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

Xin Gao  
*Department of Sociology, Tsinghua University, Beijing 100084, China*

Jar-Der Luo  
*Department of Sociology, Tsinghua University, Beijing 100084, China*

Kunhao Yang  
*Graduate School of Arts and Sciences, University of Tokyo, Tokyo 153-8902, Japan*

Xiaoming Fu  
*Institute of Computer Science, University of Göttingen, Göttingen 37077, Germany*

Loring Liu  
*Tencent Computer System Co. Ltd, Shenzhen 518000, China*

See next page for additional authors

Follow this and additional works at: [https://tsinghuauniversitypress.researchcommons.org/journal-of-social-computing](https://tsinghuauniversitypress.researchcommons.org/journal-of-social-computing)

Part of the Computer Engineering Commons, Computer Sciences Commons, and the Science and Technology Studies Commons

---

**Recommended Citation**  
Xin Gao, Jar-Der Luo, Kunhao Yang et al. Predicting Tie Strength of Chinese Guanxi by Using Big Data of Social Networks. Journal of Social Computing 2020, 1(1): 40-52.

This Research Article is brought to you for free and open access by Tsinghua University Press: Journals Publishing. It has been accepted for inclusion in Journal of Social Computing by an authorized editor of Tsinghua University Press: Journals Publishing.
Predicting Tie Strength of Chinese Guanxi by Using Big Data of Social Networks

Authors
Xin Gao, Jar-Der Luo, Kunhao Yang, Xiaoming Fu, Loring Liu, and Weiwei Gu

This research article is available in Journal of Social Computing:
https://tsinghuauniversitypress.researchcommons.org/journal-of-social-computing/vol1/iss1/4
Predicting Tie Strength of Chinese Guanxi by Using Big Data of Social Networks

Xin Gao, Jar-Der Luo*, Kunhao Yang, Xiaoming Fu, Loring Liu, and Weiwei Gu

Abstract: This paper poses a question: How many types of social relations can be categorized in the Chinese context? 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 research suggests that Granovetter measures tie strength so as to distinguish strong ties from weak ties, and the Dunbar circle theory may offer a plausible approach to calculating 5 types of relations according to interaction frequency via unsupervised learning (e.g., clustering interactive data between users in Facebook and Twitter). In this paper, we differentiate the layers of an ego-centered network by measuring the different dimensions of user’s online interaction data based on the Dunbar circle theory. To label the types of Chinese guanxi, we conduct a survey to collect the ground truth from the real world and link this survey data to big data collected from a widely used social network platform in China. After repeating the Dunbar experiments, we modify our computing methods and indicators computed from big data in order to have a model best fit for the ground truth. At the same time, a comprehensive set of effective predictors are selected to have a dialogue with existing theories of tie strength. Eventually, by combining Guanxi theory with Dunbar circle studies, four types of guanxi are found to represent a four-layer model of a Chinese ego-centered network.

Key words: tie strength; Dunbar circle theory; Chinese Guanxi theory; supervised classification model; social network

1 Introduction and Question Definition

How many types of social relations or in the Chinese term—guanxi, can a Chinese person be categorized?

* Xin Gao and Jar-Der Luo are with Department of Sociology, Tsinghua University, Beijing 100084, China. E-mail: gaoxinxg@126.com; jdluo@tsinghua.edu.cn.

* Kunhao Yang is with the Graduate School of Arts and Sciences, University of Tokyo, Tokyo 153-8902, Japan. E-mail: yang-kunhao@g.ecc.u-tokyo.ac.jp.

* Xiaoming Fu is with Institute of Computer Science, University of Göttingen, Göttingen 37077, Germany. E-mail: fu@cs.uni-goettingen.de.

* Loring Liu is with Tencent Computer System Co. Ltd, Shenzhen 518000, China. E-mail: loringliu@tencent.com.

* Weiwei Gu is with Information Science and Technology, Beijing University of Chemical Technology, Beijing 100029, China. E-mail: guweiwei@mail.bnu.edu.cn.

To whom correspondence should be addressed.

Manuscript received: 2020-05-31; accepted: 2020-07-05

Guanxi theory proposed three principles for the Chinese social interactions, i.e., rules of needs, favor exchanges, and equity[1]. However, no quantitative studies have been conducted based on these social-exchange principles to categorize the types of guanxi that actually exist. Big data helps to address this challenging question. This paper tries to combine the surveyed data with big data collected from Software A—one of the most widely used social-network platforms in China.

Reviewing social network theory, Granovetter classified tie strength into two categories including strong and weak ties[2,3]. Foremost, his studies on weak ties, which can bring heterogeneity information and opportunity, have generated significant impact and have been widely applied in a variety of areas. In addition, Granovetter[2] 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 strong and weak ties. Based on his work, many researchers developed practicable methods to measure tie strength\(^{[4-10]}\). However, most of the follow-up works focused on the indicators which were highly correlated with tie strength respecting continuum of intimacy, interaction frequency, reciprocity, and friendship duration, rather than distinguishing strong ties from weak ties.

