Analysis of the Co-purchase Network of Products to Predict Amazon Sales-Rank

Utpal Prasad(✉), Nikky Kumari, Niloy Ganguly, and Animesh Mukherjee

Indian Institute of Technology Kharagpur, Kharagpur, India
utpaldps@gmail.com

Abstract. Amazon sales-rank gives a relative estimate of a product item's popularity among other items in the same category. An early prediction of the Amazon sales-rank of a product would imply an early guess of its sales-popularity relative to the other products on Amazon, which is one of the largest e-commerce hub across the globe. Traditional methods suggest use of product review related features, e.g., volume of reviews, text content of the reviews etc. for the purpose of prediction. In contrast, we propose in this paper for the first time a *network-assisted* approach to construct suitable features for prediction. In particular, we build a *co-purchase* network treating the individual products as nodes, with edges in between if two products are bought with one another. The way a product is positioned in this network (e.g., its centrality, clustering coefficient etc.) turns out to be a strong indicator of its sales-rank. This network-assisted approach has two distinct advantages over the traditional baseline method based on review analysis – (i) it works even if the product has no reviews (relevant especially in the early stages of the product launch) and (ii) it is notably more discriminative in classifying a popular (i.e., low sales-rank) product from an unpopular (i.e., high sales-rank) one. Based on this observation, we build a supervised model to early classify a popular product from an unpopular one. We report our results on two different product categories (CDs and cell phones) and obtain remarkably better classification accuracy compared to the baseline scheme. When the top 100 (700) products based on sales-rank are labelled as popular and the bottom 100 (700) are labelled as unpopular, the classification accuracy of our method is 89.85% (82.1%) for CDs and 84.11% (84.8%) for cell phones compared to 46.37% (68.75%) and 83.17% (71.95%) respectively from the baseline method.

1 Introduction

Revenue forecasting and sales prediction have recently become very active areas of research especially in the context of box office revenues of newly released movies [1–4]. Further, in almost all of these studies, analysis of the online reviews has been shown to be very effective in such forecasting/prediction. With e-commerce platforms becoming increasingly more popular it is very important for these businesses to be able to understand and, in fact, to early identify the

P. K. Reddy et al. (Eds.): BDA 2017, LNCS 10721, pp. 197–214, 2017.
https://doi.org/10.1007/978-3-319-72413-3_13

sales impact of their different products. Amazon, for instance, maintains sales-rank of every product item which is a number with 1 to 8 digits and captures the product's relative popularity and visibility in comparison to other products in the corresponding sub-category of products. Authors, publishers, marketplace sellers, and many other people and businesses use sales-rank data as an indicator of how well their products are selling. They also analyse sales-ranks to further predict how well a product may sell and to decide whether or not to at all sell a particular product on Amazon.

While the Amazon sales-rank has been the source of much speculation by publishers, manufacturers and marketers, Amazon does not itself release the details of its sales-rank calculation algorithm. Further, it has been observed that sales-rank measures a product's popularity only in its corresponding sub-category. It does not bear any direct correlation with the absolute sales of the product; in fact, it has been conjectured that the rate of growth of the sales-rank of a product is high if the product has almost no sales history while it is very slow if the product has a long sales history[1]. An early prediction of sales-rank can enable the producers and consumers to predict the overall future acceptance of a product item in comparison to other items in the e-market. Further, it could immensely help in estimating the product's exposure on Amazon in future and enable the design of suitable early intervention mechanisms geared toward promoting the product.

In this work, we present an elegant approach to automatically distinguish early on time the popular (i.e., low sales-rank) products from the unpopular (i.e., high sales-rank) ones. Unlike traditional approaches that suggest analysis of reviews (e.g., volume of reviews, latent sentiment in reviews, interval between consecutive reviews) we propose a *network-assisted* scheme for the classification of the popular products.

Key contributions:

- We define a *co-purchase* network of products belonging to the same category where each node is a product and two nodes are connected by an edge if they are bought together. For our experiments, we construct networks for two different product categories – CDs and cell phones. In the first network all the nodes are CDs and two nodes are connected if one CD is co-purchased with another. Similarly, in the second network, each node is a cell-phone and two nodes are connected by an edge if the corresponding cell phones are co-purchased[2].
- We quantify how a product is positioned in the co-purchase network by extracting various structural properties like clustering co-efficient, betweenness, Pagerank, eigenvector and closeness centrality and community membership.

[1] See discussions on sales-rank calculation at https://kdp.amazon.com/community/message.jspa?messageID=562491.

[2] Note that this construction is much different and certainly more non-trivial than a general co-purchase network of all products in which breads might also get linked to bleaches by virtue of being bought together sometimes from the store.

- Since there is no suitable baseline available for this problem, as an additional objective, we define a very competitive baseline built on features extracted through extensive analysis of the reviews. To prepare the baseline, we consider the number of times a product is co-purchased with other products. In addition, we extract various linguistic features like the extent of anger, sadness, negative emotion in the user reviews per product as well as certain general features like the volume of reviews, percentage of fake reviews, dwell time and entropy of ratings.
- We build a binary classifier based on the network features to early classify a popular product from an unpopular one. We compare the performance of the classifier with that built from the baseline features based on online reviews.

