Exploring the Auto Model Competition Patterns in China's Auto Market based on Complex Networks Theory

Sheng Zhang
College of Artificial Intelligence
Beijing Normal University
Beijing, China
zsheng_2018@163.com

Haoyang Che
College of Artificial Intelligence
Beijing Normal University
Beijing, China
chehy@hotmail.com

Jiacai Zhang*
College of Artificial Intelligence
Beijing Normal University
Beijing, China
jiacai.zhang@bnu.edu.cn

Yucong Duan
College of Information Science and Technology
Hainan University
Haikou, China
duanyucong@hotmail.com

ABSTRACT

Understanding the competition pattern of auto models is critical for stakeholders including automakers and dealers. However, the traditional methods mainly rely on the experience and analytical dimensions of the analyst, which lack reliable methodology and ignore the value of user behavior. In this paper, we propose a novel method based on complex network theory, construct an auto model competition network with users' sales leads, and analyze the static characteristics of the network. Besides, by using different community detection algorithms and constructing predictive models, we discovered that there are six major communities in the network, and that price, popularity, model level, as well as model asset ownership, are the main factors affecting community division.

1 INTRODUCTION

China's auto sales declined for the first time in 2018 [17]. This is undoubtedly putting tremendous pressure on stakeholders, including automakers and dealers. It is extremely important to understand the competition pattern of auto models, which can help them to recognize market needs, identify emerging competitors, and develop targeted auto production and sales strategies.

In terms of the competition patterns analysis, traditional methods are often limited to strategic management and market analysis, such as SWOT analysis [8] and the Porter Five Forces model [13]. However, these methods mainly rely on the experience and intuition of analysts, and lack reliable methodology. In addition, the analysis dimension is often confined to car sales and user feedback, ignoring the value of other user behaviors. Thus, it may cause unstable performance in pattern interpretation.

At the same time, with the advent of mobile Internet, vertical auto websites (VAWs) have become an important channel for people to obtain car information and buy cars. More and more users will browse the car information on VAWs and leave their sales leads (customer's personal information, including name and phone number, for sales purposes) before purchasing a car, so that dealers can contact them to make an appointment for a test drive. After more than a decade of accumulation, leading websites have accumulated more than 500 million users, 100 million sales leads, and billions of user behavior data.

In order to solve the problems of traditional methods, we propose a novel method from the perspective of complex networks, using the sales lead data of auto models from VAWs to build an auto model competition network, and explore and analyze the auto model competition pattern of China's auto market. Figure 1 outlines our framework, which consists of three parts: data preprocessing, network construction, and competition pattern analysis. Among them, competition pattern analysis includes network visualization, characteristic analysis, and community structure analysis. Compared with the traditional method, our method has the following advantages: First, our model is based on a complex network and has a solid theoretical foundation. Second, we use the sales lead data of auto models, which is more valuable than data such as car sales. It comprehensively reflects the preferences of users and the comparison of different models. Lastly, we have established a complete analysis framework, which can improve the efficiency and reliability of the analysis.

By applying our model to 6,152,335 sales leads of 1069 auto models in January 2019, we have two main contributions:

- We constructed auto model competition networks, performed visualization and network characteristic analysis, revealing the characteristics such as intensified competition and small-world phenomenon.
- We found six major communities using community detection algorithms, and built prediction models based on them. We found that price, model level, and popularity were the main factors to affecting community division.

The rest of this article is organized as follows. Section II introduces the related work of strategic management, marketing and complex network in auto competition pattern analysis. In section III, we describe the dataset and data preprocessing steps. In section IV, we construct the auto model competition network in January 2019 and perform the network visualization. Section V analyzes the static characteristics of the network. In section VI, we divide the community structure of the network, and find the main factors affecting community division by constructing predictive models. Section VII concludes the paper.

