Predicting Tie Strength of Chinese Guanxi —by Using Big Data of Social Networks*

Jar-Der Luo[1], Xin Gao[1], Jason Si[2] Lichun Liu[3] Kunhao Yang[4], Weiwei Gu[5], Xiaoming Fu[6]

Please cited from:

Luo, Jar-Der, Xin Gao, Jason Si, Lichun Liu, Kunhao Yang, Weiwei Gu, Xiaoming Fu, 2019, "A Triadic Dialogue among Big Data Analysis, Sociology Theory, and Network Dynamics Models", invited speech, the 1st International Conference of Social Computing, Sep. 27th -28th.

ABSTRACT

This paper posts a question: How many types of social relations can a Chinese categorize? In social networks, the calculation of tie strength can better represent the degree of intimacy of the relationship between nodes, rather than just indicating whether the link exists or not. Previous researches suggest that Granovetter measures tie strength so as to categorize strong and weak ties, and the Dunbar circle theory may offer a plausible approach to calculate tie strength between individuals via unsupervised learning (e.g., clustering the interactive data between users in Facebook and Twitter). In this work, we differentiate the levels of tie strengths to measure the different dimensions of user interaction based on Dunbar circle theory. For labelling the types of ties, we also conduct a survey to collect the ground truth from the real world users and link to the surveyed data to big-data indicators collected from a widely used social network platform in China, QQ. After repeating the Dunbar experiments, we modify big-data indicators and predictive models in order to have a better fit for the ground truth. Eventually, four types of tie strength could better represent a four-layer supervised model developed from the Chinese Guanxi theory combined with Dunbar circle studies.

CCS CONCEPTS

• Applied computing → Law, social and behavioral sciences → Sociology

KEYWORDS

Tie Strength, Dunbar Circle theory, Chinese Guanxi theory, Supervised Model, Social Network

ACM Reference format:

Jar-Der Luo etc. 2019. Predicting Tie Strength of Chinese Guanxi: by Using Big Data of Social Networks. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management (CIKM) (CIKM'2019). ACM, Beijing, China, 9 pages.* https://doi.org/10.1145/1234567890

* Thanks for the financial support of Tencent Research Institute Project " Research on identification of opinion leaders based on QQ big data", project number: 20182001706 and Tencent Social Research Center project "Analyzing personal relationship on WeChat and QQ big data", project number: 20162001703.

1. Sociology Dept., Tsinghua University, Beijing, PR China.
2. Tencent Research Institute, Beijing, PR China.
3 Tencent computer system Co. Ltd, Shenzhen, PR China.
4. The University of Tokyo, Japan.
5. Beijing Normal University, Beijing, PR China.
6. University of Goettingen, Germany.

1 Introduction and Question Definition

How many types of social relations, or in the Chinese term, guanxi, can a Chinese categorize?

Guanxi theory proposed three principles for the Chinese social interactions, i.e. rules of needs, favor exchanges and equity [1]. However, there was no quantitative studies, based on these interaction principles, to categorize the types of guanxi that Chinese actually have. Big data helps solving this challenging question. Adopted from Dunbar's studies, this paper tries to combine the surveyed data with big-data collected from QQ, a most widely used social-network platform in China, to address this research topic.

Reviewing social network theory, Granovetter classified the tie strength into two categories including strong and weak ties [2, 3]. Foremost, his studies on weak tie, which can bring heterogeneity information and opportunity, has generated significant impact and wide applications in many areas. In addition, Granovetter also pointed out that tie strength could affect the flow of information and the logic of interaction between people. However, in his work there is no specific method, mathematically, to indicate whether an exact boundary exists between the strong and weak tie. Based on his work, many researchers had made efforts to develop practicable methods to measure the tie strength [4, 5, 6, 7, 8, 9, 10]. Nevertheless, most of the follow-up work focused on the continuum of intimacy, interaction frequency, reciprocity and friendship duration, rather than distinguishing the strong tie from weak tie.

It is also well known that the Dunbar theory [6, 11, 12, 13, 14, 15, 16, 17] suggested a five-category model of social relations a plausible solution for the measurement of tie strength, which defines specific five circles which have a clear boundary between two contingent circles. In contrast to his Western counterpart, within the Chinese social context, Fei [18], Yang [19] and Hwang [1] categorized Chinese tie strength (hereafter, it is called guanxi) into three types of interaction principles to describe the social

relationships of the traditional society in China. They proposed the different behavioral principles for the three types of Chinese guanxi, including: 1) rules of need - representing family or pseudo-family member, 2) rules of favor exchange - representing familiar ties, and 3) rules of equity - representing acquaintance ties. Comparing with the five circles in Dunbar theory, some guanxi among Chinese people, to some extent, may not have such clear boundaries between two adjacent relationships. Thus, a challenging question is how to compute tie strength between individuals in the Chinese context?

This work attempts to establish a model to calculate the tie strength between individuals by inputting online interaction data from a social network platform, and then output a predictive model, which categorizes Chinese guanxi. We will test the model against the ground truth collected from social surveys. To be specific, the input data is from one of the most populous online social network platforms called QQ, which has large active users, rich functions and multiple network footprints of users. Thus, the data set is a valid one for us to explore the social relationships of Chinese people nowadays.