Notable Observations:

- Our proposed method can work even in the absence of any reviews for a product. This is especially important at the early stages of a product launch when online reviews for a product are scarce.
- The network features that we propose are notably more discriminative than the baseline features based on review analysis.
- The classifier that we build, in particular, for two different product categories (CDs and cell phones) results in a remarkably better classification accuracy compared to the baseline scheme. In case we label the top 100 (700) products based on sales-rank as popular and the bottom 100 (700) as unpopular, the classification accuracy of our method is 89.85% (82.1%) for CDs and 84.11% (84.8%) for cell phones compared to 46.37% (68.75%) and 83.17% (71.95%) respectively from the baseline method.

The rest of the paper is organised as follows. In Sect. 2 we discuss the relevant past literature. We describe our dataset in detail in Sect. 3. We next define the co-purchase network and study the metrics describing the position of a node in the network that can discriminate the popular products from the unpopular ones in Sect. 4. In Sect. 5 we describe the baseline features. We present the classification results in Sect. 6. Finally, we conclude in Sect. 7 by summarizing our key contributions and outlining some future directions.

2 Related Work

Revenue forecasting and sales prediction has been extensively studied for the entertainment industry especially in the context of box office revenues for motion pictures [1–4]. Almost all of these studies focus on mining online reviews to predict the sales performance. For instance, in [1,3] the authors use online reviews to construct novel diffusion models that are capable of accurate revenue forecasting. In [2], the authors analyse online ratings, and in particular, identify the valence of user ratings as a good predictor for motion picture revenue. Further, in [4], the authors perform detailed sentiment analysis of the review text to predict the sales.

With an exponential increase in the use of e-commerce platforms, there is an increased importance for the study of sales impact of different products on these platforms. While there has been a huge volume of work to design accurate recommender systems for e-commerce platforms to help consumers choose the best products [5–7], very little attention has been paid to analyse the future impact of a product. A few works that are remotely associated to this task are as follows. [8] studies the on-line shopping behaviour to improve user engagement on e-commerce sites. Amazon data has been used to study image-based recommendations [9] and to build complimentary product networks [10]. Work has been done to analyze Amazon e-commerce reviews to understand the (i) helpfulness of the reviews [11] and the (ii) latent sentiment in the reviews [12]. [13] posits that identity-descriptive information in the reviews can improve product sales.

In contrast, to the above approaches, we present for the first time a network-assisted method based on the co-purchasing behavior of the consumers to distinguish early the popular products from the unpopular ones. Our method is unique as it can work even in the complete absence of the online reviews which is usually the case immediately after the product launch.

3 Dataset Description

We have used the Amazon product data shared by McAuley et al. [9,10] for our experiments. The dataset contains reviews and metadata pertaining to the different products on Amazon. It consists of nearly 140 million reviews spanning from May 1996 to July 2014. It includes various review related information like ratings, text, helpfulness votes as well as product metadata which includes descriptions, category information, price, brand, image features, also bought products, also viewed products, etc. There are multiple broad categories in the dataset like books, electronics, CDs, cells, clothings, etc. We consider two of the largest categories – cells and CDs – for our study. There are 3,749,004 reviews and 492,799 product metadata entries corresponding to the category of CDs. Similarly, there are 3,447,249 reviews and 346,793 product metadata entries corresponding to the category of cells. We consider only those products which have greater than 1 review per month on average from the period 2010 to 2013. There are 2,624 such CDs and 11,564 such cells. We sort the products in each category according to sales-rank and consider various windows for the purpose of classification. We take top (bottom) k products with $k = 100$, 300, 500 and 700 from the sorted list of products and call them popular (unpopular). For each of these products we extract both network-centric as well as baseline features and build the binary classifier. As we shall show later, even for a weak separation at $k = 700$, the network-centric features perform notably better classification than the baseline features. Note that in all the classification experiments, we divide the dataset into training and test examples. All the products launched in the market before the start of 2013 (and are part of the list of the top (bottom) k products) are used for training while the rest are used for testing. We compute the features for each product in the training set using data till the end of 2012

and predict the sales-rank value of the newly launched products (after 2012) at the end of July 2014. This approach strictly ensures that there is no information leakage in our prediction scheme.

4 Network-Assisted Sales-Rank Characterization

In this section, we propose a novel network-assisted approach to characterize the sales-rank of Amazon products. In the rest of this section we define the network, extract important structural properties from this network indicating how a particular node is positioned and use these to establish strong differences between the popular and the unpopular products.

4.1 Co-purchase Network

We construct a co-purchase network of the products belonging to a category using the "also-bought" information available in the metadata. The co-purchase network has products as nodes and an edge between the nodes if two products are bought with one another. We remove all nodes from the network which have very low degree ($\leq$1).

