

Kroo Case $14.95

iPhone 3G/8GB $99

Touch Penn $19.95

Portable Bose Sounddock $149

iKlear Spray Kit $24.95

Item Bundles on Amazon

Frequently Bought Together

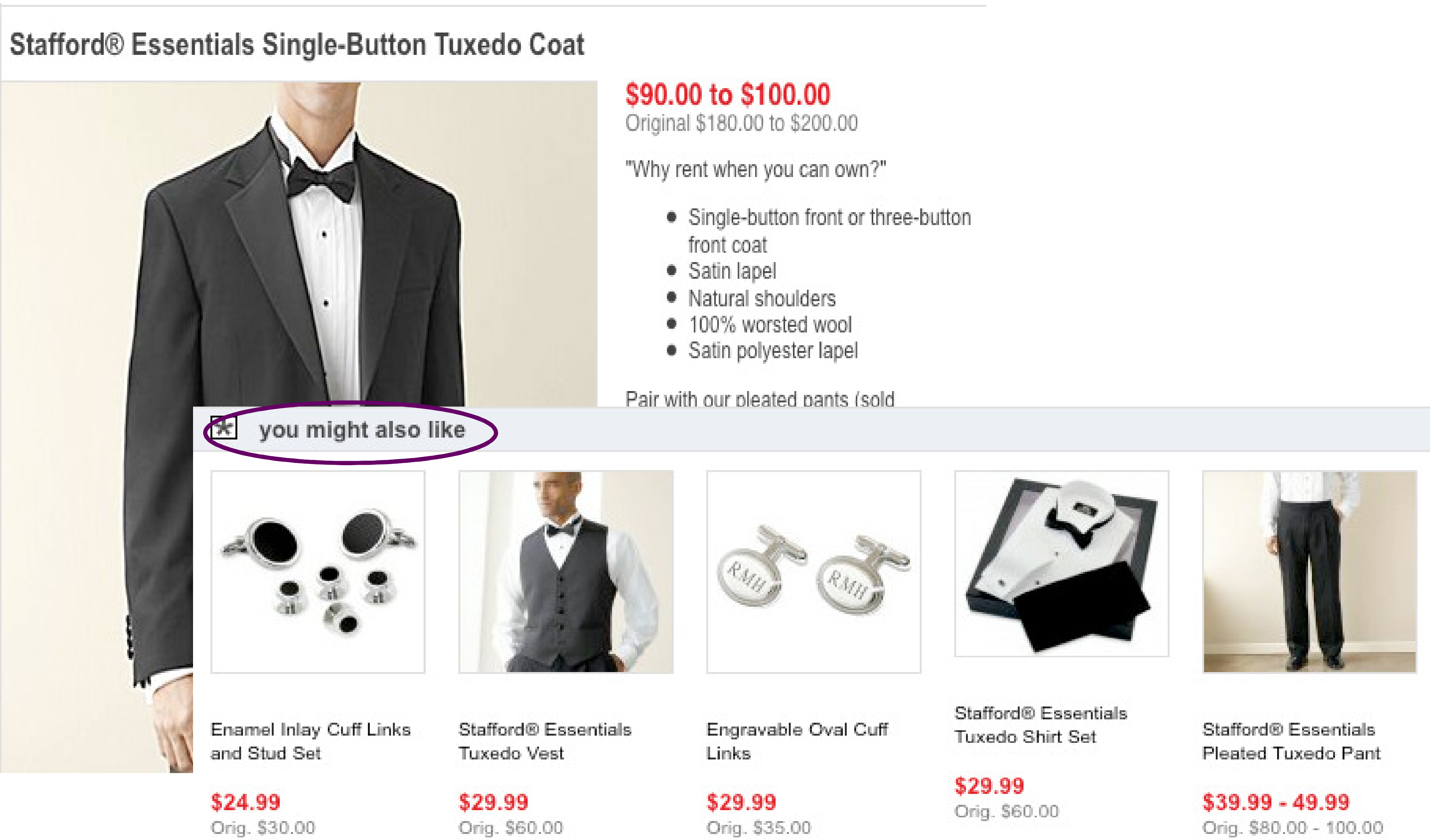
Total List Price: ~~$45.94~~
Price For All Three: **$31.77**