2 RELATED WORK

Many investigations have researched the auto market competition pattern from different aspects. In this section, we will classify

*corresponding author

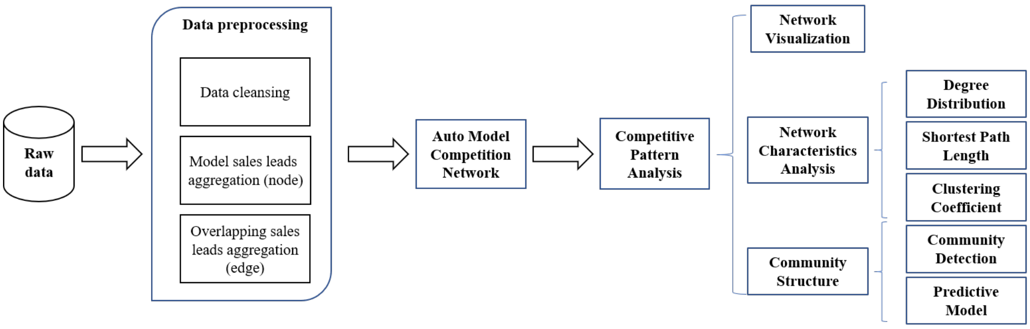


Figure 1: Framework Overview

the related work into strategic management and marketing analysis, and complex network.

Strategic management and marketing analysis are the most common methods to study the competition pattern in the auto market. Study [4] applied SWOT analysis and the five-force model, studied the competition pattern of Chery Automobile, and pointed out the huge threat posed by other brands entering the low-end model market. Study [14] analyzed the competitive environment, opportunities and challenges faced by FAW-Volkswagen's new energy models based on the PEST model and the SWOT model. However, these methods highly depend on analysts' experience and intuitions instead of a solid methodology foundation, which could perform less stable in bidding presentation and pattern interpretation.

Another method is complex networks based on graph theory. In recent years, the research of complex networks has expended from the fields of physics and computers to society and technology. Numerous theoretical studies and empirical analyses have also emerged [1, 5, 15]. In the auto field, Lijuan Zhang et al. studied the cooperation network between automakers and parts suppliers, and found the small-world phenomenon of the network [16]. Jianmei Yang et al. used the Newman fast community algorithm to divide the network of auto companies into different communities based on their product categories, and established a multi-layer network to analyze the confrontation behavior between automotive companies [10]. However, these researches are only from the perspective of automakers and suppliers, without taking user behavior data into consideration.

In summary, different from existing researches, we build an effective framework from the complex network perspective, and use massive sales leads data from VAWs to analyze the auto model competition pattern.

3 DATA AND PREPROCESSING

The original dataset is from one of China's largest VAWs, which contains 1 PB anonymous log data from January 2017 to January 2019. Each entry includes anonymous user ID, province, city, sales lead source and time, as well as the corresponding brand, model and style information, which is shown in Table 1.

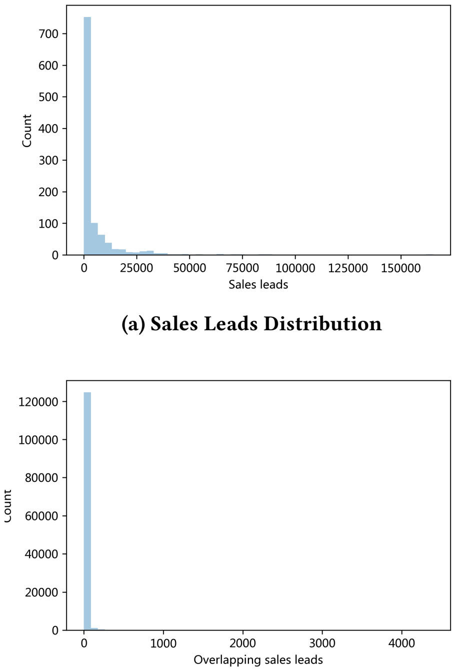
Table 1: The Original Dataset Schema

Field	Description	Example
ID	Row ID	RID_00000001
User	Anonymous user ID	UID_111111111
Province	User province	Guangdong
City	User city	Shenzhen
Source	Sales lead source type	Mobile, Web, other
Time	Sales lead time	2019-01-01 00:00:00
Brand	Brand of sales lead	Volkswagen
Model	Model of sales lead	Volkswagen Lavida
Style	Style of sales lead	Volkswagen Lavida 1.4L

As we mentioned before, sales leads refer to users' information for sales use, including names, regions and contact information of potential customers. If a user leaves his/her information on a model on a VAW, which indicates that he/she is interested in this car and could be a potential buyer. Because sales leads require the user's personal information, users will be more cautious when leaving sales leads, improving the authenticity of the data. The overlapping sales lead refers to the situation that different models have the same sales leads, which means that users may be interested in multiple models at the same time, so these models could be potential competitors. Overlapping sales leads reflect users' comparison of different models. Compared with other data, sales leads and overlapping sales leads have higher authenticity, and timely and accurately reflect the user's preferences for models (sales leads) and comparisons between different models (overlapping sales leads).

