BIG DATA MINING AND ANALYTICS
ISSN 2096-0654 03/06 pp284–296
Volume 1, Number 4, December 2018
DOI: 10.26599/BDMA.2018.9020024

Understanding the Behavioral Differences Between American and German Users: A Data-Driven Study

Chenxi Yang, Yang Chen*, Qingyuan Gong, Xinlei He,
Yu Xiao, Yuhuan Huang, and Xiaoming Fu

Abstract: Given that the USA and Germany are the most populous countries in North America and Western Europe, understanding the behavioral differences between American and German users of online social networks is essential. In this work, we conduct a data-driven study based on the Yelp Open Dataset. We demonstrate the behavioral characteristics of both American and German users from different aspects, i.e., social connectivity, review styles, and spatiotemporal patterns. In addition, we construct a classification model to accurately recognize American and German users according to the behavioral data. Our model achieves high classification performance with an F1-score of 0.891 and AUC of 0.949.

Key words: behavioral difference; online social networks; Yelp; machine learning

1 Introduction

Cultural differences are a core question of interest among sociologists. Over the past decades, the cultural differences, reasons behind these differences, and phenomena that these differences reflect, including collectivism, individualism, and social sustainability, have been intensively studied[1–3]. In these studies, data for understanding users' culture-related behaviors is obtained using traditional methods, such as questionnaires, video documentation, and other personal interview methods.

Compared with rather static online textual data (which remains untouched once published) such as news or web pages, situation-aware interactive information like user reviews and comments provides more daily life-related and accessible (typically via text) opinions and thoughts. Nowadays, Online Social Networks (OSNs)[4] have witnessed rapid growth, attracting billions of users worldwide. People contribute to profiles, social activities, and life tracks on OSNs, and deep cultural impacts exist among these behaviors. Krasnova and Veltri[5] conducted a survey on Facebook to explore the differences in individual willingness to self-disclosure between American and German users via a questionnaire to Facebook users. They concluded that American users are more active on Facebook and have higher privacy concerns than Germans. To our knowledge, Ref. [5] is the first work that has analyzed users' online behavior from a cultural perspective. However, given that most of the cultural phenomena evolved for many years and developed from generation to generation, the scale of the online survey-based research is, although larger than the survey in

• Chenxi Yang, Yang Chen, Qingyuan Gong, and Xinlei He are with the School of Computer Science, Fudan University, Shanghai 200433, China, and the Engineering Research Center of Cyber Security Auditing and Monitoring, Ministry of Education, Shanghai 200433, China. E-mail: chenyang@fudan.edu.cn.
• Yu Xiao is with the Department of Communications and Networking, Aalto University, 02150 Espoo, Finland. E-mail: yu.xiao@aalto.fi.
• Yuhuan Huang is with the Faculty of European Languages and Cultures, Guangdong University of Foreign Studies, Guangzhou 510420, China. E-mail: huangyuhuan@gdufs.edu.cn.
• Xiaoming Fu is with the Institute of Computer Science, University of Göttingen, 37077 Göttingen, Germany. E-mail: fu@cs.uni-goettingen.de.
* To whom correspondence should be addressed.
Manuscript received: 2018-03-05; accepted: 2018-03-20

the real world, still not large enough to form a cultural impact. Further, data comprehensiveness is of great consequence to cultural analysis. Apart from the answer to the questions in the survey or text posted by users, the movement pattern and Points-Of-Interest (POIs), which are closely related to the cultural impact on a user, also matter.

To this end, Location-Based Social Networks (LBSNs), such as Yelp[6], Foursquare/Swarm[7–9], Momo[10], Skout[11], and Dianping[12, 13], which allow users to undertake location-centric activities in addition to social interactions, offer a viable data source for such cultural studies. These LBSN platforms record the activity data of massive users and provide researchers a great opportunity to compare human behavior from both spatiotemporal and social networking perspectives.

The USA and Germany, which have the largest populations in North America and Western Europe, respectively, are two important culture clusters in the world. They have different languages, traditions, and geographical conditions but also share the same Anglo-Saxon origins. First, Germans care more about social stability than Americans[14], whereas Americans tend to become outstanding individuals because of their elite culture[1]. Second, in terms of collectivism, Germans prefer risk taking and uncertainty avoidance compared with Americans[15]. However, the differences between German and American users in terms of behavioral patterns are rarely studied. We aim to know whether the behavior of users on LBSNs is consistent with previous cross-cultural research results[5, 16] and if not, identify which aspects of behavior have changed online.

In this paper, we use data from a representative LBSN, Yelp, as basis for our study. We conduct a data-driven study based on the Yelp Open Dataset (https://www.yelp.com/dataset). Yelp can help people discover local businesses, a.k.a., POIs, such as restaurants, hospitals, or spas. It allows users to publish reviews or conduct check-ins in selected businesses. Yelp users completed more than 142 million reviews by the end of Q3 2017 (https://www.yelp.com/about). Moreover, Yelp serves more like an "urban guide", a review platform with location information and category preference, rather than a channel that only helps you make friends. This platform partially reflects various aspects of people's daily life, which allows the inference of the entire profile and clear social engagement of a user. Given its popularity and rich user-generated content, Yelp is selected for our user behavior study. Compared with Ref. [5], social connectivity, spatiotemporal patterns, and writing styles are combined based on the data from a much larger scale of users in our work. We select the USA and Germany as the examples to understand online behavior from a cultural perspective and make comparisons between these two representative cultural clusters in North America and Western Europe respectively. Our key contributions are summarized as follows.

We provide a comprehensive statistical and demographic analysis of American and German users' behavior based on the Yelp Open Dataset, and compare the results in a comprehensive manner. We find that American users are more influential on Yelp than German users, and their friends are scattered in more cities. Our spatiotemporal analysis shows that German users have a clearer line between daytime and nightlife than American users. On the basis of our analysis of review texts, we also prove that collectivism is important for German users, whereas individualism is a priority for American users.

In this paper we verify the feasibility of applying big data analysis in the context of cultural behavior. In particular, we employ our analysis of users' online behavior to construct a classification model that can accurately detect whether a user is from the USA or Germany. With this classification model, we achieve an F1-score of 0.891 and AUC of 0.949 for detecting whether a Yelp user is from the USA or Germany, which serves as a strong buttress of our analysis and feature selection. We find that writing style- and social graph-related features are the most distinguishing features to differentiate American and German users.

The rest of the paper is structured as follows. We first introduce Yelp and the dataset used for our study in Section 2. In Section 3, we analyze the data for both user groups and businesses on Yelp from a cultural view of USA and Germany and identify the features that are strongly related to cultural diversity. We then provide comprehensive evaluations on our classification model using various supervised machine learning algorithms, including the importance of different feature sets, in Section 4. We review the related work in Section 5 and conclude this work in Section 6.

