In the above definition, the time constraint (b) is easy to check by comparing the timestamps of two transactions. The problem is how to determine whether the item constraint (a) is satisfied, i.e. if two items from different transactions actually refer to the same good. Although *Universal Product Code (UPC)* is ideal for identifying unique goods, it is not common for items in a real-world e-commerce marketplace to include *UPC*. A straightforward approach is to compare the titles of two items. For example, if the titles of items x and y are the same or use the same set of words, then x and y are the same; otherwise they are different goods.

2.3 Identifying Resale Via Elastic Matching

However, the exact matching identification is clearly not an appropriate criterion. On the one hand, the user A may change the title of item y in order to boost sales. For example, the item x was initially not listed with a suitable title. It is highly likely that item x would be sold at a low price. User A found this fact, and immediately bought this item. (S)he later listed the same product just bought using a more descriptive title, thus had a potential to sell at a higher price and make profits. Therefore, exact matching of item titles will miss a lot of meaningful resale activities. It is much desired that an **elastic matching** of items can be used to accurately identify resales.

On the other hand, the pattern in Figure 2(a) will not capture all resale activities due to the limitation of using the single account matching. A lot of people on a real-world e-commerce marketplace have more than one account. Meanwhile, members of a family might have their own individual accounts. For example, a person can have an account for selling and another account for buying because of privacy. Another scenario might be the husband bought some products online, and later his wife resold again because she did not like them. We call this type of user behavior in reselling as **account transfer**.

Therefore, a more general identification approach is proposed to address the above two issues, which is shown in Figure 2(b). Suppose the accounts of the same person/family are linked together, and the link is illustrated using the blue dotted line; and items referring to the identical goods are linked using the red dashed line. We want to discover the resale activities that satisfy:

DEFINITION 2 (**Elastic Matching Identification [EMI]**). *A user (User A) bought an item x and another user (User B) sold an item y. The sale activity tuple (x, A, B, y) is a resale if it satisfies the following constraints:*

(a) Item Constraint: *Item x and item y should be in different transactions, but they have a similarity score which is greater than a threshold α.*

(b) Time Constraint: *The purchase time of item x is earlier than the purchase time of item y.*

(c) Account Constraint: *User A and User B are linked because they belong to the same person or family (entity).*

The definition of EMI will help identify the case that resellers change the content of listings as well as the resale activities coming through account transfer. In the *Item Constraint*, a similarity function is needed to measure the similarity of two items. On e-commerce marketplaces, there are mainly three attributes to describe items: *Titles*, *Descriptions* and *Photos*. *Descriptions* are noisy which contains many irrelevant content, such as sellers' own stories, refund policies and shipping charges. In the meantime, usually the descriptions are lengthy, thus the computational costs of description similarity matching make it infeasible in a large-scale data environment. For *Photos*, even the state-of-art image comparison techniques cannot achieve satisfactory accuracy, and they are all ineffective for massive data. Considering the above limitations, we use the Jaccard similarity of item *titles* as the similarity function, which provides a good trade-off between accuracy and efficiency. Previous studies [15][29] in a similar context, QA archives, also demonstrate that using titles rather than descriptions for the similarity measure is of highest effectiveness. Let the titles of items x and y are T_x and T_y respectively. The similarity function between items x and y is defined as:

$$similarity(x, y) = \frac{|T_x \bigcap T_y|}{|T_x \bigcup T_y|}$$

where $|T_x \bigcap T_y|$ denotes the number of common words in T_x and T_y, and $|T_x \bigcup T_y|$ denotes the number of unique words in T_x and T_y.

If the similarity score of two items is greater than a given threshold α, the items are considered to be identical. While we understand that for any two random items x and y, even their similarity score is high enough, it is not necessary that they are the identical goods, because they may be different goods but the same type of product. However, considering the account constraint and time constraint, it is required that both items should be bought/sold by the same entity, and the purchase activities are in sequential order. Thus, these ensure that items satisfying above constraints refer to the identical goods.

3. EXTRACTING RESALE ACTIVITIES

In this section, we present how to extract resale activities from large-scale data sets at an e-commerce site given the criteria of resale activity identification in the previous section.

3.1 MapReduce Framework

In order to present how to extract resale activities, we first briefly describe the MapReduce framework. MapReduce [8] is a programming framework to support computation on large-scale data sets in distributed environments. The advantages of MapReduce are (1) the ability to run jobs in parallel (2) automatic management of data replication, transfer, load balancing, etc., and (3) the standardization of Map and Reduce procedures and concepts. MapReduce has been successfully adopted by many companies to handle massive data, including Yahoo, Google, Amazon, eBay, etc.

A typical MapReduce framework mainly contains two steps: Map step and Reduce step. The details of MapReduce can be found in [8] and [18]. Large e-commerce sites usually store multi-petabyte data on distributed machines. Therefore, we use MapReduce paradigm to extract resale activities from large-scale e-commerce transaction data.

