Large Data Problems at the Long Tail:

The eBay Story

Challenges and Opportunities

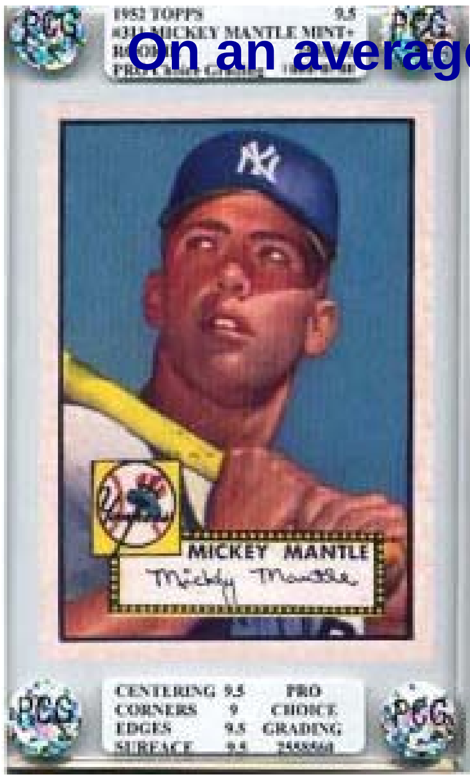
Neel Sundaresan

Research Labs

http://labs.ebay.com

What wags the Tail?

eBay users trade about $1,400 worth of goods on the site every second.



On an average day on eBay…

A vehicle sells every 2 minutes

A part or accessory sells every 3 seconds

Diamond jewelry sells every 83 seconds

A Timberland shoe sells every 10 minutes

A trading card sells every 6 seconds

Then There was One…

When asked if he understood that the laser pointer was broken, the buyer said "Of course, I'm a collector of broken laser pointers"

Divine Reward!

The Long Tail Nature

- Buyers outnumber sellers 5:1
- Seller-items sold has a power law distribution
- Seller-revenue has a power law distribution
- Buyer-items bought has a power law distribution

  For a sample period over 1 month mean is 3 and median is 1
- Buyer-spent money has a power law distribution

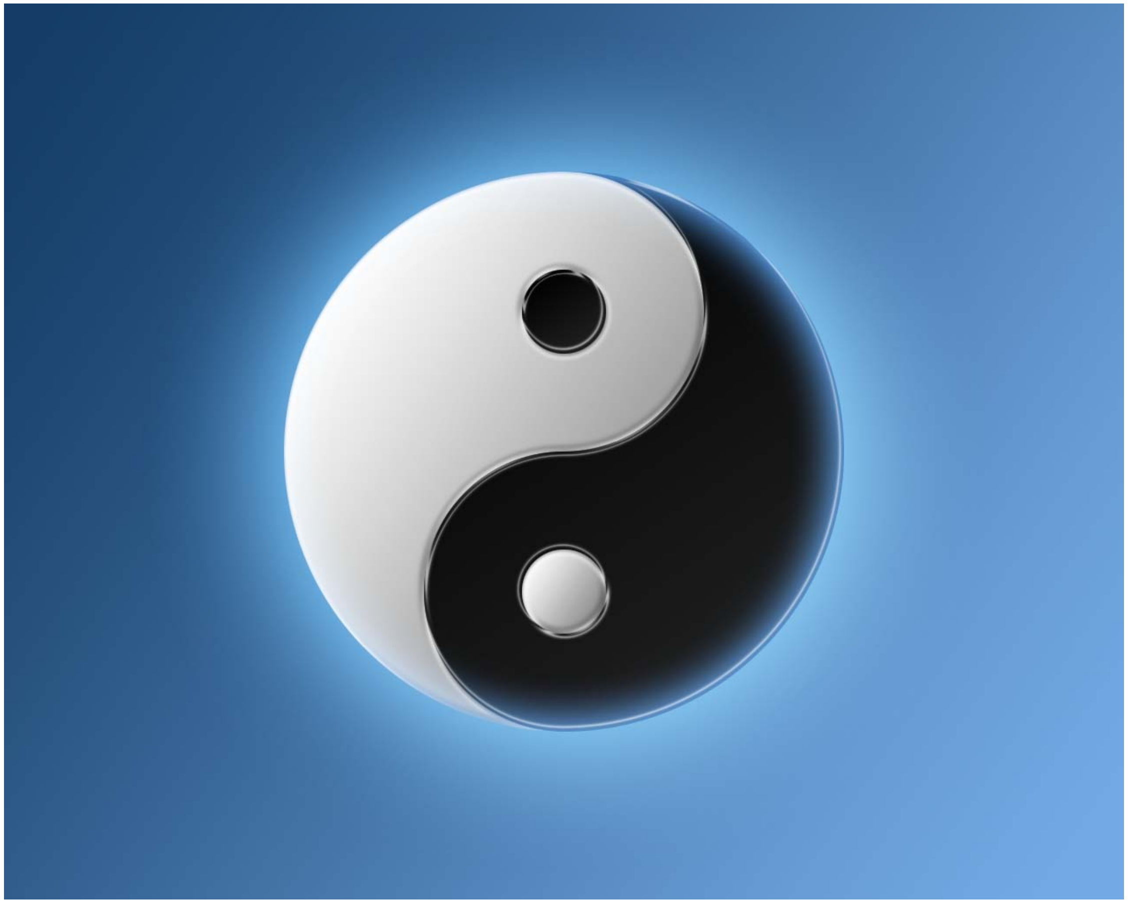
  For a sample period of 1 month mean is 98$ and median was 45$
- Categories browsed

  mean 10.5, median 5
- ~10M new items a day, most items eventually sell, items last from a day to 30 days, most items not cataloged, some auction-some fixed price

Opportunities at the Tail

- Vast majority of products appeal to small number of users
- Vast majority of products of this nature can only be carried by small number of sellers
- These account for sizable consumption
- “Selling less of more” becomes important

Algorithm vs Data



Neel Sundaresan

Research Labs

Search Challenge

- Near similar titles

  “Apple IPOD Nano 4GB Black NEW! Great Deal!”

  “Apple IPOD Nano 4GB Black NEW! Skin Great Deal!”

  What does someone querying for “ipod nano” look for?

What’s in a word

- Spelling corrections for Swarowski?

  64 of them!