In this study, our main contributions includes: 1) Inspired by Dunbar theory, we try to propose a classification model that can categorize the Chinese guanxi into distinct categories in a quantitative way. 2) This paper verifies the fitness of Dunbar circle in the Chinese context. Through theoretical exploration based on Gunaxi theory and mining results of online social network data, a temporal-contextual four-layer model is proposed, which has good predictive power in computing the Chinese tie strength. 3) The methodology applied in this paper can be extended to many other research areas. Based on social science theory and social surveys, our interviewees first label the types of their guanxi as the ground truth, which can be used to test the accuracy of our classification model.

2 Dunbar Circle Theory and Chinese Guanxi Theory

In previous research, the Dunbar circle theory [5, 11, 12, 13, 14, 15, 16, 17] was proposed to measuring the tie strength. The Dunbar circle, by definition, is a concentric structure with five circles, each of which represents a specific strength of social ties: from the innermost to the outermost, they are: 1) support clique, 2) sympathy group, 3) overnight camp group or affinity group, 4) community or active network, and 5) tribe. And there are different interactive logics and functions in various circles. In addition, the distribution of the number of each circle is also put forward, ranging from 5 in the innermost group to 300 in the outermost group (shown in Figure 1).

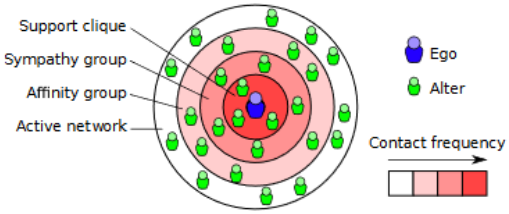

Figure 1: Dunbar Circle Illustration

In recent study, Dunbar [16] used Facebook and Twitter data to verify his theory within the arena of online social network data. He used only one dimension of interaction input, called contact frequency, and clustered the indicators of active users and their active contacts on Facebook and Twitter (Facebook and Twitter friends) [20]. This study finds five clusters, and put on the label for each cluster, which roughly matches with his theory suggested numbers of ties in each category. For example, Dunbar found from this research that each Facebook user had an average of 5.28 friends in the emotional support group. However, community and tribe groups cannot be well found in this study. It seems that this big-data analytical result did not completely confirm Dunbar's circle theory. In addition, without an actual label, it is risky to categorize a tie into one wrong circle. For example, a tie may be regarded as a support clique based solely on the high contact frequency between two people, but actually they contact so frequently just because of working relation.

The purpose of this study is stated as follows: 1) Making a localized explanation of Dunbar's five types of relationship in the Chinese context and designing a questionnaire to collect ground truth, i.e. defining and collecting the labels of real guanxi. 2) Using on-line interaction data of the interviewees from a widely used social network platform in China to repeat Dunbar's experiment and finding whether it is good fit for the Chinese context. 3) Employing the Chinese Guanxi theory and the on-line interactive indicators to revised the model. 'Guanxi' is defined as "a relationship to the extent that is high trust and relatively independent of social structure around the relationship" [1].

3 Defining Tie Strength and Ground Truth

3.1 Labeling Guanxi from Survey

Considering the difference between the two cultural contexts, five types of tie strengths need redefined, that can not only enable to repeat the Dunbar study, but also explore the model based on Chinese Guanxi theory. However, the Hwang's theory about "favor, guanxi and face" [1] proposes only the three principles of Chinese social interactions, rather than a clear boundary between various types of guanxi. According to the classification of Dunbar theory, we apply the three principles for 5 types of guanxi, which are respectively: 1) Family or pseudo-family members, good fit for rule of need, which is roughly equal to support clique. 2) Intimate familiar ties containing very strong friendship, suited for rule of favor exchange, which is roughly equal to sympathy group in which friends provide emotional support to each other. 3) Familiar ties

with long-term friendships, also following rule of favor exchange, which is roughly equal to affinity group. 4) Potential "friends", the ties that are mainly built on instrumental purpose but with the potential to have expressive features, roughly equals to active networks. 5) Acquaintances, purely instrumental ties, follow the rule of equity.

To obtain the ground truth of these categories of guanxi, a survey was conducted in the China's major first-tier cities, including Beijing, Shanghai and Guangzhou in China from May 1 to June 1, 2018. We used the sampling method by calling for volunteers to answer the questionnaire. Eventually we get the samples with age mostly ranging from 18 to 28 by using the way of offline face-to-face interviews. After ruling out the cases under 18 and above 28, we take the survey as a typical sample of Chinese large cities' young people.

Since the number of friends that a person can maintain is quite large, it is difficult for the interviewees to fill in all friends they have. Therefore, we asked interviewees to fill in at least 3 QQ friends in the most inner circle (theoretically, this number is under 5), at least 5 friends in the next three circles, and at least 8 people in the outermost circle (theoretically, it is about 300) respectively.