In order to construct the auto brand competition network, we need to process the data into the required form. First, since some cars have different ids, and/or names of brand, model or style, we need to identify and unify them. After that, duplicate, missing and erroneous entries are eliminated. And because the automakers and dealers usually analyze the auto model data monthly, we need to aggregate the sales leads data by month. Besides, shorter (such as daily) or longer periods (such as annually) may not be able to accurately or timely reflect the model competition pattern. For instance, if 100 users left their sales leads in October 2018 on Camry, the model sales leads of Camry in October 2018 are 100. Finally, we extract and aggregate the same sales leads between different models as overlapping leads. For example, if ten users left their sales leads on Camry and Jetta in October 2018, the overlapping sales leads between Camry and Jetta in October 2018 are 10. To study the recent competition pattern, we selected the data for January 2019, including 1069 brands with 6,152,335 sales leads and 1,129,919 overlapping brand sales leads.

Figure 2 illustrates the model sales leads and overlapping sales leads distribution in January 2019. In Figure 2 (a), it is obvious that most models have relatively low sales leads, but a few models such



(a) Sales Leads Distribution

(b) Overlapping Sales Leads Distribution

Figure 2: Distribution of Sales Leads and Overlapping Sales Leads

as Jetta, Lavida and Sylphy have a very high amount of sales leads, ranging from 1 to 165,199. Figure 2 (b) shows the distribution of overlapping sales leads between different brands. Similarly, most overlapping sales leads are low, while others are very high such as overlapping sales leads between Jetta and Santana (4,393 overlapping sales leads). Figure 2 indicates the number of sales leads between different models is huge, suggesting that there are different model divisions.

4 NETWORK CONSTRUCTION & VISUALIZATION

The auto model competition network is essentially a graph. By regarding the auto models as nodes (sales leads as size), and competition relationship as edges (if two nodes have overlapping sales leads) which link different models, we can abstract the auto model competition network. In the network, brands with overlapping sales leads are considered to be competitors. And the network is built with networkx Python library [7].

There are 1069 nodes and 126,650 edges in the network of January 2019. Among them, there are 28 isolated nodes (i.e., no edges). And Figure 3 shows the network of January 2019 without isolated nodes. The size and color of nodes reflect the number of sales leads for the model, and the thickness of edges represents the amount of overlapping sales leads. To be specific, if the size of the node is larger and the color of the node is redder, it has more sales leads. And if the thickness of the edge is thicker, the color of the edge is redder, there are more overlapping sales leads between the two nodes, and their competition is fiercer. In addition, the figure is drawn using Gephi and its built-in ForceAltas2 layout algorithm [2, 9]. Non-overlap option was chosen to ensure the nodes do not overlap. And all the node sizes (the number of sales leads from the nodes) and edge weights (the number of overlapping sales leads from the edges) have been rescaled for clarity.

Figure 3 gives an overview of the auto model competition network. Basically, the nodes in the middle have more sales leads and overlapping sales leads. However, there is a certain distance between the nodes with the most sales leads (such as Lavida and Jett), suggesting a potential community structure and they may belong to different communities.

5 NETWORK CHARACTERISTICS ANALYSIS

Degree distribution, average shortest path length and clustering coefficient are the most common characteristics of a network. In this section, we will analyze the characteristics of the auto model competition network in January 2019, and discuss the interpretation of these characteristics.

