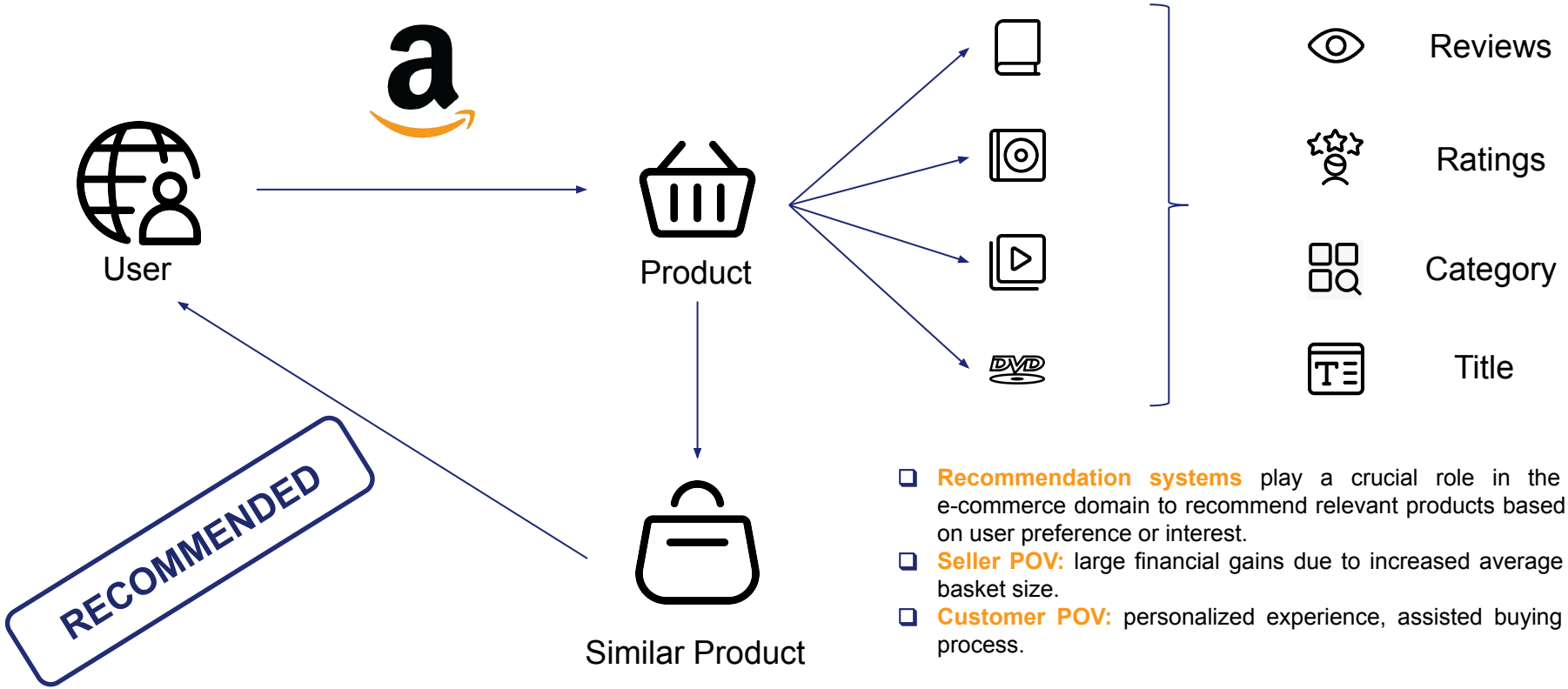
Machine Learning in Network Science Presentation

Mathieu TARDY, J´er´emie FERON, Cl´ement DE LOUBRESSE

Outline

1. Introduction & Motivations
2. Problem Definition & Data Selection
3. Methodology
4. Evaluation & Recommendations

Introduction & Motivations



- **Recommendation systems** play a crucial role in the e-commerce domain to recommend relevant products based on user preference or interest.
- **Seller POV:** large financial gains due to increased average basket size.
- **Customer POV:** personalized experience, assisted buying process.

