The Topological Characteristics and Community Structure in Consumer-Service Bipartite Graph

Lin Li[1], Bao-Yan Gu[2], and Li Chen[3]

[1] Business School, University of Shanghai for Science and Technology, Shanghai 200093, China
Lilin@usst.edu.cn
[2] gubaoyan@hotmail.com
[3] lichen0411@163.com

Abstract. We apply network analysis to study bipartite consumer-service graph that represents service transaction to understand consumer demand. Based on real-world computer log files of a library, we found that consumer graph projected from bipartite graph deviates significantly from theoretical predictions based on random bipartite graph. We observed smaller-than-expected average degree, larger-than-expected average path length and stronger-than-expected tendency to cluster. These findings motivated to explore the community structure of the network. As a result, the weighted consumer network showed significant community structure than the unweighted network. Communities picked out by the algorithm revealed that individuals in the same community were due to their common specialties or the overlapping structure of knowledge between their specialties.

Keywords: bipartite graph, consumer demand, topological features, community structure, weighted network.

1 Introduction

Complex systems have been a new paradigm for study of management, physical and technological domains [1]. A great deal of scholars making advances in different areas offers an opportunity to promote the knowledge exchange needed to reap the benefits of this basic research for problems in management, organization, and business. *Management Science* (2007,53(7)) published ten papers that use complexity theory to study the emergence, coordination, efficiency, and innovation in small groups, firms, and markets with an eye to the needs of practicing managers. One analysis tool of complex systems is network analysis which enables one to quantity the components and interactions of any different systems that have actors and relationships. Although network analysis has a long research history in graph theory and developed key concepts in social science, recent advances have shown cross talk among the different disciplines. The central idea of recent studies is to have agents interact with each other according to prescribed rules that may change over time as the agents adapt to their environment and learn from their experiences [2, 3]. Therefore we may understand

J. Zhou (Ed.): Complex 2009, Part I, LNICST 4, pp. 640–650, 2009.

the possible origins of the system and know the key variables leading cause and effect relationships by network analysis methodology.

The underlying hypothesis of marketing literature on consumer purchase behavior is that consumers naturally form cohesive subgroups with consistently correlated preferences and that consumer preferences are adequately expressed in the sales-transaction data. Data on sales-transaction can be obtained relatively easily and they are very popular in data mining. How to transform the sales transaction into a graph is the key to use network analysis tools studying consumer behavior, and the bipartite graph can do this well, which has two types of vertices and edges running only between vertices of unlike types. Many social networks are bipartite, forming what the sociologists call *affiliation networks*, i.e., networks of individuals joined by common membership of groups [4]. Recent studies have adopted the bipartite graph modeling to study sales transaction data; those findings motivated the development of a new recommendation algorithm based on graph partitioning [5].

This paper applies bipartite graph modeling to study reader borrowing behavior in the library of a university. We are interested in whether the real networks deviate from the theoretical predictions based on the generating function method which is introduced by Newman [6]. If these networks exhibit significantly different topological characteristics than expected values, we attempt to identify the underlying mechanism that governs consumer-service behavior. Here we consider the lending book as a kind of service supplied by the library institution. We hope that our research can bring about useful insights to service institutions which try to enhance their service efficiency and service quality by analyzing the interaction between consumer and service.

2 Related Research Work

2.1 Bipartite Graphs

In organizations and events, people gather because they have similar tasks, interests or share a preference for a particular thing. For instance, directors and commissioners on the boards of a corporation are collectively responsible for its financial success and meet regularly to discuss business matters. In such networks there are usually two sets of vertices, which are called *actors* and *events*, and edges connect vertices from different sets only. This type of network is called a *two-mode network* or a *bipartite graph*, which is structurally different from the one-mode network or unipartite graph, in which each vertex can be related to each other vertex. Examples of such networks include the board of directors of companies, co-ownership networks of companies, collaboration of scientists and movie actor collaboration networks. The last two are sometimes called *collaboration networks*. In the case of movie actors, the two types of vertices are movies and actors, and the net work can be represented as a graph with edges running between each movie and the actors that appear in it [6].

In many cases, graphs that are bipartite are actually studied by projecting them down onto one set of vertices or the other-so called “one mode” projections.