3.2 Proposed Algorithms

Based on Definition 2, evaluating whether a transaction is a resale transaction requires to verify the account constraint, time constraint and item constraint. For account constraint, users from the same person/family should be grouped together as a user entity. The grouping policy includes matching of names, gender, addresses and user behaviors. Since

ALGORITHM 1: Account Matching

Input: Account Linking Table: $acc = \{(entity\_id, user\_id)\}$;
Transaction Table: $tran = \{(item\_id, buyer\_id, seller\_id, item)\}$

```
Account-Matching-Map(Table acc, Table tran)
begin
    for each (entity_id, user_id) ∈ acc do
        Output key-value pair (user_id, entity_id);
    end
    for each (item_id, buyer_id, seller_id, item) ∈ tran do
        Output key-value pair (buyer_id, (tag:"buy", item));
        Output key-value pair (seller_id, (tag:"sell", item));
    end
end

Account-Matching-Reduce(Key k, Value v[1...m])
begin
    output_key = null;
    for each v ∈ v[1...m] do
        if v is of type ID then
            output_key ← v;   break;
        end
    end
    for each v ∈ v[1...m] do
        if v is of type (tag,item) then
            Output key-value pair (output_key, v);
        end
    end
end
```

ALGORITHM 2: Item Matching

Input: Output from Account Matching

```
Item-Matching-Map(Key k, Value v)
begin
    Output key-value pair (k, v);
end

Item-Matching-Reduce(Key k, Value v[1...m])
begin
    buying_list, selling_list = empty;
    for each v ∈ v[1...m] do
        if tag == "buy" then
            add v into buying_list;
        end
        if tag == "sell" then
            add v into selling_list;
        end
    end
    for each v ∈ buying_list do
        for each v' ∈ selling_list do
            if v.timestamp < v'.timestamp  &&
            similarity(v, v') > α then
                Output resale activity (v, k, v');
            end
        end
    end
end
```

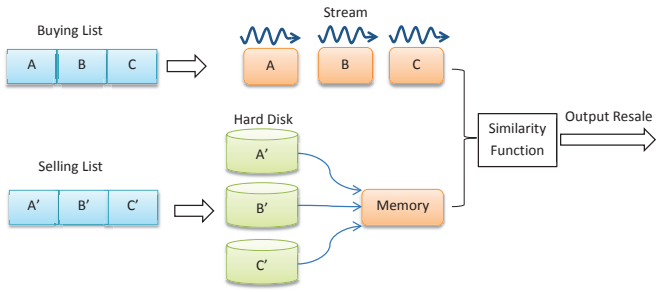

Figure 3: System Illustration: Stream-based Reducer. Lists being partitioned and read sequentially to resolve insufficient memory problem.

user grouping is not the focus of this study, we assume the user grouping data are pre-computed. All user entities have unique entity IDs and multiple accounts from the same person/family are linked to the same entity ID. For item constraint, a similarity function is applied to item titles, and the similarity score is used to determine if the two items represent the same good. Usually the transaction data on e-commerce sites are stored in different tables. For the simplicity of algorithm description, we assume all transaction data are stored in one table. This table can be regarded as the join result of a list of transaction related tables.

We propose a two-stage framework to extract resale activities. The first stage is to correlate items with the entity IDs to handle account transfer problem. The second stage is to generate resale transactions bought and sold by the same entity IDs based on elastic item matching and time constraint.

The pseudo code of the first stage is illustrated in Algorithm 1. In the Map step, the inputs are pre-computed account linking table and a table including all transaction data. A pair containing each *user_id* and its corresponding *entity_id* is sent to the reducer. In order to capture the buying and selling information, each item in transactions is mapped to two key-value pairs, i.e., a pair taking *buyer_id* as the key and a pair taking *seller_id* as the key. The type information regarding to buying or selling is stored as a tag for future processing. In the Reduce step of account matching, the *entity_id* of the user (stored in key k) is first obtained. Then we substitute the *user_id* k for the *entity_id*, and send all items associated with the same user to the next stage. Through the first stage, all items are linked to the entity ID. Thus even transactions from two different accounts belonged to the same person/family, they are aggregated in one place.

In the second stage, the idea is to create two lists of items associated with the same *entity_id* using the output of the first stage. For each entity, we collect its all purchased items and add into *buying_list*. Similarly, we get its all sold items and put into *selling_list*. The tag created in the first stage is used to decide which list a given item should be added into. It is clear that a resale activity must contain one item from *buying_list* and one item from *selling_list*. Thus, we further perform a pair-wise similarity matching of two lists, and if two items satisfy both item constraint and time constraint, they are output as resale activities. The details of the second stage can be found in Algorithm 2. Although the algorithms are described under the resale market context, we note that the proposed algorithms can be easily generalized to different applications which require matching.

3.3 Stream-based Approach

As shown in Algorithm 2, we can see that two item lists have to be stored in the memory. In a global e-commerce marketplace, it is not rare that a single user buys or sells millions of products annually. Therefore, Algorithm 2 becomes infeasible while handling such large-scale data. To make the algorithm work in practice, we extend Algorithm 2 to solve the insufficient memory problem. Inspired by [31], we introduce a stream-based MapReduce approach for Algorithm 2. We note that the proposed stream-based method is interesting in its own right as a general method for reducing memory requirement for large-scale MapReduce tasks, and may be useful for a number of different web-scale applications.

The stream-based reduce step is illustrated in Figure 3. The *buying_list* and *selling_list* in Algorithm 2 are further partitioned into blocks (*buying_list* $\rightarrow (A, B, C)$ and *selling_list* $\rightarrow (A', B', C')$). The size of blocks depends on the actual memory size of local machines. The blocks of one list are stored on the hard disk to prevent insufficient memory. In Figure 3, we store blocks of *selling_list* (A', B', C') on the hard disk. Each block from *selling_list* is read sequentially from the hard disk, and only one block can be stored in the memory at any given time. The blocks from *buying_list* (A, B, C) are sent as streams and match with the block of *selling_list* in the memory. As shown in Figure 3, blocks A', B', C' will be sequentially loaded into memory and match with block A from *buying_list*. After block A has matched with all blocks from *selling_list*, it can be safely removed from the memory, and the next block B from *selling_list* can be streamed in.

4. EXPLORING RESALE MARKET

After resale activities are collected, we explore what can be discovered from the resale market. In order to better understand the market, we apply data mining techniques and address two main problems in this section:

(1) **Observations**: What does the resale market look like? How large is the resale market? Who are doing resales? Why are they doing resales? Do different regional markets show similar patterns?