Add all three to Cart | Add all three to Wish List

Show availability and shipping details

- ☑ **This item:** The Harafish by Naguib Mahfouz
- ☑ Children of the Alley: A Novel by Peter Christopher Theroux
- ☑ The Yacoubian Building: A Novel by Humphrey T. Davies

Complimentary Items: JCPenney



JCPenney (budget b = $175)

Total price $174.97

Total price $169.98

"Star" Composite Retrieval Problem

Given a user query ***Q (central item c, a budget b)*** retrieve top-k *compatible* satellite packages s.t. *the total cost is within budget.*

Composite Retrieval Requirements

- **Properties of a single composite item**
  - **type coverage:** maximize user's exposure to as many instances of different satellite types as possible
  - **compatibility:** satellite-central or satellite-satellite
  - **validity:** total cost of central item and compatible package is within budget (e.g., price, time)
  - **maximality:** build the largest valid package

Composite Retrieval Requirements

- **Properties of all k composite items**
  - **diversity:** maximize user's exposure to as many different satellite items as possible and minimize overlap between composite items
- **Cannot be captured with a ranking function**

Maximal Star Package (budget = $350)

iPhone 3G 8GB $99

+

Kroo Case $14.95

+

Car Charger $14.95

+

Touch Penn $19.95

+

Portable Bose Sounddock $149

+

iKlear Spray Kit $24.95

= $322.8

forms a **valid** composite item with *iPhone 3G/8GB* as does any strict subset of this package.

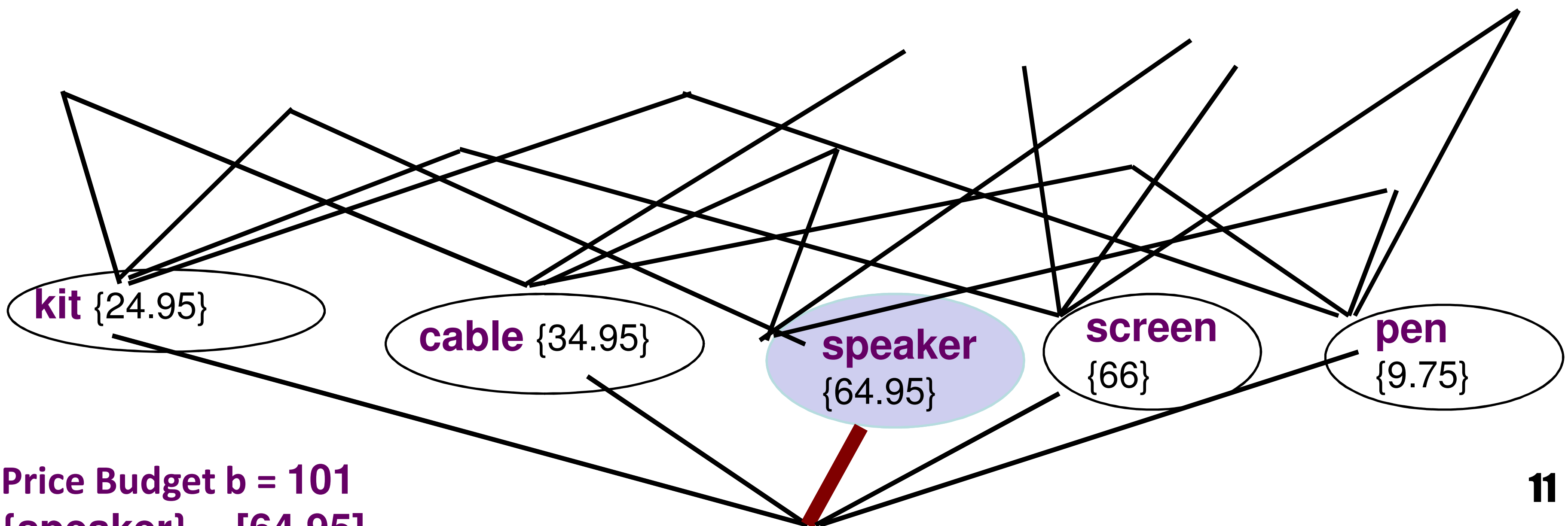
forms a **maximal** package. Addition of any new item violates validity

