Measurement and Analysis of the Reviews in Airbnb

Qian Zhou[1,2], Yang Chen[1,2], Chuanhao Ma[1,2], Fei Li[1,2], Yu Xiao[3], Xin Wang[1,2], Xiaoming Fu[4]
[1]School of Computer Science, Fudan University, China
[2]Engineering Research Center of Cyber Security Auditing and Monitoring, Ministry of Education, China
[3]Department of Communications and Networking, Aalto University, Finland
[4]Institute of Computer Science, University of Goettingen, Germany
Email: {chenyang, xinw}@fudan.edu.cn, yu.xiao@aalto.fi, fu@cs.uni-goettingen.de

***Abstract*—Airbnb, a recently emerged online lodging service that allows house and apartment dwellers to lease out their premises to short-term renters like tourists, is reconstructing the value chain of the traditional hotel industry. It works as a platform that connects hosts and travelers and facilities their interaction and exchange. Studying this service could shed light on understanding the emerging sharing economy from a user-centric perspective. In this work, we collect the profiles of 43.8 million Airbnb users, and analyze the reviews they published online. We model the interactions between Airbnb users using a review graph, and study their mobility patterns by investigating their reviews. To the best of our knowledge, our work is the first measurement study of massive Airbnb users on a global scale, and it provides insights of their activities in both cyberspace and the physical world.**

I. Introduction

Firstly launched in 2008, Airbnb becomes a popular online service for listing and renting short-term lodging in residential properties around the world today. In January 2018, Airbnb has more than 3 million listings in 65,000 cities and 191 countries[1]. Besides online booking, Airbnb also facilitates the social interaction between tens of millions of users. For example, users can communicate using a private messaging service, and can share their lodging experience through posting reviews for other users publicly. Understanding the user behavior is essential for improving user experience. However, so far there lacks a comprehensive study of Airbnb user behavior.

Each Airbnb user has a personal profile, including demographic information like home country as well as user reviews. In this paper, we adopt a data-driven approach to analyze Airbnb user behavior. We crawled the profile pages of almost all – if not all – Airbnb users (as of Nov. 8, 2015), and collected their demographic information and all the published reviews. Based on the massive amount of data we have collected, we analyze the Airbnb user behavior from the following perspectives.

First, we conduct a demographic analysis of Airbnb users based on several key fields of user profiles, including home country, verification status and their roles in apartment leasing. We find that Airbnb is getting globally recognized, although most users are still from North America and Europe.

Second, we focus on the visible interactions between hosts and guests, which are revealed by public reviews. We model the interactions with a global *review graph* G, and describe them with a number of classic graph metrics. By examining the evolution of the review graph from 2008 to 2015, we discover that more and more users have been added to a giant weakly connected component which covers at least 98% users in G.

Last but not least, we dive into the mobility patterns of Airbnb users. After studying the users' movements from both spatial and temporal aspects, we figure out the time and location preferences in users' traveling. Also, based on our results of sentiment analysis, a majority of users are satisfied with their lodging experiences.

II. Data Collection and Preprocessing

A. Data Collection

In our study, we aim to obtain a complete view of Airbnb user behavior. Therefore, instead of using such a subset of users for study, we have crawled all 43.8 million Airbnb users' personal profiles including all the published reviews. Due to the strict per IP address rate limit, it becomes challenging to crawl all the user data in a short time. We address this issue as follows. Firstly, each Airbnb user has a unique numeric UID. The UID is assigned sequentially, i.e., a user registered earlier will get a smaller UID. For each user, we can access her profile page via the URL https://www.airbnb.com/users/show/UID. When we registered a new account on Sep. 25, 2015, we got the up-to-date maximum UID, i.e., 45063045. Secondly, we divided the ID range [1, 45063045] evenly into 185 chunks, and launched 185 virtual instances on the Microsoft Azure platform to crawl the personal pages simultaneously. Each of these instances has an unique IP address. The crawling process was run from Sep. 25, 2015 to Nov. 8, 2015. Except few unused IDs, we have obtained 43.8 million users' profiles and all the published reviews. Note that we respect the privacy of Airbnb users. Only publicly accessible data are crawled.

B. Data Preprocessing

We derive the interactions between Airbnb users from the published reviews, and model the interactions with a social

[1]https://www.airbnb.com/about/about-us?locale=en

ISBN 978-3-903176-08-9

graph. We call the social graph "review graph". The reviews can be classified into two categories, including the reviews from guests and the ones from hosts. According to [8], for more than 70% of online bookings through Airbnb, the users have published reviews for the visits. Therefore, it is feasible to profile the Airbnb user behavior such as mobility patterns and social interactions based on the analysis of user reviews.

We denote the review graph by $G = (V, E)$. Each node in the node set V represents an Airbnb user. Two nodes are connected with a directed edge, if one of the users has hosted the other one and at least one of them has posted reviews. For example, if user A has stayed in user B's apartment, A might post a review on B's profile page from the guest's perspective. Meanwhile, B might post a review from a host's perspective. If either A or B has posted a review online, there will be a directed edge (v_A, v_B). All the edges form the edge set E. When building the review graph, we exclude the users who have never posted or received any review. The resulting review graph includes 19,341,495 nodes and 17,553,551 edges.