Questionnaire design-for the five types of ties:

- Please list at least three individuals whom you consider as the most intimate persons in your life, such as family or pseudo-family members.
- Please list at least five individuals with whom you have especially intimate friendship relations, such as your blood brothers.
- Please list at least five individuals who have long-term friendship favorable exchanges with you.
- Please list at least five individuals (including their relationship with you) who are not very close to you right now, but you are likely to build friendship and favor-exchange relations with him/her in the future.
- Please list at least eight individuals who are acquaintances, and you may or may not contact them in the future.

In addition to the type of tie strength between the respondents and their friends, we also ask the unique identification number of the respondents and their contacts mentioned in the social network platform.

There are totally 2012 ties were collected in the survey. We then selected the ties with active users on their both ends – the same data processing as Dunbar. Finally, there are 1502 labeled ties left. The number of labeled ties from the innermost layer to the outermost layer are 118, 147, 223, 349 and 665.

3.2 On-line Interaction Data

The on-line interaction data is collected from one of the most populous online social network platforms, QQ (hereafter called QQ data). With over 800 million monthly active users[1] in June 30, 2018, it is a good dataset for us to explore the social relationships among the Chinese people.

We collected the on-line interaction data through cooperation with Tencent. In the process of research, we strictly comply with the privacy rules: 1) we collect a user's on-line interaction information only when getting permission from the user. 2) Tencent's data collectors search for the data needed, anonymize the data, and then our analysts analyze the data with anonymization.

In the previous researches, an increasing number of indicators were developed to classify and predict tie strength [3, 4, 7, 9, 16, 17]. Based on these studies and the variables in our dataset, we figure out a series of meaningful indicators as the primary predictors for our guanxi-classification model (shown as table 1, in which all indicators are defined). Then, we compute Pearson correlation between the predictor and tie strength (shown in Table 2; note: tie strength in our guanxi-classification model is measured by the labeled guanxi collected from the survey), and only the indicators highly correlated to tie strength stay in analysis process.

These significant predictors include as follows. First, chat frequency, highly correlated to tie strength (.209**), is also used by Dunbar as the solely indicator in his model. In our data, there is no significant difference between the total number of messages from user i (the interviewee) to user j (the tie selected by the interviewee) and the number of messages from user j to user i. Thus, we merely select the messages sent from user i to user j.

Second, for one interviewee, since the total number of messages sent to different ties vary largely, the relative chat frequency (.138**) should be considered. That is, the number of messages from user i to user j divided by the total number of messages from user i.

Third, since standard deviation of chat frequency during working days and non-working days are important indicators to distinguish whether the two ends of a tie have working relationship. We also find that the interaction patterns (mean, standard deviation and distribution of chat frequency) between them show a high correlation with tie strength (.200** and .206**). Furthermore, the standard deviation of chat frequency (.384**) also contributes to measure the strength of ties.

Fourth, frequencies of offline meeting in the last three months, computed by the GPS information, are also significant variable to manifest the intimacy between two people (.160**, .106** and .150**). The different meeting time indifferent months, in holidays or non-holidays, makes it important to distinguish various relationships.

Fourth, the number of mutual friends, also known as common neighbors, is also a meaningful indicator correlate with tie strength (.132**), reflecting some information of the whole network.

The similarity between two people should affect their intimacy, therefore, the similarity in gender and professions is computed. However, the result shows that they have very little correlation with tie strength, and thus we erase off these two predictors. The Pearson correlation results present in Table 2.

Except the GPS information which can only be traced back last three months, the other indicators are calculated by a full year data, from July 2017 to June 2018. Matching the survey data (labeled

[1] https://www.ithome.com/html/it/377039.htm

Code	Indicator	Notes
RFF	Relative Chat Frequency	Chat Frequency between user i and user j divided by the total messages that user i sent during the period T0-T1.
TFF	Total Chat Frequency	Chat Frequency between user i and user j during the period T0-T1.
WFF	Total Chat Frequency during Working Time	Chat frequency between user i and user j on working days (Monday to Friday) during the period T0-T1.
NWFF	Total Chat Frequency during Non-working Time	Chat frequency between user i and user j on un-working days (Saturday and Friday) during the period T0-T1.
STF	The Standard Deviation of Chat Frequency	The standard deviation of daily chat frequency between user i and user j during the period T0-T1.
WSTF	The Standard Deviation of Chat Frequency during Working Time	The standard deviation of daily chat frequency on working days during the period T0-T1.
NWSTF	The Standard Deviation of Chat Frequency in Non-working Time	The standard deviation of daily chat frequency on nonworking days during the period T0-T1.
TRI	The number of mutual friends	The number of mutual friends between user i and user j at time T1
MF_8	Meeting Frequency in August	The frequency of user i and user j occur at the same time within a certain GPS in August
MF_9	Meeting Frequency in September	The frequency of user i and user j occur at the same time within a certain GPS in September
MF_10	Meeting Frequency on National Day	The frequency of user i and user j occur at the same time within a certain GPS on National Day
GS	The similarity of gender	The gender similarity of user i and user j
IS	The similarity of working industry	The working industry similarity of user i and user j

guanxi categories) and the QQ data (online interaction data), a data set with 1502 labeled ties is thus generated.