Maximal Packages Construction

- Maximal packages construction problem bears resemblance to maximal Frequent Itemset Mining
  - items need to be constructed subject to compatibility and budget constraints (as opposed to simply extracted from transaction database)
  - support requires scanning transaction database while budget is checked locally
  - A-priori algorithm too costly for real-time

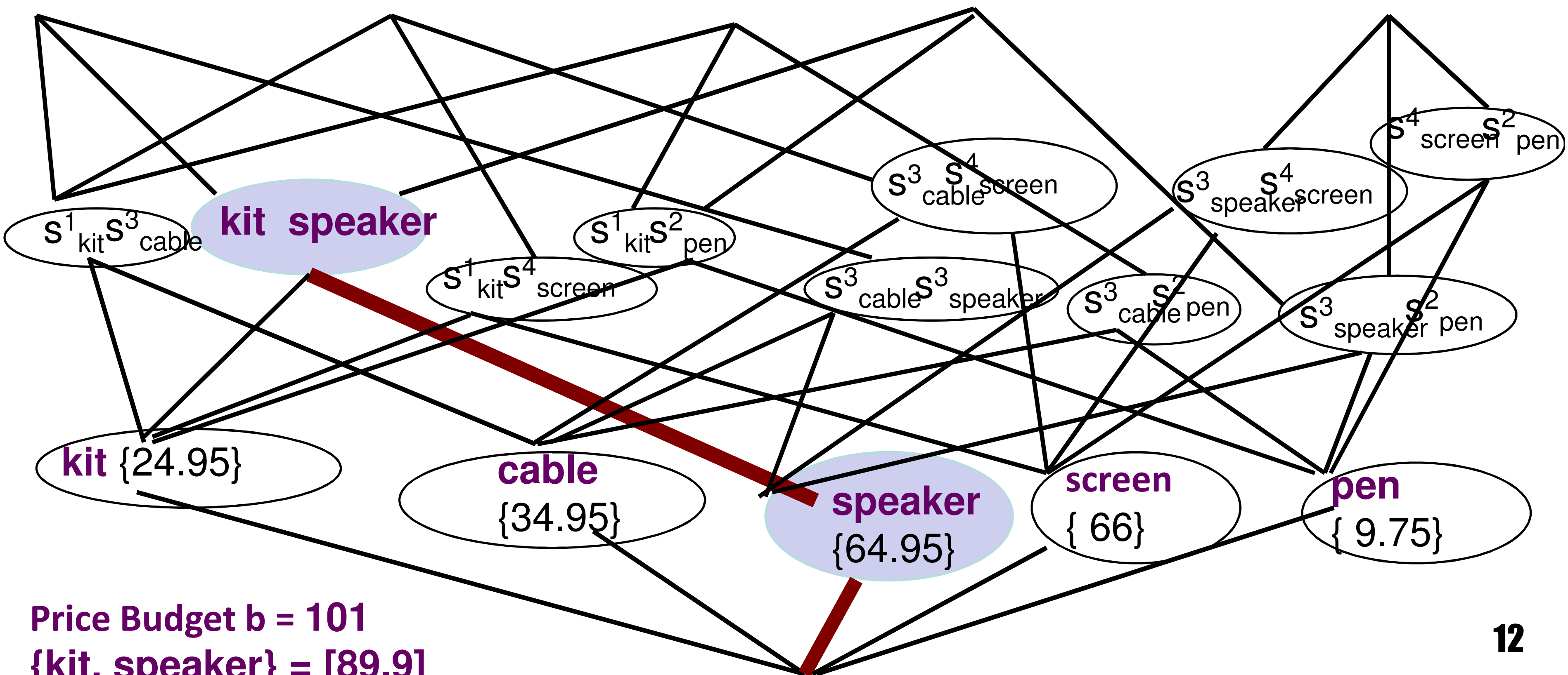
- Counting all maximal packages is #P-Complete

D. Gunopulos, H. Mannila, and S. Saluja. Discovering all most specific sentences by randomized algorithms. In F. N. Afrati and P. G. Kolaitis, editors, ICDT, volume 1186 of Lecture Notes in Computer Science, pages 215–229. Springer, 1997.

