2014 Fourth International Conference of Emerging Applications of Information Technology

Analysis of Product Purchase Patterns in a Co-purchase Network

Partha Basuchowdhuri, Manoj Kumar Shekhawat
Department of Computer Science and Technology
Heritage Institute of Technology
Kolkata, WB, INDIA 700107
Email: parthabasu.chowdhuri@heritageit.edu,
shekhawatmanojkumar@gmail.com

Sanjoy Kumar Saha
Department of Computer Science and Engineering,
Jadavpur University,
Kolkata, WB, INDIA 700032
E-mail: sks_ju@yahoo.co.in

***Abstract*—Real world social networks are often temporal in nature. They evolve with time as new nodes may appear, old nodes may cease to exist and the relationships between the entities may also change. In this paper, we have studied a publicly available dynamic network dataset, namely Amazon co-purchase network dataset. We have analyzed the network data to understand the significance of nodes with high in-degree and high out-degree. We subsequently analyze the evolution of communities in the network by observing the change of association among nodes and by observing communities to see how many members they could retain over time. We show some frequent itemsets from the co-purchase market basket in terms of the item categories and subcategories. We finally select some central entities and group of such entities from the network to recommend how promoting some of the co-purchased items may increase the sales of the selected items.**

***Keywords*—Dynamic networks, social networks, Amazon co-purchase network.**

I. Introduction

Online marketplaces have become popular due to the buyers' convenience to shop from any place. There are also other added advantages, like, shipping the product to home, buying a product without having to handle cash, buying products that are not available in the local marketplace, and many more.

Online stores like Amazon first came into business in 1995 and is now deemed as a pioneer for online business models. In India, online purchase was popularized much later, compared to USA. Presently in India, many online stores compete for supremacy with the leaders, namely, Flipkart and Amazon. These online marketplaces store the users' profile data, their buying history and their browsing history to analyze their marketplace behavior to better understand their buying pattern. This could enable the sellers, i.e., the online marketplaces to understand the potential for promoting different types of products.

Amazon uses a recommendation system to suggest other frequently co-purchased items during sale of an item. If the co-purchase is beneficial to the customer then the recommendation based marketing makes it highly likely that the extra item is also bought from the same marketplace. Providing discounted price on recommended co-purchasable items, only when co-purchased, may also provide further incentive to the customer.

These online stores acquire data from their own website and the data is not limited to co-purchase data. This paper deals with analysis of a large co-purchase network data from an online marketplace. We analyze the network data from graph theoretic perspective. Interestingly, the co-purchase network data is temporal in nature and data is available for four time-stamps. We thereby also observe the evolution of the communities and transformation of nodes' association to the communities. Based on this analysis, we recommend which products to market and how to market so that it could increase the sales.

II. Related Works

The Amazon co-purchase network was used by Leskovec et al. [1] for analyzing person-to-person recommendation in viral marketing. Overall the recommendations were found not to be very effective in inducing purchases. But they showed that the viral marketing works better when the data is initially categorized based on certain features. In another paper, Amazon co-purchase network was used to explain e-commerce demand. They claim that item categories with flatter demand distribution is influenced more by the structure of the network [2]. Clauset et al. proposed a community detection method in 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. Modularity [3] is a popularly known function for measuring overall goodness of detected communities. 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. They detect communities with a high maximum modularity value, i.e., 0.745, but the size of the communities seem very large to find any evident patterns from them. For example, the largest sized community they get consists of more than 100 thousand nodes, which is more than 25% of the total number of nodes in the networks. Luo et al. studied the local communities in Amazon co-purchase network and claimed that recommendation works better for digital media items than books [5]. Motif analysis, for 3-node and 4-node motifs, has also been performed on the Amazon co-purchase network [6]. Frequent motifs have been identified, which single-handedly can not contribute well to understanding the behavioral pattern of the co-purchase network. Recent works on finding frequent subgraphs and mining subgraphs in dynamic networks [7], [8],

978-1-4799-4272-5/14 $31.00 © 2014 IEEE
DOI 10.1109/EAIT.2014.11

[9], [10] have led to new perspective in analyzing temporal business networks.