In such a projection two actors are considered connected if they have appeared in an event together. That is to say we study relations among one kind of vertices: relation between actors or between events, but not between actors and events. The construction of the one-mode network however involves discarding some of the information contained in the original bipartite network, and for this reason it is more desirable to model networks using the full bipartite structure[6]. But in description of a bipartite network there are more complications: some structural indices must be computed in a different way for bipartite networks, for example, the concept of degree, distance and centrality of vertices. Techniques for analyzing one-mode networks cannot always be applied to two-mode networks without modification or change of meaning. So what can we do? The solution commonly used is to change the two-mode network into a one-mode network, which can be analyzed with standard techniques [7]. We also follow this projection approach in our study. Some studies have done in extracting the hidden information of bipartite network projection, such as the weighting method proposed by Zhou *et al.*[8], which provides a method for compressing bipartite network and highlights a possible way for personal recommendation. Zhang *et al.*[9] proposed a model named a stretched exponential distribution (SED) to explain the topological characteristics of many empirical collaboration networks.

Newman *et al.*[6] derived the theoretical predictions of the topological measures of the one-mode projection based on a given vertex degree distribution of the full bipartite graph using generation function method. These topological measures include average degree, average path length, and clustering coefficient. The generation function $G_0(x)$ of a unipartite undirected graph is defined

$$G_0(x) = \sum_{k=0}^{\infty} p_k x^k \tag{1}$$

where p_k is the probability that a randomly chosen vertex on the graph has degree k. The theoretical predictions of the statistical properties of the unipartite graphs projected from a bipartite graph can be derived from the two generating functions associated with the degree distributions of the two types of vertices.

In our context, we will speak in the language of "readers" and "books" in the bipartite consumer-service graph. Let be the probability distribution of the degree of readers (the number of books which readers have borrowed) and be the distribution of degree of books (the number of readers by which books have been borrowed). Two generating functions can be constructed thus:

$$f_0(x) = \sum_j p_j x^j,\ g_0(x) = \sum_k q_k x^k \tag{2}$$

Newman *et al.*[6] show that the generation function of the unipartite reader graph projected from the he bipartite consumer-service graph is given by

$$G_0(x) = f_0(g_1(x)) = f_0(\frac{g_0'(x)}{g_0'(1)}) \tag{3}$$

The corresponding theoretical predictions of average degree z_1, average path L , and triangle clustering coefficient C are giver by

$$z_1 = G_0'(1),\ L = 1 + \frac{\log(N/G_0'(1))}{\log((\frac{f_0''(1)}{f_0'(1)})(\frac{g_0''(1)}{g_0'(1)}))}$$
$$C = \frac{M}{N}\frac{g_0'''(1)}{G_0''(1)} \tag{4}$$

where M is the total number of books and N is the total number of readers. All of these results work equally well if "readers" and "books" are interchanged.

2.2 Community Structure

Social networks usually contain dense pockets of people who "stick together." Social interaction is the basis for solidarity, shared norms, identity, and collective behavior, so people who interact intensively are likely to be considered a social group. This phenomenon is called *homophily* [7]. A numbers of recent studies have focused on the statistical properties of networked systems. A few properties seem to be common to many networks: the small-world property, power-law degree distributions, and network transitivity. Another property which is found in many networks is the property of community structure, in which network nodes are joined together in tightly-knit groups between which there are only looser connections [10].

The traditional method for detecting community structure in networks is hierarchical clustering. The networks are represented a nested set of increasing large components (connected subsets of vertices) according to how closely connected the vertices are, which are taken to be the communities. But hierarchical clustering has a tendency to separate single peripheral vertices from the communities to which they should rightly belong [10]. Another deficiency of these methods can't tell us when the communities found by the algorithm are good ones. Algorithms always produce some division of the network into communities, even in completely random networks that have no meaningful community structure [11]. Girvan and Newman [10, 11] proposed an algorithm (GN) based on the iterative removal of edges with high "betweenness" scores that appears to identify community structure with some sensitivity, they also propose a measure called *modularity* for the strength of the community structure, which gives an objective metric for choosing the number of communities into which a network should be divided. As pointed out by Newman and Girvan [11], the principal disadvantage of their algorithm is the high computational demands it makes.

