Analysis of Online Product Purchase and Predicting Items for Co-purchase

Sohom Ghosh, Angan Mitra, Partha Basuchowdhuri and Sanjoy Kumar Saha

Abstract In recent years, online market places have become popular among the buyers. During this course of time, not only have they sustained the business model but also generated large amount of profit, turning it into a lucrative business model. In this paper, we take a look at a temporal dataset from one of the most successful online businesses to analyze the nature of the buying patterns of the users. Arguably, the most important purchase characteristic of such networks is follow-up purchase by a buyer, otherwise known as a co-purchase. In this paper, we also analyze the co-purchase patterns to build a knowledge-base to recommend potential co-purchase items for every item.

Keywords Viral marketing · Dynamic networks · Social networks · Recommendation system · Amazon co-purchase networks

1 Motivation and Related Works

Online market places are becoming largely popular with knowledge of internet among customers. Web-based marketplaces have been active in pursuing such customers with high success rate. For example, a large portion of the Black Friday

S. Ghosh · P. Basuchowdhuri (✉)
Department of Computer Science and Engineering, Heritage Institute of Technology, Chowbaga Road, Anandapur, Kolkata 700107, WB, India
e-mail: parthabasu.chowdhuri@heritageit.edu

S. Ghosh
e-mail: sohom1ghosh@gmail.com

A. Mitra · S.K. Saha
Department of Computer Science and Engineering, Jadavpur University, 188 Raja S. C. Mullik Road, Jadavpur, Kolkata 700032, WB, India
e-mail: anganmitra@outlook.com

S.K. Saha
e-mail: sks_ju@yahoo.co.in

A. Nagar et al. (eds.), *Proceedings of 3rd International Conference on Advanced Computing, Networking and Informatics*, Smart Innovation, Systems and Technologies 43, DOI 10.1007/978-81-322-2538-6_60

sales in USA or the Boxing Day sales in UK have gone online so that instead of waiting outside of the stores in cold for hours, the customers can sit in their home comfortably and still avail the deals by browsing through their websites. The online shopping stores have gone to great lengths to invest on high configuration servers that could seamlessly handle the requests thousands of buyers at the same time.

Such alarming growth in business in online market-places has drawn interest towards analyzing the customers' browsing or buying history. The customers' browsing or buying data is stored in the database and such data could be analyzed to observe patterns to understand the customers better. One such feature, popularly used by online market-places to recommend users for a potential sale, is the item co-purchase pattern of the customers. Customers often buy multiple products together, which may or may not be related. The frequency of a co-purchase may indicate the probability of the same co-purchase in future. The co-purchase patterns changes with time depending on many temporal factors. Understandably, this problem is a temporal problem and data from different time instances are needed to understand the dynamic behaviour of the customers.

In a study on Amazon co-purchase network Leskovec et al. [1], features of person-to-person recommendation in viral marketing were disclosed. For convincing a person to buy something, such recommendations were not very helpful. But they showed partitioning data based on certain characteristics enhances the viral marketing strategy. E-commerce demand has been explained in another paper using Amazon co-purchase network. Their claim is that item categories with flatter demand distribution is influenced more by the structure of the network [2]. A community detection method was suggested by Clauset et al. which took O (md log n) time where, n, m are the number of nodes and edges respectively and d is the number of hierarchical divisions needed to reach the maximum modularity value. Overall goodness of detected communities is popularly measured by Modularity [3]. This method is popularly known as CNM [4] and they have used Amazon co-purchase network as a benchmark data to find the communities in the network. The communities detected by their algorithm has maximum modularity 0.745. But, these communities were so large that finding out patterns from them was not of much use. For instance, the largest community having about 100,000 nodes consisted of about one-fourth of the entire network. Luo et al. made a study of the local communities in Amazon co-purchase network and claimed that recommendation yields better results for digital media items than books [5]. 3 and 4 node motifs had been analysed in Amazon co purchase network [6]. It was found that frequent motifs did not contribute much in comprehending the behaviour of the network. Recent works on detection of frequent sub-graphs [7–10] has helped us in interpreting the dynamics of temporal network. J. Han et al. devised a method popularly known as FP-growth to find frequently occurring patterns in transaction databases [11]. C. Bron and J. Kerbosch formulated an algorithm to find out maximal cliques in an network [12]. Basuchowdhuri et al. studied the dynamics of communities in amazon copurchase network [13]. They have analysed the evolving product purchase patterns.