2 Background and Dataset

In this section, we provide an overview of Yelp and then introduce the dataset used in this study.

2.1 Background of Yelp

Founded in 2004, Yelp.com has become one of the world's largest online "urban guide" and business review sites[17]. On Yelp, users can write reviews, upload photos, conduct check-ins, and rate their experiences at different types of businesses such as restaurants and hotels. Yelp allows a user to conduct a check-in at a business only when the user is close enough to the business. In addition, users are supposed to give a review of a business several days after their visit. Yelp covers 21 main categories and over 1200 sub-categories of businesses (www.yelp.com/developers/documentation-Fusion). It provides a platform for users to express their preferences over different business categories. Meanwhile, it serves as a social networking platform. Users can make friends with other users who show interests to similar business categories. Together with reviews and check-ins, the data about users' preferences and friends reflect user behavior in daily life in an informative way.

2.2 Dataset description

We study the Yelp Open Dataset, which was used in the Yelp Dataset challenge. The dataset covers over 4 700 000 reviews, 156 000 businesses, and 1 100 000 users. Each review contains text and/or rating attributes.

The dataset is composed of 11 tables. For each user, we can obtain his/her friends, the year when the user started using Yelp, average number of stars, and other comprehensive assessments of his/her reviews and tips. For each business, its location, category, reviews, tips, and check-ins are all available. Regarding each user's home city, we assume that the user belongs to the city where he or she reviews most, defining the city as the "home city" and getting the country information accordingly. Regarding users' home countries, the USA, Germany, Canada, and UK are the four main countries.

3 Data Analysis

In this section, we study the behavior of American and German users using the Yelp Open Dataset described in Section 2. Our goal is to better understand the differences and similarities of American and German users in terms of location distribution of friends, social graph characteristics, writing style of reviews, preference for business categories, rating preferences, and temporal patterns of check-ins. This section is divided into three subsections, i.e., social graph, reviews, and check-ins.

3.1 Analysis of the social graph

To understand the social behavior of American and German users on Yelp, we use the Stanford Network Analysis Platform (SNAP)[18] for social graph analysis. SNAP is a general purpose network analysis and graph mining library written in C++. Using SNAP, we then analyze some representative network metrics, i.e., degree, Clustering Coefficient (CC), PageRank, and connected components in Yelp's social network G.

3.1.1 Analysis of the social graph as a whole

The Yelp's friendship network, G, has 8 981 389 nodes and 35 444 850 edges. Figure 1a shows the Cumulative Distribution Function (CDF) of the degrees in G. The degree of a node represents the number of edges connected to the node. A higher degree in G means that the user has more friends. No nodes have zero degrees, which means that any user on Yelp has at least one friend. The average degree of G is 7.0. Compared with many other OSNs, 7.0 is a small average degree. Given that Yelp is not a website dedicated to social networking, users on Yelp are not chasing after a large number of friends. Therefore, the connection on Yelp is looser than that in most OSNs.

The CC measures the cliquishness of a typical friendship circle. A higher average CC indicates that it is more likely for nodes to form tightly knit groups. Figure 1b is the CDF of CC of the nodes in G. Over 70% of Yelp users have a CC of zero and the friendship networks' average CC is 0.055, which demonstrates a weak connection between Yelp users. This result is due to users that employ Yelp as a guide service rather than a networking site.

PageRank is a metric that quantifies the importance of different nodes in a network[19], which has been applied by the Google search engine to rank websites. The PageRank value of any node of a network is between 0 and 1. A higher PageRank value suggests that the corresponding node is more important in this network. In Fig. 1c, the CDF of PageRank displays similar results.

Figure 1d shows the sizes of the top 10 connected components. The connected component is a subgraph in which any two nodes are connected to each other by paths, and this subgraph connected to no additional nodes in the supergraph. A total of 18 512 connected components exist in network G. The largest connected

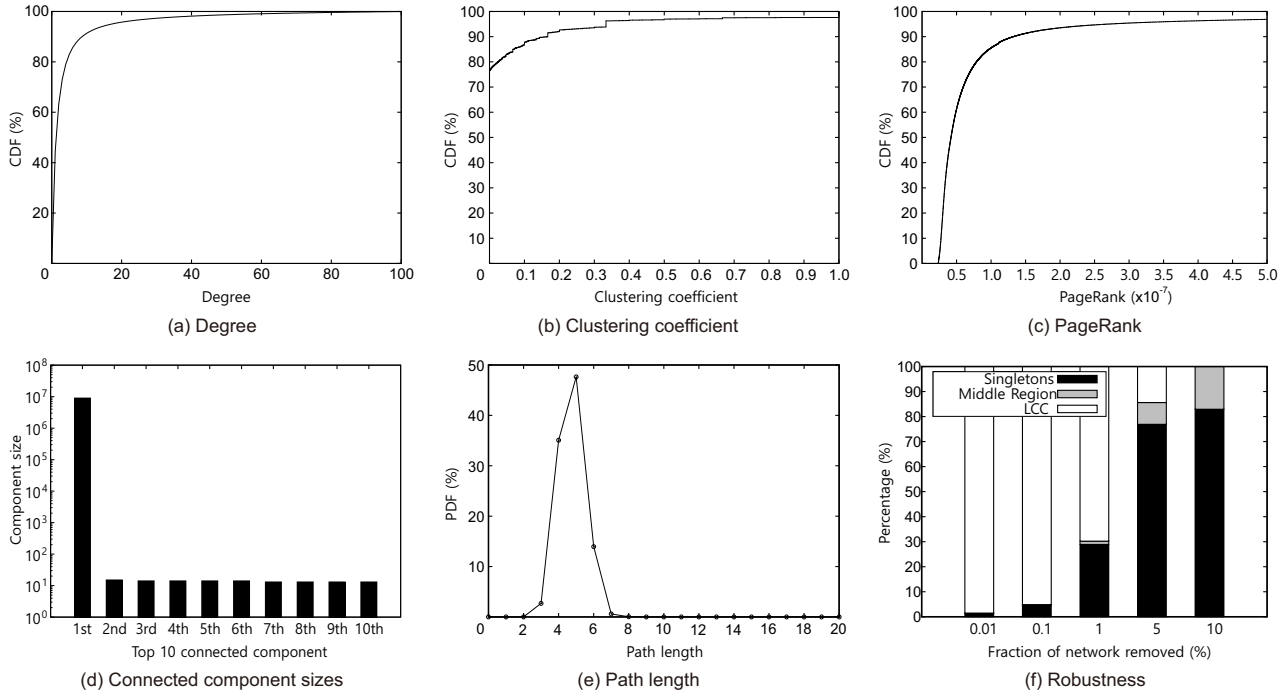


Fig. 1 Analysis of the friendship network.

component has 8 938 630 users, covering 99% users in the Yelp Open Dataset. The second largest connected components have 15 nodes and the third to sixth have 14 nodes each. We also calculate the distribution of the shortest path lengths of all node pairs in the Largest Connect Component (LCC), as displayed in Fig. 1e. The average shortest path length is 4.93, and the average CC of G is 0.055.