It is also well known that the Dunbar circle theory\(^{[6, 11-17]}\) suggested a five-category model of social relations as a plausible solution for measuring tie strength, which defined five specific circles, having a clear boundary between two contiguous circles. In contrast to the Western counterparts, within the Chinese social context, “a relationship is a guanxi tie to the extent that trust is high and relatively independent of social structure around the relationship”\(^{[18]}\). The trust in guanxi can be built on family ethics\(^{[19]}\), favor exchange\(^{[1,20]}\), and mutual obligation\(^{[21,22]}\). Liang\(^{[19]}\) argued that Chinese society is “family-ethics based”, since the social-exchange principles in guanxi were transformed from family ethics and could be applied to persons outside family. Fei\(^{[23]}\) referred this phenomenon as “the differential modes of association”, that is, a Chinese divides his/her ego-centered network into several circles and applies different social-exchange principles to the contacts in various circles. The requirement of family ethics decreases from inner to outside circles. In addition, the closer a person to the centered ego is, the more independent their guanxi to network structure is. In other words, the impacts of structural variables, such as group closure, network density, and number of common friends, etc., decrease when the tie strength of guanxi increases.

Then how many circles a Chinese can recognize in his/her ego-centered social network? Yang\(^{[20]}\) and Hwang\(^{[1]}\) categorized Chinese tie strength (hereafter, known as guanxi) into three types of social-exchange principles to describe the social relationships of the traditional society in China. Hwang\(^{[1]}\) proposed different behavioral principles for the three types of Chinese guanxi including (1) rules of need—a kind of behavioral principle used between family or pseudo-family members, (2) rules of favor exchange—being kept between familiar ties, and (3) rules of equity—being adopted between acquaintance ties.

Compared with the five circles in Dunbar circle theory, the layers of guanxi among Chinese people, to some extent, may not have such clear boundaries between the two adjacent relationships. Thus, a challenging question emerges: how to compute tie strength between individuals in the Chinese context by using social network data? In addition, which indicators computed from online data can be used to classify Chinese tie strength?

This work attempts to establish a model to calculate tie strength between individuals by inputting online interaction data from a social network platform, and then outputting a prediction model to categorize Chinese guanxi. The model will be tested against the ground truth collected from social surveys. The input data coming from one of the most populous online social network platforms in China, called Software A, have a large number of active users, detailed functions, and multiple network footprints of users, and contain records of a long period. Thus, it is a sufficient and suitable dataset to explore the social relationships of today’s Chinese people.

In this study, our main contributions include (1) computing specific indicators of online interaction for categorizing Chinese guanxi; (2) inspired by Dunbar circle theory, proposing a classification model that can categorize Chinese guanxi into distinct categories in a quantitative way; (3) modifying the Dunbar circle theory in the Chinese context. Through theoretical exploration based on Guanxi theory and mining results of online social network data, a contextual four-layer model is proposed, which has good predictive power in computing the Chinese tie strength. (4) The methodology applied in this paper can be extended to many other research areas. Based on social science theory and social surveys, our interviewees firstly label the types of their guanxi as the ground truth, which can be used to test the accuracy of the following classification models.

2 Dunbar Circle Theory and Chinese Guanxi Theory

In previous research, the Dunbar circle theory\(^{[6,11-17]}\) was proposed to measure 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. Different interaction logics and functions exist in various circles (see Fig. 1).

In a recent study, Dunbar et al. [16] used Facebook and Twitter data to verify their theory within the arena of online social network data. They used one dimension of interaction input—chatting frequency, and clustered the indicators of active users and their active contacts on Facebook and Twitter (Facebook and Twitter friends) [24]. The study showed that for the two Facebook datasets, they were best described by a four-layer structure, and for the Twitter dataset, a five-layer structure was more suitable [16]. According to their theory of the number of ties in each category, the clusters whose number of ties is roughly matched are labeled. For example, the study found that each Facebook user had an average of 5.28 friends in the emotional support group. However, community and tribe groups could not be well found in this study. It seems that this big-data analytical result did not completely confirm Dunbar circle theory. In addition, without an actual label, it is risky to categorize a tie into one specific group. For example, a tie may be regarded as a support clique based solely on the high contact frequency between two people, but actually they only interact with each other so frequently because they are part of the same workplace.

The purpose of this study is as follows: (1) make a localized explanation of Dunbar’s five types of relationship in the Chinese context and design a questionnaire to collect ground truth, i.e., to define and collect the labels of real guanxi. (2) Replicate Dunbar’s experiment to determine its validity in the Chinese context by using online interaction data of the interviewees from a widely-used social network platform in China. (3) Employ the Chinese Guanxi theory and online interactive indicators to revise the model.