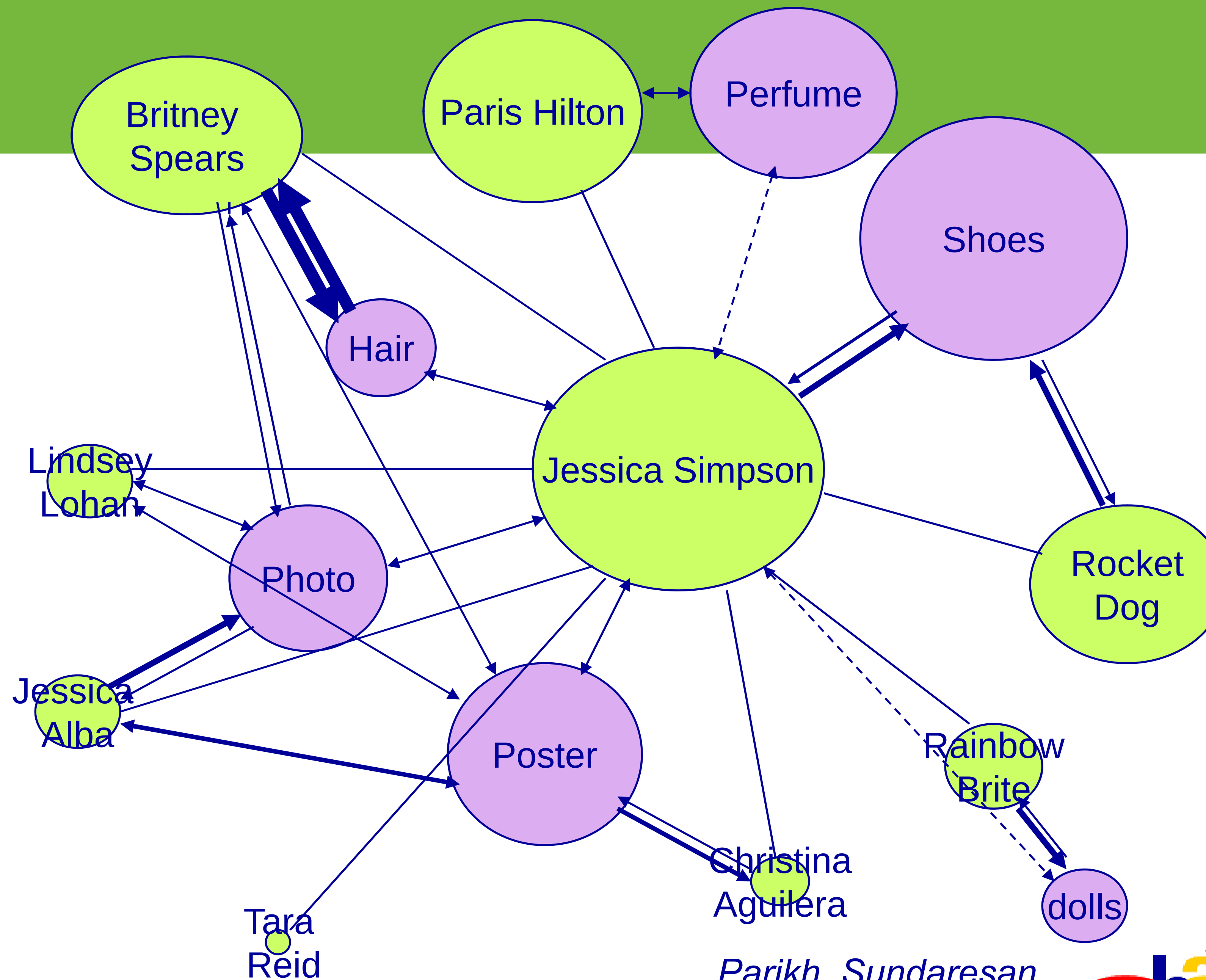
swarvoski,swaroski,swavorski,swarovsky,swarski,**swarowski**,swarvorski,swarofski,
swarvski,sworovski,swarovksi,swarosky,svarovski,swarowsky,swarkovski,swarovki,
svarowski,swaraski,swaroviski,swarovoski,swaravski,swarorski,swartski,swarovsk,
swaovski,swarvosky,sworoski,swalovski,sworvski,swavroski,swrovski,
swaorvski,swavoski,sawrovski,swarovsi,swarovaski,svaroski,swarozski,swarouski,
swarokski,swarvaski,sworvoski,swarvowski,swarosvki,svarovsky,swaravoski,swarovky,
sawarski,sarovski,swarovzki,swarocski,swarovskl,swarovsku,swarovkski,swarrovski,
swarovske,swarowvski,swarvovski,saworvski,swarosvski,swarovrski,swarivski,swarsovski

Courtesy K. Mauge’

What’s That in English?