We further study the density of the "core" of the Yelp friendship network G. From 0.01% to 10%, we remove some nodes with the highest degrees from the network and analyze the remaining nodes. As in Fig. 1f, we group the remaining nodes into three categories, i.e., the LCC, singletons, and middle region. We find that the number of singletons already surpasses that of the nodes still in LCC after we remove 5% of the nodes with the highest degree. Therefore, Yelp's friendship network is not as strongly connected as other mainstream OSNs, such as Renren[20] and Cyworld[21].

3.1.2 Comparison between American and German users

Table 1 shows the mean and variance of CC of American and German users' social graph, which indicates that the mean of CC of American users is slightly higher than that of German users. We also compute the mean and variance of degrees in the American and German users' social graph. The mean and variance of degree of American users are both larger than those of German users. In other words, some American users make a lot more friends on Yelp than German users.

In Table 1, both the mean and variance of PageRank values of German users are larger than those of American users. The PageRank values are used to define the top 0.1% users of the whole social graph as "influential users". We use one set, P, to represent the influential users of the entire graph. We also find that 0.31% of the American users belong to P, whereas for German users, that number is 0.18%. These results illustrate that both the USA and Germany have more influential users than the average level of the entire network of Yelp. Moreover, the proportion of influential users is higher in American users than in German users.

Figure 2 reflects the location distributions of friends

Table 1 Graph attributes.

Nation	Avg. CC	Var. CC	Avg. Degree	Var. Degree	Avg. PageRank	Var. PageRank
USA	0.054	0.030	7.0	2379.0	1.180×10^{-7}	4.405×10^{-8}
Germany	0.041	0.032	2.0	909.0	2.120×10^{-5}	8.360×10^{-8}

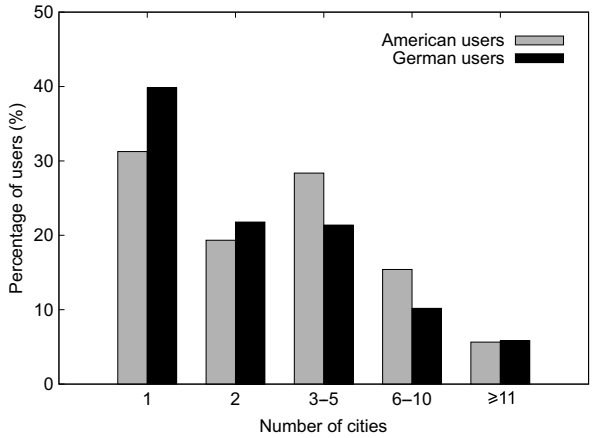


Fig. 2 Location distribution of friends.

of American and German users. The x-axis represents the number of cities where the users' friends are spread. The y-axis indicates the percentage of users who have friends in a certain number of cities. As shown in Fig. 2, the share of German users who only have Yelp friends in one or two cities is greater than that of American users. When the number of cities is three or higher, the share is about 15% larger in American users. We find that German users' friends are gathered in few cities, whereas the location distribution of American users' friends is slightly wider than that of German users'.

3.2 Analysis of reviews

Apart from the analysis of social graph in Section 3.1, we also group the reviews in accordance with the country where the businesses and the users are located. In Section 3.2.1, the text attribute of review serves as a significant part when analyzing the character trait. We then give the category preference of American and German users in Section 3.2.2. Subsequently, we present the visit and rating preference in Section 3.2.3.

3.2.1 Writing style analysis

Linguistic Inquiry and Word Count (LIWC)[22] is used to fully understand the text of reviews. It has been widely used in computerized text analysis to learn how the words we use in everyday language reveal our thoughts, feelings, personality, and motivations. In our study, given that the text of reviews comes from two languages, namely, English and German, we also use the German LIWC2001 Dictionary[23].

As shown in Table 2, the numbers represent the occurrence frequency of a specific set of words. In the first two columns, "Affect" and "Anger", which represent the affective process with the example words "happy", "ugly", and "bitter" and contents with "hate", "kill", and "pissed"[24] have something to do with indulging in anger. The average occurrence of "Affect" and "Anger" is higher among American users than among German users. As in Ref. [25], American students behave in a more affective way than German students and this conclusion conforms to our results from emotions behind writing style. With regard to the "Tenta" (representing the tentative words) and "Certain" (representing the certainty), American users prefer to write tentative-related words such as "maybe", "perhaps", and "guess"[24], whereas German users mention certainty-related words like "always" and "never" frequently. For "Swear", including words like "fuck", "damn", and "shit", German users are less likely to use swear words than their American counterparts when reviewing on Yelp. Words such as "buddy" and "neighbor", which belong to the "friends" category, appear more frequently in American users' reviews than in German users' reviews.

Table 2 Occurrence frequency of different categories of words in reviews.

Nation	Affect	Anger	Tenta	Certain	Swear	Friends
USA	6.045	0.252	2.298	1.845	0.087	0.438
Germany	4.688	0.191	0.965	2.842	0.017	0.288

In Table 3, we also determine the differences in occurrence frequency of "I", "We", and "Leisure" in the six main categories between American and German users. The frequency of writing the reviews with the pronoun "I" by American users is twice than that of German users. When talking with "We", the most possible category American users are in is "Nightlife", whereas it is "Restaurants" for German users. We

Table 3 Dimension values of american and german users.

Nation	Category	I	We	Leisure
USA	Beauty & Spas	7.02	0.53	1.16
	Health Medical	6.85	0.61	0.63
	Home Services	4.89	1.61	0.79
	Nightlife	3.72	**1.79**	**2.68**
	Restaurant	4.08	1.49	1.58
	Shopping	5.47	0.88	1.44
	Avg.	**5.34**	1.15	0.438
Germany	Beauty & Spas	4.17	0.27	1.06
	Health Medical	3.58	0.31	**1.81**
	Home Services	2.16	1.05	1.7
	Nightlife	1.73	1.29	1.52
	Restaurant	1.78	**1.34**	1.42
	Shopping	3.04	0.46	1.15
	Avg.	**2.74**	0.79	1.44

believe the frequency of the usage of "We" can be positively related to the high prevalence of people going to the particular category of businesses together. Therefore, "Nightlife" and "Restaurants" serve as the favorite category of American and German users, respectively, when they go out for social gatherings.