3 Defining Tie Strength and Ground Truth

3.1 Labeling guanxi in survey

Considering the difference between distinct cultural contexts, five types of tie strengths need to be redefined, which can not only enable Dunbar and his colleague’s study to be replicated, but also explore the model based on Chinese Guanxi theory. However, Hwang’s theory about “favor, guanxi, and face” [1] indicates there are only three principles refering to Chinese social interactions, rather than a clear boundary between various types of guanxi. According to the classification of Dunbar circle theory, we apply the three principles for 5 types of guanxi, which are: (1) family ties including real and pseudo-family members, good fit for rule of need, which are roughly equal to support clique. (2) Intimate familiar ties containing very strong friendship, suitable for rule of favor exchange, roughly equal 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 are roughly equal to affinity group. (4) Potential “friends”, the ties that are mainly built on instrumental purposes but with the potential to be friends, roughly similar to active networks. (5) Acquaintance ties, purely instrumental ties, follow the rule of equity. The design of these labels also aims to make a comparison of the three Chinese categories, respecting family ties, familiar ties, and acquaintance ties, based on the Chinese Guanxi theory [25].

To obtain the ground truth of these categories of guanxi, a survey was conducted in China’s major first-tier cities, including Beijing, Shanghai, and Guangzhou from May 1 to June 1, 2018. The snowball sampling method was used. Samples generally were within the age range of 18–28 years old and data were obtained via face-to-face interviews. After ruling out the cases under 18 and above 28, we take the survey as a typical sample of young people living in large Chinese cities, which fits with the age distribution of users of Software A.

Since the number of friends a person maintains 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 Software A 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.

Questions concerning the five types of guanxi are as follows:

- Please list at least three individuals who 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.
- Please list at least five individuals who have long-term friendships with favorable exchanges.
- Please list at least five individuals who are not very close to you right currently, but you are likely to build friendships and favor exchange relations in the future.
- Please list at least eight individuals who you have acquaintances with, and who you may or may not contact in the future.

In the face-to-face interviews, interviewers explained the meaning of these five ties to interviewees. In addition to the type of tie strength between the respondents and their friends, we also asked the Software A number of respondents and their contacts mentioned in the social network platform.

A total of 2012 ties were collected in the survey. We then selected the ties with active users on both ends—the same data processing as used by Dunbar et al.\textsuperscript{[16]} After this, there were 1502 labeled ties remaining. The number of labeled ties from the innermost layer to the outermost layer are 118, 147, 223, 349, and 665, respectively.

### 3.2 Online interaction data

As mentioned above, the online interaction data are collected from one of the most populous online social network platforms, Software A (hereafter called Software A data). With over 800 million monthly active users in June 30, 2018, Software A provides a qualified dataset for us to explore the social relationships among Chinese people, especially young Chinese people living in large cities.

We collected the online interaction data through cooperation with Tencent. During the entire research process, we strictly complied with privacy rules, particularly by collection of users’ online interaction information only after obtaining permission from the users. After searching for data, the data were made anonymous by Tencent’s data collectors, after that our researchers analyzed the anonymous data.

Previous studies have developed an increasing number of indicators to classify and predict tie strength.\textsuperscript{[3, 4, 7, 9, 16, 17]} Granovetter\textsuperscript{[13]} proposed that continuum of intimacy, interaction frequency, reciprocity, and friendship duration highly correlated with tie strength. From the perspective of the network structure, Burt\textsuperscript{[26]} suggested that relationship structure shaped tie strength between egos. Lin et al.\textsuperscript{[27]} created a micro-level social capital theory that further pointed out how similarities in terms of age, gender, and education level between two people affected the tie strength.\textsuperscript{[13]} Therefore, it is necessary to indentify the relevant big data indicators based on above-mentioned theories, which can be used for computing tie strength. 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 in Table 1, in which all indicators are defined). Then we also compute Pearson correlation between the predictors and tie strength to select the preliminary features for our prediction models.

Firstly, chatting frequency is measured by counting the frequency of chatting between two persons who add each other as “friends”, by which we can obtain each interviewee’s ego-centered friend network. The results of Pearson correlation analysis show that total chatting frequency is highly correlated with tie strength (0.209***), which is indicated by the five categories of guanxi in our study. Chatting frequency is also the sole indicator used by Dunbar et al.\textsuperscript{[24]} in their clustering model. Our data show that there is high correlation between the total number of messages from user $i$ (the interviewee) to user $j$ (the tie selected by the interviewee) and that from user $j$ to user $i$. Thus, we merely compute the undirected chatting frequency from user $i$ to user $j$. Concerning that the difference interaction pattern of each interviewee and his/her total number of messages vary largely, the relative chatting frequency should be considered. That is, the number of messages from user $i$ to user $j$ is divided by the total number of messages from user $i$ during the period $T_0 - T_1$, and this variable also shows high correlation with tie strength (0.138***). Since standard deviation of chatting frequency during working and non-working time is important indicator to distinguish whether the two ends of a tie have a working relationship, we split the variables into working days and non-working days and the results show a high correlation with tie strength (0.200*** and 0.206***). Furthermore, the standard deviation of chatting frequency (0.384***), also positively correlates with the strength of ties. The correlation diagram of these variables is shown in Fig. 2.

Secondly, friendship duration, which can be characterized by duration that the two individuals have been friends on Software A, demonstrates no significant correlation with tie strength in our data.