2 Problem Definition

Given a directed graph $G(V, E)$, where nodes are products of an e-commerce transaction database and directed edges represent the co-purchase relation, we analyze market dynamics, relation between some of the inherent properties of the graph, temporal human predilection, variation in number of reviews and build a recommendation system that would help in bolstering the revenue. The frequency of co-purchase although important, was not available in the dataset. Hence, we ignore the frequency of co-purchased items and use topological characteristics to build a recommender system.

3 Examining Features of the Co-purchase Network

We have examined the Amazon co-purchase network data in this paper. Unlike other social network datasets, temporal characteristics are present in this dataset, i.e., it comprises a network and its snapshots at four timestamps. The snapshots reveals the dynamism in the structure of the original network and enables us to study the evolving follow-up purchasing patterns. For the ease of comprehending, we represent this dynamic network as a set of time-stamp graphs G_0, G_1, G_2, G_3, where G_0 is the original network at t = 0 and the evolved versions of G_0, after one, two and three units of time are G_1, G_2 and G_3, respectively. We them tabulate them as follows in Table 1. In this table, LWCC and LSCC refers to largest weakly connected component and largest strongly connected component respectively. Considering all time stamps the number of bi-directed edges are 5,853,404 while that of distinct edges are 10,847,450. We subsequently look into a few features of the network. Firstly, we look into its extended reachability by looking at the nodes' 2-hop degree. The 2-hop degree distribution expresses the extent to which viral marketing would be able to boost up sales in the network. Next, we review two features of the network that gives us a knowledge of how good transitivity of co-purchase is maintained in the network. The last two features observed, express the distribution of review writing by the customers. We relate the review writing with buying and thereby project the buying frequency and its distribution.

Table 1 Snapshot graphs in amazon co-purchasing network

Graph	|V|	|E|	Size of LWCC	Size of LSCC	Global clustering co-efficient	Month, Year
G_0	262,111	1,234,877	262,111	241,761	0.4198	March 02, 2003
G_1	400,727	3,200,440	400,727	380,167	0.4022	March 12, 2003
G_2	410,236	3,356,824	410,236	3,255,816	0.4064	May 05, 2003
G_3	403,394	3,387,388	403,364	395,234	0.4177	June 01, 2003

3.1 2-Hop Degree Distribution of Items

Degree distribution of a typical social network is known to display a plot of scale-free nature. But, 2-hop degree of a node represents its ability to extend its reach beyond its immediate neighbors. Its significance lies in the realization of influence that a node exerts on its neighbors' neighbors. Such measures in sparse networks can be useful as average reachability between nodes is high. In sparse networks, the nodes with high 2-hop degree value are the central nodes, from which all the nodes can be reached quickly. Figure 1 shows the 2-hop degree distribution of Amazon co-purchase network in log-log scale. The distribution plot follows power law with a heavy tail.

3.2 Clustering Coefficient Distribution

Local clustering coefficient gives an essence of how good or bad a node can pulls others into formation of a dense network, a quasi-clique for example. Figure 2 shows the distribution of clustering coefficient in log-log scale. Nodes with high clustering coefficient (CC) has appeared more number of times than those with lower CC. This reveals the cliquishness of the network.