- Avon SSS 5 OZ Shave Gel NIB FS GWP

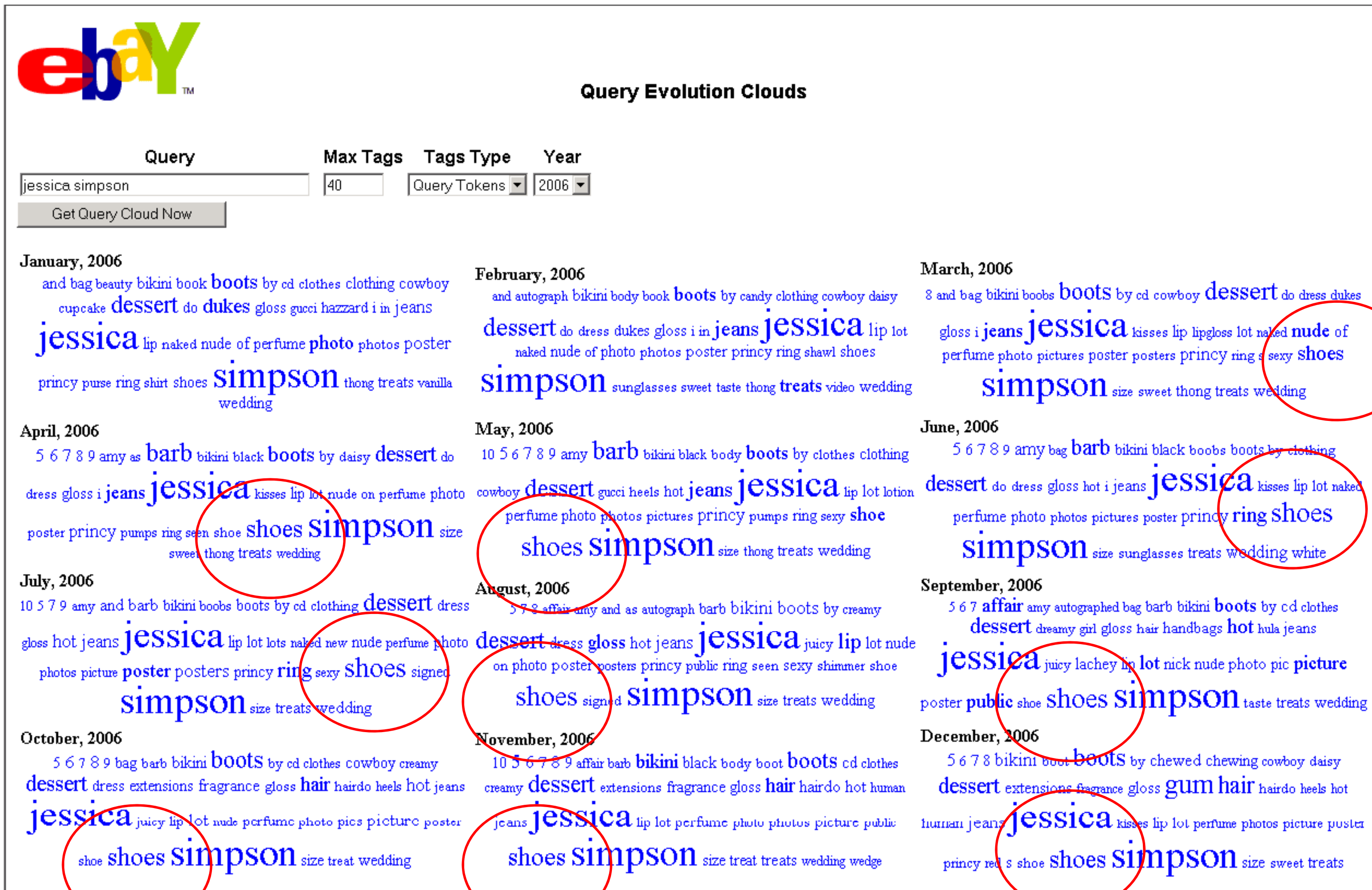
SKIN SO SOFT	NEW IN BOX	Free Shipping	Gift With Purchase



Parikh, Sundaresan

Evolution of Jessica Simpson

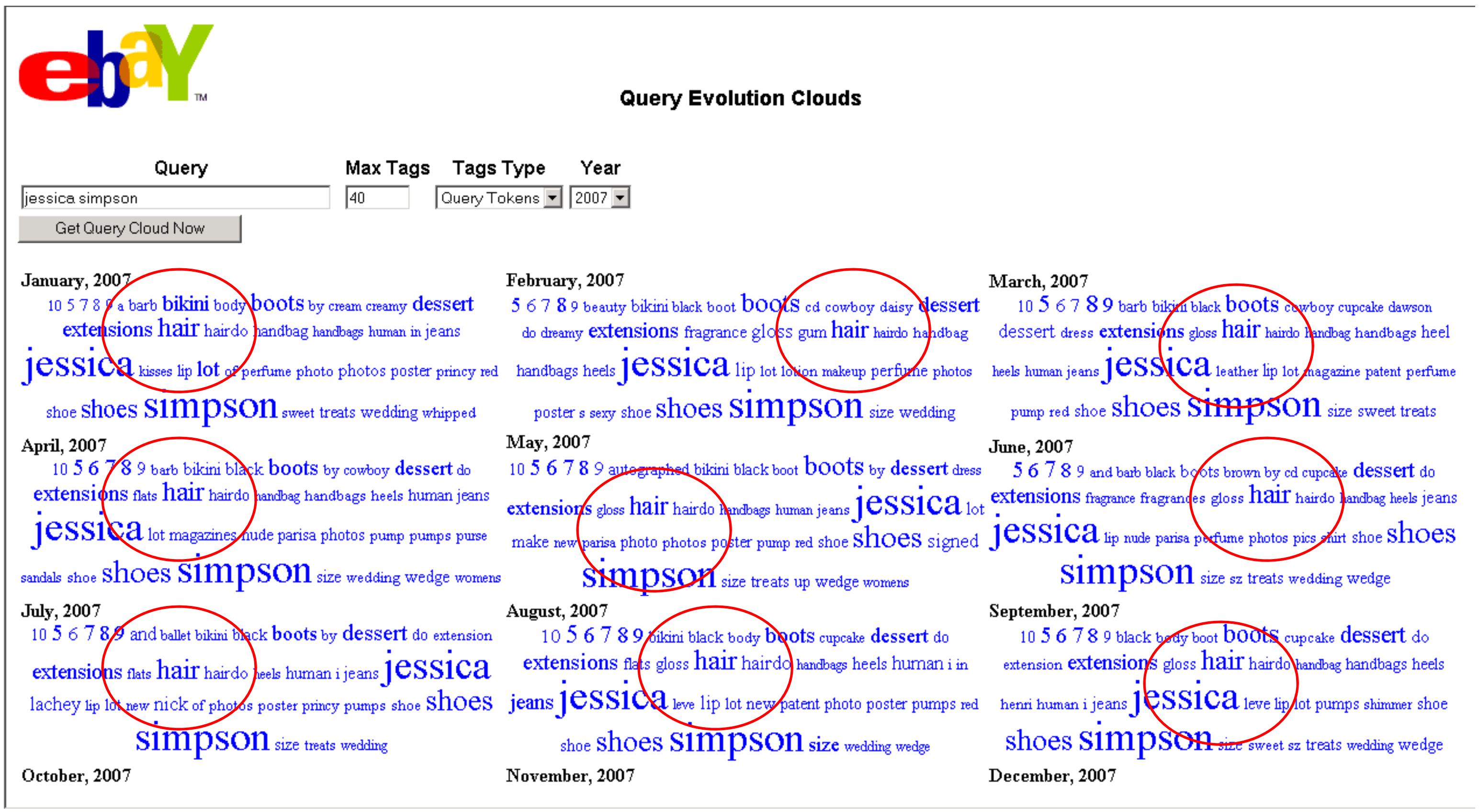
JS (contd)



Neel Sundaresan

Research Labs

JS (contd)