As we are interested in the yearly temporal evolution of the review graph, we need to know when each node and edge was created. Because the registration time (year and month) of each Airbnb user is published on the user's profile page, the creation time of each node can be obtained directly from there. The creation time of an edge depends on when the reviews are published. If a user has visited another one for several times and has posted reviews for more than one visit, we set the creation time of the edge as the year when the first review was published. We derive the year information from reviews following three steps. (1) We obtain the year information directly from the reviews when possible. There are two types of reviews, one from the guest and the other from the host. On each Airbnb user's profile page we can find the published time information (year and month) of the latest 7 reviews of each type. The reviews published earlier are listed in a reverse chronological order, but their published time information are hidden. Among all the users who have posted or received at least one review, only 3.32% of them have received more than 7 reviews from guests, and 1.74% of them have received more than 7 reviews from hosts. Still, 30.67% of reviews do not have the time information. For these reviews, we infer their published time in the following two steps. (2) If the host and the guest have made "mutual reviews", which means they have written reviews for each other, we can assume a short time interval between the reviews since Airbnb only allows a user to write a review for a trip within 14 days after checkout. To validate this assumption, we examine all the mutual reviews with timestamps. The results show that 97.8% of them were published in the same month, while 99.3% of them were published in the same year. Given a pair of mutual reviews, if one of them has a timestamp, it is very likely that the other one was published in the same year. With this feature, we are able to estimate the published year of 22.77% of all the reviews. (3) As all the reviews are listed in a reverse chronological order, we utilize this feature to estimate the range of the published year of reviews. For example, three reviews were published in order. If both the earliest and the latest ones were estimated to be published in the year of 2009, the middle one must be published in 2009 as well. With this feature, we manage to estimate the exact published year of 6.03% of edges. For the last 1.87% edges, we assign each of them a randomly generated year within the estimated time range.

Besides the author and published year of reviews, we also look into the content of each review. We conduct sentiment analysis of all the reviews written in English, which covers 92.66% of all the published reviews. We use a natural language processing (NLP) library called NLTK [1] to extract users' sentiment information from reviews. Based on the output of NLTK, we follow the VADER algorithm to calculate a sentiment score for each review [12]. VADER is designed for sentiment analysis of social media content. The sentiment score for each review ranges from -1 to 1. A score of 1 means the review is strongly positive, -1 means the review is very negative, and 0 indicates the review is neutral. Among all the reviews written in English, 97.13% of them are positive, 1.98% of them are neutral, and only 0.89% of them are negative. In other words, nearly all the reviews written in English are positive about the lodging experience.

III. Data Analysis

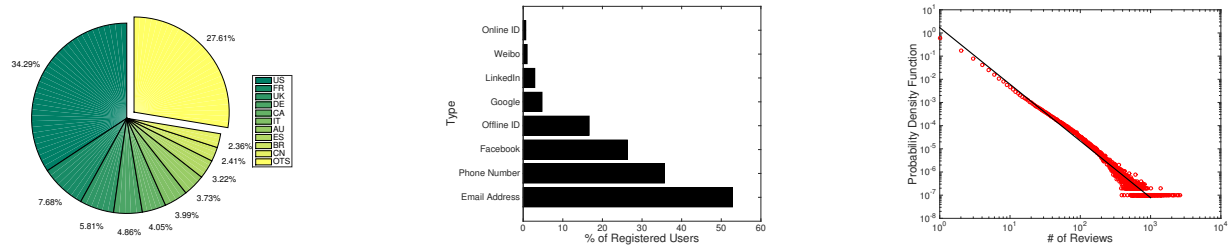
This work aims at providing insights on the Airbnb user behavior based on the analysis of personal profiles including published reviews. We analyze the crawled user data from the following three aspects. Firstly, we performance a comprehensive demographic analysis in § III-A to reveal the composition of Airbnb users. Secondly, we model the social interactions between users with a review graph, and analyze the static and dynamic characteristics of the review graph in § III-B. Thirdly, we investigate the mobility patterns of Airbnb users in § III-C.

A. Demographic Analysis

1) Statistics: The personal profile of a typical Airbnb user includes several information fields, such as location, verified ID, and "About Me". In addition, a user can request to become a verified user, in order to get a "V" badge displayed on her profile page. A small number of hosts satisfying certain requirements can also receive the "superhost" badge, which will also be shown on the user's profile page.

User Location Referring to the "location" indicated on the personal profile, we identify the home country of 86.20% of registered Airbnb users. As shown in Fig. 1(a), 34.29% of users come from the United States. In total, nearly 60% of users come from one of the 5 countries, including the United States, France, United Kingdom, Germany, and Canada. We can see that so far Airbnb is still more popular in North America and Europe than other areas in the world.

Except the reason that Airbnb is a US-based company, there may be other reasons for such user composition. In this work, we pick the top 8 countries with most Airbnb users, and try to discover the correlation between the number of Airbnb users and the social and economic factors like GDP, GDP per capita



(a) Home Country Distribution (b) Percentage of Different Types (c) Number of Reviews

Fig. 1: Analysis of User Profiles and Reviews

and population. In Table I, the figures of GDP and GDP per capita were retrieved from the website of the International Monetary Fund (IMF)[2], while the population information was obtained from the Department of Economic and Social Affairs of the United Nations[3].

We model each column of Table I as a vector, and denote the vectors by v_{user_num}, v_{GDP}, $v_{GDP\_per\_capita}$, and $v_{population}$, respectively. After that, we calculate the Pearson correlation coefficients between v_{user_num} and each of the other three vectors. The resulted correlation coefficients are denoted as r_{GDP}, $r_{GDP\_per\_capita}$, and $r_{population}$, respectively. The value of the correlation coefficient reflects the impact of the corresponding vector on the number of registered Airbnb users in each country. Note that a Pearson correlation coefficient is between -1 and 1. 1 refers to total positive linear correlation, 0 indicates no linear correlation, and -1 refers to total negative linear correlation. Concerning only the top 8 countries with most Airbnb users, r_{GDP} is 0.9927, $r_{GDP\_per\_capita}$ is 0.6124, and $r_{population}$ is 0.9873. If we extend the scope to include all the Airbnb users around the world, the values of r_{GDP}, $r_{GDP\_per\_capita}$ and $r_{population}$ are 0.8654, 0.2950, and 0.2049, respectively. Obviously, the number of registered users in a certain country is positively relevant to this country's GDP, whereas it is less relevant to the GDP per capita and the population.