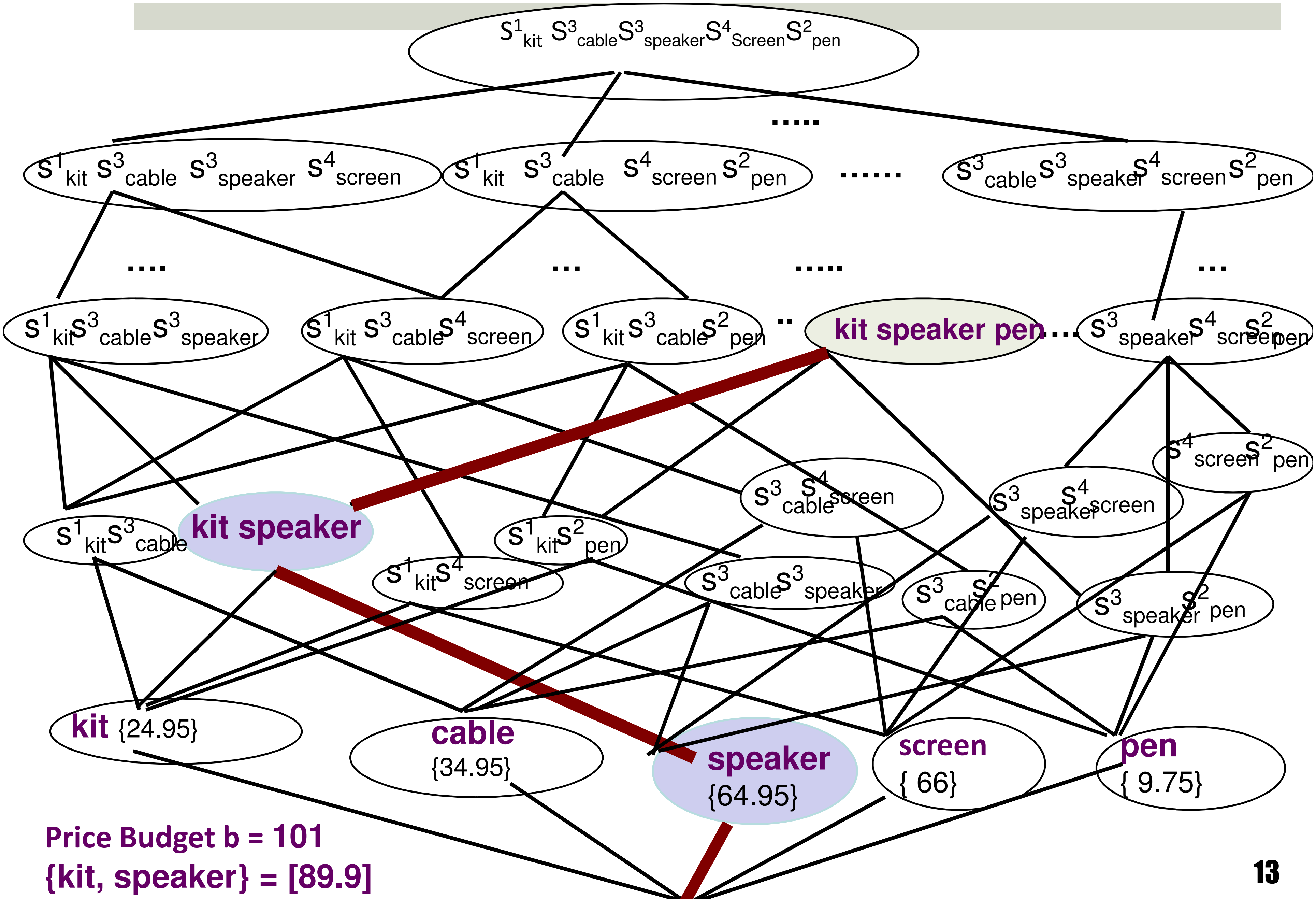


Price Budget b = 101

{speaker} = [64.95]



Price Budget b = 101
{kit, speaker} = [89.9]



Price Budget b = 101

{kit, speaker} = [89.9]

{kit, speaker, pen} = [99.65]

Maximal Packages: Random Walk

- The random walk is repeated until each maximal package found has been generated at least ***twice***

- Good Turing Test used in population studies to determine the number of unique species in a large unknown population