Thirdly, we include the frequency of user $i$ giving
Table 1  Explanations and definition of indicators.

| Theoretical meaning                  | Indicator aspect                  | Indicator definition                                                                 | Code   |
|--------------------------------------|-----------------------------------|--------------------------------------------------------------------------------------|--------|
| Chatting frequency                   | Peer-to-peer interaction indicator| Relative chatting frequency                                                         | rff    |
|                                      |                                   | Total chatting frequency                                                             | tff    |
|                                      |                                   | Total chatting frequency during working time                                         | wff    |
|                                      |                                   | Total chatting frequency during non-working time                                      | nwff   |
|                                      |                                   | Standard deviation of chat frequency                                                | stf    |
|                                      |                                   | Standard deviation of chatting frequency during working time                          | wstf   |
|                                      |                                   | Standard deviation of chatting frequency during non-working time                      | nwstf  |
| Friendship duration                  | Period of adding a friend          | Friendship duration                                                                  | fd     |
| Reciprocity                          | Gift-giving record                 | Frequency of giving monetary gift                                                   | mg     |
|                                      |                                   | Frequency of giving online-service gift                                             | sg     |
| Intimacy                             | GPS record                         | Meeting frequency in August                                                          | mf_8   |
|                                      |                                   | Meeting frequency in September                                                       | mf_9   |
|                                      |                                   | Meeting frequency during the national anniversary of China                            | mf_10  |
| Friend’s labelling term              | Intimate term to describe a friend |                                                                                      | FLT    |
| Network structural variable          | Social network structure           | Number of common friends                                                             | CF     |
|                                      |                                   | Number of common groups                                                              | CG     |
| Similarity                           | Identity recognition               | Gender similarity                                                                     | GS     |
|                                      |                                   | Age similarity                                                                        | AS     |
|                                      |                                   | Working industry similarity                                                           | WIS    |
| Other variable                       | GPS record at 00:00–05:00           | Living together or not                                                                | LT     |

Gifts, such as Software A dollar and online services, to user j in our model. This variable can be treated as a type of reciprocity.

Fourthly, two types of intimacy are included, i.e., face-to-face meeting frequency and friends labeled by interviewees, including key words as “family”, “friends”, and “colleagues”. Offline meeting frequency is also recorded. This variable is computed by the GPS information of users i and j in a three-month period. If both of them are located within a certain distance at the same time, they are recorded as having had a face-to-face meeting. Three-period records are included, i.e., August (summer holidays for students), September, and October (including the longer public holiday for the Chinese national anniversary). The records are significantly correlated to tie strength (0.160***, 0.106*, and 0.150***).

We use nature language processing software to extract keywords from Software A users’ notes (a small column that a user can describe his/her contacts) about his/her friends. In total, 96 different words were extracted and their weights of intimacy were generated in the analysis of ego-contact chatting frequency. The correlation coefficient of this variable with the tie strength is 0.25***.

Fifthly, structural indicators, such as the number of common friends, indicated by common neighbors in friend network and common Software A groups, are also important for identifying guanxi categories. The variable common friends are high correlated with tie strength (0.132***), while the analysis of the number of common groups reveals a contradictory result.

Sixthly, the similarities between two friends’ age, gender, occupation, and other such factors are provided by the social network platform. We can then identify whether the two persons are living together or not by comparing the GPS information between 00:00 and 5:00 over a certain period of time. This variable is highly correlated with tie strength (0.29***). In the process of data mining, we find that a big difference exists in the online interaction pattern between the family members.
who live together and those who do not, and thus, we add this variable in our model.

Apart from the GPS information that can only be traced back to the most recent three months, other indicators are calculated based on a full year of data, from July 2017 to June 2018. Thus, face-to-face meeting frequency is calculated through the data spanning from August 2018 to October 2018. Those indicators with a high-proportion of missing values are ruled out in the further analysis process, such as the frequency of giving gifts and the number of common groups. Matching the survey data (labeled guanxi categories as the ground truth of tie strength) and the Software A data (online interaction variables), a dataset with 1502 labeled ties is thus generated. As stated above, under the guidance of tie-strength, network structure, and social capital theories, the indicators are computed from online big data, and then those significantly associated with tie strength are analyzed and included as features in our prediction models.

The first required exploration is a preliminary list of indicators’ contributions and their theoretical meaning. Thus, we run the multivariate linear regression of these indicators on tie strength (shown in Table 2). Since there is a multicollinearity problem between the peer-to-peer interaction indicators (see Fig. 2.), we use the Standard Deviation (SD) of chatting frequency during working time (wstf) instead of SD of chatting frequency. In addition, chatting frequency during non-working time (nwff) can be used to replace total chatting frequency, so that we eliminate collinearity and preserve the heterogeneity between working time and non-working time. Finally, the variable relative chatting frequency (rff) is kept in the model as it has a low correlation with other variables.

According to the regression results, the peer-to-peer interaction indicators are the most important in terms of computing tie strength. In addition, there is different effect between the working time and non-working time variables.