Problem Definition & Data Selection (1/2)

548,552 Products

1,545,228 Edges

2.78525 Average Degree

7,781,990 Reviews

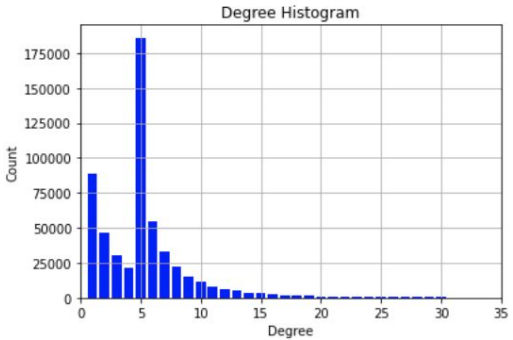
393,561 Books

103,144 Music CDs

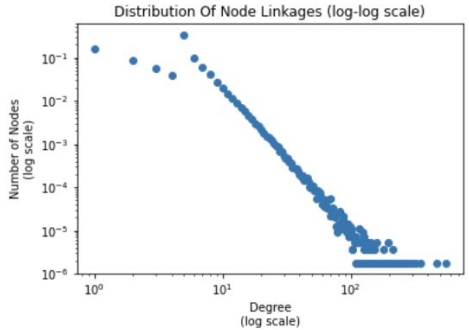
26,131 Videos

19,828 DVDs

- Amazon's co-purchasing network metadata*
- Initially presented in a text file, which we converted into a dictionary in order to create a relevant Graph
- Power-Law Degree Distribution



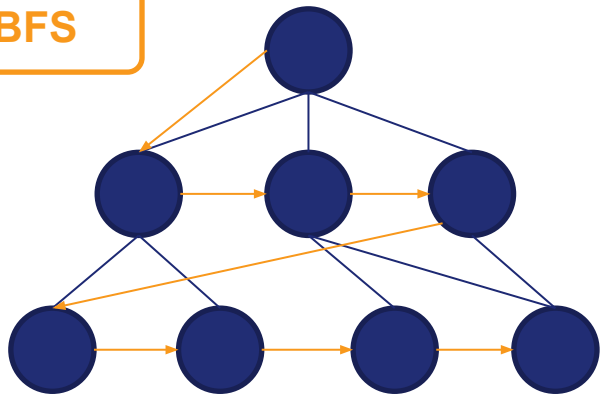
Log-log

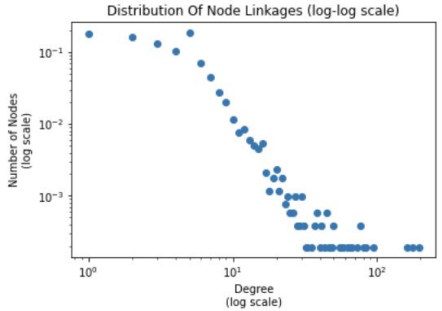


** Available at https://snap.stanford.edu/data/amazon-meta.html*

Problem Definition & Data Selection (2/2)

BFS





Our resulting graph is composed of 5,142 products, with an average degree of 2.4 and 12,343 edges.

- Due to the size of the graph, we anticipated potential **computational complexity issues**.
- Hence, we explored different techniques to **create representative samples**.
- We chose to use **BFS** as a graph sampling method because it exhibited advantages relative to random walk and other sampling techniques
- We could potentially face bias towards high-degree nodes but considering our previously shown degree distribution, we assumed that this technique was relevant.

3. Methodology

3.1 Embeddings – Key Metrics

Resource Allocation Index

- The sum over all common neighbors of nodes *u* and *v* of one over the degree of these nodes.
- $\sum_{w \in \Gamma(u) \cap \Gamma(v)} \frac{1}{|\Gamma(w)|}$
- If products A and B have a large RA index, hence there exist other products with which they are frequently bought with.