3.3 Triplet and Triangle Distributions

A triplet is a structure where a node spawns two children. A triplet provides with an opportunity of a triad closure which is the minimum unit of friendship. For each

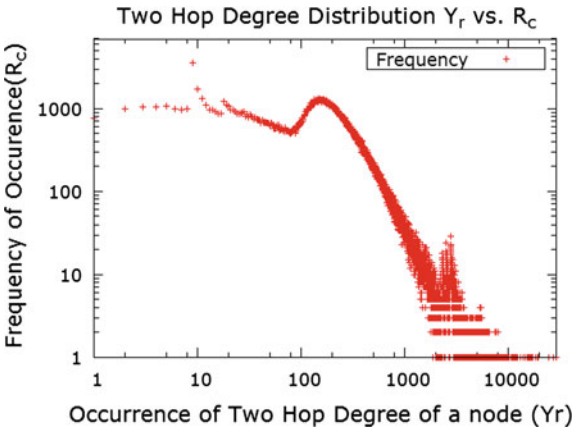

Fig. 1 2-hop degree distribution of amazon co-purchase network in log-log scale

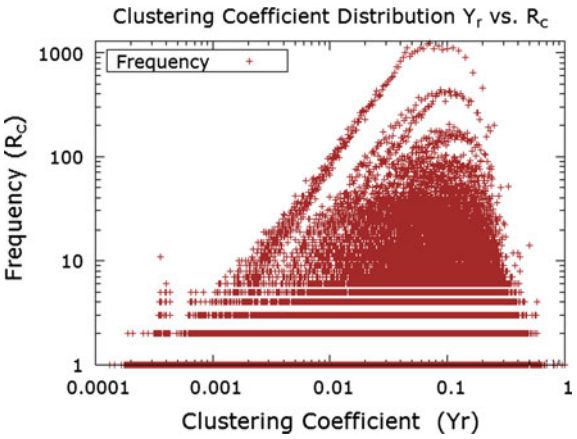

Fig. 2 Clustering coefficient distribution of amazon co-purchase network in log-log scale

node we have found out how many possible triplets exists in which it takes part. On the other hand, triangles depict closure of such triplets and is essentially the smallest structure displaying sense of community. Figure 3 shows triplet and triangle distribution in Amazon co-purchase network. It is a plot between occurrence of triplets and triangles in which a node is present and its frequency in log-log scale. Algorithm [1] states the way of discovering triangles from triplets.

3.4 Detection of Burst Mode

Burst mode is usually observed in human behavior and therefore is a characteristic of social networks. It shows a human predilection of being increasingly engrossed in a particular activity before loosing interest and settling down. A few plots below

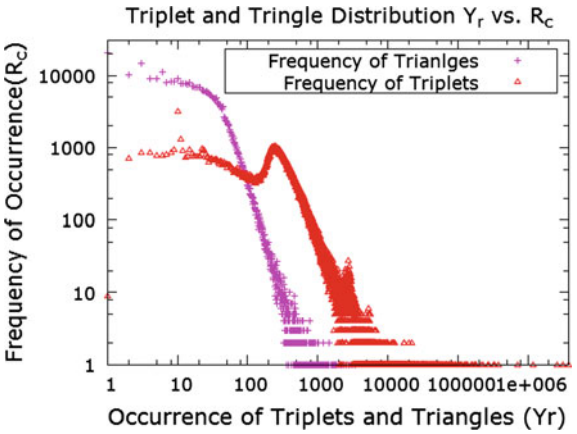

Fig. 3 Triangle and triplet distribution of amazon co-purchase network in log-log scale

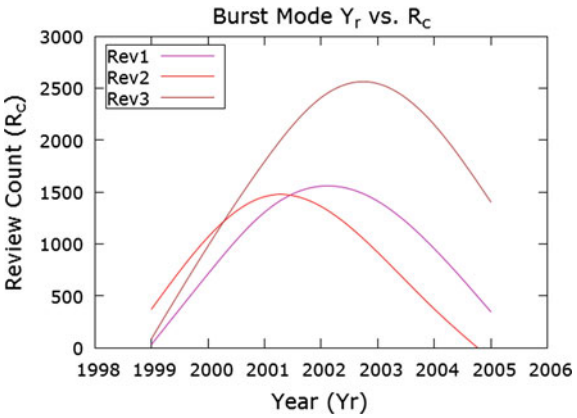

Fig. 4 A plot of burst mode characteristics as shown by three random reviewers

show how users started to review items with increasing vigor and then with a sharp descent coming down to a state of stability. Figure 4 shows the burst mode characteristics shown by three reviewers.