Verified IDs As a method of improving the trust between users, Airbnb encourages users to submit their online and offline IDs for verification. After a user adds her ID information to her personal profile, Airbnb is responsible for verifying that the user does own the ID in question. From the values of the "Verified ID" field, we can find out which types of IDs have been verified. According to the personal profiles we have collected, most users have chosen to verify their "E-mail address", "Phone Number", and "Facebook Account". As shown in Fig. 1(b), these three types cover 52.82%, 35.64% and 26.29% of all the verified IDs, respectively. These are followed by the government-issued offline IDs, such as Passport and Driver License, which takes 16.65%. A user can request to become a "verified user". Upon request, Airbnb will verify the following items, including an online ID, a government-issued offline ID, a profile photo, a phone number, and an email address. Only 19.43% of all users are verified.

Reviews Airbnb users can write reviews for their hosts or guests. Fig. 1(c) demonstrates the distribution of the number of published reviews per user. It fits nicely with the power law model, i.e., $P(k) \propto k^{-\alpha}$ [5]. To evaluate how well the model fits the distribution, we adopt the coefficient of determination, i.e., the R^2 value. The value of R^2 ranges from 0 to 1. The larger the value is, the better the fitting is. When α is set to 2.4468, the value of R^2 is 0.9557, indicating a nice fit with the distribution of the number of per-user reviews.

Among all the users who have posted at least one review, only 3.53% of them have played both guest and host roles, and 90.76% of them only act as guests. Compared with the 5.71% of users who have written reviews as hosts, the number of guests is much bigger, which means that most of people use Airbnb for searching and booking accommodation instead of leasing out their apartments.

Superhost An Airbnb usr can become a "superhost" and get a superhost badge on her profile page, if she satisfies certain requirements, including hosting at least 10 groups of guests, receiving a "5-star" for at least 80% of the reviews posted by her guests, and completing each of the confirmed reservations. According to our study, there are only 68,883 superhosts, which means about 0.16% of Airbnb users are classified as superhosts.

About Me Besides the above-mentioned fields, there is an "About Me" field in each user's profile. It allows a user to add more information about herself. Optionally, users could add their "School", "Work" and "Language" information. Among all users, 28.66% have provided the "School" information, 10.16% have added the "Work" information, and 9.36% have said something about their "Language".

2) Temporal Evolution of Airbnb Demographic: According to the registration time of each Airbnb user, we can review the growth of Airbnb in terms of the number of newly registered users in each of the past 8 years. As illustrated in Fig. 2(a), both the number of registered users and the amount of published reviews have been growing steadily. The figures grow much faster during summers, showing that people are more active in traveling in summer.

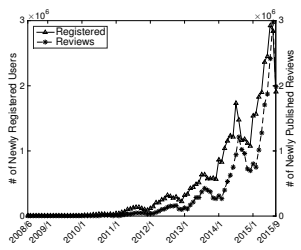
We further look into the geographical distribution of Airbnb users and measure the diversity of home countries. Here we

[2]http://www.imf.org/

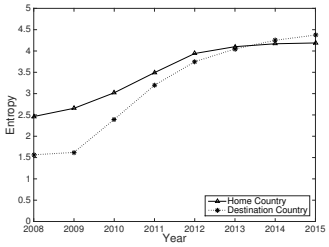
[3]https://esa.un.org/unpd/wpp/

TABLE I: Number of Registered Users v.s. GDP / GDP per Capita / Population

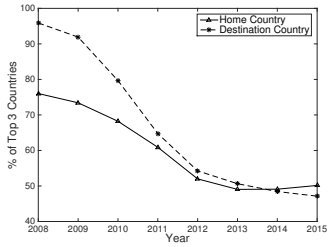
Country	Number of registered users	GDP (millions of USD)	GDP per capita (USD)	Population
United States	12,979,691	18,561,930	56,084	324,119,000
France	2,910,159	2,488,280	37,653	64,668,000
United Kingdom	2,195,446	2,649,890	43,902	65,111,000
Germany	1,834,505	3,494,900	40,952	80,682,700
Canada	1,528,011	1,532,340	43,413	36,286,200
Italy	1,507,997	1,852,500	29,867	59,801,000
Australia	1,407,956	1,256,640	51,181	24,309,000
Spain	1,215,428	1,252,160	25,843	46,065,000



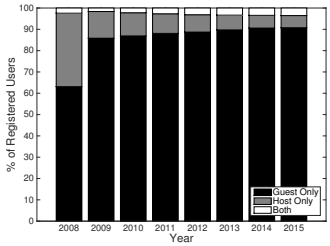
(a) Number of Registered Users and Published Reviews



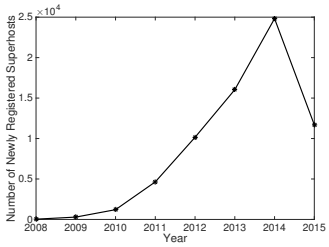
(b) Home Country Entropy



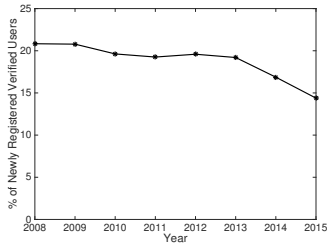
(c) Percentage of Top 3 Home/Destination Countries



(d) Distribution of Hosts/Guests/Both



(e) Number of Superhosts



(f) Percentage of Verified Users

Fig. 2: Demographic Analysis (Temporal)

introduce a metric called "home country entropy" and denote it by E_{home}. E_{home} can be calculated using the formula $E_{home} = -\sum_{i=1}^{k} p_i log_2 p_i$, where p_i refers to the fraction of users coming from the $i-th$ country. The value of E_{home} increases with the diversity of home countries. Similarly, we calculate the "destination country entropy" (E_{dest}). The information of "destination country" can be extracted from the reviews made for each trip. According to Fig. 2(b), Airbnb has become more and more globally recognized, in terms of both users' home countries and destination countries.