III. Problem Definition

Given a directed graph $G(V,\ E)$, where nodes are items or products in an online marketplace and directed edges denote the co-purchase relation, we analyze the co-purchase patterns and build a recommendation system that would help maximizing the revenue. It should be noted that, co-purchase network maintains a sense of directionality among products. If a directed edge starts from a source node and reaches a sink node, then the product that denotes the source node, is independently bought. As the product, i.e., the item at the source node is bought, the item at the sink node is co-purchased. The frequency of the co-purchase is not available and therefore, the graph is considered to be unweighted.

IV. Exploring Features of the Dataset

In this paper, we have analyzed the Amazon co-purchase network data. Unlike other social network datasets, the Amazon co-purchase network dataset has a temporal aspect to it, i.e., the dataset consists of a network and its snapshots after equal time-intervals. The snapshots display how the original network's structure changes and thereby enabling us to find the changing co-purchasing patterns. For better understanding, we represent the Amazon co-purchase dynamic network as a set of time-stamp graphs $\{G_0,\ G_1\ ,G_2\}$, where G_0 is the original network at $t = 0$ and the evolved versions of G_0, after one and two units of time are G_1 and G_2, respectively. Fig. 1 shows how a network may experience temporal evolution and how the community structures within the network keep changing.

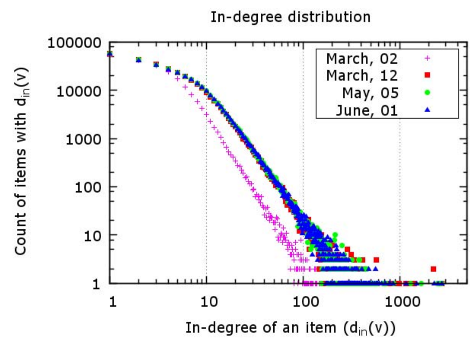
As shown in table I, the repository storing the Amazon co-purchase network has another snapshot which is much smaller than the other three. Although it has similar average clustering co-efficient as the other snapshots, we have not considered it due to disparity in size. Another argument for ignoring the network was the time difference between the ignored network and G_0, which is considerably less to allow any change.

$Graph$	$\|V\|$	$\|E\|$	$Avg.\ Clustering$ $Co-efficient$	$Month,$ $Year$
G_{-1}	262,111	1,234,877	0.4198	March 02, 2003
G_0	400,727	3,200,440	0.4022	March 12, 2003
G_1	410,236	3,356,824	0.4064	May 05, 2003
G_2	403,394	3,387,388	0.4177	June 01, 2003

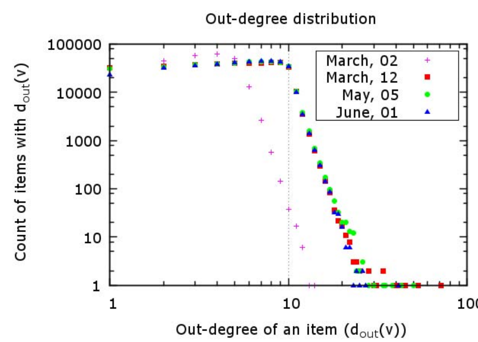
TABLE I. Snapshot graphs in Amazon co-purchasing network

A. Degree distribution of the items

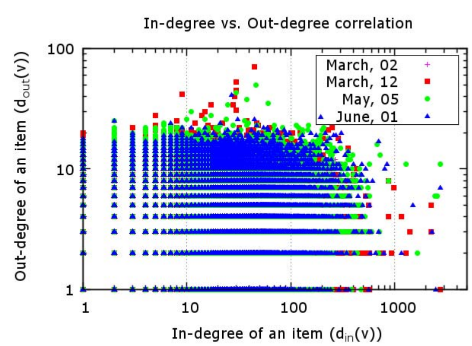
Degree of a node, in this co-purchase network, depicts the extent of an item's co-purchasability. Co-purchasing relation being a directed relation, the significance of in-degree and out-degree of a node would be different. A node with higher out-degree means when that item is bought many other items are co-purchased, which might be accessories for the original item. On the other hand, a node with higher in-degree means that the item is co-purchased with many other items. It may also indicate that such an item is a commonly used item and the co-purchase relationship with the other item does not provide a purchasing correlation but an incidental purchase of the item when the other item was also purchased. Similarly, one item being bought before the other may not necessarily mean that buying the first item leads to buying the second and the buying sequence may be incidental only.