4.2 Structural Properties of the Network

In the rest of this section we analyse the different structural properties of the co-purchase network to identify how a particular node (product) is positioned in this network. We further show how this positional information distinguishes a popular product from an unpopular one. In all the results that we present, the popular class comprises top 300 products as per sales-rank and the unpopular class comprises the bottom 300 products as per sales-rank. Note that for all the network features, these results remain very similar even when one considers top 100, 500 or 700 products as per sales-rank in the popular class and respectively bottom 100, 500 and 700 products in the unpopular class.

Clustering Coefficient. In our experiments, we use the definition of clustering coefficient based on triplets of nodes. For unweighted graphs, the clustering of a node u is the fraction of possible triangles that the node u is part of and is given by

$$c_u = \frac{2T(u)}{deg(u)(deg(u) - 1)}$$

where $T(u)$ is the number of triangles that the node u is part of and $deg(u)$ is the degree of u. A high clustering co-efficient indicates a densely connected neighborhood for a node. An interesting observation is that nodes corresponding to popular products tend to have higher clustering co-efficient, i.e., a denser neighborhood than the unpopular products in the co-purchase network (see Figs. 1 and 2 for the two product categories). This possibly indicates that popular products tend to be co-purchased with other popular products thus forming dense neighborhoods or "rich-clubs" [14] of popular products.

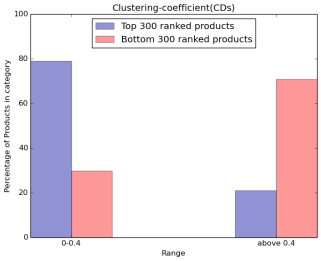

Fig. 1. Percentage of products from the two class vs clustering coefficient buckets (for CDs).

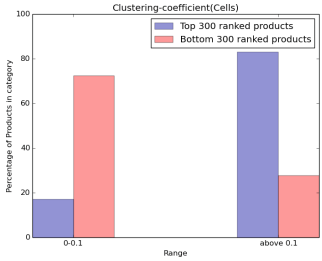

Fig. 2. Percentage of products from the two class vs clustering coefficient buckets (for cells).

Betweenness. Betweenness of a node in a graph measures the extent to which the node lies on paths between other vertices. It is equal to the fraction of the number of shortest paths from all vertices to all others that pass through that node. Betweenness is given by the equation:

$$g(v) = \frac{\sum_{s!=v!=t} \sigma_{st}(v)}{\sigma_{st}}$$

where σ_{st} is the number of shortest paths between the vertex pair (s, t) and $\sigma_{st}(v)$ is the number of those paths among these that pass through v.

Betweenness of a node in the co-purchase network is a reflection of how often the product corresponding to this node bridges two or more unrelated products. Since popular products are often bought with many other products, they tend to have higher betweenness. This is apparent from Figs. 3 and 4 where we plot for both the product categories, the percentage of popular and unpopular products in low and high buckets of betweenness. Most of the unpopular products have low betweenness while the popular ones have high betweenness.

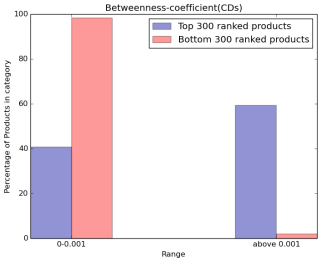

Fig. 3. Percentage of products from the two classes vs betweenness buckets (for CDs).

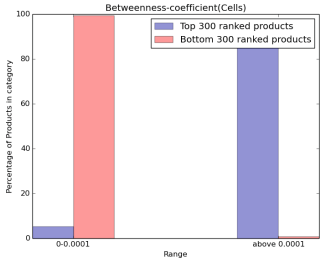

Fig. 4. Percentage of products from the two class vs betweenness buckets (for cells).

Closeness. Closeness centrality of a node u is the reciprocal of the sum of the shortest path distances from u to all $n-1$ other nodes (assuming there are n nodes in the network). Since the sum of the distances depends on the number of nodes in the graph, closeness is normalized by the sum of the minimum possible distances $n-1$. If the graph is not completely connected, we compute the closeness centrality for a node corresponding to its own connected component. Mathematically, the closeness is:

$$C(u) = \frac{n-1}{\sum_{v=1}^{n-1} d(v,u)}$$

where $d(v,u)$ is the shortest path distance between v and u.

In the co-purchase network, a popular product will be also bought with many other products and should therefore be close to most of the other nodes in the network. This is apparent from Figs. 5 and 6 where we plot for the two product categories, the percentage of popular and unpopular products in low and high buckets of closeness. For both the categories, the lower bucket of closeness has a large percentage of unpopular products while the higher bucket has a large percentage of popular products.