To verify the results of the home/destination country entropy, we check the distribution of the most popular home countries and destinations. Fig. 2(c) shows how many percentage of trips are made by the users from the top 3 most popular home countries and how many percentage of trips are made to the top 3 most popular destinations. We can see that in the first 3 years the percentages are larger than 60%. The curves started to drop in 2011, and got stabilized around 50%. These results are consistent with the values of the home/destination country entropy. In short, Airbnb is growing not only the number of registered users, but also its geographic diversity.

Regarding the number of hosts and guests, as shown in Fig. 2(d), the proportion of pure guests among all Airbnb users has grown from 63.11% to 90.76%. Meanwhile, the proportion of pure hosts keeps decreasing, while more and more users would play both the roles of hosts and guests. Compared with the growing number of travelers, the number of listed properties is growing relatively slow. If we look at the number of superhosts, users have joined the "superhost" group from time to time. Fig. 2(e) shows how many users registered in a certain year have become superhosts by end of 2015. Compared with hosts registered before 2015, fewer hosts registered after 2015 have become superhosts. This is partly due to the strict requirements of becoming superhosts, for example, superhosts must have hosted at least 10 trips.

Although the number of Airbnb users is growing steadily, the proportion of verified users has not grown. According to Fig. 2(f), among all the Airbnb users, around 20% of them are verified users. This number is smaller for those registered in 2014 and 2015. We believe more of them will apply for verification in the future.

B. Social Interaction Analysis

We utilize the review graph generated from the collected personal profiles for analyzing the social interactions between Airbnb users. We will first measure the complete review graph using the graph metrics listed below, and then analyze the

temporal evolution of the review graph and the characteristics of verified users.

- **Indegree and outdegree**: Indegree refers to the number of incoming edges a node has. The indegree of a node (user) is equal to the number of visitors the user has hosted. Outdegree refers to the number of outgoing edges a node has. The outdegree of a node (user) indicates the number of users she has visited.
- **PageRank**: PageRank is a metric that measures and ranks the importance of nodes in a graph [16]. It has been used by Google to rank the websites. We use this metric to discover "important users" in the graph.
- **Strongly connected component (SCC)**: An SCC is a subgraph where there is a path between any two nodes, while no additional node or edge can be added to this subgraph without breaking the nature of "strongly connected".
- **Weakly connected component (WCC)**: A WCC is a subgraph where there is a path between any two nodes when all edges are viewed as undirected. In addition, no additional node or edge can be added to this subgraph without breaking the nature of "weakly connected".
- **Communities**: A social network often exhibits a community structure. A community is formed by a number of nodes which are densely connected internally.

1) Review Graph: Static Analysis: ***Indegree and Outdegree*** The Cumulative Distribution Function (CDF) of indegree and outdegree among all the nodes is illustrated in Fig. 3(a) and Fig. 3(b). For comparison, we also visualize the CDF of indegree and outdegree among verified users and superhosts. The indegree and outdegree of the nodes corresponding to verified users and superhosts are relatively high, compared with other nodes. According to [15], the median indegree and outdegree of Twitter social graph are 16 and 39, respectively. Obviously, the numbers are much smaller in case of Airbnb, which means the review graph of Airbnb is rather sparse.

PageRank We use PageRank to measure the importance of each node in the review graph. We choose 1000 nodes with the largest PageRank values and compare their characteristics with those of the entire Airbnb population. None of these 1000 nodes is purely a guest. 31.1% of them are pure hosts, while 68.9% of them play both roles. We can see that hosting more is a critical indicator for becoming an important node in the review graph. In addition, the median indegree and outdegree of these 1000 nodes are 732 and 3, respectively. Both figures are much higher than those of the entire review graph.

Regarding the verification status, 100%, 89%, and 79.3% of the top 10, 100, 1000 nodes with highest PageRank values are verified users. Although verified users only cover 19.43% of the nodes in the entire G, verified users are more likely with higher PageRank values. Also, we are aware that about 14.48% of the top 1000 nodes are multi-user accounts, for example, the user name is "Alice and Bob" or "Carol & Tom". In contrast, only 0.282% of Airbnb accounts are multi-user accounts. Therefore, a viable portion of most important Airbnb accounts are operated by multiple people, for example, a couple or a family.

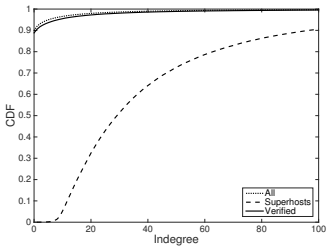
TABLE II: Percentage of Users in Top 3 Countries per Community

Community	Countries (% of Users)		
$C1$	US (66.54%)	CA (7.18%)	UK (3.26%)
$C2$	US (76.76%)	CA (3.09%)	UK (2.65%)
$C3$	FR (31.95%)	US (10.41%)	ES (8.21%)
$C4$	US (16.55%)	IT (14.92%)	FR (14.01%)
$C5$	AU (28.76%)	US (12.54%)	CN (5.56%)
$C6$	FR (22.78%)	ES (13.33%)	US (10.76%)
$C7$	DE (16.80%)	US (12.80%)	FR (10.61%)
$C8$	UK (36.81%)	US (13.16%)	FR (8.53%)
$C9$	US (54.41%)	FR (5.59%)	DE (4.83%)
$C10$	US (15.72%)	DE (13.98%)	FR (9.80%)

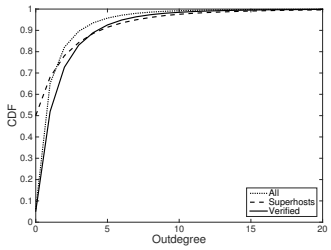
SCC and WCC We are interested in the connectivity among users in G. The sizes of the five largest SCCs are 63497, 13, 5, 4, 3, respectively. The largest SCC only covers 0.33% of all nodes, and the second largest SCC has only 13 nodes. This means very few nodes are strongly connected with each other. Differently, the sizes of the top five largest WCCs are 10969215, 15, 15, 15, and 14, respectively. We can see that the largest WCC covers 98.28% of nodes in G. Different from the small sizes of the SCCs, there is one giant WCC covering the major portion of all Airbnb users. In other words, most of the Airbnb users are weakly connected.