3.5 Annual Variation in Number of Reviews

During a year there are peak times when the sales are maximum. In the Amazon co-purchase network we have plotted the variation in number of reviews for 7 years and some common features came up. The graph begins with a peak because of the continued sales during the festive season. The graph ends with a peak which around the month of December. The number of reviews are predictably high due to the festive mood of Christmas and New Year. Another commonly occurring peak is near April which is time for Good Friday and Easter. Figure 5 gives us a glimpse of how the variation in number of reviews takes place over the years.

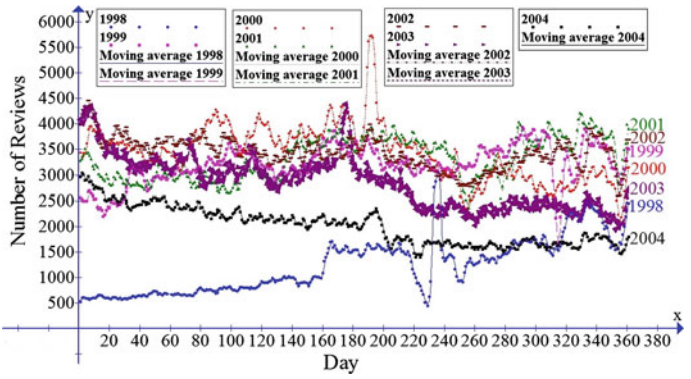

Fig. 5 Annual variation in number of reviews

Algorithm 1 Algorithm to Find all Possible Triangles from Triplets in a Graph

```
K contains the list of adjacency list of nodes
R contains the list of all possible triangles.Initialized to NULL
for all t in K do
  t is a node whose adjacency list is contained in K
  for all x in adjacency list of t do
    P is initialized to a null set
    P= Set of Nodes such that P_i belongs to Adjacency list of x − t
    N= P ∩ Adjacency List of t
    for all triplet in N do
      if Transform(triplet) not in R then
        R=R+ Transform(triplet)
      end if
    end for
  end for
end for
Transform(triplet) {
return (x1,x2,x3) such that x1 < x2 < x3 }
```

4 Recommendation System

One of the ways to boost sales is providing ease to customers for finding those products which match their tastes and choices. A certain percentage of this comfort level is achieved by a good recommender system. We design a recommender system which partly uses content based filtering and collaborative filtering.

4.1 Assigning Weights to Edges

The Amazon co-purchase network graph is essentially unweighted. An attempt has been made to make it a weighted network such that meaningless data can be eliminated before going into further stages of recommendation. This also forms the first stage of recommendation as the edges with significant weights can now be recommended with each other.

$$\text{Weight (a,b)} = |A \cap B|$$

A Set of Reviewer ID's who reviewed Item *a*
B Set of Reviewer ID's who reviewed Item *b*

Figure 6 gives a distribution of hashed edge and its corresponding weight in log log scale. Here, the edges are hashed by assigning consecutive numbers after sorting them in decreasing order of their corresponding weights i.e. edge (156,632–156,634) having the highest weight (842) is hashed as 1, followed by edge

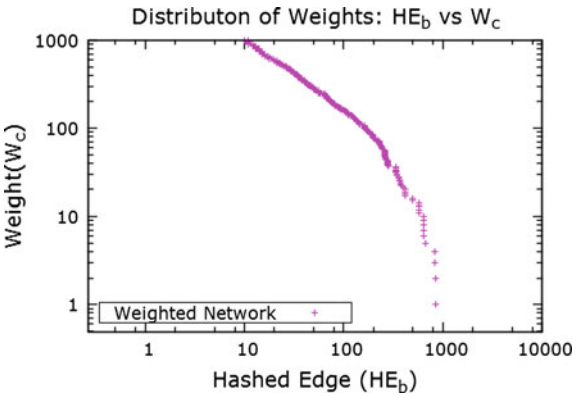

Fig. 6 A plot of hashed weights

(197,325–197,327) with weight (826) and so on. From the graph we have located the region where the graph attains steepest slope and thus the top 2000 edges are taken for recommendation. Using this method, we find that pairs like 'A Christmas Carol' and 'Jingle All the Way', 'A Bend in the Road' and 'The Smoke Jumper' and many other other items in pairs are recommended. The content wise similarity of items in each pair is worth noticing.