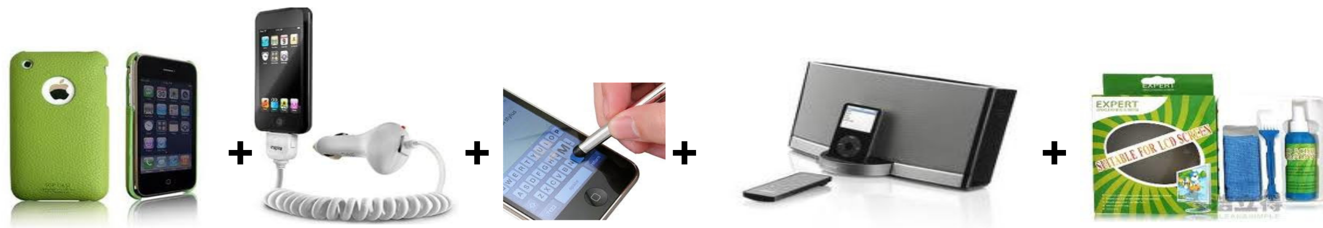
W. A. Gale and G. Sampson. Good-turing frequency estimation without tears. Journal of Quantitative Linguistics, 2(3):217–237, 1995.

  - ***Good Turing Test : $P_0 = n_1 / N$***
    - P_0 = probability that a maximal package is not visited at all (frequency of all unseen species in the original population)
    - n_1 = no of maximal packages visited ***only once*** (number of individuals that are the lone representatives of their species)
    - N = no of different maximal packages (random sample of N individuals from input population)

  - if $n_1 = 0$, $P_0 = 0$

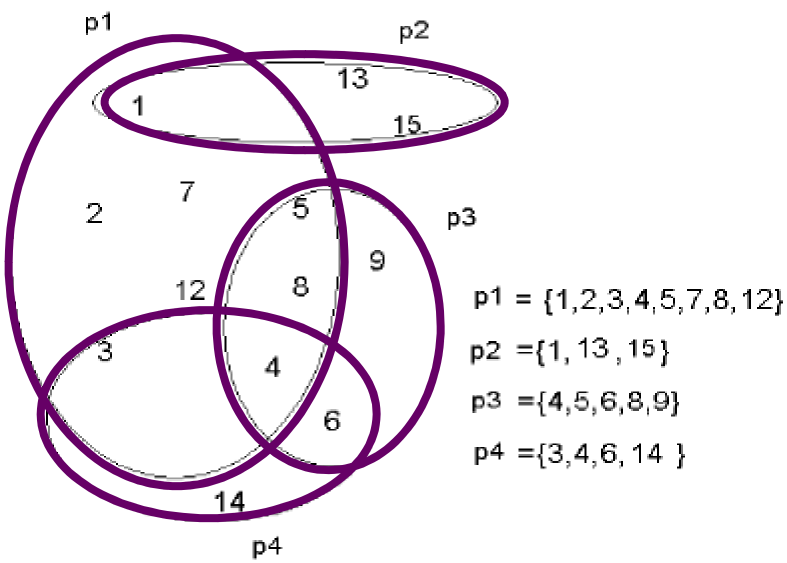
Diversity: k most representative packages

Diversity: k=2



Diversity: Select k Representatives

- ***Clustering?*** a set distance measure (e.g., Jaccard) does not capture coverage.
- Diversity formulated as **selecting maximal representatives whose power sets *maximize coverage***



- p1 consists of 255 packages
- p2 consists of 7 packages
- p3 and p4 consists of 31 packages

k= 2,

- Best summary: {p1,p3}
- 279 packages (255 + 31 – 7)
- {p2,p3} only 38 packages

NP-Completeness and Approximation Algorithms for Selecting k Representatives

- **Selecting k representatives is NP-Complete** by reduction from ***Maximum K-Set Cover***

- ***A basic greedy algorithm*** adapted directly from the greedy approximation algorithm for the ***Maximum K-Set Cover Problem,*** *with approximation factor* ***1-1/e***

  - Uses ***Inclusion-Exclusion principle*** for deriving cardinality of the union of a set of sets
  - The naïve way of calculating coverage requires summation of exponential number of terms

Fast Greedy: Top-k Style Early Stoppage of Greedy Algorithm

- ***A faster version*** *of the Greedy Algorithm* uses ***Bonferroni Inequalities*** to estimate the coverage for a given depth r, where r is an odd number and 1<r<n (n is the total no of packages)

  - picks candidate packages in round-robin

  - computes upper and lower thresholds of coverage using Bonferroni estimates, and stops when the upper is no longer larger than lower.