Newman [12] described a new algorithm for extracting community structure from networks, which has a considerable speed advantage over previous algorithms, running to completion in time that scales as the square of the network size. This allows us to study much larger systems than has previously been possible. This algorithm is based on the idea of modularity Q which is defined as follows. Let e_{ij} be the fraction of edges in the network that connect vertices in group i to those in group j, and let $a_i = \sum_j e_{ij}$. Then

$$Q = \sum_i (e_{ii} - a_i^2) = Tre - \parallel e^2 \parallel \tag{5}$$

is the fraction of edges that fall within communities, minus the expected value of the same quantity if edges fall at random without regard for the community structure. If a particular division gives no more within-community edges than would be expected by random chance we will get $Q = 0$. Values other than 0 indicate deviations from randomness, and in practice values greater than about 0.3 appear to indicate significant community structure.

Starting with a state in which each vertex is the sole member of one of n communities, we repeatedly join communities together in pairs, choosing at each step the join that results in the greatest increase (or smallest decrease) in Q. The change in upon joining two communities is given by $\triangle Q = e_{ij} + e_{ji} - 2a_i a_j = 2(e_{ij} - a_i a_j)$, we can select the best cut by looking for the maximal value of Q. The entire algorithm runs in worst-case time $O((m + n)n)$ on a network with m edges and n vertices. It is worth noting that this algorithm can be trivially generalized to weighted networks in which each edge has a numeric strength associated with it, by making the initial values of the matrix elements e_{ij} equal to those strengths rather than just zero or one[12].

3 Empirical Study

3.1 Consumer-Service Network

We constructed consumer-service networks using data sets provided by the library of a university in a seven year period from 2001 to 2006. The raw data for the networks described here is a computer log file containing lists of information, including readers, books they borrow, date, and other information such as readers' department, books' China library classification code, and so forth. Projection of reader-book network is straightforward. In the projected network, two readers are connected if they have borrowed at least one common book. We can also project book network in the similar way, but these results are not in this paper.

For the simplicity of calculation, the network only includes readers who are teachers and graduate students although data for undergraduate students are available in the database. The bipartite reader-book graph has 3205 reader vertices, 44127 book vertices and 135637 edges.

3.2 Topological Characteristics of the Network

The degree of a vertex is the number of links incident with it. The probability of a vertex with degree k (or the degree distribution) p_k is the most important topological property of the network. In bipartite reader-book graph, the degree distributions of both users p_j and books q_k which are calculated from Equation (2) are shown on logarithmic scales in Figure 1, and p_j appears to have power-law tails, and an exponentially truncated power law is a better fit for q_k. The

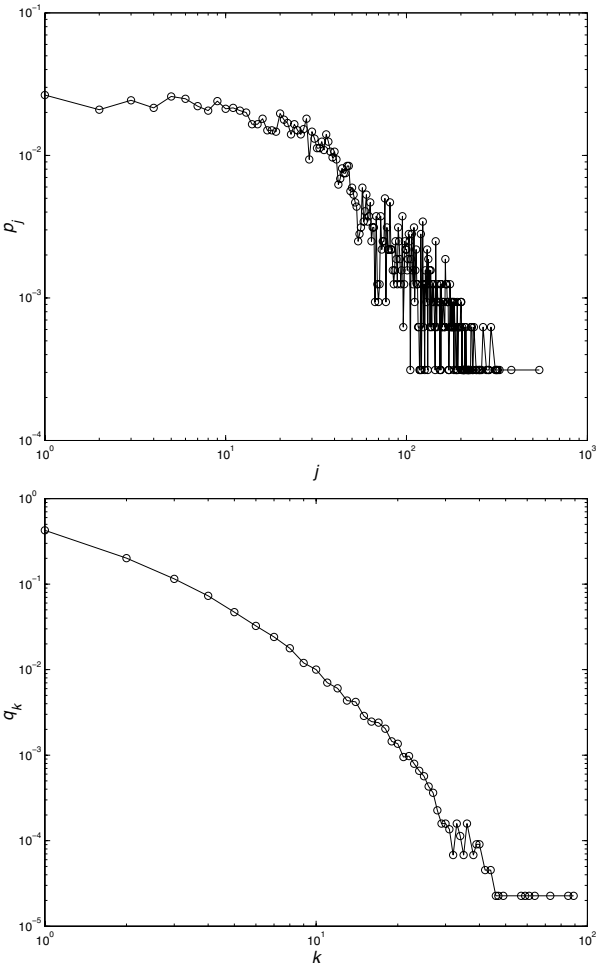

Fig. 1. The degree distributions of both readers p_j and books q_k in bipartite graph

average degree of the network is the average value of all vertices degree. The average degree of readers is 42.30 and that of books is 3.07.