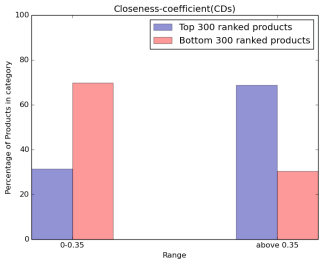

Fig. 5. Percentage of products from the two classes vs closeness buckets (for CDs).

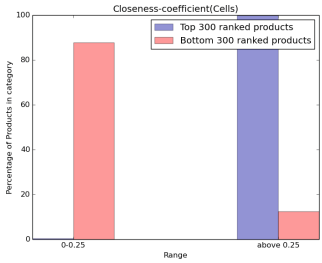

Fig. 6. Percentage of products from the two class vs closeness buckets (for cells).

Eigenvector. Eigenvector centrality is a measure of the recursive influence of a node in a network. It assigns relative scores to all nodes in the network based on the concept that connections to high-scoring nodes contribute more to the score of the node in question than equal connections to low-scoring nodes. Popular products tend to have a higher eigenvector centrality (see Figs. 7 and 8).

Pagerank. Pagerank works by counting the number and quality of links to a node (product) to determine a rough estimate of how important the product is. The underlying assumption is that more important products are likely to receive more links from other websites. Popular products are likely to have better Pagerank centrality (see Figs. 9 and 10).

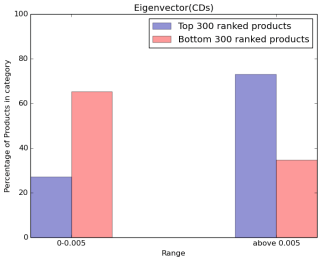

Fig. 7. Percentage of products from the two classes vs eigenvector buckets (for CDs).

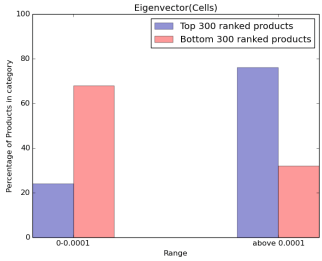

Fig. 8. Percentage of products from the two classes vs eigenvector buckets (for cells).

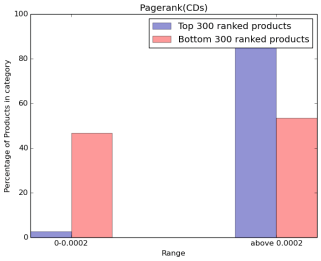

Fig. 9. Percentage of products from the two classes vs Pagerank buckets (for CDs).

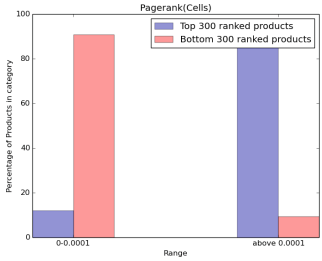

Fig. 10. Percentage of products from the two classes vs Pagerank buckets (for cells).

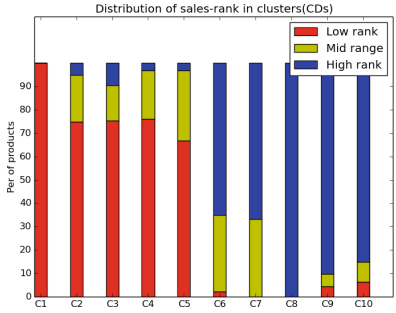

Fig. 11. Stacked bar plots showing the percentage of products from the different sales-rank classes in some representative communities (CDs).

Community Membership. We use the popular Louvain [15] community detection algorithm to find the community structure of the co-purchase network. We observe that the community memberships of the popular products are very distinct from that of the unpopular ones. Precisely, in a majority of

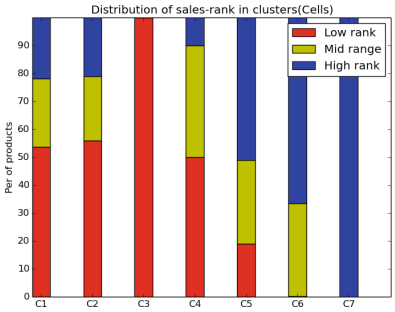

Fig. 12. Stacked bar plots showing the percentage of products from the different sales-rank classes in some representative communities (cells).

cases the communities are either mostly composed of only the higher sales-rank (i.e., the bottom class) products or mostly composed of only the lower sales-rank (i.e., the top class) products (see Figs. 11 and 12 for the two product categories respectively).

5 Baseline Features Based on Review Analysis

Since there is no baseline available in the literature for this problem, as an additional objective, we design a set of competitive baseline features through an extensive analysis of the reviews. The idea of using the reviews for building a baseline is inspired by similar approaches used for box office revenue forecasting [1–4]. The set of features can be further classified into (i) general and (ii) linguistic features. We discuss each of these in the following two subsections.

5.1 General Features

In this subsection, we shall define some general features extracted from the pattern of online reviews of the products. In all the results that we present, the popular class comprises the top 300 products as per sales-rank and the unpopular class comprises the bottom 300 products as per sales-rank. The separation weakens as more products from the top and the bottom zones of the sales-rank are included into the respective classes.