As for the variables of intimacy, apart from offline meeting frequency in September (mf.9), the offline meeting frequency in August contributes more on tie strength than the national anniversary holiday period (mf.8 > mf.10). In addition, Friend’s Labelling Term (FLT) is highly associated with measuring tie strength. From the perspective of network structure, the number of Common Friends (CF) has a significant impact on guanxi categories, while similarities in gender, age, and occupation show no significant contribution to guanxi computation. We have established a set of valid indicators collected from online interaction data and used them as the input in the prediction models.

In the process of building a prediction model, we find that Living Together (LT) is important to some types of guanxi in China. We thus explore the two-way interaction effect of the LT with the other variables to identify whether there are different interaction patterns between people who live together and those who do not. The results confirm our assumptions (shown in Table 3). Interaction effects indeed exist between the variables, such as chatting frequency, intimacy, and living together.

Furthermore, we also explore the differences of these online interaction indicators among various layers (shown in Table 4). We pay attention to the mean value of the most important indicators (high significance) and show the mean variation trend in Fig. 3. The statistics show that there is indeed a huge difference among various guanxi categories. However, the mean difference in some layers is notably different, suggesting it is

| Table 2 | Tie strength regressing on online interaction variables. |
|---------|---------------------------------------------------------|
| Model   | Chatting frequency | Intimacy | Structure: | Similarity | R² | Adjusted R² |
|         | wstf        | nwff      | rff        | mf.8 | mf.9 | mf.10 | FLT     | CF | GS | AS | WIS |         |
| 1       | 0.452***   | -0.131*** | 0.083***   |       |      |       |         |    |    |    |     | 0.1509  |
| 2       | 0.394***   | -0.157*** | 0.177***   | 0.120*** | 0.019 | 0.073* | 0.213*** |    |    |    |     | 0.2518  |
| 3       | 0.390***   | -0.158*** | 0.173***   | 0.112*** | 0.029 | 0.071* | 0.208*** | 0.086** |    |    |     | 0.2590  |
| 4       | 0.423***   | -0.172*** | 0.276***   | 0.179*** | 0.064 | 0.013 | 0.207*** | 0.099 | -0.033 | 0.122 | -0.142 | 0.3531  |

| Table 3 | Interaction multivariate linear regression. |
|---------|--------------------------------------------|
| Model   | Chatting frequency | Intimacy | Structure: CF | Other: LT | Interaction item | R² | Adjusted R² |
|         | wstf | nwff | rff | mf.8 | mf.9 | mf.10 | FLT | LT×rff | LT×wstf | LT×nwff | LT×FLT |         |
| 1       | 0.355*** | -0.177*** | 0.183*** | 0.049 | -0.004 | 0.074* | 0.199*** | 0.094*** | 0.248*** |         | 0.3118  | 0.3060  |
| 2       | 0.337*** | -0.112*** | 0.603*** | 0.046 | 0.003 | 0.068* | 0.196*** | 0.095*** | 0.256*** | -0.424* | 0.3149  | 0.3085  |
| 3       | 0.650*** | -0.104* | 0.171*** | 0.046 | 0.0007 | 0.067* | 0.197*** | 0.094*** | 0.297*** | -0.289*** | 0.3216  | 0.3152  |
| 4       | 0.336*** | 0.023*** | 0.177*** | 0.046 | 0.003 | 0.068* | 0.195*** | 0.095*** | 0.263*** | -0.127* | 0.3154  | 0.3090  |
| 5       | 0.363*** | -0.186*** | 0.182*** | 0.049 | -0.005 | 0.073* | 0.500*** | 0.090*** | 0.263*** | -0.288* | 0.3152  | 0.3087  |
reasonable to categorize the ties into different layers and some are not significant simultaneously. For instance, the difference of the indicators, which contributes dominantly to identify tie strength (i.e., tff, nwff, wstf, and FLT) between the fourth and the fifth circles, is not as significant as others. This implies that further theoretical exploration and revised classification model are needed. More detailed analysis will be provided in Section 4.

To briefly summarize, concerning the indicators to measure tie strength computed from online social interactions, the regression results in this paper show that: (1) The indicators of chatting frequency, intimacy, and network structural variables have a significant impact on classifying guanxi, as existing theories have shown. Furthermore, we also find that it is necessary to distinguish working time and non-working time for chatting frequency. Although the similarities of age, gender, and occupation have no significant impact on tie strength in regression, the accuracy can be improved when adding them into the following prediction model, as it is very important to identify the respondents of the most intimate friends, especially for young people. (2) Living together significantly moderates the effects that chatting frequency and intimacy have on tie strength. Similar to the variable similarity, living together is important for identifying family members, which will be shown in the following process of building prediction models.

4 Repeating Dunbar’s Experiment and Modelling Process

As we mentioned above, Dunbar et al.\cite{17} clustered sole contact frequency of the ties among active users by using Facebook and Twitter data\cite{18}. They 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 $k$, it was found that the best $k$ for Facebook data was 4 and for Twitter was 5. The latter result collaborated with their hypothesis.