Additionally, a preliminary indicator contribution needs to be explored first. Thus, we run the regression of these indicators on tie strength (shown in Table 3). There are problems of multicollinearity regression coefficients instead of the significance level. According to the regression results, Total Chat Frequency during Non-working Time (NWFF), The Standard Deviation of Chat Frequency (STF), Total Chat Frequency during Working Time (WFF), Total Chat Frequency (TFF), the Standard Deviation of Chat Frequency during Working Time (WSTF) are the 5 most important indicators. Some other indicators respecting frequencies of offline meeting in August, September and October (MF_8, MF_9, MF_10), the number of mutual friends (TRI), the Relative Chat Frequency (RFF) are also included in our model, since they have both theoretical relevance and significant correlation with tie strength in our data. So far, we have established a set of indicators for on-line interaction data used as the input in the following models.

Table 1: The Explanations and Computation of Indicators

Table 2: Pearson Correlation between Relationship Strength and Indicators

indicator	Pearson coefficient
NWFF	.206**
WFF	.200**
TFF	.209**
RFF	.138**
STF	.384**
WSTF	.372**
NWSTF	.313**
TRI	.132**
MF_8	.160**
MF_9	.106**
MF_10	.150**
GS	0.028
IS	0.032

Note: ** The correlation was significant at the 0.01 level (bilateral)

Table 3: Regression Results of Relationship Strength

Note: * p≤5%, ** p≤1%, *** p≤0.1%, in a two-tailed test.

Furthermore, it is important to identify the differences in these on-line interaction indicators among various layers (shown in Table 4). We pay attention to the mean of serval most important indicators and show the difference in figures (as shown in Figure 2 and Figure 3). The statistics manifest that there is indeed huge difference among various guanxi categories. However, mean difference in some layers are notable different, suggesting it is reasonable to category the ties into different layers. However, some are not significant simultaneously. For instance, the difference of the dominant indicators (i.e. TFF, WFF, NWFF, STF, WSTF, NWSTF) between the fourth and the fifth circles is not significant, which implies that further theoretical exploration is needed, and we will provide more detailed analysis in Section 4th.

Table 4: The mean value of indicators in different layers

Intimacy	1	2	3	4	5
NWFF	0.92	5.32	23.78	103.01	370.07
WFF	2.97	14.41	70.21	245.27	1054.08
TFF	37.88	60.53	398.81	1454.35	5532.27
RFF	0.0001	0.0018	0.0017	0.005	0.0201
STF	0.5655	1.2386	4.9100	10.6793	13.1091
WSTF	0.5383	1.0753	4.1804	9.4951	12.2046
NWSTF	0.2423	0.7139	2.7707	9.5284	10.7601
TRI	7.66	9.63	12.04	15.2	11.23
MF_8	1.18	2.36	2.44	3.48	3.65
MF_9	1.13	2.05	2.6	2.48	2.58
MT_10	0.28	0.48	0.72	0.86	0.69

Note: 1 represents the outermost layer and the lowest intimacy, 5 represents the innermost layer, i.e. family and pseudo-family members. On a scale of 1 to 5, the degree of tie strength decreases.

Indicators	Regression coefficients	Robust standard errors
NWFF	-0.664***	.039
WFF	0.278*	.001
TFF	0.276***	.000
RFF	0.043	.000
STF	0.322*	1.078
WSTF	0.132	.016
NWSTF	0.020	.013
TRI	0.083***	.009
MF_8	0.079**	.002
MF_9	0.038	.006
MF_10	0.079**	.006