Volume of Reviews. Volume of reviews is expressed as the total number of reviews for a product in the span 2010–2013. Analysis of this feature for both the classes of products shows that popular products have a higher number of reviews as compared to unpopular products (see Figs. 13 and 14 for the two product categories).

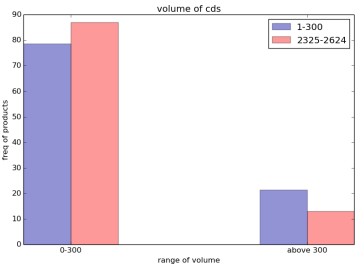

Fig. 13. Percentage of products from each class vs the volume of reviews (for CDs).

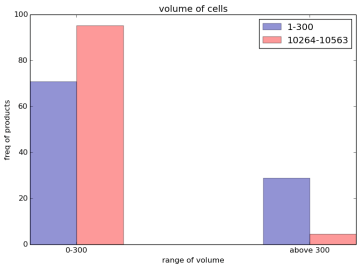

Fig. 14. Percentage of products from each class vs the volume of reviews (for cells).

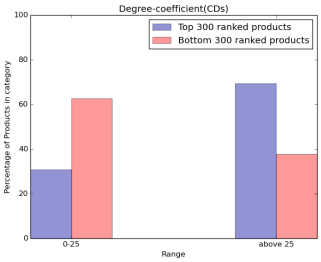

Fig. 15. Percentage of products from the two class vs co-purchase count buckets (for CDs).

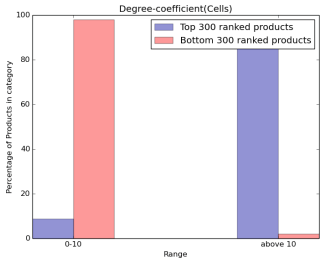

Fig. 16. Percentage of products from the two class vs co-purchase count buckets (for cells).

Number of Co-purchases. This feature counts the number of products with which a particular product is co-purchased. Note that this is also the degree of a node in the co-purchase network and we treat the same as a baseline feature to specifically show later that this trivial measure is not as good an indicator of sales-rank as the other non-trivial network measures described in the previous section. Figures 15 and 16 shows for the two product categories that a larger percentage of popular products fall in the high co-purchase count bucket while a larger percentage of unpopular products fall in the low co-purchase count bucket.

Percentage of Fake Reviews. We calculate the percentage of fake reviews by first classifying the reviews as real or fake using a Naïve-Bayes supervised learning approach based on standard tf-idf features. For training the classifier, we have used a proxy dataset available from Yelp that has a huge collection of review text which are already marked fake [16]. In Figs. 17 and 18 we observe that the percentage of fake reviews is more for an unpopular product as compared to a popular product. A possible reason for this is that certain users (for instance, the sellers themselves) might have vested interest in promoting an unpopular product.

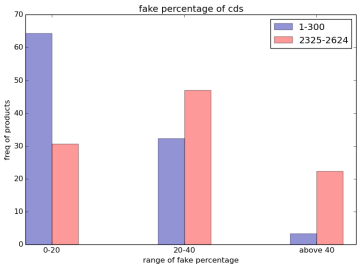

Fig. 17. Percentage of products from each class vs volume of fake reviews (for CDs).

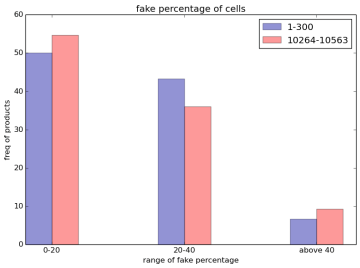

Fig. 18. Percentage of products from each class vs volume of fake reviews (for cells).

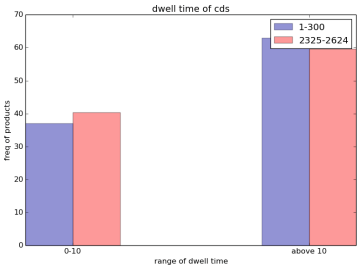

Fig. 19. Percentage of products from each class vs dwell time (for CDs).

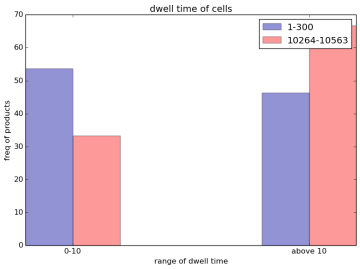

Fig. 20. Percentage of products from each class vs dwell time (for cells).

Dwell Time. In the context of the current problem, we define dwell time as the continuous stretch in number of months for which a product is receiving reviews. We observe from our analysis that average dwell time is more for popular products (see Figs. 19 and 20 for the two product categories).