(a) In-degree distributions for the four available snapshot networks.



(b) Out-degree distributions for the four available snapshot networks.



(c) In-degree vs. Out-degree correlation plot

Fig. 2. In-degree, out-degree distributions and their correlation for four snapshots of Amazon co-purchase network.

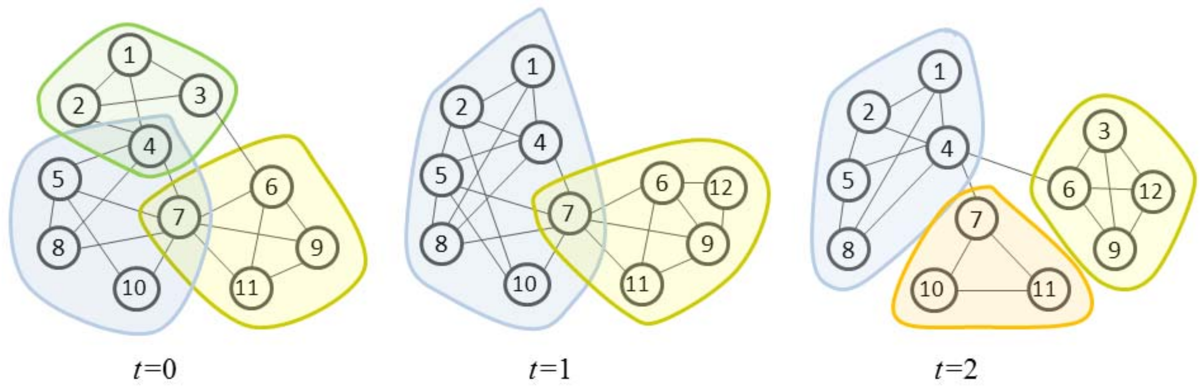


Fig. 1. An example of a network and its evolving communities shown at different snapshots.

In the Amazon co-purchase networks, we found the in-degree range to be larger than that of the out-degree. The in-degree distribution follows a power law with a heavy tail, as can be seen in figure 2(a). From figure 2(b), we can say that out-degree distribution also follows a similar distribution, which is typical of small-world scale-free networks. Note that, in-degree distribution of nodes in G_{-1} is similar to that of the others, but it does not overlap with the other plots because the nodes in G_{-1} are much lesser in number than the other snapshot networks. In figure 2(c), the nodes with both high in-degree and out-degree are mostly found in G_2 and G_3.

V. Analysis of the Purchase Patterns

A. Items as part of communities

Accuracy of a network based analysis may be affected by the network's size. Larger the network size, it becomes more difficult to analyze it. Furthermore, larger the network, more difficult it is to find the areas of interest and focus on those parts. In case of Amazon co-purchase network, insignificant transactions may be deemed as noise and they contribute to inaccuracy in the analysis of the network. Amazon co-purchase network snapshots are very large in size and therefore it is necessary the process the data first before we start looking into the data for co-purchase patterns. We have performed a few steps to coarsen the data and separate the part we want to study, from the rest of the network.

Node	$t = 0$	$t = 1$	$t = 2$
0	C_{327}	C_1	C_1
21245	C_{542}	C_{147}	C_{82}
383678	-1	C_{27}	C_{82}
410235	-1	C_{1784}	-1