Query Evolution Clouds
Query
Max Tags
Tags Type
Year
jessica simpson
40
Query Tokens
2007
Get Query Cloud Now
January, 2007
10 5 7 8 9 a barb bikini body boots by cream creamy dessert extensions hair hairdo handbag handbags human in jeans jessica kisses lip lot of perfume photo photos poster princy red shoe shoes simpson sweet treats wedding whipped
February, 2007
5 6 7 8 9 beauty bikini black boot boots cd cowboy daisy dessert do dreamy extensions fragrance gloss gum hair hairdo handbag handbags heels jessica lip lot lotion makeup perfume photos poster s sexy shoe shoes simpson size wedding
March, 2007
10 5 6 7 8 9 barb bikini black boots cowboy cupcake dawson dessert dress extensions gloss hair hairdo handbag handbags heel heels human jeans jessica leather lip lot magazine patent perfume pump red shoe shoes simpson size sweet treats
April, 2007
10 5 6 7 8 9 barb bikini black boots by cowboy dessert do extensions flats hair hairdo handbag handbags heels human jeans jessica lot magazines nude parisa photos pump pumps purse sandals shoe shoes simpson size wedding wedge womens
May, 2007
10 5 6 7 8 9 autographed bikini black boot boots by dessert dress extensions gloss hair hairdo handbags human jeans jessica lot make new parisa photo photos poster pump red shoe shoes signed simpson size treats up wedge womens
June, 2007
5 6 7 8 9 and barb black boots brown by cd cupcake dessert do extensions fragrance fragrances gloss hair hairdo handbag heels jeans jessica lip nude parisa perfume photos pics shirt shoe shoes simpson size sz treats wedding wedge
July, 2007
10 5 6 7 8 9 and ballet bikini black boots by dessert do extension extensions flats hair hairdo heels human i jeans jessica lachey lip lot new nick of photos poster princy pumps shoe shoes simpson size treats wedding
August, 2007
10 5 6 7 8 9 bikini black body boots cupcake dessert do extensions flats gloss hair hairdo handbags heels human i in jeans jessica leve lip lot new patent photo poster pumps red shoe shoes simpson size wedding wedge
September, 2007
10 5 6 7 8 9 black body boot boots cupcake dessert do extension extensions gloss hair hairdo handbag handbags heels henri human i jeans jessica leve lip lot pumps shimmer shoe shoes simpson size sweet sz treats wedding wedge
October, 2007
November, 2007
December, 2007

Recommender Systems – At Scale

- MovieLens – 10M ratings, 10K movies, 70K users
- Netflix Prize Challenge – 100M ratings, 10K movies, 500K users
- eBay Challenge – 2+M txns/day, 10M+ new items, 200+M buyers and sellers

Recommender Systems – Known Ways

- Collaborative Filtering

  Neighborhood models

  - compute relationships between items or users

  Latent Factor Models

  - Use Matrix Factorization. Work well with large users, item factors

Matrix Factorization

- Users and items are mapped as a latent factor space (of dimensionality σ)
- User-item interactions are modeled as inner products

Each item v is represented by a vector q_v in R^σ representing the extent to which the item has the different features in R^σ

Each user u is represented by a vector p_u in R^σ representing the extent to which the user is interested in different features in R^σ

Preference for item v by user u r_{uv} is computed as the dot-product $q_v^T p_u$

The major task is computing the mapping of each item and user to the corresponding factor vectors in R^σ

Ref. Koren et al. IEEE 09

Matrix Factorization (contd.)

- SVD raises difficulties due to high sparsity

  In Netflix sparsity is 1:100, eBay sparsity is even higher (1:10K)

  Instead of SVD, directly modeling on the observed data is preferred

  To learn the factor vectors q_v *and* p_u

  We minimize the regularized square error on the known observations:

$$\min_{q,p} \sum_{(u,v)\in\kappa} (r_{uv} - q_v^T p_u)^2 + \lambda(\|q_v\|^2 + \|p_u\|^2)$$

  where r_{uv} is known for κ

  This minimization equation is solved using for e.g. Stochastic Gradient Descent or Alternating Least Squares

Ref. Koren et al. IEEE 09

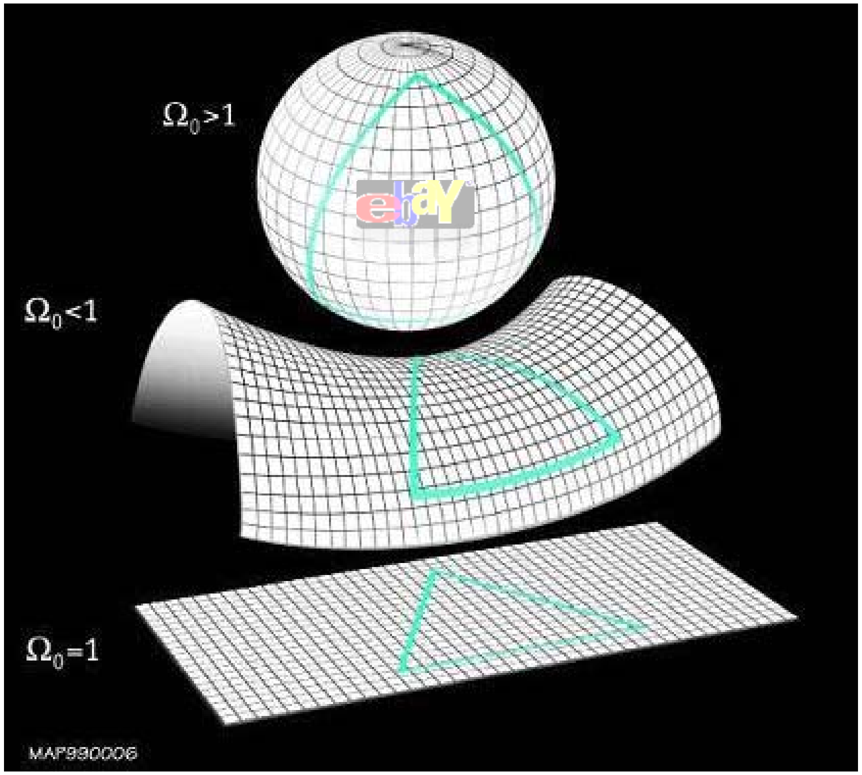
Matrix Factorization – the eBay Challenge