Entropy of Ratings. Entropy of ratings is calculated as $-\sum_1^5 p_i \log p_i$, where p_i denotes the probability of a rating i across all the reviews of a product. A high entropy would indicate that the product receives diverse ratings from users while a low entropy would indicate similar ratings from all users. We see that unpopular products have higher entropy and thus users have more diverse/mixed opinion about them (see Figs. 21 and 22 for the two product categories).

5.2 Linguistic Features

In this section we perform extensive analysis of the review text to design various features based on the language structure. One again, the popular class comprises the top 300 products as per sales-rank and the unpopular class comprises the bottom 300 products as per sales-rank. The separation weakens as more products from the top and the bottom zones of the sales-rank are included into the respective classes.

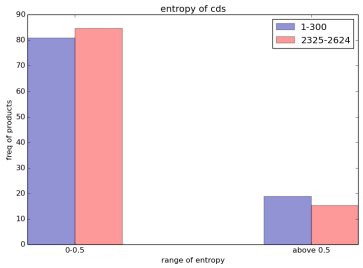

Fig. 21. Percentage of products from each class vs entropy of ratings (for CDs).

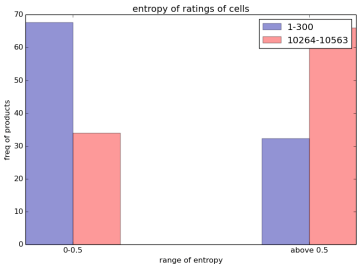

Fig. 22. Percentage of products from each class vs entropy of ratings (for cells).

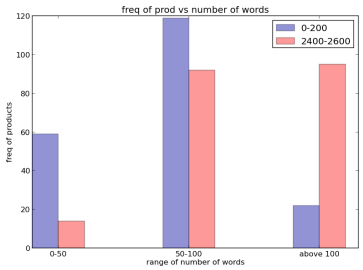

Fig. 23. Number of products from each class vs number of words (for CDs).

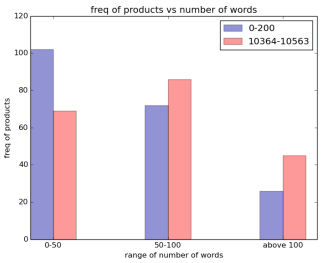

Fig. 24. Number of products from each class vs number of words (for cells).

Number of Words. We have taken the number of words in a review as one of the initial linguistic features. We observe that unpopular products have more lengthy reviews (see Figs. 23 and 24) and this might be due to the fact that people are usually not very satisfied with the product which compel them to give lengthy comments so that the product can be improved.

Word Diversity. Word diversity is defined as the entropy of fraction of words in a particular review which is calculated using the formula $-\sum_1^N p_i \log p_i$ where p_i denotes the count of a particular word i in a review divided by total length of the review i.e., N. We see that for popular products, entropy is less (see Figs. 25 and 26) as compared to the unpopular products which signify that reviews for popular products are more well-formed and linguistically better structured.

Anger. We next investigate some of the interesting linguistic factors using the LIWC[3] (Linguistic Inquiry and Word Count) text analysis tool [17]. The tool provides, as output, percentage of words in different categories for an input text. The categories are broadly divided into linguistic (21 dimensions like pronouns,

[3] http://liwc.wpengine.com/.

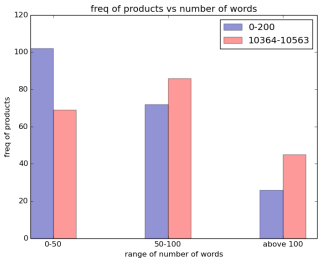

Fig. 25. Number of products from each class vs diversity of words (for CDs).

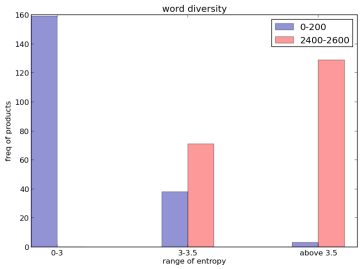

Fig. 26. Number of products from each class vs diversity of words (for cells).

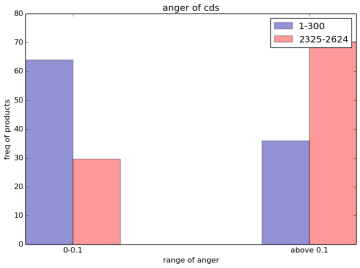

Fig. 27. Percentage of products from each class vs the extent of LIWC anger feature (for CDs).

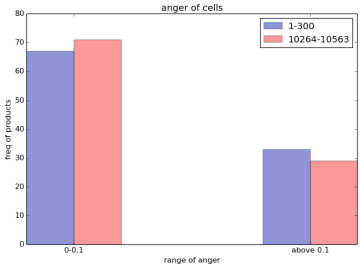

Fig. 28. Percentage of products from each class vs the extent of LIWC anger feature (for cells).

articles etc.), psychological (41 dimensions like affect, cognition etc.), personal concern (6 dimensions), informal language markers and punctuation apart from some general features like word count, words per sentence etc. The first factor that we find well differentiates between popular and unpopular products is the extent of anger in the reviews. While popular products have low anger content in their reviews, the unpopular ones have high anger content (see Figs. 27 and 28 for the two product categories.).