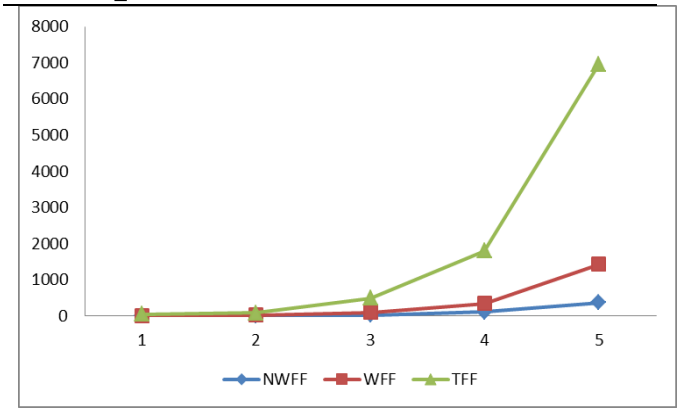

Figure 2: The difference of chat frequency in different layers

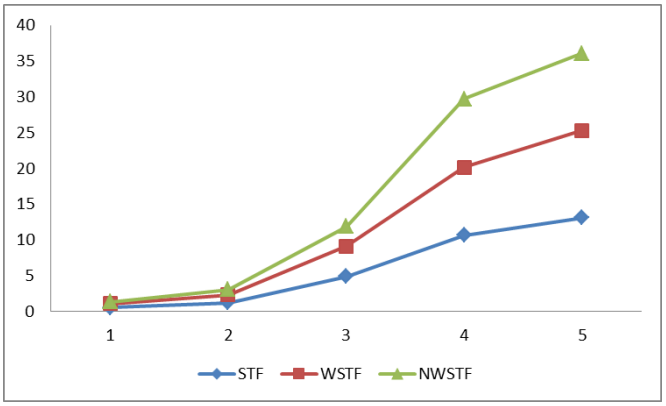

Figure 3: The difference of in different layers

4 Repeating Dunbar's Experiment and Modelling Process

As we mentioned above, Dunbar [17] clustered solely contact frequency of the ties among active users by using Facebook and Twitter data[18]. He respectively used k-means to cluster all ties into different k, ranging from 1 to 20. By calculating the Akaike Information Criterion index (AIC) of the models with varying k, he found that the best k for Facebook data was 5, that collaborates with his hypothesis. However, for Twitter's data, k was 6 in which an additional innermost circle was found.

We first repeat Dunbar's experiment to test the fitness of his method for our data by using K-Means to cluster the chat frequency indicators. One thing is different from him is that we have the label of these tie, which can be used to verify the accuracy of this method. We evaluate the accuracy of this model in two ways. The first one is labeling the clustering groups based on the principle of maximizing the accuracy. Briefly saying, clusters found in data will be defined as certain guanxi categories. According to the label appearing most frequently in one cluster, which will be named after the label. For example, if the label 'family tie' is mentioned 200 times in a cluster of 300 surveyed ties, it will be named after 'family tie' cluster. Then, after all clusters are labeled, we can calculate the accuracy of this model to evaluate the difference between the analytical results of Dunbar's method and the real guanxi categories. According to our analysis, the accuracy can only reach the level of 32%.

We also compute the Normalized Mutual Information (NMI) value to compare the clustering results with the ground truth. That is a method usually to detect the difference between the two clusters.

$$\mathrm{NMI} = \frac{-2\sum_{i=1}^{c_{Cluster}}\sum_{i=1}^{c_{label}} C_{ij}\cdot\log(\frac{C_{ij}\cdot N}{C_i\cdot C_j})}{\sum_{i=1}^{c_{Cluster}} C_i\cdot\log(\frac{C_i}{N}) + \sum_{j=1}^{c_{label}} C_j\cdot\log(\frac{C_j}{N})}$$

where N is the number of ties, C is the confusion matrix, and C_{ij} in the matrix indicates the number that the nodes belongings to the i cluster which get right label as well. NMI ranges from 0 to 1, and the greater is the value, the more accurate is the method. The result shows that the NMI value is 0.0473.

The results above imply that using an unsupervised learning method does not predict the real accurately in our data. Lacking of survey and labels, it is difficult to give an exact definition to each cluster and well explain the clustering result. Thus, an improved model should be built so as to take more on-line interaction features and ground truth of Chinese guanxi categories into consideration.

The differential mode of association theory, proposed by Chinese indigenous sociologist Fei [18], interprets the special characteristics of Chinese social relations. The Guanxi theory following Fei's theory [1, 18, 19] pointed out that there are three social exchange principles among Chinese people, as stated in the section of questionnaire design.

Hwang's theory [1] attempted to explain these rules of social exchange in China, i.e. rules of need, favor exchange and equity. Rule of need can be applied to the most inner circle of a centered ego, for example, family and pseudo-family members. This rule emphasizes the unconditional and mutual supporting of each other. It is very similar to what Dunbar called "support clique". Outside the most-inner circle, there are some especially intimate familiar ties, adopting the rule of favor of exchange to provide emotional support to each other. This is similar to "sympathy group". In the next circle, familiar ties mix expressive and instrumental motivations, which requires both sides of the 'guanxi' to conduct long-term social exchanges in various ways. This is roughly equal to "affinity group". The outer-most two circles are composed of mainly instrumental ties following the rule of equity. One of the two circles has the possibility to develop friendship relations, and one is pure instrument ties.

The differential mode of association theory is the most widely cited theory of Chinese guanxi. Thus, it is necessary to explore the other revised models based on Guanxi theory. Since there is no significant difference between some layers as shown in Figure 2, we try to propose some more flexible ways to categorize the guanxi layers in the 5th section.

5 Revising Model by Theory

Following the discussion in the last chapter, first, we use supervised machine learning methods, including Support Vector Machine (SVM), decision tree, Random Forest and Gradient Boosting Decision Tree (GBDT), to build the classification model according to the Dunbar circle theory, and make a comparison with the K-means clustering analysis.

We divide the test set into 10, 20, 30, and 40 percent of the data and both accuracy and recall rates are reported in Figure 3. The input data is the on-line interaction indicators selected in section 3 and the output are 5 categories according to Dunbar's theory preliminarily. The highest accuracy rate was 54.3% and recall is 32.67%, as shown as Figure 4. The accuracy of the unsupervised learning is 42.34%, which is even smaller than the lowest accuracy, 46.51%, in all supervised Machine Learning methods.