- **Degree distribution:** The degree of a node refers to the number of edges connected to the node.
  The degree distribution is shown in Figure 4 (a). As we can see, the number of nodes decreases as the degree increase and decreases almost constantly, except for the beginning part. Since the degree represents the number of connected edges of a node, that is, the number of directly adjacent nodes, which means the degree of a node represents the number of direct competitors of the model it represents. Therefore, Figure 4 (a) illustrates that as the number of competitors increases, the number of models decreases. Among them, the node with the highest degree is Lavida (with 809 competitors), instead of the node with the most sales leads—Jetta. On the contrary, there are also 28 nodes with a degree of 0, that is, isolated nodes without competitors. And these models are excluded in the following discussion. Besides, the average degree is 236.95, which shows that there are nearly 240 competitors per model, reflecting the fierce market competition.
- **Average shortest path length and diameter:** The average shortest path length is the average distance between all pairs of nodes (if the graph is connected). And diameter describes the maximum path length in a network.
  Due to the large difference in weight between nodes, and the weighted shortest path length cannot be used to describe the small-world phenomenon of the network, we will ignore the weight of the connected edges (i.e. regarded as a binary network). And as we mentioned before, since the original network is not connected, we choose the largest giant component (LGC network), which is exactly the original network after removing all isolated nodes.
  The average shortest path length of the LGC network is 1.82, and the diameter is 4, which are really small compared to the number of nodes (1041 nodes). Figure 4 (b) shows the distribution of the shortest path length between all node pairs in the network. Obviously, most nodes have direct competition (the shortest path length is 1, 23.4%) or common competitors (the shortest path length is 2, 71.1%). Only less than 0.5% of the shortest path length equals to the diameter of the network (length = 4).
- **Clustering coefficient:** The clustering coefficient measures the situation of interconnection between neighbor nodes of nodes in the network.

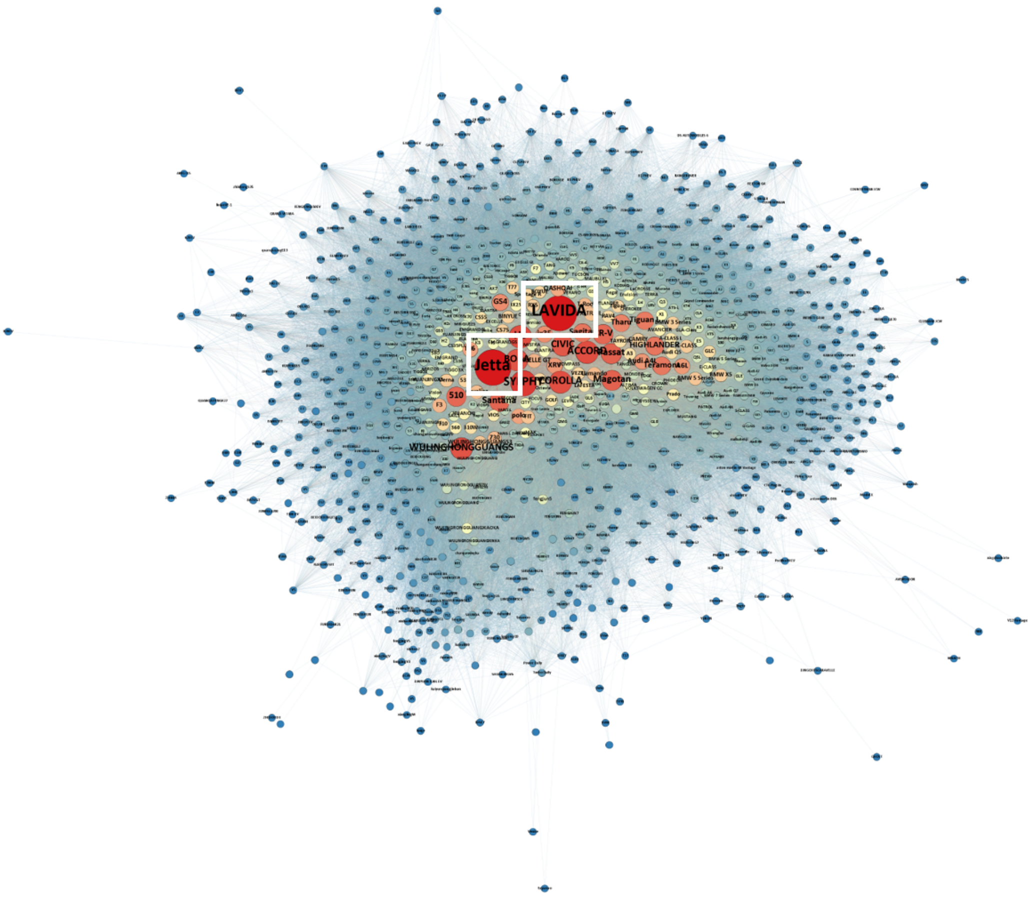


Figure 3: Auto Model Competition Network

Figure 4 (c) depicts the distribution of clustering coefficient in the network. The clustering coefficient of most nodes is between 0.4 to 0.8, which indicates that most of the models with common competitors are also competitors, and there is a relative obvious clique effect. And the average clustering coefficient is 0.64, significantly higher than corresponding random network.