- The *q* space of items is extremely large and volatile.
  - There's very little overlap between buyer interests/purchases and actual items
- However, the query space is fairly static (even though in the millions)
  - The user space can be mapped to queries and the item space can be mapped to queries

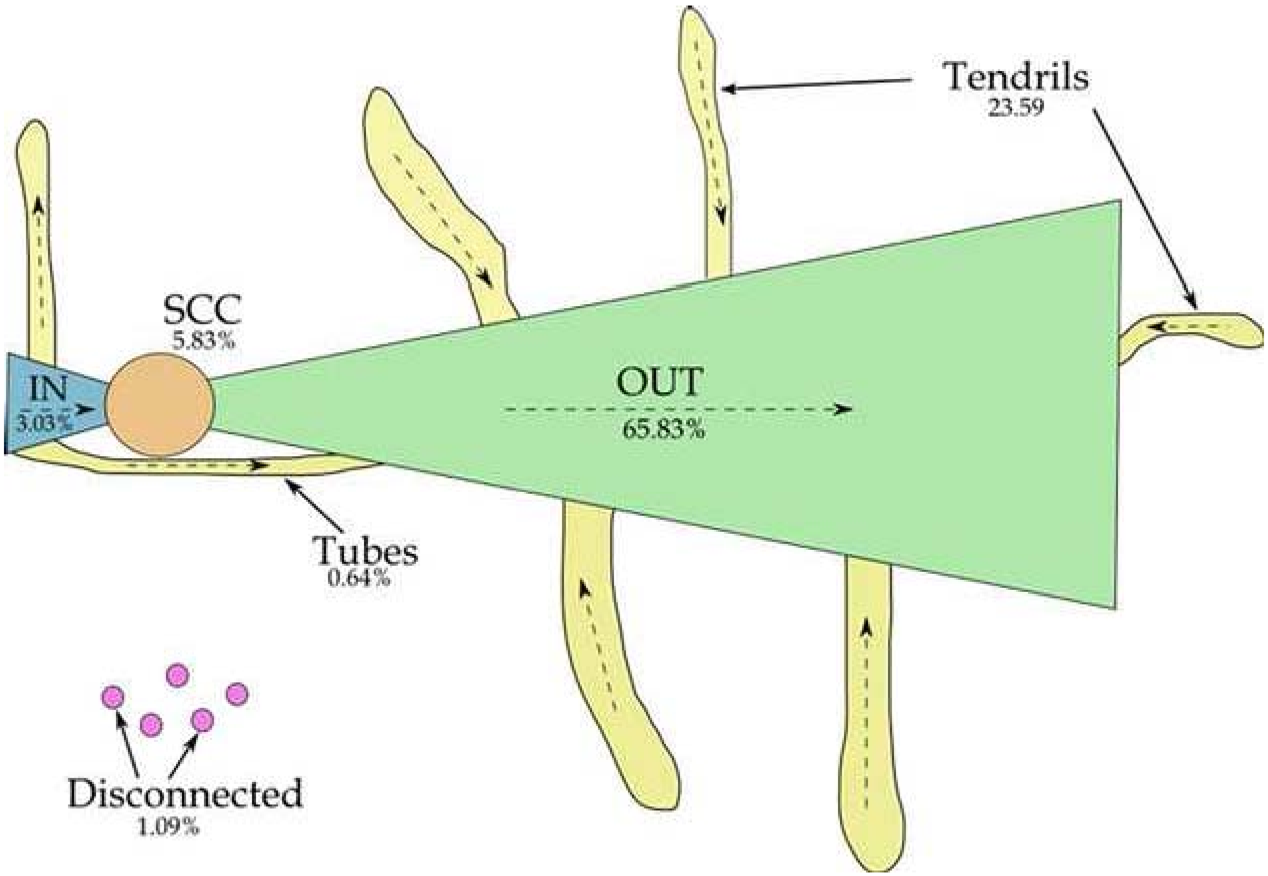
    One could use explicit feature vectors on the user and item space q_v p_u then optimizing to discover a weight matrix W in $q_v W p_u$

    Alternatively, Latent Topic Models (e.g. Latent Dirichlet Allocation - LDA) can be used to map users to latent topics, and latent topics to search queries

What Shape is the Universe of eBay?