(2) **Prediction**: What factors lead to a successful (profitable) resale? Given a list of features, can we predict if a resale activity will be profitable?

4.1 Experiment Setup

We use an open source implementation of MapReduce, Hadoop[1], to extract resale activities. The Hadoop cluster stores over 10 petabytes transaction data from the marketplace. For the extracted resale activities satisfying pattern tuple (x, A, B, y) in Figure 2(b), we further obtain a list of attributes associated with users A, B and items x, y from the Hadoop cluster to analyze. As listed in Table 1, 24 attributes of resale activities from multiple sources are captured. Attributes related to x and A are listed in the type *Buying Related*, and attributes related to B and y are listed in the type *Selling Related*. Feedback scores are used to measure the overall rating of users. Generally speaking, a user with high feedback score is often an experienced seller/buyer with good reputation. Each user has two feedback scores: *feedback as seller* and *feedback as buyer*, which represent the user's selling and buying performance, respectively.

4.2 Validation

In order to study the effectiveness of the resale activity identification by elastic matching identification (EMI), we measure the correctness of matched resale activities and test the sensitivity of the parameter of EMI, *i.e.,* the similarity threshold α. We compare our method against the exact matching approach in Section 2.2 as the baseline. Since we do not directly have the ground truth, we use items sold on eBay which are associated with the eBay product catalog[2] as the validation data set. The product information of those items is either manually added by sellers or identified by a combination of UPC, brand, model and ISBN (for books). For each resale activity as tuple (x, A, B, y), we examine if items x and y belong to the same product in the catalog, and calculate the accuracy score to evaluate the proposed elastic matching method. One should note that we do not use the catalog information to obtain resale data due to its limited coverage. We only use the catalog information for evaluation purpose.

[1]http://hadoop.apache.org/

[2]http://pages.ebay.com/help/sell/product-details.html

Table 1: Extracted Attributes Related to Resale That Cover Feedback, Title, Time, Price, Categories, etc.

Type	Attributes
Buying Related	the original seller id, his/her feedback as seller, his/her feedback as buyer, buyer's entity id, user id, his/her feedback as seller, his/her feedback as buyer, item id, item title, purchase date, item price, site id, leaf category name, meta category name
Selling Related	seller's user id, his/her feedback as seller, his/her feedback as buyer, item id, item title, purchase date, item price, site id, leaf category name, meta category name

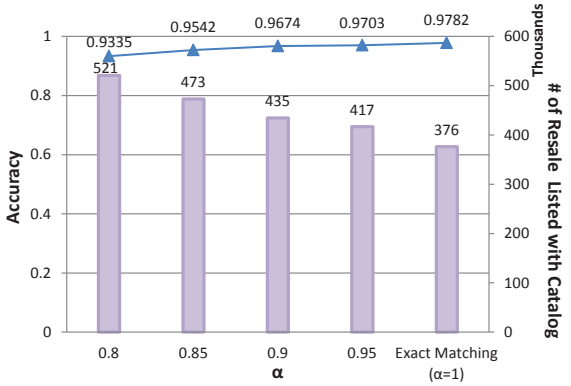

Figure 4: Accuracy Results with Threshold α. The line shows the accuracy scores (left Y-scale) by varying α and the bars represent the corresponding number of resale activities (right Y-scale) in thousands.

The results are reported in Figure 4. The threshold α is illustrated on the X-axis, the accuracy of EMI is illustrated on the left Y-axis and the number of resale activities with catalog information is presented on the right Y-axis. The line shows the accuracy scores by varying α and the bars present the corresponding number of resale activities. α being set to 1 corresponds to the exact matching approach, which is shown as the baseline. From the figure, one can observe that the number of resale activities is 376k for exact matching. For EMI, the number goes up to 521k while setting α to 0.8, which is 36.4% more than that of exact matching. From the effectiveness perspective, the accuracy score of exact matching is 0.9782, because of the noises in the real data set. Some sellers manually typed wrong product information, which causes mismatch even the titles of buying and reselling items are exactly the same. We further notice that the accuracy of EMI is over 93% while setting α is 0.8. This clearly shows the proposed elastic matching approach is highly accurate, and its obtained resale activities are indeed meaningful. We can also observe that the accuracy score further gains with the increase of α, while the number of resale activities decreases. This is quite nat-

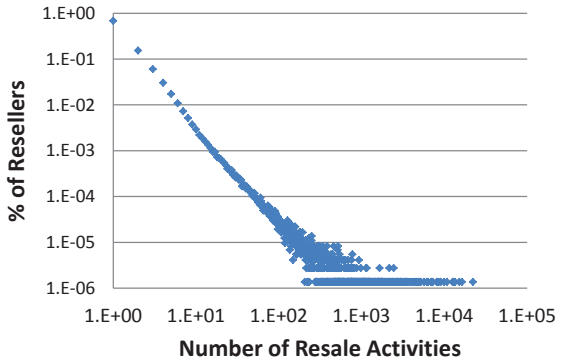

Figure 5: Distribution of the Number of Resale Activities. We note that the number of resale activities follows a power-law distribution with a long tail.

Figure 6: Top Keywords Added to Resale Listings. These frequently added keywords reflect the most common attributes that resellers think users care about.

ural since a tighter constraint on the title similarity reduces the chances of mismatch, but it also misses many reselling. By increasing α from 0.8 to 0.9, the improvement in accuracy is 3.39% but the number of obtained resale activities reduces 16.51%. Considering the tradeoff between accuracy and quantity, we use $\alpha = 0.8$ in the following analyses. Under this setting, we obtain a sample that includes over 13 million resale activities from the Hadoop cluster.

4.3 Observations and Analyses

The sample of resale activities is from over 30 regional sites, which cover 35 meta categories. In the following study, we analyze the resale market from various key perspectives and provide some interesting observations related to online resale.