The projected reader network contains only 3175[1] vertices and 269,721 edges. There is only one giant component in the network, i.e., the network is completely connected. In the reader network, degree of vertex means the number of neighbors with which the reader borrowed the one or more same books. The degree distribution ρ_k of the projected reader network is shown in figure 2, and it shows more fluctuation.

[1] If the degrees of some readers' adjacent (books) vertices are 1, these reader vertices are lost in the projected reader network [8].

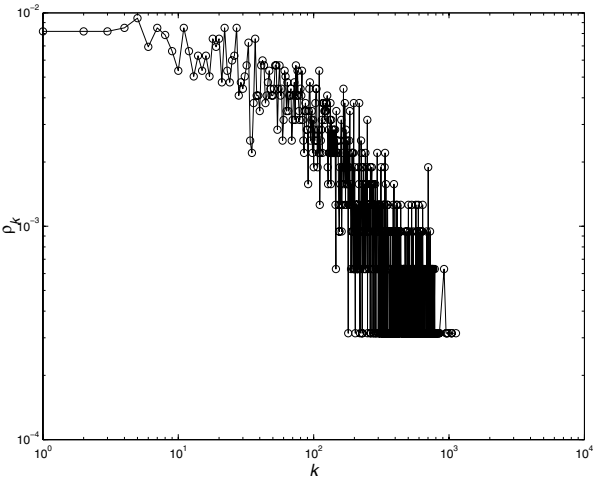

Fig. 2. The degree distribution ρ_k in projected reader graph

A fundamental concept in graph theory is the 'geodesic' or shortest path of vertices and edges that links two given vertices. With the concept of distance, we can define closeness centrality. The closeness centrality of a vertex is based on the total distance between one vertex and all other vertices, where larger distances yield lower closeness centrality scores. We use breadth-first search [12] to calculate exhaustively the lengths of the shortest paths from every vertex on the network and averaged theses distances to find the mean distance between any pair of readers.

An interesting idea circulating in the social networks community currently is that of "transitivity," which describes symmetry of interaction among trios of actors [14]. It refers to the extent to which the existence of ties between actors A and B and between actors B and C implies a tie between A and C. The transitivity is that fraction of connected triples of vertices which also form "triangles" of interaction. Here a connected triple means an actor who is connected to two others. This quantity is usually called the clustering coefficient, and can be written

$$C = \frac{3 \times number\ of\ triangles\ on\ the\ graph}{number\ of\ connected\ triples\ of\ vertices} \tag{6}$$

We calculated the actual topological measures of the projected reader network, such as average degrees, average path lengths, and clustering coefficients. We used equation (4) to calculate the expected average degree, average path length, and clustering coefficients of the projected network with given two-mode network degree distribution. These bipartite consumer-service graph degree distributions were computed directly from the borrowing log file. The deviation of the actual values from the expected values of the three topological measures would indicate that the relationship between readers in projected network is not only determined

by the degree distribution of the two-mode network, in other words, there is some underlying mechanism to cause the deviation.

In Table 1 we show values of the three actual topological measures of the reader network and theoretical predictions calculated from Equation (4). The reader network exhibits a substantially larger average path length and higher clustering coefficient than those of random networks, while the average degree measure is smaller than that of its random counterpart. The percentage errors of predictions from actual vales are given in the last row of Table 1. As the table shows, these are all quite large: they vary from 36% for average degree to 90% for cluster coefficient. These findings strongly suggest that the consumers' demand is not random, and we conjecture intuitively that the reader network may show group structure. Communities appear in networks where vertices join together in tight groups that have few connections between them. Dense connections in groups can produce high clustering coefficient and not very high average degree, and looser connections between groups cause larger average path length. One might well imagine that reader network would divide into groups representing particular areas of research interest or specialty. However, this assumption must be empirically confirmed.