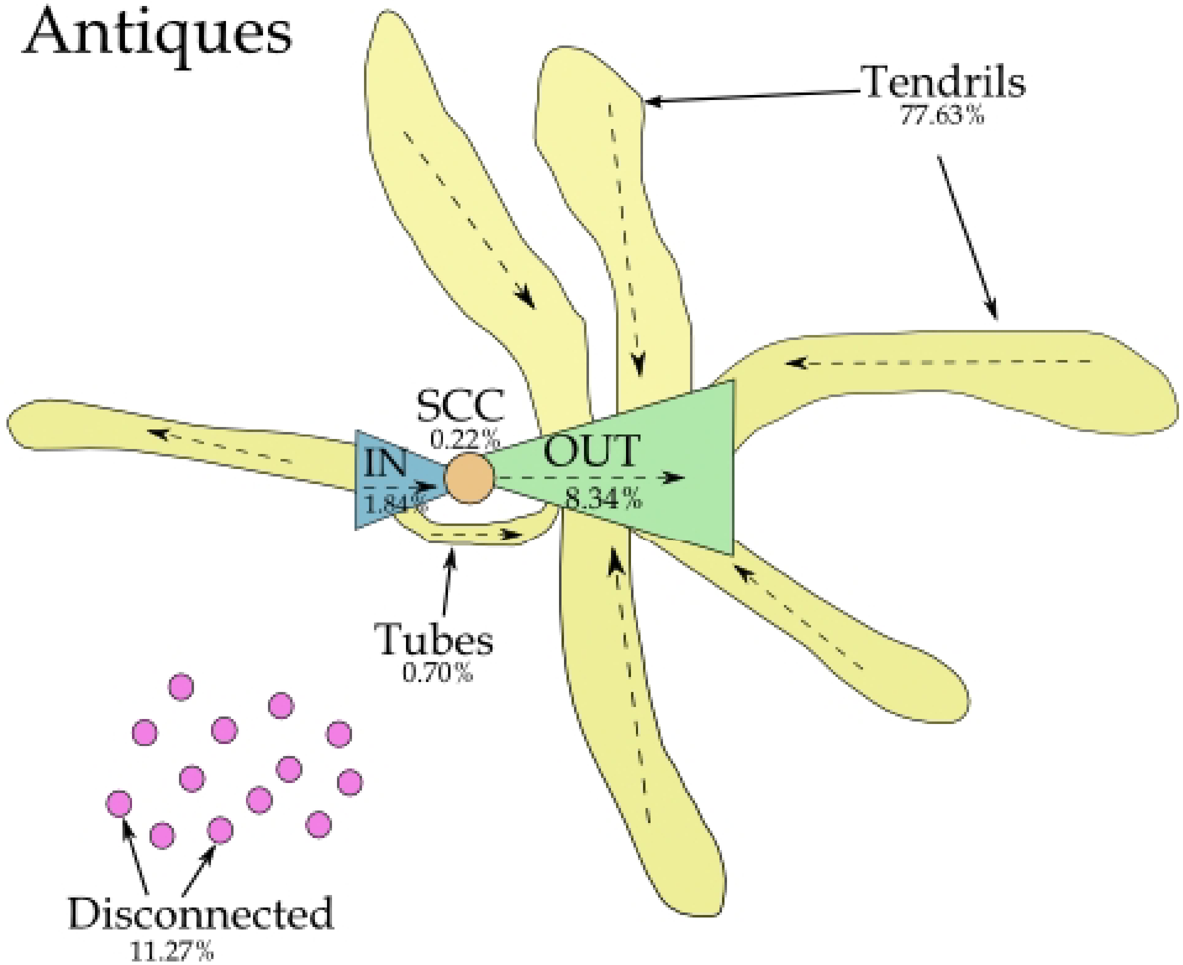


eBay - The Neck-Tie

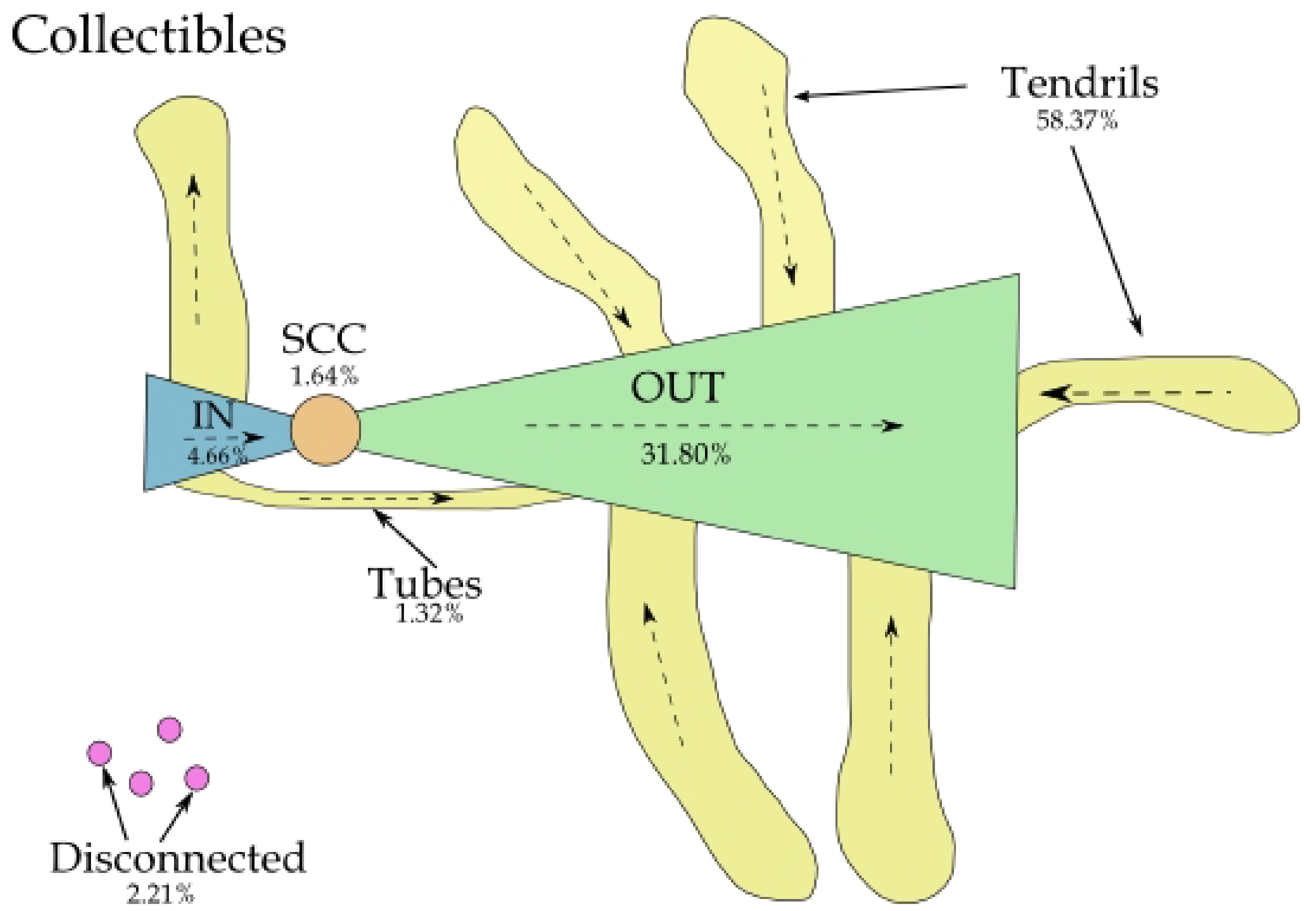


Shen, Sundaresan

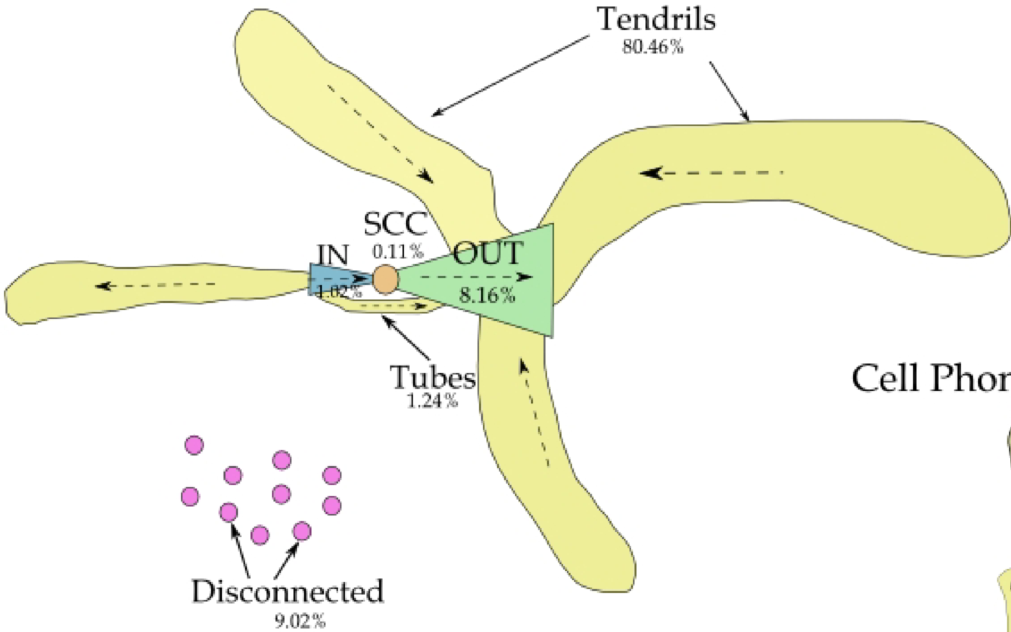