Who Is Reselling? We first investigate the number of resellers and their resale patterns. We plot the distribution of the number of resale activities in Figure 5 on a log-log scale. The X-axis represents the number of resale activities per user, whereas the Y-axis shows the proportion of resellers who did the corresponding number of resales. From the figure, one can observe that resale follows a heavy-tailed distribution. Interestingly, top 10% resellers contribute 43% of resale activities. 68% of the resellers have only a single resale and they provide 18% of all resale activities. The figure shows that a large portion of users that resell very few times and a significant number of resale activities come from a small group of top highly active resellers. In addition, the trend interestingly follows a power-law distribution:

$$f(x) \propto x^{-\beta}$$

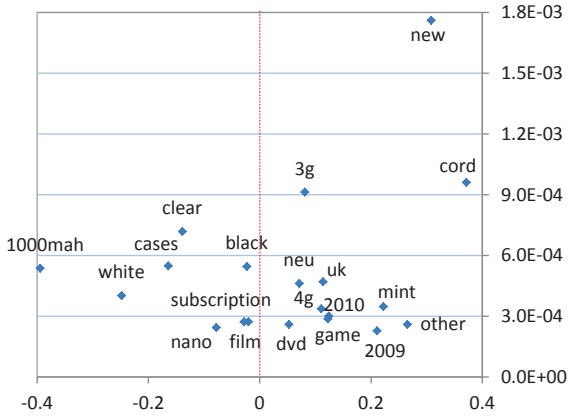

Figure 7: Average Relative Price Difference Versus Word Frequency. Note that X-axis and Y-axis denote the average relative price difference (ARPD) and word frequency, respectively. The higher the ARPD score, the more likely it is for resale to have profits by adding the given keyword.

where x represents the number of resale activities and β is computed to be 2.233.

What Content Do Resellers Change? On e-commerce sites, search results are mostly based on titles. Hence, many experienced sellers optimize titles to achieve better sales. We perform text analysis on the content that resellers added to new listings. For each resale activity, we obtain the added words from the reselling transactions, remove stop words and summarize them into an aggregated word cloud view in Figure 6. We observe that adding the keyword "new" is the most frequent as the largest tag cloud item. In summary, the most common words resellers added are the descriptions of product conditions ((brand) new, sealed, mint, etc.), colors (white, black, pink, etc.), appealing functions (game, dvd, clear, etc.), detailed specifications (1000amh, 3g, 4g, etc.) and years (2009, 2010, etc.). Actually these keywords correspond to the most common attributes that resellers think users care about.

In addition, to study the effectiveness of added words, we present a case study on the top 20 words with their frequencies and *average relative price differences* in Figure 7. The *average relative price difference* (**ARPD**) of a given word w is defined as

$$\frac{\sum_{(x,A,B,y)\in S, w\notin T_x, w\in T_y} \frac{P_y - P_x}{max(P_y, P_x)}}{|\{(x,A,B,y)\} : (x,A,B,y) \in S, w \notin T_x, w \in T_y\}|}$$

where the set S contains all resale activities of the pattern tuple (x, A, B, y). T_x and T_y represent the titles of items x and y, and P_x and P_y denote the corresponding prices. A positive ARPD score means the resale price is greater than the buying price, and vice versa. Clearly, higher ARPD represents greater resale values achieved by adding the word w. From the figure, we observe that "new" is the most frequently added words by resellers, and it is indeed quite effective in terms of resale values. Similarly, "mint" and "neu" (new in German) also have positive effects to increase resale prices. The word "cord" has a very high ARPD score. The reason is that it is usually associated with resales on accessaries such as adapters, cables and power supplies. Many such items were first bought from sellers in China and resold at higher prices to buyers in USA and UK. We further observe that

Table 2: Top 10 Resale Categories. Most significant categories are related to books or electronics.

Category	% of Total Resale
Clothes, Shoes & Accessories	11.1%
Used Books	9.7%
Computers & Networking	7.1%
Consumer Electronics	6.2%
Video Games	5.9%
Phones	5.8%
Home & Garden	4.3%
Cell Phones & PDAs	4.1%
Books	3.3%
Movies	3.0%

Table 3: Resale Activities on Regional Sites and Cross-sites. We observe that resale is an international and cross-border phenomenon.

Description	% of Total Resale
Same Site: USA	37%
Same Site: UK	34%
Same Site: Germany	16%
Same Site: Other	7%
USA ⇒ Other	3%
Other ⇒ USA	1%
Other ⇒ Other	2%

some other effective words are related to appealing functions, detailed specifications and years. Interestingly, adding words on colors do not increase resale prices in many cases. We note that "1000mah" especially has a low ARPD score. The reason is that this word is strongly tied with batteries which are considered as consumables, and batteries are priced fairly lower if used.

What Are the Top Categories of Resale? We list the top 10 resale categories in terms of percentile in Table 2. "Clothing, Shoes & Accessories" is the largest category in the resale market, and followed by "Used Books". In the meanwhile, many categories related to electronics appear in the top 10 list, including "Cell Phones & PDAs", "Computers & Networking", "Consumer Electronics", "Phones", etc.

Is Resale an International Activity? Table 3 shows the resale activity distribution on different regional sites. We can learn from the spatial analysis that most resold items are purchased and sold on the same regional sites. The three largest resale markets are USA, UK and Germany, probably because of the popularity of online shopping in these three countries. We also notice cross border resale activities. 5% of resale activities are purchased on one site but resold on another site. 3% of resale activities are bought on the USA site and resold on other sites.