3.2.2 Preference for business categories

Experimentally, our results indicate the distinction of category preference between American and German users. We analyze the distribution of reviews in 10 main categories. Figure 3 displays the category pattern of American and German users, with the y-axis describing the logarithmic coordinates of review percentage of a certain category. For "Food", "Nightlife", and "Shopping" categories, American and German users share similar preferences in general. For other categories, visible differences are found. We find that German users show great interests to "Restaurants" as German users' percentage is 49%, which is much larger than that of American users' (41%). American users visit businesses in the category of "Beauty & Spas" (including subcategories like "Barbers", "Hair Loss Centers", and "Day Spas"), "Health Medical", "Home Services", and "Automotive" (including subcategories such as "Auto Detailing", "Registration Services", and "Car Wash") much more frequently than German users, except in "Public Services".

3.2.3 Visit & rating analysis

For the rating section, we also compute the star preference for businesses of American and German users. In the Yelp dataset, the average stars of American and German users are 3.73 and 3.78, respectively, with similar variance (1.18 and 1.01, respectively). As shown in Fig. 4, the star does not follow a normal distribution and most of the users have a rating of 3–5 in a five-point scale. German users prefer to grade businesses with a high point but not full. By contrast, American users are more likely to give 1 out of 5 or 5 out of 5 compared with German users, who adopt a milder rating style.

Figure 5 displays the CDF of the percentage of users' review out of their home city (we call it *outer review percentage* in the rest of this paper) of users whose review count is larger than zero. In general, American users have a larger outer review percentage than German users. Considering the friend distribution in Section 3.1.2, the location distributions of friends and reviews of American users are both wider, whereas those of German users' are both centralized to few cities. Our results also exhibit the strong and positive correlations of visit and friendship distribution as displayed in Ref. [26].

3.3 Analysis of check-ins

Temporal distribution of check-ins has been widely

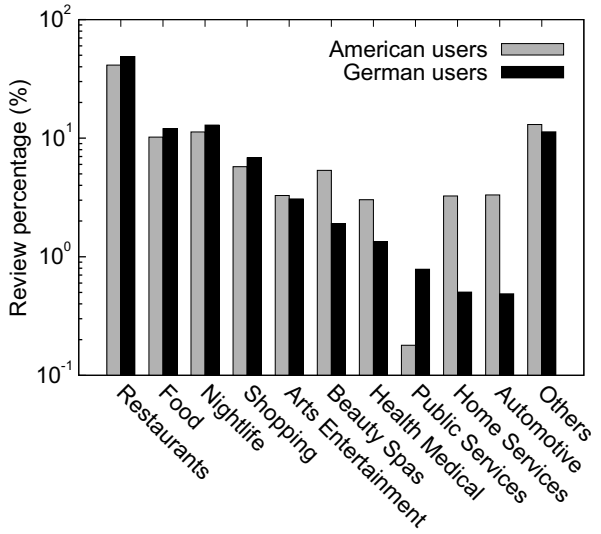


Fig. 3 Category pattern.

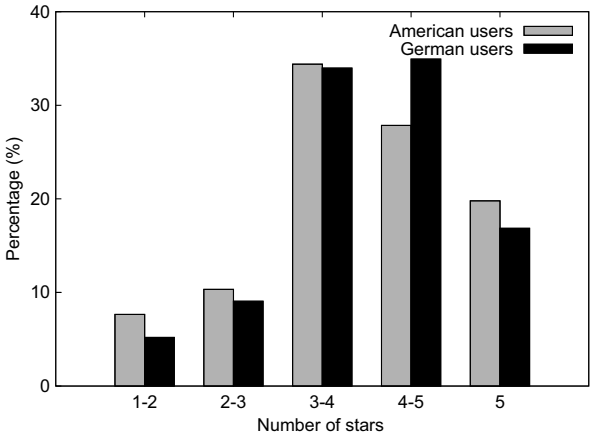


Fig. 4 Distribution of the number of stars.

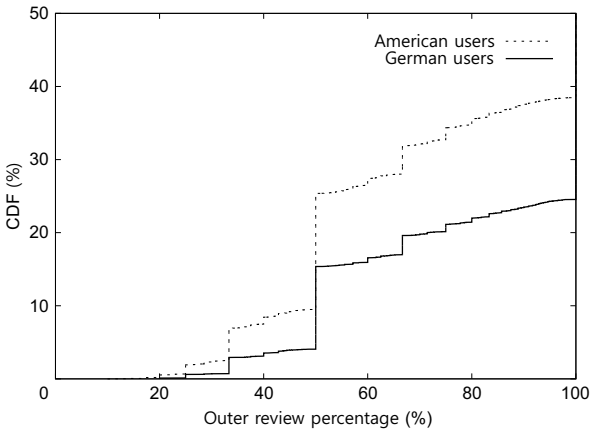


Fig. 5 CDF of outer review percentage of users whose review count is larger than zero.

used to characterize the behavior of LBSN users[27]. We analyze the temporal patterns of check-ins at businesses in Fig. 6. The y-axis represents the percentage of check-ins during each hour of a week. Figure 6 illustrates that American and German users both conduct more check-ins between Monday and Saturday and much less on Sunday. Simultaneously, the differences in German users' everyday noon peak (the first peak of a day) and night peak (the second peak of a day) can reach 6%, which is much more explicit than that of American users (0.03%). Given the exact time of everyday, we find that the lunch and dinner peaks of German users are around 11 a.m. and 6 p.m., respectively, whereas those of American users are around 1 p.m. and 10 p.m., respectively.

4 Implementation of the Country Classification Model

After comparing the behavior of social graph, reviews and check-ins between American and German users on Yelp, we establish a broad sense of the behavioral differences between these two groups of users and attempt to understand the cultural influence behind these behavioral differences. From an integrated view, we implement a model based on the features extracted from the analysis results in Section 3. This model aims to predict a user's home country and evaluate the classification performance of our approach and other related proposals. In addition, we investigate the importance of each feature to examine the extent these various features can influence the behavior of American or German users on Yelp.