In summary, low average shortest path length and high clustering coefficient imply the network possesses the small-world phenomenon. It means that although most nodes are not connected to each other, most nodes can be reached in a few steps. And it is likely to contain cliques or sub-networks, which implies that the network may contain multiple communities, and this will be discussed in section VI.

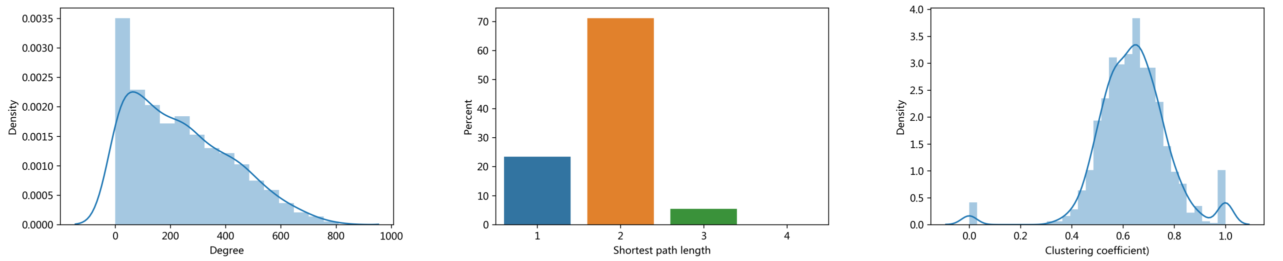
In conclusion, the auto model competition network presents the differences in degree distribution and small-world phenomenon. Corresponding to the real world, they illustrate the fierce market competition, and potential multiple communities.

6 COMMUNITY STRUCTURE AND PREDICTION

In fact, the auto models already have different classifications according to auto brand, usage, nationality, price range and so on. However, these classifications can only represent the model's own attributes, and cannot comprehensively reflect the users' evaluation and actual division in the auto market. For automakers and dealers, it is important to understand the actual division of auto models in the auto market, identify current and even potential competitors, and assist them in formulating future production and marketing strategies. Besides, we have initially determined that there is a certain community structure in the auto model competition network. Therefore, in this section, we will first detect the community structure of the network, and then build prediction models based on the communities to find key features that affect community division and users' choice.

6.1 Community Structure Detection

The community structure was proposed by Girvan and Newman in 2002 [6]. Generally, a community represents a group of nodes with similar characteristics, and there may be multiple communities in a network. According to the definition, the nodes within a community are more closely connected, while the nodes of different communities are loosely connected. At present, many community detection algorithms have been proposed, such as the GN algorithm [6], the fast Newman algorithm [11], and the Louvain algorithm [3]. At the same time, Newman et al. also proposed a modularity function to evaluate the quality of community structure division in the network [12]. This value is between [-1/2, 1], and the closer is it to 1, the better the community division effect. In fact, the value in practical applications is generally between 0.3 to 0.7 [12].



(a) Degree Distribution **(b) Shortest Path Length Distribution** **(c) Clustering Coefficient Distribution**

Figure 4: Distribution of Degree, Shortest Path Length & Clustering Coefficient

Table 2: Comparison of Community Detection Algorithms

Algorithm	Modularity Score	Number of clusters	Computation Time (s)
Fast Newman	0.031	282	7.652
Louvain	0.329	6	4.412

Table 3: Features to Predict Community Division

Fields	Description	Example
Num_leads	Number of sales leads of the model	1000
Price_high	The highest price of the model	16.28 (in 10,000 CNY)
Price_low	The lowest price of the model	11.08 (in 10,000 CNY)
Model_level	The model classification	Minicar (14 kinds in total)
Country_name	The country of the model	Germany (10 countries in total)
Country_class	Model asset ownership	Domestic (or imported/ joint venture)
Brand_name	The brand of the model	Volkswagen, . . . (130 brands in total)

Table 4: Model Prediction

Model	Accuracy	Precision	Recall	F1 Score
Random Forest	0.8119	0.8290	0.8120	0.8036
XGBoost	0.8220	0.8420	0.8220	0.8203

In this section, we use the Fast Newman and Louvain algorithms for community detection, both of which are greedy algorithms based on modularity maximization. And the algorithms' results are shown in Table 2 (edge weights are considered here). Obviously, the Louvain algorithm performs better, not only has a higher modularity score, but also has a shorter computation time. Besides, the interpretability of 6 clusters of 1041 nodes is significantly higher than that of 282 clusters. Figure 5 shows the community detection results of the Louvain algorithm, where different colors represent different communities. Although the number of nodes in each community is different, the nodes within the same community are all in proximity. A detailed interpretation of the community division will be in the next part.