What is the Correlation Between User Reputation and Price? In order to analyze the relationship between reputation and price, we plot the average relative price difference versus user feedback in Figure 8. We normalize the user feedback to range $[0:1]$, and a higher score represents the given user has a more positive rating. The *average relative price difference* with respect to feedback is defined as

$$\frac{\sum_{f(B)\in[r_i,r_{i+1})}\sum_{(x,A,B,y)\in S}\frac{P_y-P_x}{max(P_y,P_x)}}{|\{(x,A,B,y)\}:(x,A,B,y)\in S, f(B)\in[r_i,r_{i+1})\}|}$$

where $f(B)$ is the feedback score of user B and $[r_i, r_{i+1})$ is a given feedback score range. Higher ARPD represents greater

Figure 8: Average Relative Price Difference Versus User Feedback. Users in an online world with higher reputation tend to resell with larger profits. The red line illustrates a linear function that the data roughly fit.

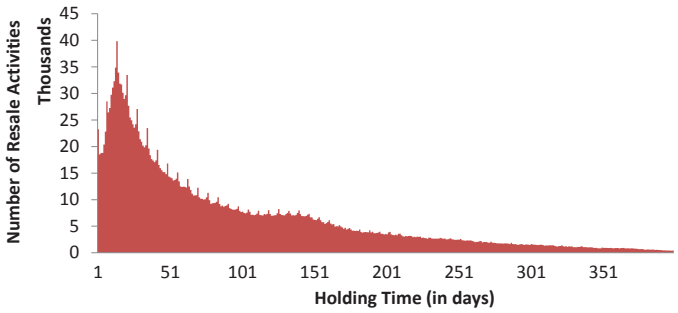

Figure 9: How Long Do Resellers Hold Items Before Resell. We note there is a peak in the range from 10 to 30 days.

profits achieved. There is an obvious trend that users with higher feedback scores achieve larger price increase, mostly because of the positive effects of online trust built in an online world. Although markets are considered stochastic, we can still observe that the data well fit a linear function (R^2 of 0.755).

How Long Do Resellers Hold Items? We show the temporal analysis that resellers hold items before resell in Figure 9. Clearly, most resellers hold items for less than 6 months. In particular, there is a significant peak in the range from 10 to 30 days. This indicates most resale activities happen almost immediately after the buyers receive the merchandise. It can also be concluded that very few resellers hold items for over one year to resell.

4.4 Prediction

As we have found many interesting insights of the resale market from the analysis above, we focus on the fact of successful resales in this subsection. From Figure 8, some resellers are making profits through resale activities, whereas some other users lose money. In order to have a quantitative understanding of what lead to a successful resale, we first analyze the features related to the profitability of resale. Then we use classification models to predict if a resale activity will be profitable given a set of features.

Table 4: Base Features Obtained for the Prediction Task

Type	Feature	Description
Feedback	ORG_SELLER_FDBK_AS_SLR	the *feedback as seller* of the original seller who sells the good to resellers
	ORG_SELLER_FDBK_AS_BYR	the *feedback as buyer* of the original seller
	BUY_FDBK_AS_SLR	the *feedback as seller* of the account reseller used to buy the good
	BUY_FDBK_AS_BYR	the *feedback as buyer* of the account reseller used to buy the good
	SELL_FDBK_AS_SLR	the *feedback as seller* of the account reseller used to resell the good
	SELL_FDBK_AS_BYR	the *feedback as buyer* of the account reseller used to resell the good
Title	BUY_AUCT_TITL	the title in the buying transaction
	SELL_AUCT_TITL	the title in the reselling transaction
Time	BUY_DATE	the purchase date of the buying transaction
	SELL_DATE	the purchase date of the reselling transaction
Site	BUY_SITE_ID	the buying transaction's listed regional site
	SELL_SITE_ID	the reselling transaction's listed regional site
Category	BUY_LEAF_CATEG_ID	the leaf category of the buying transaction
	SELL_LEAF_CATEG_ID	the leaf category of the reselling transaction
	BUY_META_CATEG_ID	the meta category of the buying transaction
	SELL_META_CATEG_ID	the meta category of the reselling transaction
Photo	BUY_PHOTO_COUNT	the number of photos in the buying transaction
	SELL_PHOTO_COUNT	the number of photos in the reselling transaction
Shipping	BUY_SHIPPING_FEE	the shipping fee in the buying transaction
	SELL_SHIPPING_FEE	the shipping fee in the reselling transaction
Quantity	BUY_QUANTITY	the number of items in the buying transaction
	SELL_QUANTITY	the number of items in the reselling transaction
Price	BUY_ITEM_PRICE	the purchase price in the buying transaction

4.4.1 Features Comparison

In order to build the classification models, we obtain a set of related features. For the problems of resale activities classification and prediction, the class labels are set to be binary, which are either *TRUE* (profitable) or *FALSE* (not profitable). For the classification purpose, we have obtained a list of base features from heterogeneous sources including transactions, user profiles, categories, sites, etc. They are listed in Table 4. There are 23 base features in total, which are related to 9 types:

- **Feedback**. The feedback is a score to measure the users performance. Can a user with higher feedback score sell with high profitability? Note that each resale activity is associated with two transactions: a buying transaction and a reselling transaction. Thus, separate features are created for both buying transactions and reselling transactions.
- **Title**. As we have shown in Figures 6 and 7, many resellers changed the listing titles for reselling. Does the length of title affect sale prices? Are titles the longer the better?
- **Time**. Do resale activities vary by seasons? Is there a 'best time' for reselling?
- **Site**. Do different regional markets show similar resale patterns? Which regional market is best for reselling?
- **Category**. We obtain both meta category and leaf category of each item. For example, for an iPod, its meta category is "Consumer Electronics", and its leaf category is "iPod & MP3 Players". Different categories may have different resale performances.
- **Photo**. Does the number of photos affect resale? "A picture is worth a thousand words." Can uploading more photos have a better presentation of products and thus boost sales?
- **Shipping**. The shipping charge is how much consumers need to pay for the shipping of the merchandise. Does free shipping help sales?
- **Quantity**. The number of goods sold in one transaction.
- **Price**. The price of resellers paid to buy the goods. The price of resellers resold is used to generate class labels, hence it is not listed as a feature.