Sad. Next we report the extent of sadness in the review text from the suite of LIWC features. Figures 29 and 30 show that products with low average sadness values in their reviews are more probable to belong to the popular class while the opposite is true for the unpopular class.

Negative Emotion. Another discriminating LIWC feature is the extent of negative emotions present in the review text. Figures 31 and 32 show that a product with low average negative emotion value is more probable to belong to the popular class of products. The opposite is true for the unpopular class.

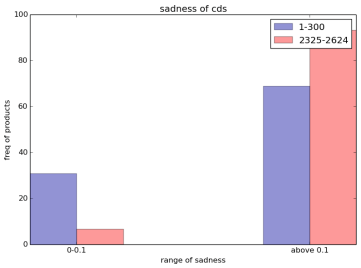

Fig. 29. Percentage of products from each class vs the extent of LIWC sadness feature (for CDs).

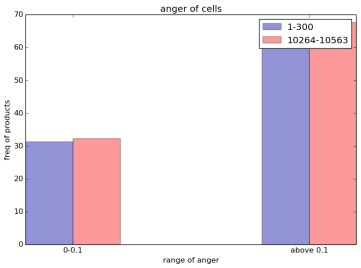

Fig. 30. Percentage of products from each class vs the extent of LIWC sadness feature (for cells).

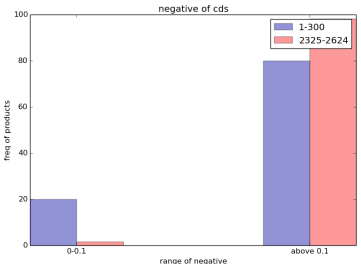

Fig. 31. Percentage of products from each class vs the extent of LIWC negative emotion feature (for CDs).

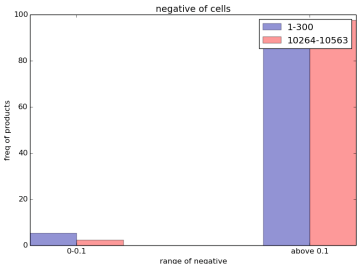

Fig. 32. Percentage of products from each class vs the extent of LIWC negative emotion feature (for cells).

All the results in this section together clearly portray that none of the review based features are as distinguishing as the network-centric features proposed in the previous section.

6 Classification

In this section, we design a binary classifier (linear SVM) to early predict the class of a product using the network features described earlier. We assign positive and negative labels (popular/unpopular) against the products by choosing appropriate thresholds for defining popularity. The top (bottom) k $(100, 300, 500, 700)$ products are classified as popular (unpopular). We split the products of a category into training and test sets. All the products launched before the start of 2013 are used for training while the rest are used for testing. We compute the features for each product in the training set using data till the end of 2012 and predict the sales-rank value of the newly launched products (i.e., since 2013 beginning) at the end of July 2014. This ensures a fair distribution of training set across the two classes. We present our results for two different product categories and different values of $k = 100, 300, 500, 700$ (see Table 1).

Table 1. Accuracy of classification (baseline and our method) for the two product categories.

Accuracy (CDs)		
Number of products	% Accuracy (baseline features)	% Accuracy (network features)
100	46.37	**89.85**
300	61.58	**92.68**
500	69.23	**86.15**
700	68.75	**82.10**
Accuracy (cells)		
Number of products	% Accuracy (baseline features)	% Accuracy (network features)
100	83.17	**84.11**
300	71.95	**84.15**
500	68.56	**81.44**
700	71.95	**84.8**

In parallel, we also learn an SVM classifier using the baseline features outlined above. Once again, we produce results for both the product categories as well as different values of k. In all cases, our network assisted approach overwhelmingly outperforms the baseline scheme.

Note that the classification based on network features by far outperforms the baseline features. Even for a weak separation of $k = 700$ we obtain an accuracy improvement of $\sim$19% for CDs and $\sim$17.8% for cells. Note that this result also shows that the trivial feature of the count of co-purchases (also the degree of a node in the co-purchase network) used as a part of the baseline features is not as effective in predicting the popularity class as the more non-trivial features based on network properties.

Importance of the network features: We further perform a χ^2 test to identify the importance of the individual network features in the classification for both the product categories. We find (see Table 2) that **community membership** is a very discriminative feature for both the products, **closeness** is more discriminative for cells while **clustering coefficient** is more discriminative for CDs.

Network + baseline features: A final question that one might ask is whether, the performance of the classifier improves if the baseline features are available and are used in addition to the network features. To answer this question, we report in Table 3 the accuracy we obtain by using both the network and the baseline features. As one would expect, in all the case, the improvements resulting from this combination is only marginal.

Table 2. χ^2 ranking of the network features.