In our research, we firstly repeat Dunbar’s experiment to test the appropriateness of the method for our data by using K-Means to cluster the chatting frequency indicators. Unlike Dunbar, for our data, we have the labels of these ties, which can be used to verify the actual accuracy of this method. The accuracy of this model is evaluated in two ways. The first way concerns labeling the clustering groups based on the principle of maximizing accuracy. In brief, clusters found in data will be defined as certain guanxi categories. According to the label appearing most frequently in one cluster, the label will be attached to the cluster, which ensures the maximum overall accuracy. For example, if the label “family tie” is mentioned 200 times in a cluster of 300 surveyed ties, the cluster will be named “family

Table 4 Differences of online interaction indicators among five types of guanxi layer.

| Layer | tff   | nwff  | nwstf | FLT   | mf_8  | CF   |
|-------|-------|-------|-------|-------|-------|------|
| 1     | 39.4  | 0.962 | 0.561 | 1.53  | 1.05  | 6.78 |
| 2     | 63.4  | 5.57  | 1.13  | 1.67  | 2.21  | 8.75 |
| 3     | 415   | 24.9  | 4.29  | 1.93  | 2.11  | 11.1 |
| 4     | 1524  | 106   | 9.26  | 2.03  | 3.17  | 14.1 |
| 5     | 5704  | 382   | 12.5  | 2.4   | 3.72  | 10.7 |

Note: On a scale of 1–5, the degree of tie strength decreases, with 1 being the lowest degree and 5 the highest.
tie” cluster. Then once 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 period accuracy can only reach a level of 32%.

Another method to evaluate the accuracy is computing the Normalized Mutual Information (NMI) value to compare the clustering results with the ground truth. This is a method usually used to detect the difference between the two clusters. NMI ranges from 0 to 1, and the greater the value, the more accurate the method. The result shows that the NMI value is only 0.0473.

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

The differential mode of association theory, proposed by Chinese sociologist Fei[23], interprets the special characteristics of Chinese social relations. According to Fei’s theory[23], the Guanxi theory[1, 18, 19] posits that there are three main social exchange principles among Chinese people (as stated in the questionnaire design section).

Hwang’s theory[11] attempted to explain these rules of social exchange in China, i.e., rules of need, favor exchange, and equity. Rules of need can be applied to the most inner circle of a centered ego, e.g., family and pseudo-family members. This rule emphasizes unconditional and mutual support. It is akin 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 for each other. This is similar to the “sympathy group”. In the next circle, familiar ties mix expressive and instrumental motivations, which require both sides of the guanxi to conduct long-term social exchanges in various ways. This is roughly equal to the “affinity group”. The outermost 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 tie.

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 seen in Fig. 3, we try to propose some more flexible ways to categorize the guanxi layers in Section 5 of this paper.

5 Revising Model by Theory

Following the discussion in the last chapter, firstly, 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% of the data and both accuracy and recall are reported in Fig. 4. The input data are the online interaction indicators selected in Section 3 and the output data are the 5 categories according to Dunbar’s theory preliminarily. The highest accuracy is 0.543 and recall is 0.3267, as shown in Fig. 4.

The accuracy of the unsupervised learning model by clustering all the variables is 0.4234, which is even smaller than the lowest accuracy, 0.4651, in all supervised machine learning methods. Although the five-layer supervised model has a better performance than Dunbar’s unsupervised model, there remains large room for improvement.

According to stated Guanxi theory, we can 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 — merging intimate friends and familiar ties into one category, and potential friends and acquaintances into one category as instrumental ties. The highest accuracy is 0.788 and recall is 0.3849, as shown in Fig. 5. For the five-layer model, the accuracy of random guest is 0.2, and for the three-layer model, it is 0.33. However, the improved accuracy between the two models can reach to about 0.13 (0.458–0.330), as shown in Table 5. Thus, the three-layer model displays good performance in computing tie strength in Software A. It suggests that the categorization of guanxi according to Hwang’s social exchange principles[11] has the higher explanatory ability than that of the 5-layer model.

Considering the indicators’ huge mean difference (Fig. 2), we cannot 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, and the Δs (the difference of SSE) between k=3 and k=4 is
accuracy and recall of the five-layer model.

Fig. 4 Accuracy and recall of the five-layer model.

Table 5 Comparison of the highest accuracy of different models.

| Underlying theory                                      | Accuracy of random prediction | Accuracy of supervised learning | Improved accuracy |
|--------------------------------------------------------|------------------------------|--------------------------------|-------------------|
| Five-layer model (Dunbar circle theory)                | 0.2000                       | 0.5298                         | 0.3298            |
| Three-layer model (Chinese Guanxi theory)              | 0.3300                       | 0.7880                         | 0.4580            |
| Four-layer model (revised Dunbar model according to Chinese Guanxi theory) | 0.2500                       | 0.7748                         | 0.5248            |

much bigger than the others, as seen in Fig. 6. That leads us to make further exploration to verify whether there is a better four-layer model.