TABLE II. Sample node movement w.r.t. time

One important part of analyzing dynamic networks is to track the evolution of communities. For this purpose, we apply widely used Clauset-Newman-Moore (CNM) [4] method for finding disjoint community structures. We use the publicly available implementation for CNM to initially find out the communities for all the snapshot networks. Running CNM once on each of these datasets takes around 6 hours. For every dataset we run it twice to get the cover, i.e., the set of communities, with maximum modularity. Say, for i-th time-stamp the cover with maximum modularity value is given by $\{C_{i1},C_{i2},C_{i3},\ldots\}$, and is found for every network G_i where i = -1, 0, 1, 2. In table II, we show how a node changes its association for different i values. Here, -1 signifies that the node was not a part of the network at the i-th snapshot. It is important to observe that although the community label of a node changes in different time-stamps, it may still be part of the same community where the community has been labeled with different community labels in different time-stamps. Like, in table II, we see that node 0 has been a member of community 327 in G_0, community 1 in G_1 and community 1 in G_2.

To understand whether C_{327} at t=0, C_1 at t=1 and C_1 at t=2 are same or not, we aim to find the correlation between all the communities of one time-stamp and the communities from the subsequent time-stamp. Correlation between community structures is measured by calculating similarity between community structures C_{ik} and C_{jl}. The similarity between C_{ik} and C_{jl} is measured in the following way -

$$S(C_{ik}, C_{jl}) = \frac{|C_{ik} \cap C_{jl}|}{|C_{ik} \cup C_{jl}|} \quad \forall\, C_{ik} \in C_i\ ,\ C_{jl} \in C_j \qquad (1)$$

, where i, j are time-stamps and C_{ik} is the k-th community from the cover C_i found by CNM method from G_i and similarly, C_{jl} is the l-th community from the cover C_j found by CNM method from G_j. This is essentially equivalent to finding Jaccard co-efficient between C_{ik} and C_{jl}. This similarity measure shows us how similar two communities are from two different time-stamps. For sake of simplicity we can assume j to be i+1. Say, C_{ik} has split into multiple communities at i+1 time-stamp, we say that C_{ik} has evolved into a community where largest percentage of nodes from C_{ik} are preserved. In table III, we show the how some of the communities at t=1 turn out to be and list of communities from t=0, which contributed towards forming the new communities. For example, the results show that C_{327} from G_0 contributes to quite a few number of new communities and hence we can conclude that C_{327} has been broken into many parts and formed many communities for which the majority of the node were contributed by C_{327}, like C_1, C_2 and C_3 obtained from

G_1.

$t = 1$	Size	Contri-butors	New(%)	Contributors from $t = 0$ **(CF, Jaccard (%), Contribution%)**
C_1	90735	1301	18.5221	[(C_{327}, 48.1264, 23.8927), (C_{542}, 24.3729, 14.0828), (C_{921}, 22.4736, 10.6199), (C_{930}, 25.3803, 9.2125), (C_{265}, 19.9612, 1.1352), (C_3, 36.897, 1.0536), (C_{66}, 36.3112, 0.8332), (C_{424}, 24.2903, 0.8299), . . .]
C_2	23	8	0	[(C_{327}, 0.0244, 47.8261), (C_6, 13.7931, 17.3913), (C_{14}, 28.5714, 8.6957), (C_{542}, 0.0038, 8.6957), (C_{67}, 0.1812, 4.3478), (C_{82}, 1.2346, 4.3478), (C_3, 0.0386, 4.3478), (C_{16}, 14.2857, 4.3478)]
C_3	193	30	26.4249	[(C_{327}, 0.0688, 16.0622), (C_{542}, 0.0572, 15.544), (C_{921}, 0.07, 15.544), (C_{930}, 0.0364, 6.2176), (C_{141}, 0.3381, 2.0725), (C_{265}, 0.0775, 2.0725), (C_{128}, 0.1548, 2.0725), (C_8, 0.2629, 1.5544), . . .]

TABLE III. CONTRIBUTION OF COMMUNITIES FROM THE FINAL COVER AT t=0 IN COMMUNITIES FROM THE FINAL COVER AT t=1

So, it is essential to check how many nodes are common in C_{ik} and C_{jl} out of total number of nodes in those communities. Instead of calculating the number of common elements, we find out the exact contribution, i.e., the number of communities C_{ik} of C_i in terms of percentage of contribution in the creation of each community C_{jl} of C_j. The percentage of newly added nodes, which were not present at the previous time-stamp, is also important and is shown in table III.