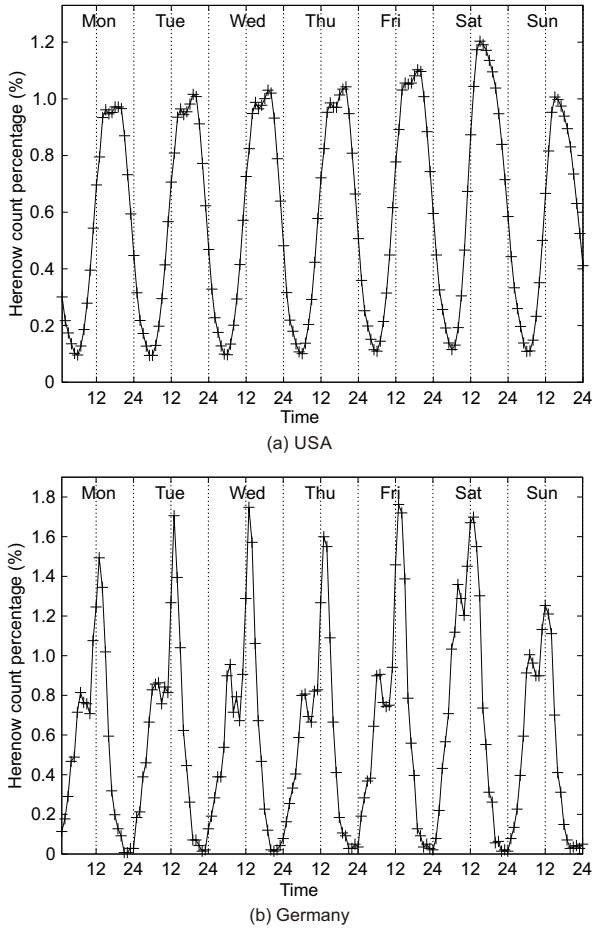


Fig. 6 Check-in patterns.

In this section, we include the implementation details and evaluation process of the classification model. First, we present a brief introduction to the tools and methods used in our study. We then describe the training and testing sets. Finally, we describe the importance of each feature set and assess the performance during country classification. To better implement the machine learning algorithms for the classification model, we adopt XGBoost[28] and Weka[29]. Specifically, XGBoost is a scalable machine learning system for tree boosting, which is widely used in machine learning contests. Weka supports a collection of machine learning algorithms for data mining tasks and it is implemented in Java. The algorithms in Weka include but are not limited to Random Forest (RF), Support Vector Machine (SVM), and C4.5 Decision Tree (J48).

To construct a training and validation dataset, we randomly select 700 American users and 700 German users from the Yelp dataset. We adopt four representative metrics, namely, precision, recall, F1-score, and AUC, to evaluate the performance of our classification model. Precision refers to the fraction of predicted German users who are really German users. Recall represents the fraction of German users who are accurately identified. F1-score is defined as the harmonic mean of precision and recall. AUC is short for "Area Under Receiver Operating Characteristic (ROC) Curve". Its value is equivalent to the probability that a randomly chosen positive example is ranked higher than a randomly chosen negative example[30]. We adopt several classic machine learning algorithms to train and validate our model using ten fold cross validation. For each algorithm, we apply grid search to find the "best" parameters, during which our goal is to achieve a high F1-score. After parameter tuning, we randomly select another 350 American users and 350 German users for testing and use the trained model to detect each users' home country. Table 4 shows our classification

Table 4 Comparison of different supervised machine learning algorithms for the classification model.

Algorithm	Parameter	Precision	Recall	F1-score	AUC
Random Forest	maxDepth=13, numFeatures=5	0.893	0.891	0.891	0.949
XGBoost	learning_rate=0.01, min_child_weight=1, max_depth=4, gamma=0.0, subsample=0.95, lamba=1, alpha=0, colsample_bytree=0.75, boost=gbtree, objective=binary:logistic	0.891	0.890	0.890	0.901
Decision Tree (J48)	confidenceFactor C=0.2, Instance/Leaf M=4	0.878	0.878	0.878	0.899
SVMp	Degree=3, Cost parameter=20.0	0.853	0.853	0.853	0.853
BayesNet	default	0.862	0.862	0.862	0.923

performance.

4.1 Evaluating classification model as a whole

As in Fig. 7, we first include all the 25 features as a whole for classification. We divide these features into four sets as in Table 5. Seven features represent the review counts of seven main categories in business-related feature sets, four represent social graph-related feature sets, ten are in writing style-related feature sets, and four comprise visit and rating related-feature sets. Table 4 compares several supervised machine learning algorithms including XGBoost, Random Forest (RF), C4.5 Decision Tree, SVM, and BayesNet. In particular, we consider SVMp (with polynomial kernel). Using McNemar's test[31] to compute statistical significance, we find that the prediction performance of nearly every two classifiers is significantly different ($p < 0.005$, McNemar's test). The only exception is when we consider the RF and XGBoost classifiers, as the difference between them is unremarkable ($p > 0.2$, McNemar's test). These two classifiers have much better prediction performance than the other ones, implying that they both can be used for classifying cultural belonging in practice. With the overall consideration of F1-score and AUC, all the F1-scores in our results from different algorithms are larger than 0.850. Among them, RF performs the best with an F1-score of 0.891 and AUC of 0.949. We adopt RF in the following subsection to compare the contributions of

Table 5 Subsets of features of the classification model.

Category	Description
Business	Restaurants and Food
	Nightlife
	Event Planning & Services
	Hotel & Travel
	Art & Entertainment
	Beauty & Spas
	Health & Medical
Social Graph	Number of cities having friends
	Clustering coefficient
	Degree
	PageRank
Writing Style	Number of words per sentence
	Frequency of occurrence of preposition
	Frequency of occurrence of pronoun
	Frequency of occurrence of "Anger"
	Frequency of occurrence of "Leisure"
	Frequency of occurrence of "Sad"
	Frequency of occurrence of tentative words like "maybe", "perhaps"
	Frequency of occurrence of certain words like "always", "never"
	Frequency of occurrence of "Friends"
	Frequency of occurrence of swear words
Visit & Rating	Number of reviews
	Number of visited cities
	Percentage of visit out of home city
	Average star

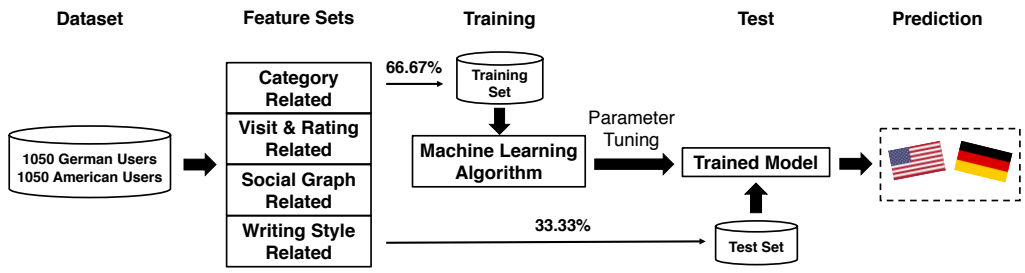


Fig. 7 Overview of the classification model.

different features.