Due to similarity means between variables in the fourth and fifth layers and difference between the first and second layers, 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 (see Fig. 7). The highest accuracy is 0.7748 and recall is 0.4292. Similarly, we also make a comparison with the five-layer and three-layer models.

Compared with the five-layer model, the accuracy and the recall of the four-layer model show significant improvement. For five-layer model, the accuracy of random guest is 0.2, and for the four-layer model, it is 0.25. The improved accuracy between the two models can be about 0.2 (0.525–0.330), as shown in Table 5. Thus, the four-layer model displays a much better performance in predicting tie strength than the others.

In addition, compared with the three-layer model, the accuracy and the recall of the four-layer model also show better performance. For the three-layer model, the

Fig. 6 $K$-Means clustering the sum of squared errors of different $k$. 
Fig. 7  Accuracy and recall of the four-layer model.

accuracy of random guest is 0.33, whilst for the four-layer model, it is 0.25. The improved accuracy between two models is about 0.07 (0.525–0.458).

We also try to compare the accuracy of other four-layer classification models with the baseline model. For instance, we merge the family/pseudo-family members and intimate friends into one category based on the five-layer model. Four types of four-layer models are thus generated. However, the other three models show worse performance than the baseline model (the accuracy of the other three models: 0.5574 when merging the five-layer model’s third and fourth layers, 0.5453 when merging the second and third layers, and 0.5279 when merging the second and innermost layers). Also, the baseline four-layer model stated above displays better performance in computing tie strength than the five-layer and three-layer models.

We can conclude that it is more suitable to divide the Chinese large cities’ 18–28 years old young Software A users into four categories in the Chinese guanxi context. This finding of four-layer ego-centered network structure supports Dunbar’s new discovery regarding the inner circle based on data from Twitter[24].

One more interesting finding is that in our four-layer model, when the variable living together is included, the accuracy of identifying the first layer, i.e., the family members, is hugely increased. Also, when age similarity is put into the model, the accuracy for predicting the second layer, i.e., intimate familiar ties, is improved.

We also adopt one of the methods—confusion matrix to better evaluate the performance of the predicted models, which can clearly figure out the percent and the different distributions of error samples of each category. Take one of the 4-layer predicted models as an example (shown in Fig. 8). The accuracy of this 4-layer model is 0.777. The performance of this model is good since the confusion matrix shows that the proportion of error samples roughly decreases with the increasing intimate distance between the different layers from an overall perspective. And for each layer, the majority of samples can be predicted accurately. However, for layer 3, there are a large quantity of samples predicted into layer 1. In addition, for the ties in layer 2, there are a large quantity of samples predicted into layer 3. So, it reveals that some intimacy familiar and familiar can not be distinguished from each other and some ties in intimate familiar act as acquaintance in social network context. Also, it is likely that some digital traces are not included in our features to distinguish these kinds of ties.

6 Conclusion and Future Work

Inspired by Dunbar’s study, this research attempts to propose a classification model that can categorize Chinese guanxi by using big data analysis of Software A active users. We firstly follow the Chinese social social-exchange principles[23,28] to design our questionnaire and we employ social survey methods to collect labeled categories of guanxi as the ground truth. Then various types of online interaction data are collected and
There are two types of familiar ties, one is especially challenge for our future studies. Unfortunately, this circle can not be distinguished from acquaintances ties in our computing process. Can we create another guanxi theory so that we can design a better question for the fourth circle? This posits a big challenge for our future studies. The employment of online interaction data is indeed very helpful for us in studying such a challenging research topic of this nature. In the several runs of dialogues between analysis guided by theories and the building of prediction models, we obtained higher accuracy in predicting tie strength, measured by the categories of guanxi. By combining survey data as the ground truth and online interaction data, we can gain fruitful research results. This study is only the beginning, and the combined methods of survey and big-data analysis within the realm of social science theory provide us with a bright road ahead for future studies.

Acknowledgment
We are grateful for the financial support of Tencent Research Institute Project "Research on identification of opinion leaders based on QQ big data", project number: 20182001706, as well as the support of Tsinghua-Göttingen Student Exchange Project IDS-SSP-2017001.