4.2 Detecting Cliques in Five Nearest Neighbour Network (5NN)

The Amazon Meta data file contains information not only about the products but also gives a set of 5 similar products for each product. This similarity is based on sub category. We form a 5-NN network from this data. The largest connected component (LCC) is found out and Bron-Kerbosch algorithm [12] is implemented on this LCC to find maximal cliques. It comes up with 3–7 membered cliques. The 5-NN network is made out of similar items and thus the cliques formed share the similarity essence among themselves and thus can be recommended. The number of 3, 4, 5, 6 and 7 membered cliques are 79833, 32211, 7878, 872 and 4 respectively. Thus, we can see with increase in clique size, its frequency decreases.

Products like "Ballroom Dancing", "Much Ado About Ballroom Dancing", "The Complete Idiot's Guide to Ballroom Dancing" are recommended together. Thus, we see that similar products are suggested for co-purchase.

4.3 Collaborative Filtering

In collaborative filtering, predictions about a person's tastes are made from his network of collaboration. The underlying assumption of this approach is that, if a

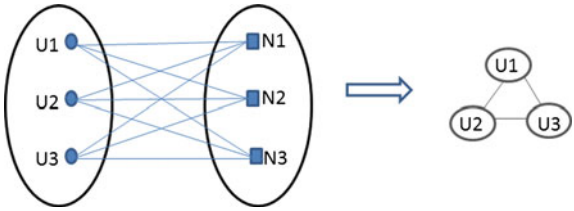

Fig. 7 Edge reduction

person A has the same opinion as that of person B on an issue, then A is more likely to follow B's opinion on a different issue than to follow the opinion of a person chosen randomly. On the contrary, it is very unlikely that two reviewers will have equal taste in every aspect. So, for an item, instead of recommending A's collection for B, it is better to find a neighbour of B with whom it has maximum similarity regarding the sub category of that item. Figure 7 shows a sample edge reduction.

Lemma 1 *For reducing number of edges of a complete bipartite graph* (n_1, n_2, E) *by converting to a one mode network the condition that must be satisfied is* $n_1 - 1 < 2\ n_2$ *or,* $n_2 - 1 < 2\ n_1$

Proof The total number of edges in the bipartite graph (n_1, n_2, E) is $n_1\ n_2$. For reduction the number of edges in the new network should be less than that of the bipartite network. So,

${}^{n_1}C_k < n_1\ n_2$ or, ${}^{n_2}C_k < n_1\ n_2$

On solving the inequalities, the following result is obtained:

$n_1 - 1 < 2n_2$ or, $n_2 - 1 < 2n_1$ □

The bipartite graph consists of two sets one of which is the set of items and the other, the set of reviewers. Here, we have tried to project this two mode network into a one mode network by graphically linking reviewer to reviewer if there exists an item that has been reviewed by both. While transforming the edges are not given direct weights rather a set of data, generic form of which is:

$W(a, b) = \{x, y \mid x$ belongs to common sub categories between a, b and y is the frequency of common occurrence of $x\}$.

Here, Table 2 shows a sample edge. It is clear that these reviewers have a greater match of taste when it comes to the genres *Style*, *Directors* and *Series* in Music, Video and Book respectively.

Table 2 Description of an edge between two reviewers (a, b)

Common subcategories reviewed (x)		Frequency of occurrence (y)
Category	Sub category	
Music	Styles	4153
Video	Directors	1555
Music	Rap	2
Book	Guides and reviews	1

Table 3 Validating the recommendation system

Node number	Original	Recommended	Intersection	Recall	Precision
372,787	6	5	5	0.83	1
255,803	9	7	6	0.67	0.86
392,440	5	6	4	0.80	0.67

4.4 Efficiency of Recommendation System

A recommendation system is useless unless and until a strong evidence of it working fine is provided. Though the transaction table between reviewers and items were not available, different path was chosen to resolve this issue. For each and every time stamp largest connected component (LCC) was found out and a similarity study was made to see how much of our recommendations were present in these LCC.