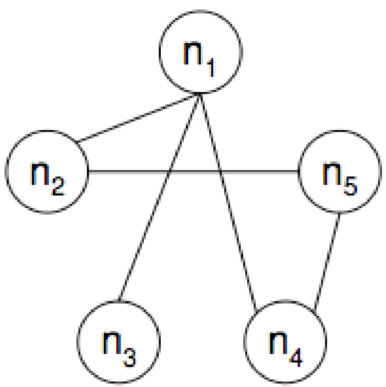
Visual Effect Diversity (k = 4)

Visual Effect Diversity (k = 4)

Visual Effect Optimization

LEMMA 3. *The* visual effect optimization *problem is NP-Complete for m satellite types, where m is bounded by n, the number of packages.*

NP-Completeness is proved by reducing the **Hamiltonian Path** problem to ours



	S_1	S_2	...
p_1	s^1_1	s^2_2	
p_2	s^2_1	s^3_2	
p_3	s^2_1	s^1_2	
p_4	s^2_1	s^4_2	
p_5	s^1_1	s^1_2	

EnhanceVE: Visual Effect Optimization Algorithm

$$p_1 = (s^1_{case}, s^1_{charger}, s^1_{kit}, s^1_{cable}, s^1_{speaker}, s^2_{screen}, s^1_{pen}),$$
$$p_2 = (s^1_{case}, s^1_{charger}, -, s^3_{cable}, -, s^1_{screen}, s^1_{pen}),$$
$$p_3 = (s^1_{case}, s^4_{charger}, -, s^2_{cable}, s^3_{speaker}, -, s^1_{pen}),$$
$$p_4 = (s^2_{case}, s^4_{charger}, -, s^2_{cable}, s^3_{speaker}, s^1_{screen}, s^1_{pen}),$$

$G_S{}^1{}_{case} = \{p_1, p_2, p_3\}$

$G_S{}^2{}_{case} = \{p_4\}$

Complexity $O(mn^2)$
m: number of satellite types
n : number of packages

Algorithm optimal when there is only one satellite type (m=1)

- Optimizes visual effect for the highest priority satellite type
- breaks ties by looking into the lower priority satellite types

Observation : Optimal ordering cannot be generated if p_3 is picked first

Experiments

- Pre-processed Yahoo! Shopping dataset
  - Items with extreme prices filtered out (below $2 and above $1000)
  - 10 product types are created
  - Central items are within price range $550 - $1000
- A total of 101,271 items, of which 2,222 are central items
- Compatibility is binary using a threshold = co-purchase, co-browse, co-rate.
- On average, we have 11,005 items per satellite type

Experiments

- User Studies using Amazon's Mechanical Turk infrastructure to evaluate
  - Selection of k representative packages by maximizing coverage
  - Visual diversity

- Users generally found these approaches useful

- Extensive performance experiments to verify the efficiency of the proposed algorithms.

What have we learned?

- **There is value in composing items in recommendation systems**
  - different facets of the problem (maximality, coverage, diversity)
- **We learned that composite retrieval is a hard problem**
  - Relevance is complex to define
  - Retrieval algorithms are challenging
  - Everything depends on the "shape" of the composite item being suggested
- **Currently investigating other "shapes" of composite items**
  - **"chains":** interactive itinerary planning (at ICDE 2011)
  - **"snowflakes"**

Interactive Itinerary Planning (ICDE 2011)

- A user interested in exploring Barcelona picks a central item and sees *compatible* travel packages *within budget*
  - Compatibility: geographic distance between central item and satellite item
  - Cost: visit time of satellite item
  - Travel itineraries have to be **chains**

Interactive Itinerary Planning (ICDE 2011)

User has 5 hours to visit Barcelona

10mn →

15mn →

30mn →

Your hotel
(central item)

Sagrada Familia
1hr
(satellite item)

San Pau Hospital – 30mn
(satellite item)

Parc Guell – 2hrs
(satellite item)

Thank You !

Outline

- Motivation
- **Problems**
  - **Maximal Package Construction**
  - **Summarization**
  - **Visual Effect Optimization**
- Algorithms
- Experiments
- Summary