Information Sciences 376 (2017) 54–70

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Overlapped user-based comparative study on photo-sharing websites

Dongyuan Lu[a], Ruoshan Wu[a], Jitao Sang[b,*]

[a] School of Information Technology and Management, University of International Business and Economics, 100029, China
[b] National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, 100190, China

ARTICLE INFO

Article history:
Received 18 June 2016
Revised 25 September 2016
Accepted 2 October 2016
Available online 6 October 2016

Keywords:
Photo-sharing websites
User behavior analysis
Overlapped user
Multi-OSN participation

ABSTRACT

Along with the development of Web2.0 technologies and the prevalence of digital capture devices, there has been an increased popularity of various photo-sharing websites, e.g., Flickr, Instagram, and Pinterest. More and more people today are creating and sharing millions of personal photos using these photo-sharing websites. Previous studies have shown that an individual may register and participate with several online social networking websites, and is referred to as an "overlapped user". In this work, we use Flickr and Instagram as test platforms to conduct a comparative study on the behaviors of overlapped users on different photo-sharing websites, including their temporal distribution, location distribution, photo annotation, and social attributes. We theoretically show that observations based on the overlapped users are more significant than those that are based on independent subject groups. Moreover, the derived observations are not only helpful in understanding the multi-online social networking (OSN) participation phenomenon in the context of photo-sharing behaviors, but they provide practical instructions to photo-sharing-based user modeling applications.

1. Introduction

1.1. Background and motivation

Nowadays, people have become used to recording their daily lives, with personal photos, and there has been the prevalence of digital capture devices. According to Yahoo statistics, up to the end of 2014, the total number of digital photos exceeded 800 billion, and is expected to reach up to 1270 billion in 2017. This explosion in the number of personal photos has led to calls for effective photo-sharing and photo-management tools.

The development of WEB2.0 techniques has provided solutions to address these needs. The photo-sharing functions available on photo-sharing websites, such as Flickr, Pinterest, and Instagram, can help people to easily upload their personal photos online for both sharing and photo management. It is reported that up to December, 2014, the number of registered users of Pinterest, Flickr, and Instagram exceeded 40, 92, and 187 million, respectively.[1] These massive amounts of data have opened up possibilities to investigate tasks of image understanding and image retrieval [28,40] from alternative perspectives. According to Wikipedia, there exist up to 24 active photo-sharing websites, of which ten are ranked in the top

* Corresponding author.
E-mail address: jtsang@nlpr.ia.ac.cn (J. Sang).

[1] http://www.thesocialmediahat.com/active-users.

http://dx.doi.org/10.1016/j.ins.2016.10.005
0020-0255/

Fig. 1. The multi-OSN participation illustration on photo-sharing websites.

1500 and five are ranked in the top 200 in Global Alexa.[2] The five most popular photo-sharing websites are Instagram, Pinterest, Imgur, Flickr, and DeviantArt. Because all of these websites offer similar functions for online photo management and sharing, the question arises regarding how to interpret this common prosperity phenomenon. A comparative study of people's behaviors on different photo-sharing websites is needed to understand their characteristics and to shed some light on this question [41].

From the perspective of people (or "users" in the context of social media), the same individual may be simultaneously engaged with multiple online social networking (OSN) websites. For example, the same individual may communicate with his/her friends on Facebook, follow real-time hot events on Twitter, subscribe to and watch videos on YouTube, and share and discuss favorite restaurants on Yelp [44]. Global Web Index 2015 has reported that within the 50 OSNs that were investigated, each individual possesses user accounts on an average of 5.54 OSNs, and actively participates in 2.82 OSNs.[3] This phenomenon is called "multi-OSN participation," and the persons holding accounts in multiple OSNs are called "overlapped users." Through preliminary data analysis, we find that the "multi-OSN participation" phenomenon is also significant in photo-sharing websites (as illustrated in Fig. 1). This enables us to conduct a comparative study between different photo-sharing websites by analyzing the overlapped users' behavior patterns, which are important in helping us to understand the characteristics of each OSN and the multiple-OSN participation phenomenon in the context of photo-sharing websites. In the rest of this section, we first review the related literatures, after which we present an overview of this work.

1.2. Related work

Two research lines are related to this work, i.e., photo-sharing website analysis and a multi-OSN comparative study.