4.2 Evaluating the contribution of different feature sets

To better understand the importance of different kinds of features in the model, we list the χ^2 (Chi-Square) statistics[32] of the top nine features. As shown in Table 6, the most discriminating feature is "Pronoun", which represents words like "I" and "You". Meanwhile, the features from writing style analysis such as "Preps (preposition)", "Tentat (tentative)", and "Certain (certainty)" are more important than other features. The writing style-related feature set plays an important role in distinguishing between American and German users on Yelp. However, social graph-related features such as "CC", "Friend_City_Num", and "PageRank" are also of importance, ranking the fifth, sixth, and seventh, respectively, in Table 6. To understand more details about other feature sets, we also evaluate four feature sets independently and compare their performance. Results are shown in Table 7. With the 10 writing style-related features, we achieve an F1-score of 0.878 and AUC of 0.937. For the social graph-related feature set, F1-score is 0.741 and AUC is 0.823. For the other two feature sets, F1-score is 0.617 and 0.601, and AUC is 0.660 and 0.619 of business-related features and visit and rating-related features, respectively. Corresponding to Chi-Square statistical analysis, the differences in writing style and social graph are the two most distinguishing attributes between users from USA and Germany on Yelp. Furthermore, preferences for business and visit and rating have a relatively slight effect on classification.

Table 6 Feature importance: χ^2 analysis.

Rank	χ^2	Feature	Category
1	969.876	Pronoun	Writing Style
2	650.939	Preps	Writing Style
3	366.716	Tentat	Writing Style
4	268.432	Certain	Writing Style
5	199.665	CC	Social Graph
6	99.615	Friend_City_Num	Social Graph
7	85.471	PageRank	Social Graph
8	73.282	Swear	Writing Style
9	60.701	Beauty & Spas	Business

Table 7 Contribution analysis of each feature subset in classification model.

Category	Presicion	Recall	F1-score	AUC
Writing Style	0.879	0.878	0.878	0.937
Social Graph	0.741	0.741	0.741	0.823
Business	0.623	0.612	0.617	0.660
Visit & Rating	0.602	0.603	0.601	0.619

5 Related Work

Combining the features of social networks with geographic information sharing, LBSNs are becoming increasingly popular. Some recent works focused on exploiting online social interactions among individuals to explain social phenomena. For example, Topa[33] generated a stationary distribution of unemployment which exhibited positive spatial correlations. Wal and Boschma[34] applied social network analysis in economic geography. For friendship distribution and user mobility, Liben-Nowell et al.[35] introduced a model capturing user behavior in real-world social networks and found that the probability of befriending a particular person and the number of closer people are inversely proportional. Cho et al.[26] discovered that long-distance travels are more likely influenced by social network ties compared with short-ranged travels. Those studies provided strong evidence that data on LBSNs can be utilized to analyze groups of user behavior. However, no studies have leveraged LBSNs for understanding cultural differences between the behaviors of users from different countries.

In the past decade, Yelp has become a worldwide online business review site, which records millions of reviews and business preferences of users from different countries. With the Yelp Open Dataset, numerous researchers have conducted studies on Yelp. Byers et al.[6] studied the correlation between the Groupon behavior of a business and the user rating distribution to this business. On the Yelp Open Dataset, fake reviews or malicious reviewers were filtered out by analyzing the texts of reviews[36–38]. Moreover, Refs. [17, 39] leveraged the rating and category or business preference extracted from the keywords or sentences in the text to analyze users' appetite and understand users' feedback. Furthermore, text and rating were also used together to understand users' feedback[40]. In this work, we combine the text, rating, and reviews on Yelp to form a comprehensive understanding of user behaviors from two cultural clusters.

To understand user behavior from a cross-cultural perspective, a preliminary work[41] reported that different cultural backgrounds have impacts at the

individual level on IT acceptance. To our knowledge, Ref. [5] is the first work that applied cultural analysis methods while studying the self-disclosure behavior on OSNs. They conducted a survey on Facebook and explored the differences in individual willingness to self-disclosure between American and German users. Reference [42] also analyzed the daily habits of school pupils in Germany and China. Unfortunately, both studies[5,42] relied on data collected from online surveys, which were quite limited. Garcia-Gavilanes et al.[43] considered the text-based "big data" on Twitter to ascertain whether a strong relationship exists between users' behavior on Twitter and traditional cultural theories. Even though Ref. [43] explored a large scale of users, aforementioned works[5,42,43] lacked rich features in their data, which are largely based on either answers to a questionnaire or the "free text" on Twitter. In addition, for Ref. [43], they cared about the relationship between users' online behavior and real culture phenomenon but did little about the differences in users' behavior between certain culture clusters, i.e., cross-cultural behavior. To narrow the gap between comprehensive data and online cross-cultural behavior, our work combines social connectivity, spatiotemporal patterns, and writing styles of users on Yelp to understand the differences and similarities between American and German users. We build on these past works to study the feasibility of using review-based websites to analyze cultural differences between certain cultural clusters.

6 Conclusion

By referring to the Yelp Open Dataset, we use the behavioral information of massive users to explore the differences between American and German users. We find that the major differences are category preference, mobility pattern, friendship distribution, rating preference, and writing style. We utilize the results to extract several human behavior patterns and generalize their features. In addition, we build a classification model to detect where a certain user comes from based on the extracted features. With this model, we validate our analysis results and gain a better understanding of the importance of various feature sets in forming a human behavior pattern on Yelp.

The cultural causes of user online behavior should be studied in future work. Expanding the variety of social platforms, for example, to other LBSNs like Foursquare, would be an important study for us to analyze the cultural differences, in combination with the present study using Yelp. We plan to use cross-OSN links[44] to obtain a user's activity data from different OSN sites. Meanwhile, an offline user study is also in our plans. We aim to build an overall behavior pattern of cultural consequences which can be applied to people with different cultural backgrounds.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (Nos. 61602122 and 71731004), the Natural Science Foundation of Shanghai (No. 16ZR1402200), Shanghai Pujiang Program (No. 16PJ1400700), EU FP7 IRSES MobileCloud project (No. 612212), and Lindemann Foundation (No. 12-2016).

References

[1] H. C. Triandis, R. Bontempo, M. J. Villareal, M. Asai, and N. Lucca, Individualism and collectivism: Cross-cultural perspectives on self-ingroup relationships, *J. Personal. Soc. Psychol.*, vol. 54, no. 2, pp. 323–338, 1988.

[2] R. Gumbrell-McCormick and R. Hyman, Embedded collectivism? Workplace representation in France and Germany, *Ind. Relat. J.*, vol. 37, no. 5, pp. 473–491, 2006.

[3] M. Ehrgott, F. Reimann, L. Kaufmann, and C. R. Carter, Social sustainability in selecting emerging economy suppliers, *J. Bus. Eth.*, vol. 98, no. 1, pp. 99–119, 2011.