Furthermore, we obtain 12 more features derived from base features. We show them in Table 5. The feature SAME_USER_ID is used to model how account transfer can affect resales. 5 derived features are binary, e.g. SAME_TITL, SAME_SITE, etc, which represent if there are some changes of attributes between buying transactions and reselling transactions. Two features related to the lengths of titles are used to represent the effects of titles. Furthermore, we extract month-of-the-year from date as another type of feature, which helps the model analyze the temporal effects. Other derived features present the numeric differences of attributes, such as time and photo count.

For the 35 features including base features and derived features, it is important to test which features have the most discriminative power in terms of profitability. Hence we measure the importance of features based on three criteria, namely Gain Ratio (GR), Information Gain (IG) and Chi-squared statistic (Chi). The top 15 discriminative features for each criterion are illustrated in Figure 10. The higher GR/IG/Chi score is, the higher discriminative power the corresponding feature has. In addition, we link the same features across different criteria to illustrate the similarities of the outputs.

From the figure, it is clear that the features selected by the three criteria highly overlap. The results on Informa-

tion Gain and Chi-squared statistic are especially close. We further observe that the feedback scores and leaf categories are the most discriminative features, which are agreed by all three criteria. It is quite natural that feedback scores are important. The reason is that buyers trust sellers with high feedback scores and sellers with high feedback scores are more likely to have good selling skills. The leaf categories are also discriminative. This suggests that products from different categories have different resale potentials. For example, reselling products in Consumer Electronics and Art are easier to get better offers, whereas reselling in Clothing and Shoes is much more difficult to being profitable. Besides feedback scores and categories, TITL_LEN_DIFF and SAME_TITL are also two discriminative features. Clearly it represents a carefully rewritten title will benefit resales. Overall, the most important features are related to category, feedback and title. Besides the features listed in Figure 10, BUY_MONTH and SELL_MONTH are also two interesting features worth to mention. Through our further analysis of data, we find that usually buying in the summer and selling in December and January will make the best profits. The reason might be a lot of people are on vacation during summer time, so it is a low season with fewer buyers and bidders. And December and January are the holiday seasons that bring more people to shop online.

4.4.2 Model Comparison

Next, we investigate the prediction of resale activities using the base and derived features. We construct a number of feature sets corresponding to feature types, and build classifiers for each feature type as baselines. We further compare those classifiers with the classifier built on all features. We split the data into 75% as the training set and 25% as the testing set. SVM is used as the classification model. We evaluate the performance on four metrics: **Precision**, **Recall**, **F-score**, and **ROC**. For all these four measure, the higher scores represent the better performance.

The classification results on various feature sets are illustrated in Table 6. Among all separate feature sets, Category achieves the best performance, with an accuracy score at 65.3%. It suggests the profitability of resale depends largely on what categories the items are listed in. This exactly matches our previous feature analysis in Figure 10. We further discover that the model with all features outperforms the baseline models with separate feature sets. These results demonstrate that different features provide different insights on predictions. Data mining techniques can automatically learn the importance of different feature types and output satisfactory results with all features considered. We further test 4 additional classification models and list the results in Table 7. The overall best classifier we test is the Decision Tree, with an accuracy score of 75.4% and an F-score of 0.826. It gives an 8% improvement over the SVM model and a 17% improvement over the Naïve Bayes model. All the results are clear evidences that the selected base and derived features are indeed essential, and data mining techniques can generate satisfactory performance for resale market predictions.

Table 6: Effectiveness of Various Feature Sets. The model that combines all features outperforms baselines with standalone feature sets on all measures.

Feature	Accuracy	Precision	Recall	F1	ROC
Feedback	58.2%	0.640	0.582	0.610	0.581
Title	56.6%	0.567	0.566	0.565	0.566
Time	54.6%	0.547	0.546	0.545	0.546
Site	54.3%	0.543	0.543	0.543	0.543
Category	65.3%	0.653	0.653	0.653	0.653
Photo	53.5%	0.564	0.536	0.550	0.537
Shipping	58.4%	0.588	0.584	0.580	0.584
Quantity	51.8%	0.608	0.518	0.559	0.520
All	**69.8%**	**0.755**	**0.668**	**0.709**	**0.686**

5. APPLICATIONS

So far, we have analyzed the resale market and studied the prediction on the profitability of resale activities. In this section, we discuss the eBay applications of the findings in previous sections. We show that, by incorporating the classification results, we can improve the efficiency of the resale market and improve both buyer and seller experience on web marketplaces.

From the classifiers we analyzed in the prediction section, we use the decision tree as the model to demonstrate the application scenarios. The advantages of using the decision tree model are its accuracy and a list of classification rules it can generate. Those rules can be easily interpreted by human beings and thus applied directly as business logics in an e-commerce system. The decision tree model generates over 18,000 rules in total for the resale activity classification task.

Since many rules generated by decision tree are either not statistically significant or not accurate enough for a real-world system, we further set two filters to select highly reliable rules:

- the minimum support of selected rules is S_{min}
- the minimum confidence of select rules is C_{min}

Using these selected highly reliable rules filtered by the minimum support and minimum confidence, we perform instance matching for each resale activity. For each incoming resale activity, if it satisfies one of the rules, the predicted label should be extremely accurate, since all selected rules are filtered based on support and confidence.

In the rest of this section, we give three example scenarios on improving a real production e-commerce system with the deployment of selected rules. We note that all these applications become possible and effective because of the extremely large data set and accurate prediction models.