1.2.1. Photo-sharing website analysis

There is evidence that an increasing number of persons actively share their personal photos online using various photo-sharing websites. The popularity of photo-sharing websites has provided a wealth of image data as well as test platforms for multimedia studies. Over the past decade, there has been much focus on exploiting these user-contributed image data to address well-known issues such as the semantic gap and intent gap in multimedia understanding problems [26]. However, only few works have focused on studying the characteristics of photo-sharing websites. In this subsection, we will review some of these works on analyzing the motivation for user participation, users' behavior patterns, and social networking topologies.

Ames and Naaman were among the first to analyze users' motivations for adding annotations to their uploaded photos on photo-sharing websites [5]. The annotation motivations were identified as being caused by socialization or functionality. The authors also proposed suggestions for designing new photo organization functions and annotation-based applications. Following this work, in [37], the authors investigated different ways in which users participate in photo-sharing communities. According to the theory of motivation, the motivations for a user's participation were classified as extrinsic or intrinsic. It was also observed that a user's tenure and social status in the community significantly affect the user's motivation to participate.

With respect to users' behavior analysis, Negoescu et al. performed pilot works by analyzing a Flickr group [34]. They proposed to represent and organize the Flickr groups using user-contributed photos and a collection of group tags. In [29], the authors conducted a comprehensive analysis of users' preferred behaviors on Flickr. The relationship between photos uploaded by users' socially-connected friends and their preferred actions was examined, and based on the results, a personalized image-recommendation solution was developed. In another work reported by Naaman's group, Ahern et al. proposed that there be emphasis on the privacy issues related to photo-sharing websites [2]. They analyzed the potential for identifying users' demographic and social networking information from the photo-sharing and annotation behaviors.

[2] http://en.wikipedia.org/wiki/ List of photo-sharing websites.
[3] http://www.globalwebindex.net/blog/internet-users-have-average-of-5-social-media-accounts.

Some implications on the design of future media-sharing applications were presented to improve protection against privacy breaches [35].

With regards to social-networking topology analysis, Mislove et al. conducted a large-scale link analysis between users on photo-sharing websites, and found that the constructed social networks follow a typical power-law distribution [33]. In [11], the authors investigated the information propagation range and speed on the Flickr social network, and they examined the extent to which the micro user-user information exchange contributes to the information propagation in macro networks.

From the above review, it is easy to deduce that most of the existing works related to photo-sharing website analysis were conducted on a single OSN, with the aim of summarizing and understanding the OSN's characteristics. In this article, we present a comparative study that aims to distinguish between different photo-sharing websites. Moreover, we further validate the discovered characteristics and conclusions by performing theoretical analysis and user modeling applications.

1.2.2. Multi-OSN comparative study

Another related area of research is on the area of multi-OSN comparisons. The success of WEB2.0 has led to the development of various social media services and websites. Recently, researchers started to compare different OSNs with respect to user behavior patterns, social networking topology, and their propagation characteristics.

With respect to the comparative study on user behavior patterns, Guo et al. investigated how users participate in and contribute to blogging, social-bookmarking, and question-answering OSNs [17]. With different user behavior patterns discovered in the three types of OSNs, this work laid out an analytical foundation for further multi-OSN user behavior analysis. In [12], the authors studied the motivational factors of users who participate in various social media conversations, and they observed that in different OSNs, users enjoy mixtures of intrinsic and extrinsic motivational factors. The authors of [42] examined users' social-networking activities and privacy settings across Facebook, Twitter, and Foursquare. The results showed that the user activities were highly correlated among these OSNs, with the potential for increased information leakage. Very recently, the authors in [8] addressed the problem of application design and development among multiple OSNs.

Regarding the social networking topology, Ahn et al. calculated the degree distribution, clustering co-efficiency, and degree correlation of Cyworld, Myspace, and Orkut [3]. They then summarized and compared the structural characteristics of the three OSNs. In [30], the traditional social network analysis (SNA) measures, such as the degree centrality and shortest path, were extended under the multi-OSN circumstances. Different microblogging OSNs were compared based on the extended measures. Buccafurri et al. recently proposed a view on user-connection analysis in a social internetworking scenario [7]. As a preliminary step, they also designed an effective solution to crawl the social graph underlying social internetworking systems [9].