B. Category based analysis of items

$t = 0$	**size**	**Book**	**Music**	**DVD**	**Video**
C_3	2591	1861	496	106	128
C_8	1141	819	220	34	68
C_{327}	45043	32525	8637	1696	2185
C_{335}	3399	2485	627	132	155
C_{542}	52424	38270	9879	1868	2407

TABLE IV. DISTRIBUTION OF PRODUCT CATEGORIES IN SOME OF THE LARGE COMMUNITIES FROM THE FINAL COVER AT t=0

From our previous discussions, we can see that how communities from one time-stamp network evolve and may lead to changed communities in successive time-stamps. If we try to track evolution of all the communities, there will be many possible combinations involving the items. Investigating so many combinations would be a difficult task. Hence, we select a threshold based on the percentage of contribution so that only a subset of the combinations form a smaller candidate set for tracking the evolution of communities. We put threshold based on the percentage of contribution from a community belonging to the previous snapshot network.

1) Lineage of a community at present time-stamp: In this part we observe which community from previous time-stamp contributes most to the present communities. For a large sized community less than two percent contribution means negligible amount of contribution, we ignore any community to community node contributions, which is less than 2%. After such coarsening of data, we display partial results in table III. This is a sample observation where the communities at t=1 are shown in terms of the nodes from the communities at t_2. For communities C_1, C_2 and C_3 from t=1 is observed in terms of communities at t=2. Say, for C_2 at t=1, there are a total of 23 nodes in it, which come from 8 different communities from the previous snapshot network. Also, it has no new node joining newly at t=1. The percentage of newly added nodes for the communities is important, because it reveals whether it is continuation of an existing group with regular inclusion of alien items or it is a new group of items, where the number of existing items is very less. In case of t_1 to t_2 transition for the network communities, addition of new nodes is considerably less and hence not much can be concluded by analyzing the new nodes in this case. In the next part, the top eight communities leading to the formation of C_2 at t=1 is shown. For an example, C_{327} at t=0 contributes to 47.8% of the nodes in C_2 at t=1 and also, the Jaccard co-efficient is measured to be 0.244. During the calculation of Jaccard co-efficient, only member nodes that are present in both the time-stamps are considered.

2) Sub-category based analysis: Each node of Amazon co-purchase network consists of an attribute value for attributes category and sub-category. So, we have calculated product category distribution for each time-stamp data as shown in table IV. Every time books turn out to be in the majority and the proportions between the number of representative items in those categories do not differ by a large margin. Also, if we try to observe the itemsets co-purchased together based on item IDs, the frequency of the patterns will be too less in number to understand if the items sold together have a pattern. Hence, we turn to the sub-category attribute and represent the network in terms of the sub-category value of the items. In this way, each snapshot network now turns into co-purchase network within sub-categories of items. From this network we extract frequencies of the co-bought items and based on that we may use association rule based techniques to find frequent itemsets. Here, we use FP-growth algorithm [11] to find out frequent itemsets based on sub-category value of items. For example, in table V, frequent itemsets based on sub-categories of items for the communities C_{327} at t=0, C_1 at t=1 and C_{82} at t=2 has been shown. A few frequently co-purchased itemsets of size three were displayed in table V. All the items are from item category - books, but with different sub-categories, such as, non-fiction, children's book, religion & spirituality, literature & fiction and business & investing.

Frequent Itemsets	Occurrences		
	T_1	T_2	T_3
['B_Non', 'B_CB', 'B_RS']	26	85	197
['B_Non', 'B_CB', 'B_LF']	46	95	196
['B_Non', 'B_CB', 'B_BI']	33	80	215

TABLE V. TOP THREE FREQUENT ITEMSETS FOR COMMUNITY PATTERN [327 1 82], WHEN NODES ARE REPRESENTED BY SUB-CATEGORIES.