Improving Reselling: Based on added words shown in Figures 6 and 7, a recommender systems can be applicable to suggest useful keywords according to frequencies and ARPD scores to make the resale title more descriptive. Besides text, we are also able to provide suggestions on category, transaction time, photo, shipping fee, user feedback, etc. Through our study of extracted rules, we find that many of them can be directly incorporated into the system and make resales more effective.

Improving General Buying: The selected reliable rules can not only help resellers, but also improve the experience of online shoppers. For example, when a buyer is browsing an item, the system can match his/her profile and the item with selected rules, and provide more personalized suggestions if the product has a high chance of resale. In general, useful information can be inferred from the classification model and be suggested to users. This feature can

Table 5: Derived Features Created for the Prediction Task

Type	Feature	Description
Account	SAME_USER_ID	binary: if account transfer occurred in the resale activity
Title	BUY_AUCT_TITL_LEN	the length of title in the buying transaction
	SELL_AUCT_TITL_LEN	the length of title in the reselling transaction
	TITL_LEN_DIFF	the difference between buying and selling title lengths
	SAME_TITL	binary: if resellers change the titles
Time	BUY_MONTH	the purchase month of the year in the buying transaction
	SELL_MONTH	the purchase month of the year in the reselling transaction
	TIME_DIFF	the waiting period to resell = reselling time - buying time
Site	SAME_SITE	binary: if the buying and selling transactions are on the same site
Category	SAME_LEAF_CATEG	binary: if resellers change the leaf categories
	SAME_META_CATEG	binary: if resellers change the meta categories
Photo	PHOTO_COUNT_DIFF	the difference between buying and selling photo counts

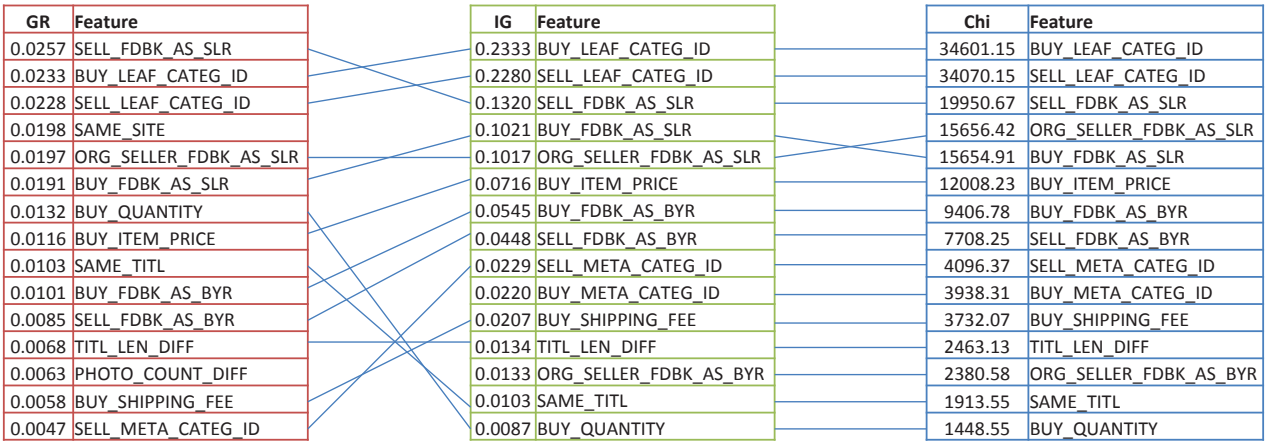

Figure 10: Feature Analysis: Gain Ratio (GR), Information Gain (IG) and Chi-squared statistic (Chi)

Table 7: Effectiveness of Various Models. The decision tree model achieves the best performance on four out of five measures.

Model	Acc.	Pre.	Rec.	F1	ROC
Naïve Bayes	64.6%	0.650	0.646	0.644	0.710
Log Regression	66.7%	0.669	0.667	0.666	0.710
Decision Tree	**75.4%**	**0.753**	**0.914**	**0.826**	0.718
Nearest Neighbor	70.9%	0.709	0.709	0.709	**0.721**

be further integrated with commercial "Trade-in Programs" provided by many online markets[3]. This will also potentially increase the revenue of the whole marketplace.

Improving General Selling: After analyzing the selected rules, we notice that many under-priced sales are due to incorrectly or inappropriately listed categories. To address incorrectly listed items, we may use an accurate classifier to suggest users other possible options. For those items that are listed inappropriately, such as in "Everything Else" or "Other" category, the system may infer the actual categories of the items based on titles, descriptions and photos, and then advise the seller to sell at a more specific category rather than the vague "Everything Else" or "Other" category.

6. CONCLUSIONS AND FUTURE WORK

In this paper, we propose the first systematic framework to mine and analyze a large-scale online resale market. We develop a stream-based MapReduce approach to process peta-scale data, and discover millions of resale activities through elastic matching identification (EMI) with high accuracy ($> 93\%$). We discover that resale activities follow a power law distribution with a 'long tail' phenomenon, where a large portion of them are contributed by a small number of highly active resellers. In the meantime, resale is an international and cross-border behaivor and appears in all different categories of products. We observe that adding useful keywords and building user trust in the online world have positive effects towards resale values. We further utilize data mining models to empirically evaluate a number of features from different sources and predict the profitability of resale activities. Finally we propose three application scenarios to demonstrate how to incorporate the above models to a real-world e-commerce marketplace. It will not only increase revenues in terms of resale but also improve both general buyer and seller experience on web markets.

In our future work, we will consider developing a similarity matching method by incorporating the entities and semantics of listing items that can scale to web-scale data sets. We will also consider applying the proposed system to other user-to-user web applications.

7. ACKNOWLEDGMENT

The first author was a summer intern at eBay Research Labs while this work was done.