Table 1. Summary of the actual and predictive three topological characteristics for the projection reader network studied here

	z_1	L	C
actual	169.903	2.149	0.292
predictive	266.035	1.388	0.153
percentage error of prediction(%)	36.135	54.827	90.850

3.3 Community Structure Analysis

To analyze the community structure we use fast algorithm proposed by Newman [12]. The algorithm is based on the concept of modularity, which is described in section 2.2 of this paper. The algorithm proceeds as follows:

1. Starting with a state in which each vertex is the sole member of n communities.
2. Calculating the change Q in upon joining of a pair of communities between which there are edges. If more than one $\triangle Q$ highest values, then one of them is chosen at random.
3. Incorporating communities resulting in the greatest increase (or smallest decrease) in Q and step 2 is repeated until one community remains.
4. Selecting the best division by looking for the maximal value of Q.

The result derived by feeding the reader network into the algorithm turns beyond expectation. The peak modularity is only $Q = 0.155$ which is too small to indicate significant community structure ($Q > 0.3$). We doubt whether the community structure assumption is correct or there is important details we missed.

Then we bring to mind the fact that in the reader network construction two readers are connected by just one link, although they may borrow more than one same book. The data sets used here present more information than in the simple networks we have constructed from them. In particular, we can count quantity of book each pairs of readers have borrowed during the period of the study. We can use this information to make an estimate of the strength of relations.

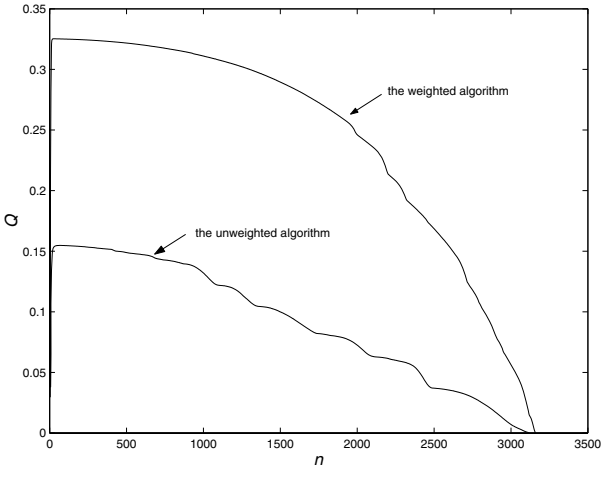

Fig. 3. Plot of the modularity Q versus community number n with unweighted and weighted algorithms in projected reader graph

We introduce weighted reader network, which allows for this by including a measure w_{ij} as the strength of interaction, which is the number of same books they have borrowed. The algorithm of detecting community structure is generalized trivially to weighted networks in which each edge has a numeric strength associated with it, by making the initial values e_{ij} of equation (5) equal to those strengths. The analysis reveals that the network consists of about 26 communities, with a high peak modularity of $Q_w = 0.325$, indicating significant community structure. In figure 3 we show the modularity Q versus community number n with unweighted and weighted algorithms. As we can see, $Q_w = 0.325$ is steeper than Q at the beginning of joining two communities and then the slope of Q_w is gentle, i.e. the weighted algorithm joins the tightly connected communities rapidly.

Eleven of the communities found by weighted algorithm are large, containing between them 89% of all the vertices, while the others are small-see Table 2. There appears to be a strong correlation between the community found by the algorithm and the department division related to the readers. The largest department component of each community is bold font style.