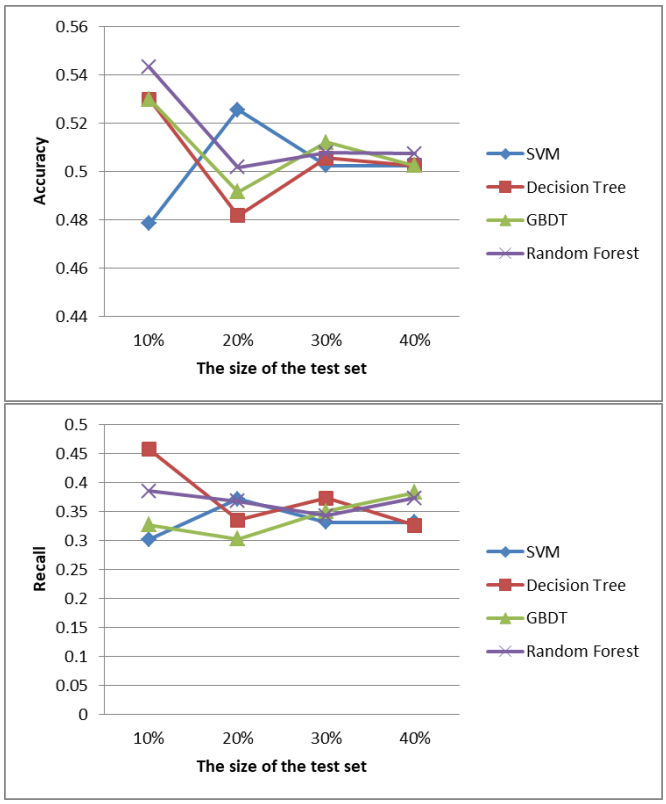

Figure 4: Accuracy and recall of the five-layer mode

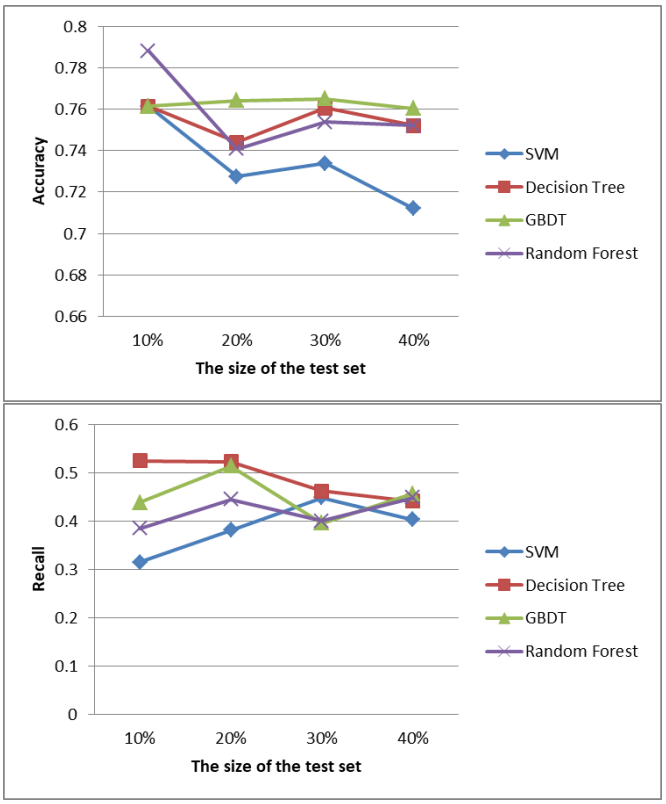

Figure 5: Accuracy and recall of the three-layer model

Although the five-layer supervised model has a better performance than Dunbar's unsupervised method, there is large room for improvement.

According to Guanxi theory stated in the last chapter, we may propose a three-layer model to present the three principles of Chinese social exchange. In the new model, we keep the layer of family ties, merges intimate friends and familiar ties into one category, and potential friends and acquaintances into one category as instrumental ties. The highest accuracy rate was 78.8% and recall rate is 38.49%, as shown as Figure 5. For the five-layer model, the accuracy of random guest is 20%, and for the three-layer model, it is 33%. However, the improved accuracy between two models can reach to 13% (0.458-0.330), as shown in Table 5. Thus, the three-layer model displays good performance in computing tie strength in QQ. It suggests that the categorization of guanxi according to Hwang's social exchange principles has the higher explanatory power than 5-layer model.

Considering the indicators' huge mean difference shown in Figure 2, we can't ignore the different interactive patterns between the intimate friendships and general familiar ties. This study thus computes sum of squared errors (SSE) of K-means clustering under different k, the Δs (the difference of SSE) between k=3 and k=4 is much bigger than the others, as shown in Figure 6. That leads us to make further exploration to verify whether there is a four-layer model.