[3]http://instantsale.ebay.com, http://www.amazon.com/Trade-In, http://www.bestbuy.com/site/Misc/Buy-Back-Program

8. REFERENCES

[1] http://www.kronikmedia.co.uk/blog/content-curation-benefits/4433.

[2] C. C. Aggarwal, Y. Zhao, and P. S. Yu. On clustering graph streams. In *SDM*, pages 478–489, 2010.

[3] C. C. Aggarwal, Y. Zhao, and P. S. Yu. Outlier detection in graph streams. ICDE '11, pages 399–409, 2011.

[4] P. Bajari and A. Hortacsu. The winner's curse, reserve prices, and endogenous entry: empirical insights from ebay auctions. *RAND Journal of Economics*, pages 329–355, 2003.

[5] S. Bukhchandani and C. Huang. Auctions with resale markets: An exploratory model of treasury bill markets. *Review of Financial Studies*, 2(3):311–339, 1989.

[6] P. Chinloy. An empirical model of the market for resale homes. *Journal of Urban Economics*, 7(3):279–292, 1980.

[7] P. Courty. Some economics of ticket resale. *The Journal of Economic Perspectives*, 17(2):85–97, 2003.

[8] J. Dean, S. Ghemawat, and G. Inc. Mapreduce: simplified data processing on large clusters. 2004.

[9] I. Dhillon, Y. Guan, and B. Kulis. A fast kernel-based multilevel algorithm for graph clustering. In *KDD*, pages 629–634, New York, NY, USA, 2005.

[10] Q. Duong, N. Sundaresan, N. Parikh, and Z. Shen. Modeling seller listing strategies. *Agent-Mediated El. Comm*, 2010.

[11] S. Guo, M. Wang, and J. Leskovec. The role of social networks in online shopping: information passing, price of trust, and consumer choice. In Y. Shoham, Y. Chen, and T. Roughgarden, editors, *ACM EC*, pages 157–166, 2011.

[12] P. Haile. Auctions with private uncertainty and resale opportunities. *Journal of Economic Theory*, 108(1):72–110, 2003.

[13] A. Hosios and J. Pesando. Measuring prices in resale housing markets in canada: Evidence and implications*. *Journal of Housing Economics*, 1(4):303–317, 1991.

[14] D. Houser and J. Wooders. Reputation in auctions: Theory, and evidence from ebay. *Journal of Economics & Management Strategy*, 15(2):353–369, 2006.

[15] J. Jeon, W. B. Croft, and J. H. Lee. Finding similar questions in large question and answer archives. CIKM '05, pages 84–90, New York, NY, USA.

[16] H. Kautz, B. Selman, and M. Shah. Referral web: combining social networks and collaborative filtering. *Commun. ACM*, 40:63–65, March 1997.

[17] T. Kudo, E. Maeda, and Y. Matsumoto. An application of boosting to graph classification, 2004.

[18] R. Lämmel. Google's mapreduce programming model - revisited. *Sci. Comput. Program.*, 68:208–237, October 2007.

[19] J. Leskovec, K. J. Lang, and M. Mahoney. Empirical comparison of algorithms for network community detection. WWW, pages 631–640, New York, NY, USA, 2010.

[20] D. Lucking-Reiley. Auctions on the internet: Whatąś being auctioned, and how? *The Journal of Industrial Economics*, 48(3):227–252, 2000.

[21] D. Lucking-Reiley, D. Bryan, N. Prasad, and D. Reeves. Pennies from ebay: The determinants of price in online auctions. *The Journal of Industrial Economics*, 55(2):223–233, 2007.

[22] H. Marvel and S. McCafferty. Resale price maintenance and quality certification. *The RAND Journal of Economics*, pages 346–359, 1984.

[23] M. Melnik and J. Alm. Does a sellerąś ecommerce reputation matter? evidence from ebay auctions. *The journal of industrial economics*, 50(3):337–349, 2002.

[24] S. Pandit, D. Chau, S. Wang, and C. Faloutsos. Netprobe: a fast and scalable system for fraud detection in online auction networks. In *WWW*, pages 201–210, 2007.

[25] P. Resnick, R. Zeckhauser, J. Swanson, and K. Lockwood. The value of reputation on ebay: A controlled experiment. *Experimental Economics*, 9(2):79–101, 2006.

[26] A. Roth and A. Ockenfels. Last minute bidding and the rules for ending second price auctions: evidence from ebay and amazon auctions on the internet. *American Economic Review*, 92(4):1093–1103, 2002.

[27] G. Shaffer. Slotting allowances and resale price maintenance: a comparison of facilitating practices. *The RAND Journal of Economics*, pages 120–135, 1991.

[28] Z. Shen and N. Sundaresan. ebay: an e-commerce marketplace as a complex network. WSDM '11, pages 655–664, New York, NY, USA, 2011.

[29] A. Shtok, G. Dror, Y. Maarek, and I. Szpektor. Learning from the past: answering new questions with past answers. WWW '12, pages 759–768.

[30] P. Shuchman. Profit on default: An archival study of automobile repossession and resale. *Stan. L. Rev.*, 22:20, 1969.

[31] R. Vernica, M. J. Carey, and C. Li. Efficient parallel set-similarity joins using mapreduce. SIGMOD '10, pages 495–506, New York, NY, USA, 2010.

[32] G. Wang, Y. Zhao, X. Shi, and P. S. Yu. Magnet community identification on social networks. KDD '12, pages 588–596.

[33] Y. Zhao, C. C. Aggarwal, and P. S. Yu. On graph stream clustering with side information. In *SDM*, 2013.

[34] Y. Zhao, X. Kong, and P. S. Yu. Positive and unlabeled learning for graph classification. In *ICDM*, pages 962–971, 2011.