[4] L. Jin, Y. Chen, T. Y. Wang, P. Hui, and A. V. Vasilakos, Understanding user behavior in online social networks: A survey, *IEEE Commun. Mag.*, vol. 51, no. 9, pp. 144–150, 2013.

[5] H. Krasnova and N. F. Veltri, Privacy calculus on social networking sites: Explorative evidence from Germany and USA, in *Proc. 43rd Hawaii Int. Conf. System Sciences* (*HICSS*), Honolulu, HI, USA, 2010, pp. 1–10.

[6] J. W. Byers, M. Mitzenmacher, and G. Zervas, The groupon effect on yelp ratings: A root cause analysis, in *Proc. 13th ACM Conf. Electronic Commerce*, Valencia, Spain, 2012, pp. 248–265.

[7] M. A. Vasconcelos, S. Ricci, J. Almeida, F. Benevenuto, and V. Almeida, Tips, dones and todos: Uncovering user profiles in foursquare, in *Proc. 5th ACM Int. Conf. Web Search and Data Mining*, Seattle, WA, USA, 2012, pp. 653–662.

[8] Y. Chen, Y. X. Yang, J. Y. Hu, and C. F. Zhuang, Measurement and analysis of tips in foursquare, in *Proc. 2016 IEEE Int. Conf. Pervasive Computing and Communication Workshops*, Sydney, Australia, 2016.

[9] Y. Chen, J. Y. Hu, H. Zhao, Y. Xiao, and P. Hui, Measurement and analysis of the swarm social network with tens of millions of nodes, *IEEE Access*, vol. 6, pp. 4547–4559, 2018.

[10] T. Chen, M. A. Kaafar, and R. Boreli, The where and when of finding new friends: Analysis of a location-based social discovery network, in *Proc. 7th Int. AAAI Conf. Weblogs and Social Media*, Cambridge, MA, USA, 2013.

[11] R. Xie, Y. Chen, S. H. Lin, T. Y. Zhang, Y. Xiao, and X. Wang, Understanding skout users' mobility patterns on a global scale: A data-driven study, *World Wide Web J.*, doi: 10.1007/s11280-018-0551-8.

[12] Y. Huang, Y. Chen, Q. Zhou, J. Zhao, and X. Wang, Where are we visiting? Measurement and analysis of venues in Dianping, in *Proc. 2016 IEEE Int. Conf. Communications* (ICC), Kuala Lumpur, Malaysia, 2016.

[13] Q. Y. Gong, Y. Chen, X. L. He, Z. Zhuang, T. Y. Wang, H. Huang, X. Wang, and X. M. Fu, DeepScan: Exploiting deep learning for malicious account detection in location-based social networks, *IEEE Commun. Mag.*, 2018. (in press)

[14] S. M. Lipset, Some social requisites of democracy: Economic development and political legitimacy, *Am. Polit. Sci. Rev.*, vol. 53, no. 1, pp. 69–105, 1959.

[15] E. U. Weber, C. K. Hsee, and J. Sokolowska, What folklore tells us about risk and risk taking: Cross-cultural comparisons of American, German, and Chinese proverbs, *Organ. Behav. Hum. Decis. Process.*, vol. 75, no. 2, pp. 170–186, 1998.

[16] M. Clyne, Cultural differences in the organization of academic texts: English and German, *J. Pragmat.*, vol. 11, no. 2, pp. 211–241, 1987.

[17] A. Hicks, S. Comp, J. Horovitz, M. Hovarter, M. Miki, J. L. Bevan, Why people use Yelp.com: An exploration of uses and gratifications, *Comput. Hum. Behav.*, vol. 28, no. 6, pp. 2274–2279, 2012.

[18] J. Leskovec and R. Sosic, SNAP: A general-purpose network analysis and graph-mining library, *ACM Trans. Intell. Sys. Technol.*, vol. 8, no. 1, p. 1, 2016.

[19] L. Page, S. Brin, R. Motwani, and T. Winograd, *The PageRank citation ranking*: *Bringing order to the Web*, Technical Report, Stanford InfoLab, 1999.

[20] X. H. Zhao, A. Sala, C. Wilson, X. Wang, S. Gaito, H. T. Zheng, and B. Y. Zhao, Multi-scale dynamics in a massive online social network, in *Proc. 2012 Internet Measurement Conf.*, Boston, MA, USA, 2012, pp. 171–184.

[21] Y. Y. Ahn, S. Han, H. Kwak, S. Moon, and H. Jeong, Analysis of topological characteristics of huge online social networking services, in *Proc. 16th Int. Conf. World Wide Web*, Banff, Canada, 2007, pp. 835–844.

[22] J. W. Pennebaker, R. J. Booth, R. L. Boyd, and M. E. Francis, *Linguistic Inquiry and Word Count*: *LIWC2015*. Austin, TX, USA: Pennebaker Conglomerates, 2015.

[23] M. Wolf, A. B. Horn, M. R. Mehl, S. Haug, J. W. Pennebaker, and H. Kordy, Computergestützte quantitative textanalyse: Äquivalenz und robustheit der DEUTSCHEN version des linguistic inquiry and word count, *Diagnostica*, vol. 54, no. 2, pp. 85–98, 2008.

[24] J. W. Pennebaker and M. E. Francis, Cognitive, emotional, and language processes in disclosure, *Cogn. Emot.*, vol. 10, no. 6, pp. 601–626, 1996.

[25] K. D. Roach and P. R. Byrne, A cross-cultural comparison of instructor communication in American and German classrooms, *Commun. Educ.*, vol. 50, no. 1, pp. 1–14, 2001.

[26] E. Cho, S. A. Myers, and J. Leskovec, Friendship and mobility: User movement in location-based social networks, in *Proc. 17th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, San Diego, CA, USA, 2011, pp. 1082–1090.

[27] R. Xie, Y. Chen, Q. Xie, Y. Xiao, and X. Wang, We know your preferences in new cities: Mining and modeling the behavior of travelers, *IEEE Commun. Mag.*, 2018. (in press)

[28] T. Q. Chen and C. Guestrin, XGBoost: A scalable tree boosting system, in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, San Francisco, CA, USA, 2016, pp. 785–794.

[29] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, The WEKA data mining software: An update, *SIGKDD Explor. Newsl.*, vol. 11, no. 1, pp. 10–18, 2009.

[30] T. Fawcett, An introduction to ROC analysis, *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, 2006.

[31] Q. McNemar, Note on the sampling error of the difference between correlated proportions or percentages, *Psychometrika*, vol. 12, no. 2, pp. 153–157, 1947.