Communities The concept of community structure is widely used to study complex networks. If the network has a "community structure", the nodes can be split into different communities. Nodes from the same community are densely connected with each other, while nodes from different communities are sparsely connected. To study the communities in the Airbnb network, we adopt the widely used Louvain algorithm [2]. This algorithm is initially designed for undirected graphs. Following the practices in [11], we convert the review graph into an undirected graph by simply considering each edge as undirected. Louvain algorithm can assign each node of the network to one and only one community. It optimizes a metric known as "modularity". The value of modularity is between -1 and 1. Normally, if this value is larger than 0.4 [7], we can conclude that the network has a significant community structure. For G, the corresponding modularity value is 0.66, which means that the Airbnb network has a viable community structure. Also, our results show that there are 81308 communities among all nodes in G. Fig. 3(c) shows that sizes of the largest 30 communities. Among all communities, top 10 of them have covered 44.19% of nodes in G, and top 30 of them have covered 56.99% of nodes in G. In particular, the country composition of the top 10 communities are shown in Table II. We find that each of these communities has only one or very few dominant countries.

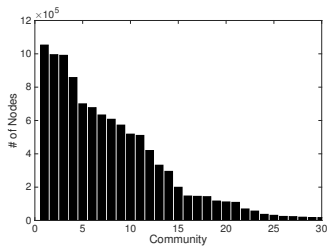
2) Temporal Evolution of the Review Graph: We are not only interested in the up-to-date structure of the review graph, but also how this graph has been constructed gradually. In this subsection, we study the temporal evolution of the review graph, taking the creation time of each node and edge into account. According to Fig. 4(a) and Fig. 4(b), the average indegree and outdegree of nodes in G grow steadily, as the



(a) Indegree

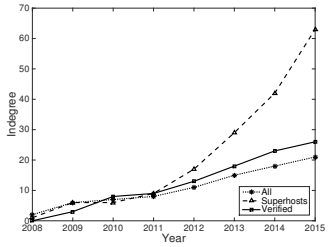


(b) Outdegree

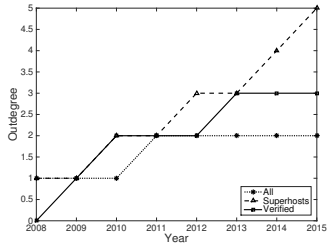


(c) Community Size Distribution

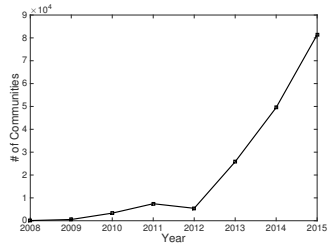
Fig. 3: Static Analysis of the Review Graph



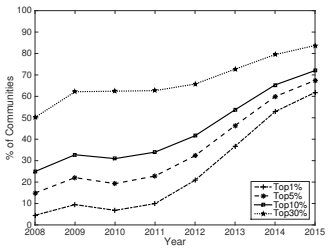
(a) Indegree



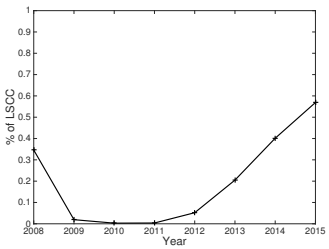
(b) Outdegree



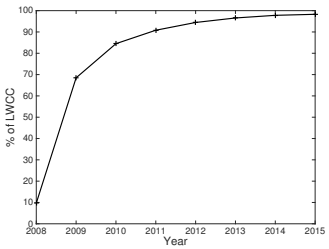
(c) # of Communities



(d) Top Communities



(e) LSCC



(f) LWCC

Fig. 4: Dynamic Analysis of the Review Graph

platform is developing rapidly. In 2008, the 80th percentile values of the indegree and outdegree are 2 and 1, respectively. In 2015, these two values become 21 and 2. Consequently, the review graph becomes denser and denser. More and more people are linked with each other through Airbnb. In Fig. 4(c), we can see the number of communities also grows year by year. Meanwhile, as shown in Fig. 4(d), the fractions of nodes within the top 1%, 5%, 10% and 30% are becoming larger and larger. We also pay attention to the fraction of the largest strongly connected component (LSCC) and the largest weakly connected component (LWCC) of G. For the LSCC (Fig. 4(e)), we can see it decreases for the first few years, and grows since 2011. However, the percentage of the LSCC is very small (less than 0.4%) all the time. Differently, we can see the percentage of LWCC (Fig. 4(f)) increases year to year. In 2008 about 10% users belong to the LWCC. This number increases yearly. Finally, in 2015, more than 90% users are involved in the LWCC. Thus most of the users are weakly connected now.

C. Mobility of Airbnb Users

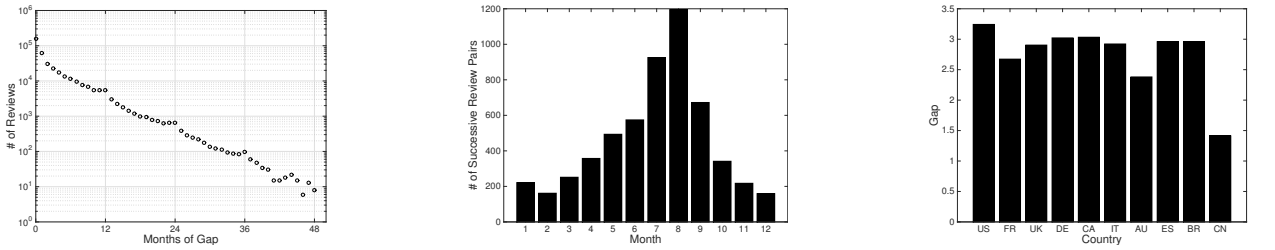
1) Spatial-Temporal Analysis: Understanding the spatial-temporal characteristics is important for an online lodging service. Thanks to the near real-time nature of review publishing, we can infer the users' mobility patterns by referring to published reviews.