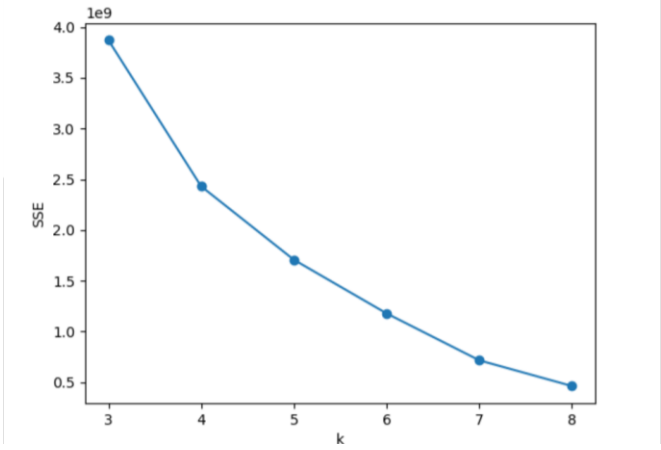

Figure 6: K-means clustering the sum of squared errors of different k

Then, based on the three-layer model, we divide the familiar ties into the intimate friends and general friends, and propose a four-layer supervised model. Again, different machine learning methods are used to calculate their accuracy and recall rates (shown in Figure 7). The highest accuracy rate is 77.48% and recall is 42.92%. Similarly, we also make a comparison with the five-layer and three-layer models.

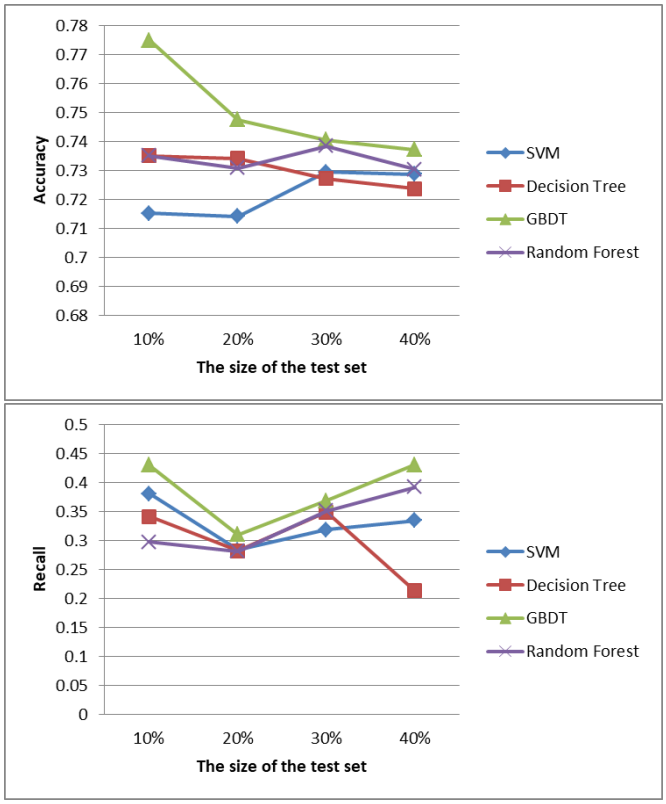

Figure 7: Accuracy and recall of the four-layer model

Comparing with the five-layer model, the accuracy and the recall of the four-layer model shows significant improvement. For five-layer model, the accuracy of random guest is 20%, and for the four-layer model, it is 25%. The improved accuracy between the two models can be about 20% (0.525-0.330), as shown in Table 5. Thus, the four-layer model displays a much better performance in predicting tie strength.

In addition, comparing with the three-layer model, the accuracy and the recall of the four-layer model also shows better results. For three-layer model, the accuracy of random guest is 33%, and for the four-layer model, it is 25%. The improved accuracy between two models is about 7% (0.525-0.458). It shows that the accuracy indeed increases.

We also try to compare other four-layer classification with the baseline model. For instance, we merge the family/pseudo-family members and intimate friends into one category. However, the others show worse performance than the baseline model. Also, the baseline four-layer model stated above displays better performance in computing tie strength than five-layer and three-layer models

We may conclude that it is more suitable to divide the Chinese large cities' 18-28 young QQ users into four categories in the Chinese guanxi context.

One more interesting finding is that we can predict accurately the family members who are often in the same position, which can be called the "family members living together". This finding collaborates with Dunbar's new discovery about the inner circle based on Twitter's data [7].

In the baseline 4-layer model, we rank the importance of on-line interaction features and show the rank order as TFF, RFF, WFF, NWFF and TRI.

Table 5: The comparison of the accuracy of different model

	five-layer model	three-layer model	four-layer model
Random prediction	0.2	0.33	0.25
Supervised learning	0.5298	0.788	0.7748
Improved accuracy	0.3298	0.458	0.5248

6 Discussions and Future Work

Inspired by Dunbar's study, this research tries to propose a classification model that can categorize Chinese guanxi by using the big data analysis of QQ users. We first follow the Chinese social interaction principles [19, 23, 24] to design our questionnaire, so that we employ social survey to collect labeled categories of guanxi as the ground truth. Then, various types of on-line interaction data are collected and various supervised machine learning methods are adopted to do the big-data analysis. In the next, we build various types of 3-layer, 4-layer and 5-layer models and test them against the ground truth to get their predictive power. Finally, in the comparison of the accuracy rates of all models, we select out the best model, learning method and on-line interaction features.