[32] Y. M. Yang and J. O. Pedersen, A comparative study on feature selection in text categorization, in *Proc. 14th Int. Conf. Machine Learning* (*ICML*), San Francisco, CA, USA, 1997, pp. 412–420.

[33] G. Topa, Social interactions, local spillovers and unemployment, *Rev. Econom. Stud.*, vol. 68, no. 2, pp. 261–295, 2001.

[34] A. L. J. Ter Wal and R. A. Boschma, Applying social network analysis in economic geography: Framing some key analytic issues, *Ann. Reg. Sci.*, vol. 43, no. 3, pp. 739–756, 2009.

[35] D. Liben-Nowell, J. Novak, R. Kumar, P. Raghavan, and A. Tomkins, Geographic routing in social networks, *Proc. Natl. Acad. Sci. U.S.A.*, vol. 102, no. 33, pp. 11623–11628, 2005.

[36] A. Mukherjee, V. Venkataraman, B. Liu, and N. Glance, What yelp fake review filter might be doing? in *Proc. 7th Inte. AAAI Conf. Weblogs and Social Media*, Cambridge, MA, USA, 2013.

[37] M. Luca and G. Zervas, Fake it till you make it: Reputation, competition, and yelp review fraud, *Manage. Sci.*, vol. 62, no. 12, pp. 3412–3427, 2016.

[38] Y. S. Yao, B. Viswanath, J. Cryan, H. T. Zheng, and B. Y. Zhao, Automated crowdturfing attacks and defenses in online review systems, in *Proc. 2017 ACM SIGSAC Conf. Computer and Communications Security*, Dallas, TX, USA, 2017, pp. 1143–1158.

[39] W. Ariyasriwatana and L. M. Quiroga, A thousand ways to say 'Delicious!'–Categorizing expressions of deliciousness from restaurant reviews on the social network site Yelp, *Appetite*, vol. 104, pp. 18–32, 2016.

[40] J. McAuley and J. Leskovec, Hidden factors and hidden topics: Understanding rating dimensions with review text, in *Proc. 7^{th} ACM Conf. Recommender Systems*, Hong Kong, China, 2013, pp. 165–172.

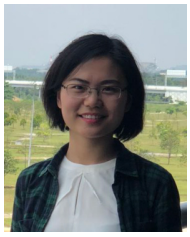
[41] M. Srite and E. Karahanna, The role of espoused national cultural values in technology acceptance, *MIS Quart.*, vol. 30, no, 3, pp. 679–704, 2006.

[42] X. M. Fu, H. Huang, X. Y. Li, H. S. Tan, and J. Tang, A comparative analysis of school pupils' daily habits in Germany and China, in *Proc. 10^{th} Int. Workshop on Hot Topics in Pervasive Mobile and Online Social Networking*, Honolulu, HI, USA, 2018.

[43] R. Garcia-Gavilanes, D. Quercia, and A. Jaimes, Cultural dimensions in twitter: Time, individualism and power, in *Proc. 7^{th} Int. AAAI Conf. Weblogs and Social Media*, Cambridge, MA, USA, 2013.

[44] Q. Y. Gong, Y. Chen, J. Y. Hu, Q. Cao, P. Hui, and X. Wang, Understanding cross-site linking in online social networks, *ACM Trans. Web*, 2018. (in press)

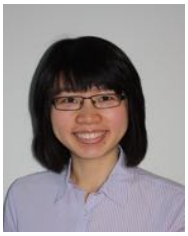
Chenxi Yang is an undergraduate student in the School of Computer Science at Fudan University, China. She has been a research assistant in the Mobile Systems and Networking (MSN) group since 2016. She visited Peking University as a research intern in 2017. Her research interests include massive data analytics, machine learning, and edge computing.



Qingyuan Gong received the BS degree from Shandong Normal University in 2012. She is now a PhD candidate at Fudan University, China. Her research interests include social network analytics, network security, and machine learning. She published referred papers in *ACM Transactions on the Web*, *IEEE Communications Magazine*, and IEEE ICPP. She has been a visiting student at the University of Göttingen (2015) and Tsinghua University (2016).

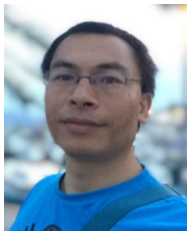
Yang Chen is an associate professor with the School of Computer Science at Fudan University, China. He leads the Mobile Systems and Networking (MSN) group since 2014. Before joining Fudan, he was a postdoctoral associate at the Department of Computer Science, Duke University, USA, where he served as a senior personnel in the NSF MobilityFirst project. From September 2009 to April 2011, he has been a research associate and the deputy head of Computer Networks Group, Institute of Computer Science, University of Göttingen, Germany. He received the BS and PhD degrees from Tsinghua University in 2004 and 2009, respectively. He visited Stanford University (in 2007) and Microsoft Research Asia (2006-2008) as a visiting student. His research interests include online social networks, Internet architecture and mobile computing. He is serving as an editorial board member of *Transactions on Emerging Telecommunications Technologies (ETT)* and *IEEE Access*. He is a senior member of the IEEE.

Xinlei He received the BS degree from Fudan University in 2017. He is now a master student at Fudan University. He has been a research assistant in the Mobile Systems and Networking (MSN) group since 2015. His research interests include social computing and data mining.



Yu Xiao received the doctoral degree (with distinction) in computer science from Aalto University in 2012. Before that, she got the master and bachelor degrees in computer science and technology from Beijing University of Posts and Telecommunications, China in 2007, and 2004, respectively. She is currently an assistant professor in Department of Communications and Networking, Aalto University where she leads the mobile cloud computing group. Her research interests include edge computing, mobile crowdsensing, and energy-efficient wireless networking. Her work has received 3 best paper awards from IEEE/ACM conferences. She is also a recipient of the 3-year postdoc grant from Academy of Finland.

Yuhuan Huang is an assistant professor with the Faculty of European Languages and Cultures at Guangdong University of Foreign Studies. She received the doctoral degree (with distinction) in intercultural German studies from the University of Göttingen, Germany in 2017. Before that, she received the master and bachelor degrees in German studies from Beijing Foreign Studies University, China in 2011 and 2008, respectively. Her research interests include German language and culture as well as intercultural communication.



Xiaoming Fu is a full professor of computer science at the University of Göttingen. He received the PhD degree from Tsinghua University in 2000. He was then a research staff at TU Berlin before moving to Göttingen in 2002, where he has been a professor and head of computer networks group since 2007. His research interests lie in networked systems and applications, including mobile and cloud computing, social networks, and big data analysis. He currently serves on the editorial boards of *IEEE Communications Magazine*, *IEEE Transactions on Network and Service Management*, and *Computer Communications*.