Antiques



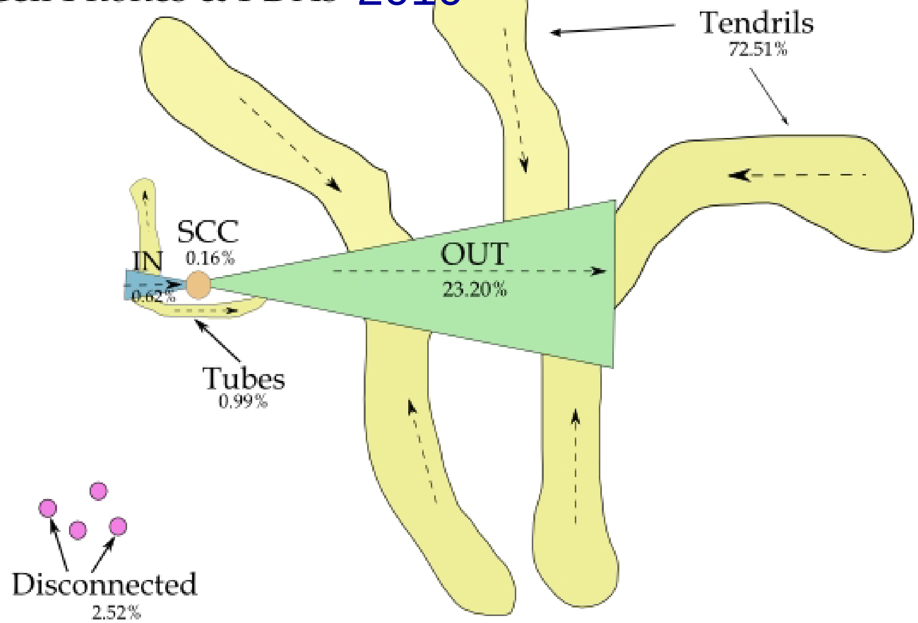
Shen, Sundaresan


Collectibles
Tendrils
58.37%
SCC
1.64%
IN
4.66%
OUT
31.80%
Tubes
1.32%
Disconnected
2.21%