We first explore the distribution of the time gap between two successive reviews published by the same user in Fig. 5(a), on a monthly base. We can see that when the time gap becomes larger, the number of corresponding successive review pairs become fewer. However, if the gap value can be divided by 12 months, there is a viable "peak". It shows that some travelers undertake their travels on a yearly base.

We further study the case with a time gap of one year, and categorize these review pairs according to the published month of the first review of them in Fig. 5(b). The x-axis denotes the published month, and the y-axis shows the number of successive review pairs with a time gap of one year. We can see most of the yearly travels take place in July and August. In Fig. 5(c), we can see the average time gap of successive review pairs of the users coming from the top 10 countries.

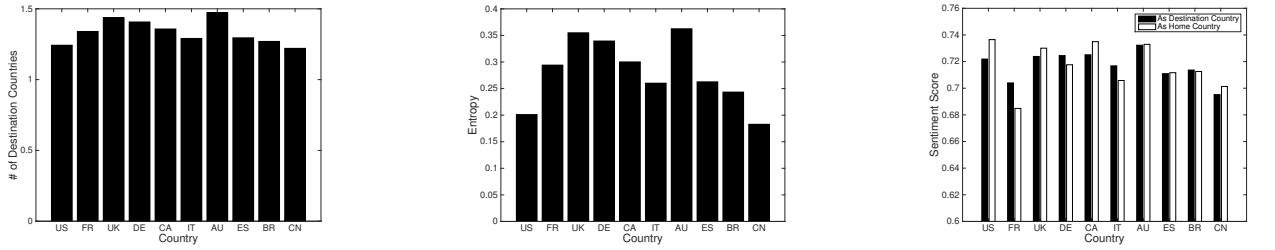
TABLE III: Fraction Distribution of "Home - Destination" County Pairs

Home \ Dest.	US	FR	UK	DE	CA	IT	AU	ES	BR	CN	$\sum_j H_{ij}$
US	0.3958*	0.0221*	0.0170	0.0082	0.0187	0.0236*	0.0044	0.0134	0.0029	0.0013	0.5073
FR	0.0161	0.0268*	0.0083	0.0041	0.0030	0.0127	0.0013	0.0097	0.0009	0.0002	0.0831
UK	0.0205*	0.0130	0.0445*	0.0056	0.0024	0.0109	0.0036	0.0099	0.0009	0.0003	0.1116
DE	0.0173	0.0078	0.0066	0.0167	0.0021	0.0091	0.0021	0.0077	0.0007	0.0002	0.0703
CA	0.0230*	0.0050	0.0036	0.0016	0.0293*	0.0050	0.0011	0.0032	0.0004	0.0002	0.0725
IT	0.0047	0.0042	0.0033	0.0021	0.0003	0.0087	0.0004	0.0031	0.0002	0.0001	0.0271
AU	0.0160	0.0072	0.0070	0.0026	0.0019	0.0072	0.0343*	0.0035	0.0005	0.0002	0.0805
ES	0.0039	0.0031	0.0028	0.0017	0.0003	0.0026	0.0002	0.0062	0.0002	0.0001	0.0210
BR	0.0037	0.0015	0.0009	0.0006	0.0005	0.0011	0.0001	0.0006	0.0030	0.0000	0.0121
CN	0.0048	0.0014	0.0012	0.0006	0.0004	0.0013	0.0009	0.0005	0.0000	0.0032	0.0144
$\sum_i H_{ij}$	0.5059	0.0921	0.0951	0.0437	0.0591	0.0823	0.0485	0.0578	0.0096	0.0058	1.0000



(a) Distribution of the Time Gap (b) Gap: 12 Months (c) Avg. Time Gaps of Different Countries

Fig. 5: Distribution of the Time Gap Between Two Successive Reviews



(a) Avg. Number of Visited Countries (b) Entropy (c) Sentiment Analysis

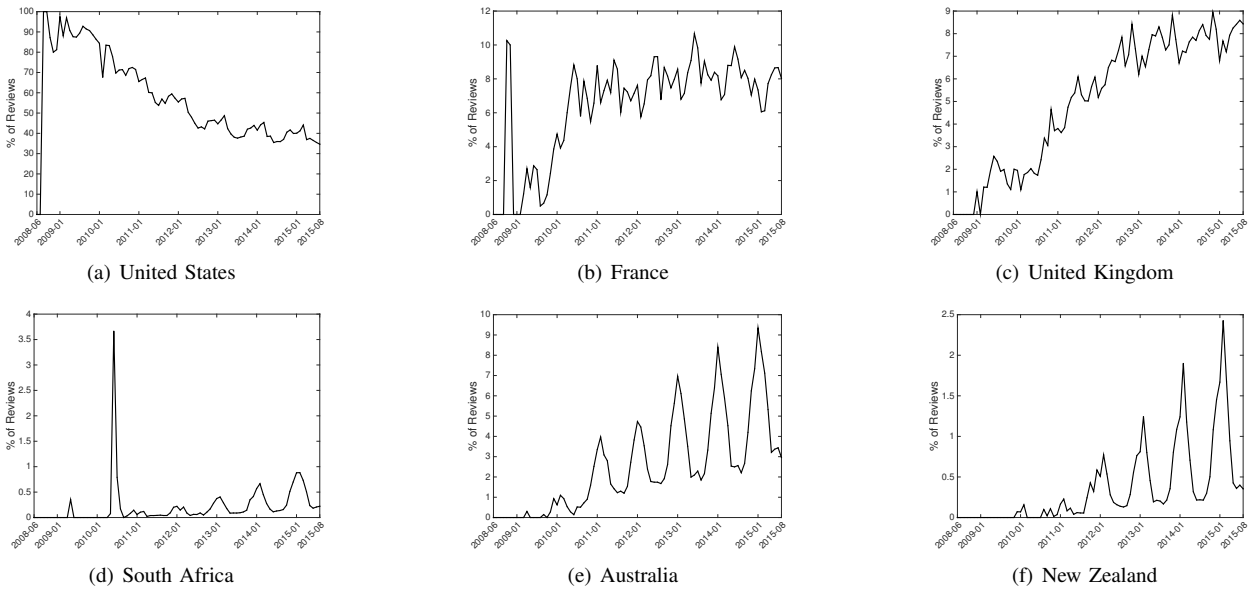
Fig. 6: Global Mobility of Users From Different Countries