C. Finding stable community cores

In this part we try to find groups of nodes which show tendency to remain together. We found that some of the communities from time-stamp t=0 got split into parts and contributed to formation of new communities at time-stamp t=1. Some of the communities at t=0 got merged together at time-stamp t=1. Also, we find some groups of nodes, which are part of a single community at t=0 but at t=1 they move from the community where its majority of the previous community's members are located and join some other group. In the next time-stamp, i.e., at t=2, they again merge back to a single community. In order to generalize the dynamic behavior of nodes of a community we categorize them into two types - group of nodes that are always together in a community from the first snapshot network to the last, and those which start in one community and also end in the same community but move to other communities back and forth during the intermediate time-stamps.

1) Group of inseparable nodes: These are the nodes which always remain together in all time-stamps. In table VI, we select some community patterns, i.e., a set of communities from three time-stamps and see the nodes which are always present in that set of communities. For an example, the first row entry from table VI says five nodes (the set is also provided) are present in C_1 at t=0, which subsequently remain in C_1 even at t=1 and t=2.

Pattern	# Nodes	Node List
1 1 1	5	[260, 261, 262, 263, 264]
2 1 1	8	[417, 744, 745, 746, 747, 748, 1053,1757]
2 1 82	5	[910, 912, 1003, 1004, 1005]
3 1 1	264	[215, 326, 373, 982, 983, 1046, 1047,1048, 1049, 1142, 1547, 1548, 1549, 1550, 1615, 1616, 1617, 1856, 1928, 1929, 2133, 2197, 2228, 2229, 2230, 2231,2232, 2482, 2483, 2496, 2497, 2498, 2499,2762, 2763, 2764, 3335, 3336, 3337, 3347, 3348,3353, 3390, 3652, 3653, 3654, 3655, 3887, 3888, 3933, 3938, 4450, 6234, 6529, . . .]
3 1 2	134	[1007, 2189, 2427, 3205, 3206, 3313, 3314, 4115, 4117, 4118, 6651, 6652, 6653, 9660, 10386, 10387, 10760, 10931, 10932, 10933, 11489, 12329, 13685, 14239, 14241, 14263, 14264, 14282, 17228, 17684, 17685, 17686, 17687, 18267, 18455, 18677, 20049, 21236, 21237, 22153, 22381, 22382, 23222, 24326, 25221, 25854, . . .]

TABLE VI. GROUP OF NODES BELONGING TO A PATTERN OF COMMUNITIES.

2) Group of nodes starting and ending together: In table VII, we can see how communities from the Amazon co-purchase network change over time. Some sample communities have been chosen to display the behavior of split and getting merged back. Here, based on similarity measure some communities from the first and last time-stamps are picked. If the source community, i.e., the community at the initial snapshot is very similar to the destination community, i.e., the community at the initial snapshot, then we can say that the community has remained intact. But the intermediate evolution of the dynamics of the community remains unknown to us. That is what necessitates the study for finding the communities which were formed at the intermediate stages. The causes are not so easily understood and need further investigation. The first entry of table VII says that C_3 at t=0 eventually leads to C_1 at t=2, but in the intermediate time-stamp, its nodes get divided into five communities. The list of the communities has also been provided.

Community at t=0	Community at t=2	Number of divisions	Split communities at t=1
3	1	5	['1', '27', '129', '147', '176']
3	2	4	['1', '27', '147', '176']
3	82	6	['1', '4', '27', '129', '147', '176']
3	205	4	['1', '27', '147', '176']
8	1	4	['1', '27', '147', '176']
8	2	4	['1', '27', '147', '176']
8	82	5	['1', '4', '27', '147', '176']
8	205	3	['1', '147', '176']

TABLE VII. SHOWS COMMUNITY DIVISIONS IN A_2 NETWORK FOR SOURCE-DESTINATION PAIR (A_1,A_3)

Book Type	Book Title
Nonfiction	Society: The Basics
Nonfiction	American Government: The Essentials
Business & Investing	Achieving Excellence Through Customer Service
Business & Investing	Smart Business Solutions: Direct Marketing and Customer Management
Literature & Fiction	Fathers and Sons
Literature & Fiction	One Flew Over the Cuckoo's Nest
Children's Books	Perrault's Complete Fairy Tales (Puffin Classics)
Children's Books	Tropical Fish Coloring Book
Religion & Spirituality	Awake, My Heart: Daily Devotional Studies for the Year
Religion & Spirituality	Victory in Singleness: A Strategy for Emotional Peace