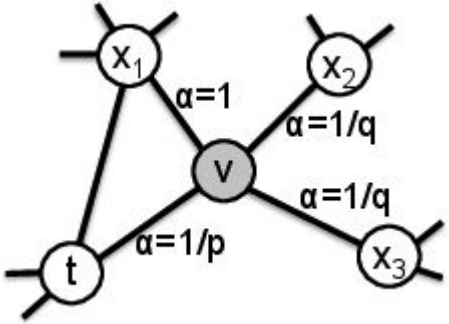
Adamic-Adar Index

- Very similar to the Resource Allocation Index.
- We divide by the log of the degree.
- $\sum_{w \in \Gamma(u) \cap \Gamma(v)} \frac{1}{\log|\Gamma(w)|}$

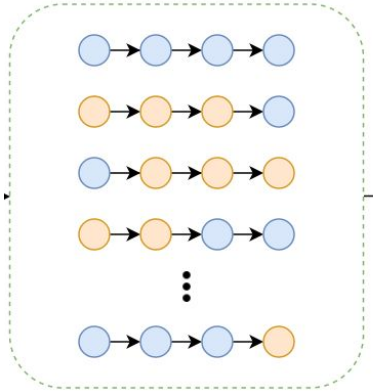
3.1 Embeddings - Node2vec

Input:

- The full graph with nodes and edges.
- Transition probabilities between nodes.
- Random walk generated according to these.



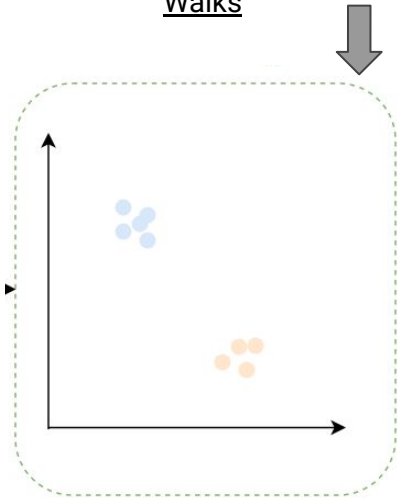
Transition probabilities



Walks

Model:

- Node2vec computes embeddings following the SkipGram framework.



Node2vec Representations

Put simply it will adjust the embeddings of nodes that regularly appear together in walks map to be similar and the ones that don't will be further spread apart.

The notion of similarity is defined by the cosine between the obtained embeddings.

3.1 Embeddings - GraphSage

Input:

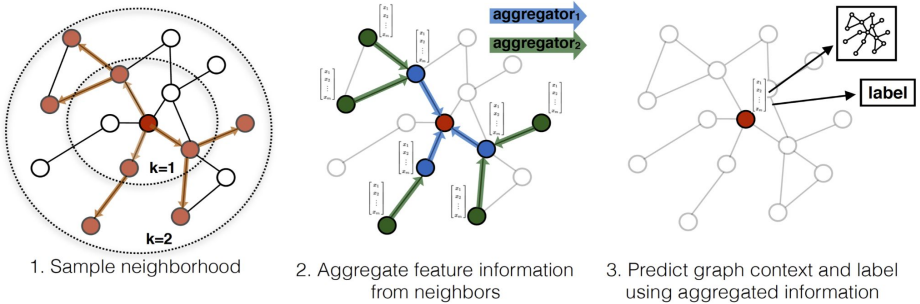
- The full graph with the previously computed node embeddings (via node2vec ect…).
- The “inverse” or “negative” graph containing the same set of nodes and only “false” links.

Model:

- Two layer GraphSage (see picture)
- Dot product predictor

Training:

- Loss: Binary Cross Entropy



1. Sample neighborhood
2. Aggregate feature information from neighbors
3. Predict graph context and label using aggregated information

Snap Stanford GraphSage framework description

The network adjusts its weights (and hence the hidden representation of the nodes) so it minimizes the the difference between the link prediction distribution on both the positive and negative graphs and the real one (1s for the positive and 0s for the negative)

3.2 Prediction

Set-up:

- We treat the problem as binary classification, for a couple of nodes (u,v) predict the probability that they are linked.
- Split train/ test

Features:

- The input features are the results of our embeddings step.
- Adamic-Adar and Resource Allocation index are given to the model as is as they are computed on a (u,v) pair.
- For Node2vec and GraphSage for every (u,v) pair we test against we computed the cosine between their respective representations.

Models:

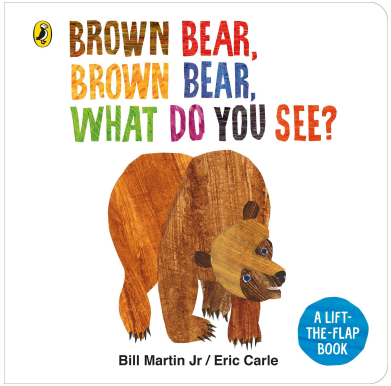
- We compared several models using different combination of features and metrics.
- 11 models used: KNeighborsClassifier, LinearSVC, GaussianNB; XGBClassifier, GradientBoostingClassifier, LinearSVC etc.