We find that users from China has the smallest average time gap, while the users from the United States has the largest average time gap.

For each review, we can extract the home country of the publisher. Meanwhile, we know the country that she visits, which is known as the "destination country". Therefore, each review has a corresponding "home-destination country pair". In Table III, we use a matrix H to quantify the fraction distribution of home-destination country pairs. For simplicity, we only consider the users coming from the top 10 countries. We select elements with a value more than 0.02 and mark them with "*". In this matrix, we can see that most of the users have paid more visits to their home countries. In terms of the number of reviews, the three most popular destination countries are United States, United Kingdom and France.

In Fig. 6, we can see the global mobility of users from top 10 countries. We use two metrics, i.e., the number of visited countries, and the destination country entropy. The first metric can simply count the number of destination countries a user have visited. From Fig. 6(a), we can see that users from Australia and United Kingdom have visited more countries. From Fig. 6(b), we show the diversity of visited countries by calculating the entropy of destination countries. Similarly, we can see a higher diversity of destination countries for users coming from Australia and United Kingdom. We also calculate the average sentiment score of each home country and destination country, and show the results of the top 10 countries in Fig. 6(c). We find that there is very little difference among these countries. In average, users from the United States are slightly happier. Meanwhile, as a destination, Australia can make more people happy.

To understand where the users go from a temporal aspect, we also conduct a country-level analysis from a destination country's perspective. We can see the evolution of the visitor population over time, and we have examined the top 30 countries according to the user population. Due to the page limit, we pick six representative countries for our study. On one hand, we select the United States, France, and United



(a) United States (b) France (c) United Kingdom

(d) South Africa (e) Australia (f) New Zealand

Fig. 7: Temporal Behavior of Different Destination Countries

Kingdom, as they have the largest user population. Since these three countries are in the Northern Hemisphere, we select three countries in the Southern Hemisphere, i.e., South Africa, Australia and New Zealand. The results are shown in Fig. 7. For each country, we show the user popularity from a temporal view. The x-axis denotes the time information, and the y-axis represents the percentage of reviews a destination country has received in a certain month. We can see all the six countries have shown a seasonal periodicity. One significant difference between countries in the two hemisphere is the peak period of a year. In the countries of the Southern Hemisphere, there is always a peak in January and a valley in July. Differently, we observe an almost opposite trend for countries of the Northern Hemisphere. As we mentioned earlier, Airbnb is widely spread around the world; accordingly we can see the share of the reviews of United States-based apartments is going down. Another interesting finding is about South Africa, there is a significant and unusual peak in June and July of 2010. We believe that this is because the FIFA World Cup 2010, which has attracted numerous soccer fans from around the world.

2) Prediction: In this subsection, we investigate the predictability of user movements. In particular, we are interested in whether a user will travel aboard. In our study, we focus on users who have conducted at least one trip in 2015. Moreover, we exclude the users who have completed less than 7 trips on Airbnb, since they do not have enough historical data for the prediction. Among the rest of users, we group them into two categories, i.e., users whose latest trip is an international trip, and users whose latest trip is a domestic trip. We call the first group users "international users", and the second group of users "domestic users". We randomly pick 24,000 international users and 24,000 domestic users to form a training dataset. We also randomly select 6,000 international users and 6,000 domestic users to form a test dataset.

We select a number of key features to distinguish between these two types of users. These features belong to four categories: (1) the ratio between international and domestic trips; (2) the time interval between each two successive trips; (3) the number of trips within a certain time interval; (4) the demographic information. Given the training dataset and the selected features, we apply different supervised machine learning algorithms to train prediction models to predict whether a user is an international user. The algorithms we study include XGBoost [4], C4.5 decision tree [18], Random Forest [3] and Bayesian Network [9]. We use the test dataset to evaluate the classification accuracy of these models. Three classic metrics are introduced, i.e., precision, recall, and F1-score. Precision means the fraction of identified international users who have really traveled abroad for their latest trips. Recall indicates the fraction of international users who have been accurately detected. F1-score represents the harmonic mean of precision and recall. According to Table IV, the XGBoost algorithm outperforms other algorithms and the overall F1-score is as high as 0.766. Therefore, the selected features could accurately distinguish international users from domestic users. To evaluate the importance of each feature, we use χ^2 (Chi square) statistic to measure each feature's discriminative power [20]. The results are shown in Table V.