Regarding the information propagation between OSNs, Leskovec et al. conducted pioneering studies on the cite correlation between different social blogging networks [24]. Large-scale blog linking and propagation graphs were constructed based on the observations. With respect to social blogging networks, extensive studies have examined the diffusion and evolution patterns of news media, multimedia sharing networks, and social-networking sites (SNS) [4,16]. Among these, in [23], cross-OSN diffusion was discussed together with macro user behaviors, and they examined the influence of user behaviors on information propagation on Twitter and Digg. Recently, Kim et al. presented an interesting work that measured three metrics, namely activity, reactivity, and heterogeneity, in order to understand the diffusion patterns between social media networks on the same trending topics [20].

To the best of our knowledge, there has so far been no work on analyzing different photo-sharing websites. More importantly, existing works along this line have performed data analysis and derived observations by sampling users from the whole population in each OSN separately. Because it is impossible to examine all the involved users, the sampling strategy is critical to the accuracy of data analysis results and conclusions. The non-representative user groups examined from different OSNs inevitably lead to sampling biases. In this work, we require that users examined from different OSNs have explicit correspondence to the same group of individuals, i.e., our comparative study is based on the overlapped users. This will effectively reduce the sampling biases and improve the confidence of comparative data observations. The statistical confidence of the overlapped user-based comparative study will be discussed in Section 4.

1.3. Overview and contributions

In this work, using *Flickr* and *Instagram* as the test platforms, we conduct a comparative study to distinguish between the two OSNs. Released in 2005 and originally designed for desktop users, Flickr is one of the first social media services that focus on online photo sharing and management. Because of its openness and professionalism, Flickr has remained popular up to the present. Instagram is considered a rising star in the photo-sharing family. It was released in 2010, and by focusing on mobile users, Instagram has acquired more photo uploads and registered users compared to Flickr within only ten months. We focus on the overlapped users between Flickr and Instagram, and we obtain their photo-sharing behavior data on each OSN. We perform a comparative study using various characteristics, including the photo-sharing temporal distribution, photo-capture location, social popularity, photo annotation, and cross-posting behaviors. From the data analysis results, we discuss the differences and similarities of the two OSNs at both macro and micro levels, and we try to understand the multi-OSN participation motivation simultaneously. From a theoretical perspective, we then analyze the advantages of exploiting the overlapped users for an OSN comparative study, and we examine the statistical significance of the overlapped user-based data observations. Finally, based on the different user behavior patterns and characteristics, we develop user

Table 1
Statistics of the collected Flickr_Instagram dataset.

	# user	#photo	Time period
Flickr	8,243	5,642,908	~ July 1st, 2015
Instagram		5,387,879	~ July 1st, 2015

modeling solutions by utilizing the behavior data from different OSNs. The experimental results validate that user data on different OSNs perform differently in terms of the prediction of user attributes.

The major contributions of this work are summarized as follows:

- We conduct a comprehensive comparative study of two photo-sharing websites, Flickr and Instagram. We examine and compare various characteristics of the photo-sharing behaviors at both the macro and micro level.
- The data analysis and comparative study is based on the same group of overlapped users. The overlapped user-based data observations are shown to be more statistically confident than those of independent subject groups.
- We apply the observations from the comparative study to user modeling applications. The user modeling experimental results validate the observations that the behavior data on respective OSNs perform differently in a specific user modeling task.

The rest of the article is organized as follows. In Section 2, we first introduce the dataset used for the comparative data analysis. Then, in Section 3, we present details of the comparative data analysis, observations, and conclusions. In Section 4, we theoretically analyze the statistical confidence of the overlapped user-based comparative study. In Section 5, we explain the photo-sharing behavior-based user modeling solutions and the experimental results obtained. Finally, in Section 6, we conclude the article by discussing some open research problems.

2. Data collection

To conduct the overlapped user-based comparative study, the first issue is the collection of multi-OSN user accounts corresponding to the same group of individuals. Currently, there are three ways of doing this: (1) Using a self-disclosed profile. An increasing number of persons are voluntarily disclosing their user accounts online, whether by filling in SNS registration information (such as Facebook, Google+) or maintaining aggregated profiles on services such as About.me and Friendfeed [43]. (2) Using shared user accounts. Many IT giants share identical accounts across different OSNs, or allow third-party services to access their user base, such as their Google account for YouTube and Google+, and Facebook's open platform. (3) Using data mining. Considering the trend where netizens are using a multitude of OSNs, many researchers are devoted to the problem of user account-linked identification, which has achieved satisfactory levels of accuracy [10].