TABLE VIII. FREQUENT PRODUCTS

VI. CONCLUSION

The study of co-purchase network reveals the purchasing trend of people and dependency of one item being bought with another. It can also be used to form strategies for increasing sales. We have used market basket analysis for that purpose. Generating frequent itemsets have been used to understand the buying interest of the customers and correlation between the product categories of interest. From table VIII, we get a clear idea of Amazon customers' purchasing trend and this can be utilized for prediction of future purchase and making strategies accordingly to thereby increase sales of items that are correlated to those frequently bought itemsets. If the frequent patterns from one time-stamp are looked in all successive time-stamp networks then it can give us the behavior of the nodes, i.e., if a set of patterns are found to be frequent in k-1 time-stamps networks then we can say that there is high probability of repetition of the frequent itemsets in k-th time-stamp. So, this type of information can be used for the purpose of building a prediction model. Finding association among the nodes has been significant for our study. We find the set of nodes which are moving together over time and generated subgraphs for those nodes for each time-stamp network. Frequent pattern sets were extracted from each subgraph. As shown in table V, we have found some patterns, based on the sub-category of the items, which are frequent in all the three time-stamp networks. On the basis of such results, we can conclude that at time-stamp t=3 these patterns with repeat with high probability. A Markov chain model might be useful for implementing a generalized recommendation system for such prediction. In this paper, we discuss about the areas, which may act as features in such models and experimentally show that the set of features we have mentioned plays an important role in understanding the dynamics of a co-purchase network.

REFERENCES

[1] J. Leskovec, L. A. Adamic, and B. A. Huberman, "The dynamics of viral marketing," *ACM Trans. Web*, vol. 1, no. 1, 2007.

[2] G. Oestreicher-Singer and A. Sundararajan, "Linking network structure to e-commerce demand: Theory and evidence from amazon.com's co-purchase network," *TPRC 2006. Available in SSRN*, 2006.

[3] M. Newman, "Modularity and community structure in networks," *Proceedings of the National Academy of Sciences*, vol. 103, no. 23, pp. 8577–8582, 2006.

[4] A. Clauset, M. E. J. Newman, and C. Moore, "Finding community structure in very large networks," *Physical Review E*, vol. 70, p. 066111, 2004.

[5] F. Luo, J. Z. Wang, and E. Promislow, "Exploring local community structures in large networks," *Web Intelli. and Agent Sys.*, vol. 6, no. 4, pp. 387–400, 2008.

[6] A. Srivastava, "Motif analysis in the amazon product co-purchasing network," *CoRR*, vol. abs/1012.4050, 2010.

[7] P. Bogdanov, M. Mongiov, and A. K. Singh, "Mining heavy subgraphs in time-evolving networks." in *ICDM*, D. J. Cook, J. Pei, W. W. 0010, O. R. Zaane, and X. Wu, Eds. IEEE, 2011, pp. 81–90.

[8] M. Lahiri and T. Y. Berger-Wolf, "Structure prediction in temporal networks using frequent subgraphs." in *CIDM*. IEEE, 2007, pp. 35–42.

[9] B. Wackersreuther, P. Wackersreuther, A. Oswald, C. Böhm, and K. M. Borgwardt, "Frequent subgraph discovery in dynamic networks," in *Proceedings of the Eighth Workshop on Mining and Learning with Graphs*, ser. MLG '10. New York, NY, USA: ACM, 2010, pp. 155–162.

[10] M. Lahiri and T. Y. Berger-Wolf, "Periodic subgraph mining in dynamic networks." *Knowl. Inf. Syst.*, vol. 24, no. 3, pp. 467–497, 2010.

[11] J. Han, J. Pei, Y. Yin, and R. Mao, "Mining frequent patterns without candidate generation: A frequent-pattern tree approach," *Data Mining and Knowledge Discovery*, vol. 8, no. 1, pp. 53–87, 2004.