A Machine Learning Approach Towards Validating: Machine Learning Approach brings a flavor of past experiences while taking current decisions. We have tried to implement the same essence. We have taken two time stamps namely t_1 and t_2 for learning a model of co-purchase predilections and have tested it on the time stamp t_3. The data set provided the edge list of all co-purchase relations between items for time stamp t_3.

Precision and Recall: For a recommender system, precision and recall are the two parameters which give a measure of the accuracy of the system. Precision (i.e. positive predictive value) is the fraction of retrieved instances that are relevant, while recall (also known as sensitivity) is the fraction of relevant instances that are retrieved. Both precision and recall are therefore based on an understanding and measure of relevance shown in Table 3. Here, original refers to the number of items that has been co-purchased with the particular node. Recommended refers to the number of items that our recommendation system has provided for that node. Intersection refers to the number of items common between Original and Recommended.

"The Music of Jerome Kern" (372,787) has high precision and recall value. "Southern Harmony and Musical Companion" (392,440) has high recall value.

5 Conclusion

The study of co-purchase network reveals how human tendencies can shape up co-purchase patterns which bears with it temporal effects. These motifs, if studied carefully can be used to develop strategies to increase sales. We made an attempt to make recommendation systems based on nearest neighbor model and graph topology based collaborative filtering. Since testing recommendations for all the

items in the network is computationally expensive, we picked up random samples to reveal how the recommendation system performs for them. We can clearly see that the recommendation can be good for a large part of the dataset but can not guarantee a highly precise recommendation for every item.

References

1. Leskovec, J., Adamic, L.A., Huberman, B.A.: â€œThe dynamics of viral marketing. ACM Trans. Web **1**(1), (2007)
2. Oestreicher-Singer, G., Sundararajan, A.: Linking network structure to e-commerce demand: theory and evidence from amazon.coms copurchase network. In: TPRC 2006. Available in SSRN (2006)
3. Newman, M.: Modularity and community structure in networks. Proc. Nat. Acad. Sci. **103** (23), 85778582 (2006)
4. Clauset, A., Newman, M.E.J., Moore, C.: Finding community structure in very large networks. Phys. Rev. E **70**, 066111 (2004)
5. Luo, F., Wang, J.Z., Promislow, E.: Exploring local community structures in large networks. Web Intell. Agent Syst. **6**(4), 387400 (2008)
6. Srivastava, A.: Motif analysis in the amazon product co-purchasing network. In: CoRR, vol. abs/1012.4050 (2010)
7. Bogdanov, P., Mongiov, M., Singh, A.K.: Mining heavy subgraphs in time-evolving networks. In: Cook, D.J., Pei, J., 0010, W.W., Zaane, O.R., Wu, X. (eds.) ICDM, pp. 8190. IEEE (2011)
8. Lahiri, M., Berger-Wolf, T.Y.: Structure prediction in temporal networks using frequent subgraphs. In: CIDM, p. 3542. IEEE (2007)
9. Wackersreuther, B., Wackersreuther, P., Oswald, A., Bohm, C., Borgwardt, K.M.: Frequent subgraph discovery in dynamic networks. In: Proceedings of the Eighth Workshop on Mining and Learning with Graphs, ser. MLG 10, pp. 155–162. ACM, New York (2010)
10. Lahiri, M., Berger-Wolf, T.Y.: Periodic subgraph mining in dynamic networks. Knowl. Inf. Syst. **24**(3), 467497 (2010)
11. Han, J., Pei, J., Yin, Y., Mao, R.: Mining frequent patterns without candidate generation: a frequent-pattern tree approach. Data Min. Knowl. Disc. **8**(1), 5387 (2004)
12. Bron, C., Kerbosch, J.: Algorithm 457: finding all cliques of an undirected graph. Commun. ACM **16**(9), 575–577 (1973)
13. Basuchowdhuri, P., Shekhawat, M.K., Saha, S.K.: Analysis of Product Purchase Patterns in a Co-purchase Network. EAIT (2014)