χ^2 values		
Feature	% CDs	% Cells
Community membership	437.62	6.55
Eigenvector	3.31	0.27
Pagerank	0.1	0.01
Closeness	0.65	35.77
Clustering coefficient	14.40	1.40
Betweenness	0.94	0.91

Table 3. Accuracy of classification (network + baseline features) for the two product categories.

Accuracy (CDs)	
Number of products	% Accuracy (network + baseline features)
100	84.05
300	92.64
500	90.08
700	86.93
Accuracy (cells)	
Number of products	% Accuracy (network + baseline features)
100	83.17
300	84.48
500	84.06
700	84.62

7 Conclusion

In this paper, we presented a network-assisted method to early predict the popularity class of a product on Amazon. In particular we made the following contributions:

- We defined a co-purchase network and computed various positional information about individual nodes; these positional information turn out to be strong indicators of popularity of a product.
- Since there was no standard baseline for this work, we proposed a baseline contrived from co-purchase count, reviews and ratings feature of a product.
- We devised a classifier based on the network properties and showed that it outperforms the baseline by large performance margins. In specific, even for a weak separation between the popular and the unpopular products the improvement in accuracy is quite high.

– Among the network features, community membership, closeness and clustering coefficient metrics are found to be quite discriminative.

Such an early prediction could be extremely helpful for the entire business including the Amazon group, the sellers, investors and marketers portraying a clear picture of the sales impact of a product. This would also facilitate the design of suitable intervention measures to promote certain products to enhance the eventual sales figure as well as to decide if some product should be withdrawn from the marketplace. Through rigorous experiments we show that our results remarkably outperform the baseline approach built from traditional review analysis.

In future, we wish to perform similar analysis for other similar e-commerce platforms and identify if the network-assisted method has universal implications. Further, we would also like to investigate if such network-centric methods could be leveraged to study other market characteristics such as purchasing behavior of the customers, selling behavior of the sellers, the speed of sales etc.

References

1. Dellarocas, C., Awad, N., Zhang, X.M.: Using online reviews as a proxy of word-of-mouth for motion picture revenue forecasting. SSRN Electron. J. (2004)
2. Dellarocas, C., Awad, N., Zhang, X.M.: Using online ratings as a proxy of word-of-mouth in motion picture revenue forecasting. Working Paper (2005)
3. Dellarocas, C., Zhang, X.M., Awad, N.: Exploring the value of online product reviews in forecasting sales: the case of motion pictures. J. Interact. Mark. **21**, 23–45 (2007)
4. Yu, X., Liu, Y., Huang, J.X., An, A.: Mining online reviews for predicting sales performance: a case study in the movie domain. IEEE TKDE **24**, 720–734 (2012)
5. Schafer, J.B., Konstan, J., Riedl, J.: Recommender systems in e-commerce. In: Proceedings of the 1st ACM Conference on Electronic Commerce, pp. 158–166 (1999)
6. Sarwar, B., Karypis, G., Konstan, J., Riedl, J.: Analysis of recommendation algorithms for e-commerce. In: Proceedings of the 2nd ACM Conference on Electronic Commerce, pp. 158–167 (2000)
7. Wei, K., Huang, J., Fu, S.: A survey of e-commerce recommender systems. In: International Conference on Service Systems and Service Management, pp. 1–5 (2007)
8. Sharma, N.V., Khattri, V.: Study of online shopping behavior and its impact on online deal websites. Asian J. Manag. Res. **3**(2), 394–405 (2013)
9. McAuley, J., Targett, C., Shi, Q., van den Hengel, A.: Image-based recommendations on styles and substitutes. In: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 43–52. ACM (2015)
10. McAuley, J., Pandey, R., Leskovec, J.: Inferring networks of substitutable and complementary products. In: Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794. ACM (2015)
11. Mudambi, S.M., Schuff, D.: What makes a helpful review? A study of customer reviews on amazon.com. MIS Q. **34**(1), 185–200 (2010)

12. Jo, Y., Oh, A.H.: Aspect and sentiment unification model for online review analysis. In: Proceedings of the Fourth ACM International Conference on Web Search and Data Mining, pp. 815–824. ACM (2011)
13. Forman, C., Ghose, A., Wiesenfeld, B.: Examining the relationship between reviews and sales: the role of reviewer identity disclosure in electronic markets. Inf. Syst. Res. **19**(3), 291–313 (2008)
14. Colizza, V., Flammini, A., Serrano, M.A., Vespignani, A.: Detecting rich-club ordering in complex networks. Nat. Phys. **2**, 110–115 (2006)
15. Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. J. Stat. Mech: Theory Exp. **10**, P1000 (2008)
16. Mukherjee, A., Venkataraman, V., Liu, B., Glance, N.: What yelp fake review filter might be doing? In: Proceedings of the 7th International AAAI Conference on Weblogs and Social Media, pp. 409–418. ACM (2013)
17. Pennebaker, J.W., Boyd, R.L., Jordan, K., Blackburn, K.: The development and psychometric properties of liwc2015. UT Faculty/Researcher Works (2015)