6.2 Community Prediction

Based on the community structure detected in the previous section, we constructed several predictive models to find the key features that affect community division.

First, we need to propose several features that may influence the community division of the auto model competition network, including the number of sales leads, the highest price of the model, the lowest price of the model, the model classification, the country of the model, the model asset ownership and the brand of the model, as shown in Table 3. Among all these features, the first three features are numerical variables, and one-hot encoding is used on the rest four features.

Then Random Forest and XGBoost with 5-folds cross-validation are applied to these features and community labels. The metrics and performances are shown in Table 4, which are all mean values with 5-folds cross-validation. Obviously, the XGBoost has better performance in all metrics. And we find that the most important features are price ('price_low' and 'price_high'), popularity ('num_leads'), model level ('model_level'), and model asset ownership ('country_class').

Therefore, according to the community division and key features, we can summarize the characteristics of all the 6 communities, illustrating in Table 5. To be specific, community 1 is mainly imported or joint-venture SUV with price ranging from 170K to 240K CNY. Community 2 is mainly popular compact cars between 120K to 170K CNY. Community 3 is basically some cheap cars, including mini cars, compact cars, small cars and SUVs. Community 4 has the most models, which are all expensive cars, such as SUVs, medium cars, medium and large cars, large cars, luxury cars and sports cars. Community 5 does not include sedans, but MPVs, trucks, pickups, vans, buses and so on. Finally, Community 6 is mainly domestic SUVs.

Combined with Figure 5 and the community characteristics above, we have several findings: First of all, the compact cars within 120k to 170k in China are the most popular ones (i.e. community 2) with the highest average sales leads. Second, SUV is the most popular model classification, appearing in almost every community. And domestic SUVs and imported and joint venture SUVs are in different communities. Finally, we find that the high price community (community 4) has the largest number of models, but the average number of sales leads is the minimum in sedans (excluding community 5).

Table 5: Community Characteristics

Community Number	Color	Number of Models	Average Number of Sales Leads	Characteristics	Example
1	Dark Green	104	10646.30	Mainly SUVs (imported or joint-venture)	CR-V
2	Pink	48	20834.48	Mainly popular compact cars	Lavida
3	Light Green	204	5598.01	Low price (mini/compact/small cars and SUVs)	Jetta
4	Violet	312	4861.70	High price (mainly SUV/medium/medium and large/large cars, Luxury cars, and Sports cars)	Accord
5	Blue	203	2271.48	Not sedan (MPV/truck/pickup/van/bus…)	WulingHongguangS
6	Orange	149	6187.33	Mainly domestic SUV	Haval H6

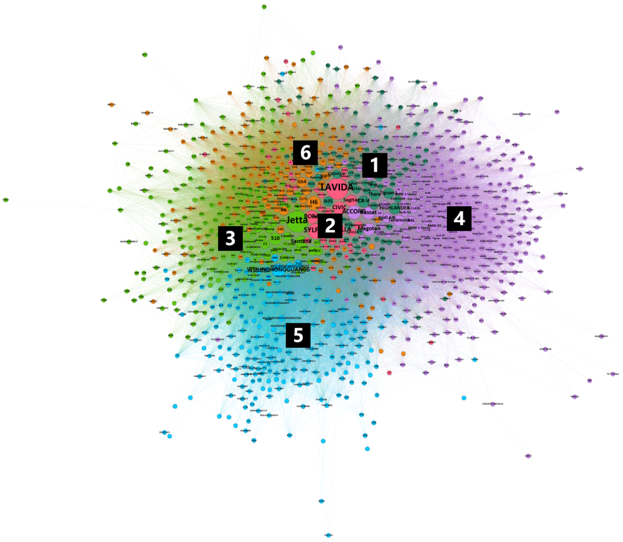


Figure 5: Community Detection in the Auto Model Competition Network

In summary, we use the Louvain algorithm to find 6 communities in the auto model competition network, and construct the XGBoost predictive model to find key features that affect community division and users' choice, and summarize the characteristics of the 6 communities.