References
[1] K. K. Hwang, Face and favor: The Chinese power game, American Journal of Sociology, vol. 92, no. 4, pp. 944–974, 1987.
[2] M. S. Granovetter, The strength of weak ties, American Journal of Sociology, vol. 78, no. 6, pp. 1360–1380, 1973.
[3] M. S. Granovetter, The impact of social structure on economic outcomes, The Journal of Economic Perspectives, vol. 19, no.1, pp. 33–50, 2005.
[4] P. V. Marsden and K. E. Campbell, Measuring tie strength, Social Forces, vol. 63, no. 2, pp. 482–501, 1984.
[5] N. Eagle, A. S. Pentland, and D. Lazer, Inferring friendship network structure by using mobile phone data, Proceedings of the National Academy of Sciences of the United States of America, vol. 106, no. 36, pp. 15274–15278, 2009.
[6] S. G. B. Roberts and R. I. M. Dunbar, Communication in social networks: Effects of kinship, network size, and emotional closeness, Personal Relationships, vol. 18, no. 3, pp. 439–452, 2011.
[7] T. Raeder, O. Lizardo, D. Hachen, and N. V. Chawla, Predictors of short-term decay of cell phone contacts in a large scale communication network, Social Networks, vol. 33, no. 4, pp. 245–257, 2011.
[8] V. Palchykov, K. Kaski, J. Kertész, A. L. Barabási, and R. I. M. Dunbar, Sex differences in intimate relationships, Scientific Reports, vol. 2, no. 1, p. 370, 2012.
[9] M. I. Akbas, K. N. Avula, M. A. Bassiouni, and D. Turgut, Social network generation and friend ranking based on mobile phone data, in Proc. 2013 IEEE Int. Conf. Communications, Budapest, Hungary, 2013, pp. 1444–1448.
[10] V. Arnaboldi, A. Guazzini, and A. Passarella, Egocentric online social networks: Analysis of key features and prediction of tie strength in Facebook, Computer Communications, vol. 36, nos. 10&11, pp. 1130–1144, 2013.
topics in social network studies including social capital, trust, guanxi circle. Chinese indigenous management researches, such as guanxi and social network analysis in big data, self-organization process, and network dynamics in China: A link between individual and organizational levels, Journal of Business Ethics, vol. 67, no. 4, pp. 375–392, 2006.

[22] Y. Mao, K. Z. Peng, and C. S. Wong, Indigenous research on Asia: In search of the emic components of Guanxi, Asia Pacific Journal of Management, vol. 29, no. 4, pp. 1143–1168, 2012.

[23] X. Fei, From the Soil, the Foundations of Chinese Society. Beijing, China: Beijing University Press, 1998.

[24] R. I. M. Dunbar, V. Arnaboldi, M. Conti, and A. Passarella, The structure of online social networks mirrors those in the offline world, Social Networks, vol. 43, pp. 39–47, 2015.

[25] K. S. Yang, Chinese social orientation: An integrative analysis, in Chinese Societies and Mental Health, T. Y. Lin, W. S. Tseng, and E. K. Yeh, eds. Hong Kong, China: Oxford University Press, 1993, pp. 19–39.

[26] R. Burt, Structural Holes: The Social Structure of Competition. Cambridge, UK: Harvard University Press, 1992.

[27] N. Lin, W. M. Ensel, and J. C. Vaughn, Social resources and strength of ties: Structural factors in occupational status attainment, American Sociological Review, vol. 46, no. 4, pp. 393–405, 1981.

[28] X. L. Shi, L. A. Adamic, and M. J. Strauss, Networks of strong ties, Physica A, vol. 378, no. 1, pp. 33–47, 2007.

Xin Gao is a PhD candidate at Department of Sociology, Tsinghua University. She received the bachelor degree from Northeastern University in 2018. Her research interests include social network studies and organization behavior.

Jar-Der Luo is a professor at Department of Sociology, Tsinghua University, president of Chinese Network for Social Network Studies, and chairman of Tsinghua Social Network Research Center. He received the PhD degree from State University of New York at Stony Brook in 1993. He researches numerous topics in social network studies including social capital, trust, social network analysis in big data, self-organization process, and Chinese indigenous management researches, such as guanxi and guanxi circle.

Xiaoming Fu received the PhD degree in computer science from Tsinghua University, Beijing, China in 2000. He was then a research staff at Technical University of Berlin until joining the University of Gottingen, Germany in 2002, where he has been a professor in computer science and heading the computer networks group since 2007. He has spent research visits at Cambridge, Columbia, UCLA, Tsinghua University, Uppsala, and UPMC, and is an IEEE senior member and distinguished lecturer. His research interests include Internet-based systems, applications, and social networks. He is currently an editorial board member of IEEE Communications Magazine, IEEE Transactions on Network and Service Management, Elsevier Computer Networks, and Computer Communications, and has published over 150 peer-reviewed papers in renowned journals and international conference proceedings.
Kunhao Yang is a PhD candidate in network science and computational cognitive science at University of Tokyo. He received the master degree in sociology from Tsinghua University, Beijing, China in 2018. He has participated in some interdisciplinary programs and is skilled in the usage of mixed statistics, sociology theory, mathematic modeling, and big data mining to solve theoretical problems encountered in social research.

Loring Liu is a senior engineer of data mining of Tencent, head of data mining team of social network business group, architect of the first generation of QQ music recommendation system, and builder of Tencent’s customer lifecycle management system. He received the bachelor degree from Sun Yat-sen University in 2007. He is dedicated to the combination of data mining technology and business for many years. He has rich practical experience and project management experience in the fields of big data analysis and mining and credit information on the Internet. Currently, he focuses on Internet credit, user-based portraits, recommendation systems, and text mining.

Weiwei Gu is an assistant professor at Information Science and Technology, Beijing University of Chemical Technology. She received the PhD degree from Beijing Normal University in 2020. Her research areas include network representation, network dynamic learning, and social network analysis.