IV. Related Work

Analysis of online service users always starts with the collection of user data. A straightforward way is to obtain all the data directly from the back-end servers. For example, Zhao et al. [21] have explored the evolution of the Renren network using the data obtained from the back-end. However, very few online service providers have opened their data for public research. Furthermore, many of them have applied

TABLE IV: Prediction of "International Users"

Algorithm	Parameter	Precision	Recall	F1-Score
XGBoost	learning rate=0.09, max_depth = 6, gamma = 0.2, seed = 2	0.792	0.741	0.766
Random Forest	247 trees, max_depth = 8	0.746	0.767	0.757
C4.5(J48)	Instance/leaf M=1,Confidence factor C=0.006	0.779	0.735	0.756
BayesNet	4 children,4 parents	0.782	0.726	0.753

TABLE V: χ^2 statistic

Rank	Feature	χ^2
1	Fraction of International Trips	15227.842064
2	Fraction of International Trips in 2015	12138.063148
3	Number of International Trips	11985.024126
4	Whether the 2nd Latest Trip is International	10841.01484
5	Home Country's GDP	9103.609107

mechanisms such as per-IP address rate limit to prevent large-scale data crawling. As in [6], we apply a distributed data crawling approach to collect all the personal profiles of Airbnb users, which allows us to conduct a comprehensive analysis.

Quattrone et al. [17] have crawled the Airbnb data of the city of London, and have studied the problem of regulating Airbnb. Their work investigated Airbnb from a socio-economic angle, and conducted a series of temporal-spatial analysis of Airbnb properties and demands in London. Ma et al. [14] focused on the Airbnb hosts, and studied how hosts describe themselves in their profile pages. Their study was based on 67,465 hosts coming from 12 cities in the United States. Differently, our work focuses on the interactions between users, and have extended the scope to the entire set of Airbnb users.

Conventionally, a "social graph" models a number of users and the "friendship" connections among them. The connection between users does not necessarily reflect the real interactions between them. To solve this issue, Wilson et al. [19] proposed to describe the interactions with an "interaction graph", and demonstrated through a data-driven study that the interaction graph can describe the user activities more efficiency than the social graph relying on social links only. Jiang et al. further studied latent interactions in the Renren social network [13] based on the profile visit histories. Their study also demonstrated that latent interactions are more meaningful than social links. Similarly, we construct our review graph based on user interactions. Our review graph models the user mobility and interactions on a global scale.

V. Conclusions

In this paper, we conduct a comprehensive user behavior analysis of Airbnb, a leading online lodging service. Our study covers different aspects, including the user composition, the interactions between users, and the cross-country mobility patterns of the users. To the best of our knowledge, our study presents the first comprehensive and evolutionary analysis of Airbnb users on a global scale. In the future, we plan to analyze the Airbnb users' online behavior and offline activities as an integrated whole. Also, we aim to detect the spam accounts using deep learning technologies [10].

Acknowledgement

This work is sponsored by National Natural Science Foundation of China (No. 61602122, No. 71731004), 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). Yang Chen is the corresponding author.

References

[1] S. Bird. NLTK: The Natural Language Toolkit. In *Proc. of COLING/ACL*, 2006.

[2] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast Unfolding of Communities in Large Networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008.

[3] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[4] T. Chen and C. Guestrin. XGBoost: A Scalable Tree Boosting System. In *Proc. of ACM KDD*, 2016.

[5] A. Clauset, C. R. Shalizi, and M. E. J. Newman. Power-Law Distributions in Empirical Data. *SIAM Review*, 51(4):661–703, 2009.

[6] C. Ding, Y. Chen, and X. Fu. Crowd Crawling: Towards Collaborative Data Collection for Large-scale Online Social Networks. In *Proc. of ACM COSN*, 2013.

[7] S. Fortunato and M. Barthlemy. Resolution limit in community detection. *Proceedings of the National Academy of Sciences*, 104:36–41, 2007.

[8] A. Fradkin, E. Grewal, D. Holtz, and M. Pearson. Bias and Reciprocity in Online Reviews: Evidence From Field Experiments on Airbnb. In *Proc. of ACM EC*, 2015.

[9] N. Friedman, D. Geiger, and M. Goldszmidt. Bayesian network classifiers. *Machine Learning*, 29(2):131–163, 1997.

[10] Q. Gong, Y. Chen, X. He, Z. Zhuang, T. Wang, H. Huang, X. Wang, and X. Fu. DeepScan: Exploiting Deep Learning for Malicious Account Detection in Location-Based Social Networks. *IEEE Communications Magazine*, 2018.

[11] D. Hric, R. K. Darst, and S. Fortunato. Community detection in networks: Structural communities versus ground truth. *Phys. Rev. E*, 90:062805, Dec 2014.

[12] C. J. Hutto and E. Gilbert. VADER: A parsimonious rule-based model for sentiment analysis of social media text. In *Proc. of AAAI ICWSM*, 2014.

[13] J. Jiang, C. Wilson, and et al. Understanding Latent Interactions in Online Social Networks. In *Proc. of ACM IMC*, 2010.

[14] X. Ma, J. Hancock, K. L. Mingjie, and M. Naaman. Self-disclosure and Perceived Trustworthiness of Airbnb Host Profiles. In *Proc. of ACM CSCW*, 2017.

[15] S. A. Myers, A. Sharma, P. Gupta, and J. Lin. Information Network or Social Network?: The Structure of the Twitter Follow Graph. In *Proc. of WWW '14 Companion*, 2014.

[16] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, November 1999.

[17] G. Quattrone, D. Proserpio, and et al. Who Benefits from the "Sharing" Economy of Airbnb? In *Proc. of WWW*, 2016.

[18] J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1993.

[19] C. Wilson, B. Boe, and et al. User Interactions in Social Networks and Their Implications. In *Proc. of ACM EuroSys*, 2009.

[20] Y. Yang and J. O. Pedersen. A comparative study on feature selection in text categorization. In *Proc. of ICML*, 1997.

[21] X. Zhao, A. Sala, C. Wilson, X. Wang, S. Gaito, H. Zheng, and B. Y. Zhao. Multi-scale Dynamics in a Massive Online Social Network. In *Proc. of ACM IMC*, 2012.