7 CONCLUSION

In this paper, we studied the competition pattern of auto models in China's auto market based on the sales leads data with the complex network theory. Our investigation involved 6,152,335 sales leads with 1069 models from a vertical auto website, and China's auto model competition network was established based on the models as nodes, and the competition relationship as edges. There are two important contributions. First, we constructed auto model competition networks of January 2019, performed visualization and network characteristic analysis, revealing the characteristics such as intensified competition and small-world phenomenon. Second, we discovered that there are 6 communities in the network, and built predictive models to find that price, popularity, model level and model asset ownership are the key features to determine the model community structure. In conclusion, with the decline in car sales, the competition between models has become increasingly fierce. And among the 6 communities in the auto model network, the compact models within 120K to 170K CNY are the most popular. SUVs occupy a pivotal position in the entire auto model market.

Our research solves the problems in previous auto competition pattern analysis: the lack of solid theoretical foundation, the lack of comprehensive data, and lack of a complete analysis framework. However, this paper only researches the characteristics and community structure of the auto model competition network in January 2019 in detail, and the subsequent work will further study the dynamic characteristics and community structures.

ACKNOWLEDGMENTS

This research was funded by the National Key Technologies RD Program (2017YFB1002502), and the General Program (61977010) of Nature Science Foundation of China. This work was also supported by the project of Beijing Advanced Education Center for Future Education (BJAICFE2016IR-003). We would also like to thank Mr. Ran Feng for his suggestion on data preprocessing.

REFERENCES

[1] Albert-László Barabási and Réka Albert. 1999. Emergence of scaling in random networks. *science* 286, 5439 (1999), 509–512.

[2] Mathieu Bastian, Sebastien Heymann, and Mathieu Jacomy. 2009. Gephi: an open source software for exploring and manipulating networks. In *Third international AAAI conference on weblogs and social media.*

[3] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment* 2008, 10 (2008), P10008.

[4] Faen Chen and Yukio Kodono. 2012. SWOT analysis and five competitive forces of chery automobile company. In *The 6th International Conference on Soft Computing and Intelligent Systems, and The 13th International Symposium on Advanced Intelligence Systems.* IEEE, 1959–1962.

[5] Paul Erdős and Alfréd Rényi. 1960. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci* 5, 1 (1960), 17–60.

[6] Michelle Girvan and Mark EJ Newman. 2002. Community structure in social and biological networks. *Proceedings of the national academy of sciences* 99, 12 (2002), 7821–7826.

[7] Aric Hagberg, Pieter Swart, and Daniel S Chult. 2008. *Exploring network structure, dynamics, and function using NetworkX.* Technical Report. Los Alamos National Lab.(LANL), Los Alamos, NM (United States).

[8] Terry Hill and Roy Westbrook. 1997. SWOT analysis: it's time for a product recall. *Long range planning* 30, 1 (1997), 46–52.

[9] Mathieu Jacomy, Tommaso Venturini, Sebastien Heymann, and Mathieu Bastian. 2014. ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. *PloS one* 9, 6 (2014).

[10] YANG Jianmei ZHOU Lian ZHOU Lianqiang. 2013. Competitive Relationships of Auto Industry and Rivalry Actions of Car Community Enterprises in China. *Chinese Journal of Management* 1 (2013).

[11] Mark EJ Newman. 2004. Fast algorithm for detecting community structure in networks. *Physical review E* 69, 6 (2004), 066133.

[12] Mark EJ Newman and Michelle Girvan. 2004. Finding and evaluating community structure in networks. *Physical review E* 69, 2 (2004), 026113.

[13] Michael E Porter and Competitive Strategy. 1980. Techniques for analyzing industries and competitors. *Competitive Strategy. New York: Free* (1980).

[14] Long Sun. 2019. *Research on the development strategy of new energy vehicles for FAW-Volkswagen.* Master's thesis. Jilin University, Changchun, China.

[15] Duncan J Watts and Steven H Strogatz. 1998. Collective dynamics of 'small-world'networks. *nature* 393, 6684 (1998), 440.

[16] Li-juan ZHANG and Chang-hong LI. 2007. Study on cooperative networks in enterprises——An analysis of automobile manufacturing. *Science-Technology and Management* 4 (2007).

[17] Jie Zheng. 2019. Negative growth dust of the auto market in 2018 is settled. *Automobile Watch* 01 (2019), 18–19.