Cell Phones & PDAs 2005

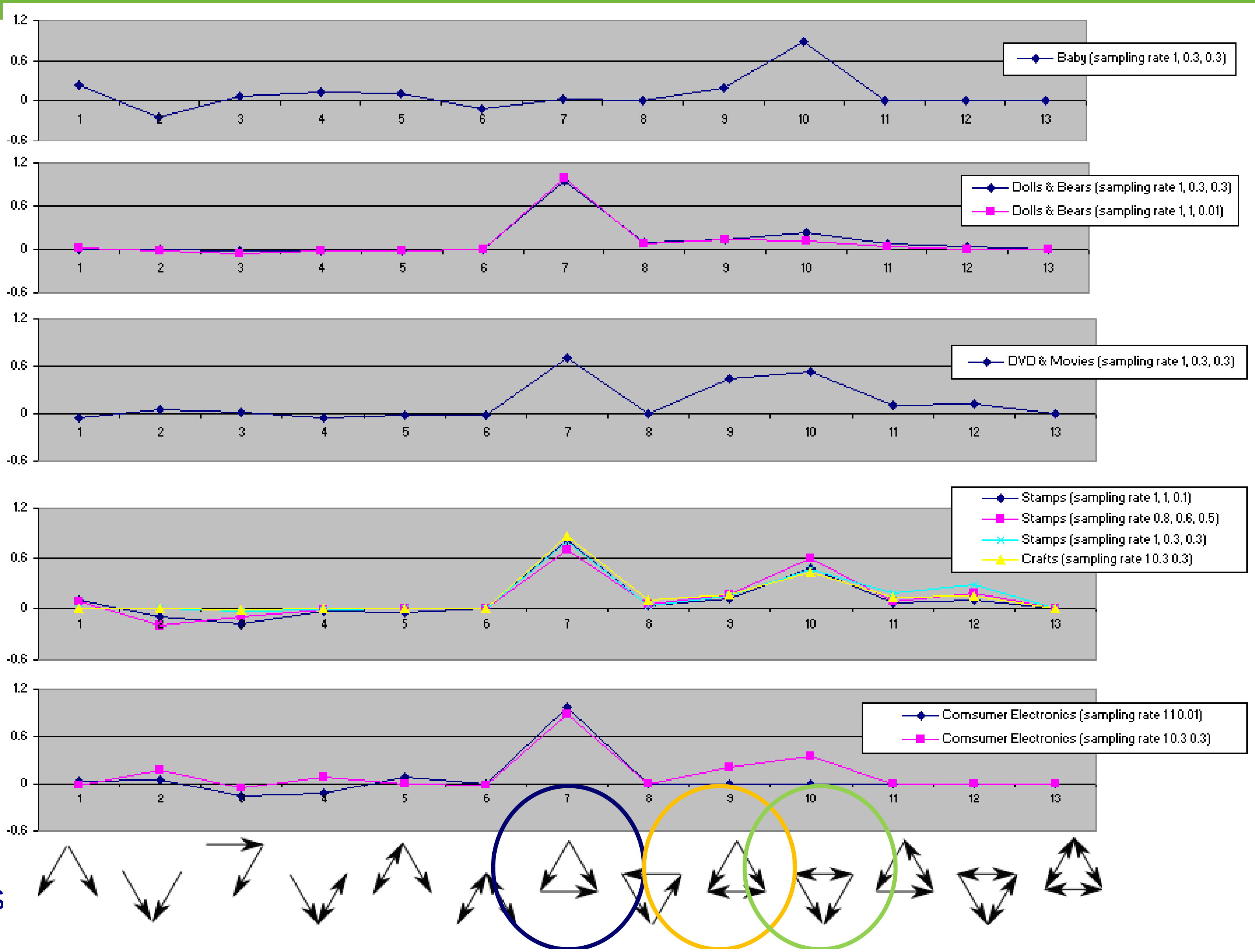


Cell Phones & PDAs 2010



Neel Sundaresan

Significance Profiles

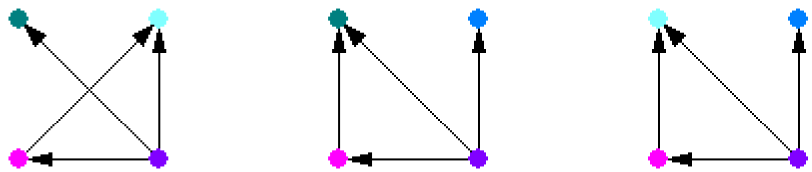


Neel S

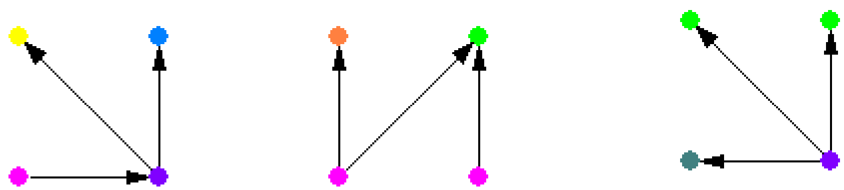
Labs

Motifs: Feedback Enhanced Triad Distribution Across Categories

- Stamps



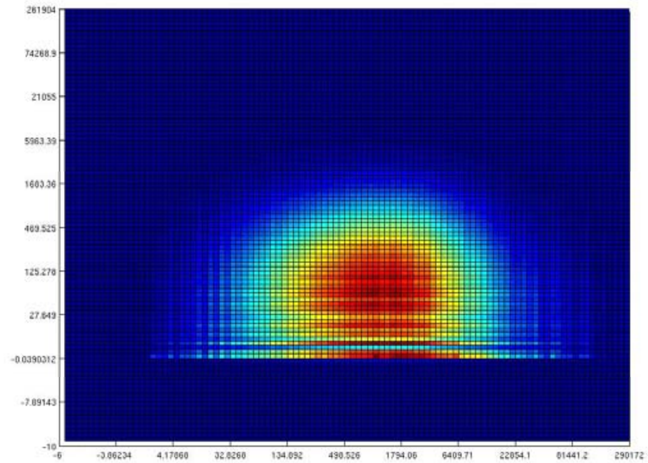
- Antiques



-1	0	5	20	100	500	1000	5000	>5000

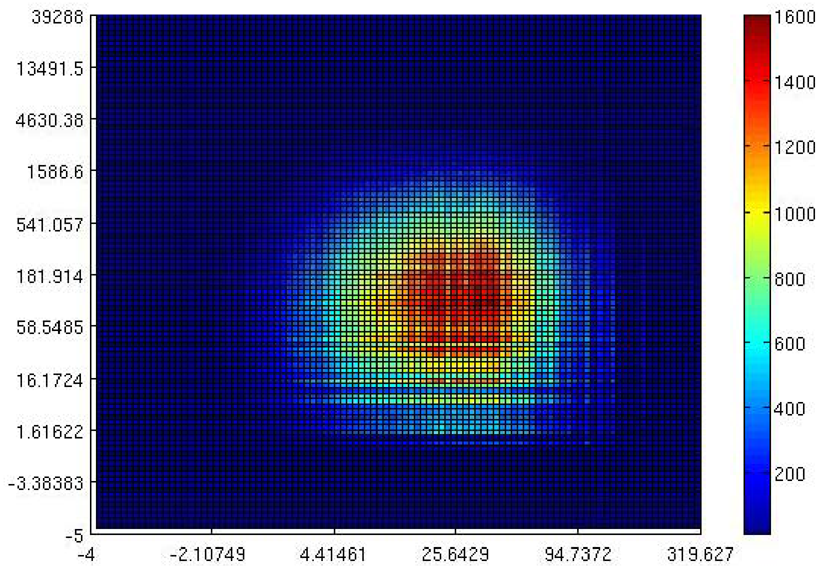
Shen, Sundaresan

Assortative Mining – Auroral Diagrams to Measure Preferential Attachment

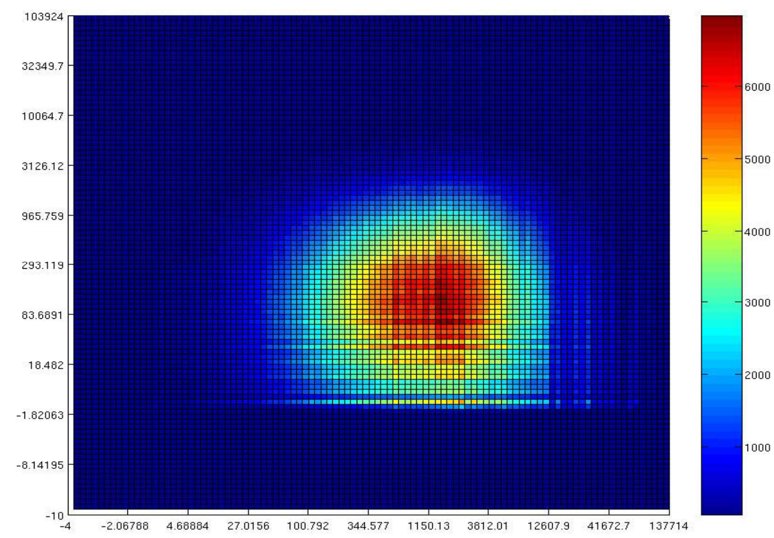


Global

Arts and Crafts



Collectibles



Neel Sundaresan

Shen, Sundaresan

Research Labs

Questions?

ArtAmnesia