Our study on the dataset from QQ users found some new evidence for the Chinese guanxi categories. First, the boundaries among the five circles based on Dunbar's theory are not distinguishable in the Chinese context. Second, Hwang's guanxi theory tell us the three social exchange principles, but we still don't know how many types of guanxi a Chinese has. In this study, several runs of big data analyses reveal that four layers of guanxi may be clearly marked out. For our typical cases, Chinese large cities' young people, four types of guanxi can have a better explanatory power in predicting tie strength. There are two types of familiar ties, one is especially intimate friends, and another one is general familiar ties conducting long-term and wide-range favor exchanges.

However, this point is not the end of our study. There are still some ways to improve the accuracy rate of predictive models. For example, we may collect more labeled guanxi from interviewees and their contacts as ground truth, obtain longer time on-line interaction data so that we can compute the dynamic change of these guanxi, and so on. So far, we are not able to exclude the possibility that four-layer model may be out-competed by other models. Furthermore, we sample our cases in a small area and narrow age range, so we must be cautious to generate this 4-layer model to other Chinese social context.

The employment of on-line interaction data is indeed very helpful for us in studying a challenging research topic. In the several runs of big-data analysis, dialogue between theory and analytical results and the construction of predictive models, we get higher and higher accuracy rates in predicting tie strength. Combing survey data and on-line interaction big data especially provides fruitful research results for our study. Although this study is only a beginning, the combined methods of survey and big-data analysis under the guidance of social science theory provide us with a bright road in the future studies.

REFERENCES

[1] K. Hwang (1987). Face and Favor: The Chinese Power Game. *American Journal of Sociology*, 92(4), 944-974.

[2] M. S. Granovetter (1973). The strength of weak ties. *American journal of sociology*, 78(6), 1360-1380.

[3] M. S. Granovetter (2005). The impact of social structure on economic outcomes. *The Journal of economic perspectives*, 19(1), 33-50.

[4] P. V. Marsden, K. E. Campbell (1984). Measuring tie strength. *Social Forces*, 63(2), 482-501.

[5] N. Eagle, A. S. Pentland, D. Lazer (2009). Inferring friendship network structure by using mobile phone data. *Proceedings of the national academy of sciences*, 106(36), 15274-15278.

[6] S. G. B. Roberts, R. I. M. Dunbar (2011). Communication in social networks: Effects of kinship, network size, and emotional closeness. *Personal Relationships*, 18(3), 439-452.

[7] T. Raeder, O. Lizardo, D. Hachen, N. V. Chawla (2011). Predictors of short-term decay of cell phone contacts in a large scale communication network. *Social Networks*, 33(4), 245-257.

[8] V. Palchykov, K. Kaski, J. Kertesz, R. I. Dunbar (2012). Sex differences in intimate relationships. *Scientific Reports*, 2(1), 370-370.

[9] M. I. Akbas, R. N. Avula, M. A. Bassiouni, D. Turgut (2013). Social network generation and friend ranking based on mobile phone data. *Proceedings of International Conference on Communications*. IEEE, Budapest, Hungary, 1444-1448.

[10] V. Arnaboldi, A. Guazzini, A. Passarella (2013). Egocentric online social networks: Analysis of key features and prediction of tie strength in Facebook. *Computer Communications*, 36(10), 1130-1144.

[11] R. I. M. Dunbar (1992). Neocortex size as a constraint on group size in primates. Journal of human evolution, 22(6), 469-493.

[12] R. I. M. Dunbar (1993). Coevolution of neocortical size, group size and language in humans. *Behavioral and brain sciences*, 16(04), 681-694.

[13] R. I. M. Dunbar, M. Spoors (1995). Social networks, support cliques, and kinship. *Human Nature*, 6(3), 273-290.

[14] R. A. Hill, R. I. M. Dunbar (2003). Social network size in humans. *Human nature*, 14(1), 53-72.

[15] W. X. Zhou, D. Sornette, R. A. Hill, R. I. M. Dunbar (2005). Discrete hierarchical organization of social group sizes. *Proceedings of the Royal Society of London B: Biological Sciences*, 272(1561), 439-444.

[16] T. V. Pollet, S. G. B. Roberts, R. I. M. Dunbar (2011). Use of social network sites and instant messaging does not lead to increased offline social network size, or to emotionally closer relationships with offline network members. *Cyberpsychology, Behavior, and Social Networking*, 14(4), 253-258.

[17] S. G. B. Roberts, R. I. M. Dunbar, T. V. Pollet, T. Kuppens (2012). Exploring variation in active network size: Constraints and ego characteristics. *Social Networks*, 31(2), 138-146.

[18] X. Fei (1998), From the Soil, the Foundations of Chinese Society. Beijing university press, Beijing, China.

[19] G.Yang (2004). Chinese psychology and behavior: a study of localization. Renmin university of China press, Beijing, China.

[20] R. I. M. Dunbar, V. Arnaboldi, M. Conti, A. Passarella (2015). The structure of online social networks mirrors those in the offline world. Social Networks, 43: 39-47.