The weighted algorithm seems to find two types of communities: Teachers and students grouped together by similarity research background (community

Table 2. Crosstabulation between the community found by the algorithm and the department divisions related to the readers

Community	Department code									Total
	A	B	C	D	E	F	G	H	+38 smaller departments	
1	**493**	13		6	8		4	36	31	591
2	57	7	42	35	2	**140**	3	5	40	331
3	5	16	**196**	10		7	2	14	39	289
4	1	**234**	1	5	1		16	2	11	280
5		41	10	**168**	1		3	1	14	238
6	26	8	5	8	**154**	3	4	2	24	234
7	**127**	10	6	7	6	5	2	2	31	196
8	**152**	10	2	3	1	5	3	1	16	193
9	18	**116**	8	1	10	3	15		7	178
10	6	26	**60**	46		8	4		18	168
11	11	**46**	8	6	3		26	10	22	132
+15 smaller communities	66	63	38	42	7	14	12	20	83	345
Total	962	599	376	337	193	185	94	93	336	3175

one, four, seven and eight), or by same foundation courses (community two, ten), and farther analysis shows these courses can be concerning foreign language or computer technology. From the view point of department, Readers of a department are mainly distributed a few communities, especially in the large size departments (department A, B). The community structure presents the divisions running along disciplinary lines as well as the mark of interdisciplinary research or knowledge.

4 Conclusions

In this paper we have applied network analysis to study consumer demand in a library setting. We represent service transaction as a bipartite graph, and study the topological characteristics of the reader graph projected from the consumer-service graph. The results show that the topological characteristics of real reader graph deviate significantly from the theoretical predictions based on a random bipartite graph. The graph exhibits smaller-than-expected average degree, larger-than-expected average path length and stronger-than-expected tendency to cluster.

We assume the existence of community structure in the reader network, and use the algorithm for detecting community structure to confirm the assumption. We have found that the unweighted reader network doesn't present community structure, while the network show clear community structure with a simple weighting method. This is consistent with the statement that projection unipartite graph from a bipartite graph causes information loss, and weighting method

is proper way to retain the original information [8]. The community structure analysis reveals that this construction captures essential ingredients of disciplinary interactions between readers.

Owing to the fact that knowing the interrelation of consumer's demand is the important thing when any organization wants to improve service quality or enhance service efficiency, we hope that the ideas and methods presented here will prove useful in the analysis of many other types of consumer-service networks.

References

1. Amaral, L.A.N., Uzzi, B.: Complex Systems-A New Paradigm for the Integrative Study of Management, Physical, and Technological Systems. Management Science 53, 1033–1035 (2007)
2. Epstein, J.M., Axtell, R.: Growing Artificial Societies: Social Science from the Bottom Up. The MIT Press, Cambridge (1996)
3. Wolfram, S.: A New Kind of Science. Wolfram Media, Champaign (2002)
4. Newman, M.E.J.: The structure and function of complex networks. SIAM Review 45(2), 167–256 (2003)
5. Huang, Z., Zeng, D.D., Chen, H.: Analyzing Consumer-Product Graphs: Empirical Findings and Applications in Recommender Systems. Management Science 53, 1146–1164 (2007)
6. Newman, M.E.J., Strogatz, S.H., Watts, D.J.: Random graphs with arbitrary degree distributions and their applications. Phys. Rev. E 64, 26118 (2001)
7. Nooy, W., de Mrvar, A., Batagelj, V.: Exploratory Social Network Analysis with Pajek. Cambridge University Press, Cambridge (2005)
8. Zhou, T., Ren, J., Medo, M., Zhang, Y.C.: Bipartite network projection and personal recommendation. Phys. Rev. E 76, 046115 (2007)
9. Zhang, P.P., Chen, K., He, Y., Zhou, T., Su, B.B., Jin, Y.D., Chang, H., Zhou, Y.P., Sun, L.C., Wang, B.H., He, D.R.: Model and empirical study on some collaboration networks. Phys. A 360(2), 599–616 (2006)
10. Girvan, M., Newman, M.E.J.: Community structure in social and biological networks. Proc. Natl. Acad. Sci. USA 99, 7821–7826 (2002)
11. Girvan, M., Newman, M.E.J.: Finding and evaluating community structure in networks. Phys. Rev. E 69, 026113 (2004)
12. Newman, M.E.J.: Fast algorithm for detecting community structure in networks. Phys. Rev. E 69, 066133 (2004)
13. Newman, M.E.J.: Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. Phys. Rev. E 64, 016132 (2001)
14. Newman, M.E.J.: Scientific collaboration networks. I. Network construction and fundamental results. Phys. Rev. E 64, 016131 (2001)
15. Gleiser, P., Danon, L.: Community structure in jazz. Preprint cond-mat/0307434 (2003)