4. Evaluation

Embedding/ Agorithms	Best Model	F1	Accuracy	AUC
Neighbor features alone	KNeighborsClassifier	79.1%	79.3%	81.4%
Node2vec alone	LinearSVC	98.9%	98.9%	98.8%
GraphSage alone	GaussianNB	78.9%	78.9%	79.3%
Neighbor + Node2vec	XGBClassifier	99.0%	99.0%	98.9%
Neighbor + GraphSage	GradientBoostingClassifier	89.0%	89.0%	89.3%
Node2vec + GraphSage	LinearSVC	98.9%	98.9%	98.8%

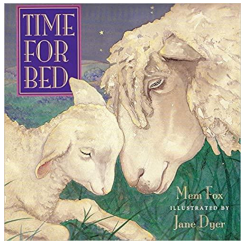
4. Recommendations - Concrete - Full Graph (½)

We provide here a more concrete visualization of our recommendation system.

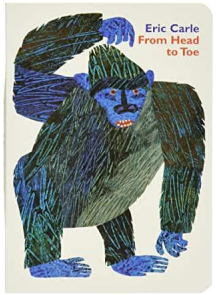


User Purchases Product
ID: 0805047905

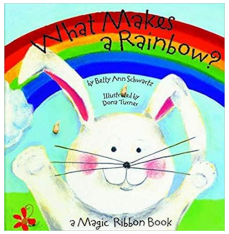
RECOMMENDED



ID: 0152010661

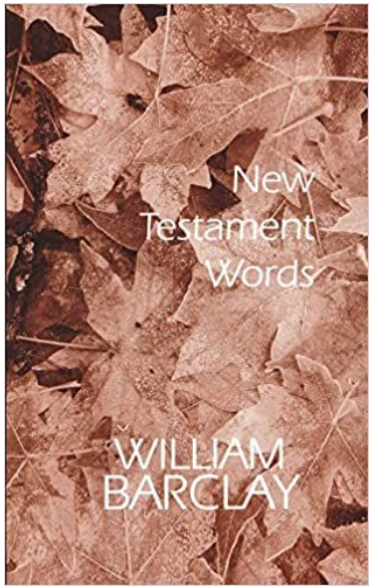


ID: 0694013013

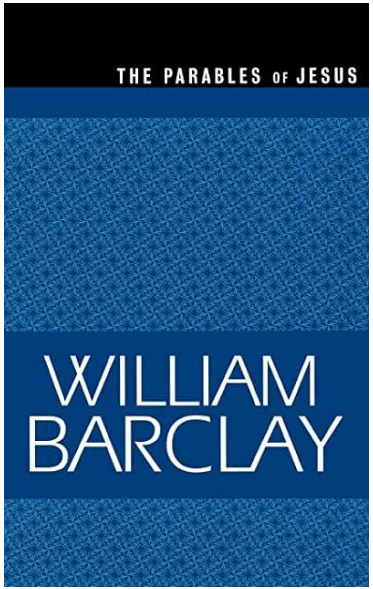


ID: 1581170769

4. Recommendations - Concrete - BFS Graph (2/2)



Here again we note that given the purchased product on the left, our recommendation appears to be highly relevant.




```
{'Id': '248350',
 'Title': 'New Testament Words',
 'Categories': 'b general christian book religion bibl z english author s
william text dictionari refer guid studi spiritu thesaurus testament greel
 'Group': 'Book',
 'CoPurchased': '0664258158 0664258166 066425828X 0664258069 0664258263',
 'SalesRank': 145639,
 'TotalReviews': 3,
 'AvgRating': 5.0,
```

```
{'Id': '341025',
 'Title': 'The Parables of Jesus (The William Barclay Library)',
 'Categories': 'b refer studi subject spiritu barclay author chr
 'Group': 'Book',
 'CoPurchased': '0664258158 0664258166 0664258263 0664221920',
 'SalesRank': 155178,
 'TotalReviews': 3,
 'AvgRating': 3.5,
```

Thank you !