In this work, we adopt the first method of collecting multi-OSN user accounts. *About.me* is a personal web hosting service that offers persons a one-page profile with which they can link their user accounts from different OSNs. We randomly collected a set of 180,000 registered About.me user profiles. By decoding the user profiles, we obtained their self-disclosed user accounts on different OSNs. We observed that a large number of users provided their accounts on photo-sharing websites, e.g., of the examined 180,000 users, the number of accounts provided on Instagram, Flickr, and Pinterest were 53,091, 26,570, and 16,500, respectively. This validates the popularity of different photo-sharing websites and the multi-OSN participation phenomenon.

To construct the overlapped user set, we selected 8243 active users who hold both Instagram and Flickr accounts. Using the Flickr and Instagram API, we searched for all of the publicly available data for these users before July 1st, 2015, and this resulted in a total of 5,642,908 Flickr photos and 5,387,879 Instagram photos. Table 1 summarizes the basic statistics of the collected dataset. In this work, we performed a comparative data analysis based on this *Flickr_Instagram* dataset.

3. Comparative data analysis

3.1. Photo-sharing temporal distribution

Timing is significant when describing behaviors in daily lives. The photo capturing and uploading time provides a way of understanding how users behave on each OSN. In this subsection, we compare the capturing and uploading time of the photos on Flickr and Instagram. Because the time information provided by Flickr and Instagram APIs both possess a Unix timestamp, we need to convert it to local time before performing the data analysis. On Flickr, the Unix timestamp is directly converted according to the uploader's time-zone information. On Instagram, for each photo, we first estimate the time zone according to the photo's geographical information, and we then convert the Unix timestamp to a local timestamp.

First, we examine the distribution over a 24-h period. Fig. 2 shows the distribution of a user uploading photos to Flickr and Instagram on weekdays (Monday to Friday) and weekends, respectively. We observe that: (1) both Flickr and Instagram exhibit a higher level of photo uploading from late afternoon to early evening, and a lower level from midnight to morning. This is consistent with personsÆ daily activities. (2) Flickr shows a similar temporal distribution over a 24-h period on

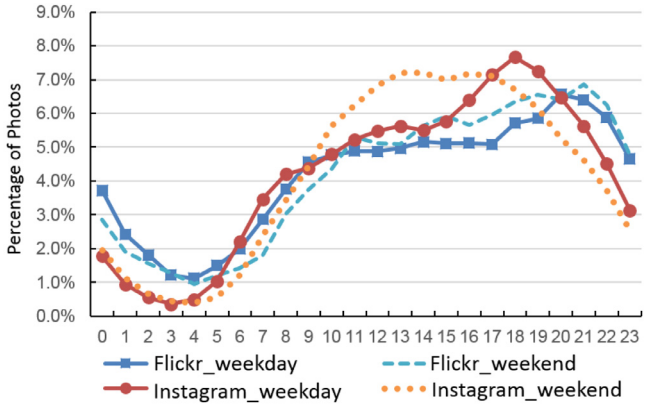


Fig. 2. Distribution of photo uploading time over a 24-h period.

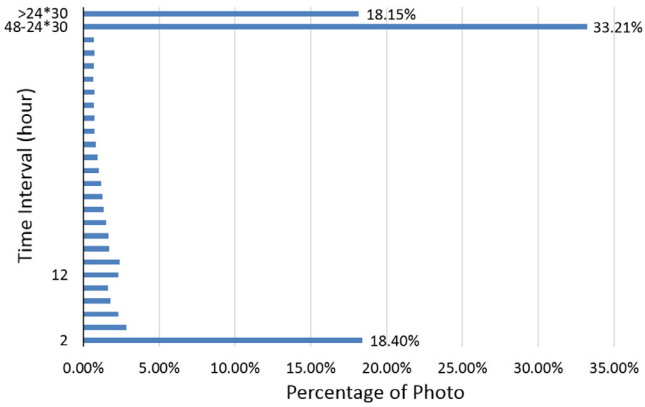


Fig. 3. The time intervals between photo capturing and uploading on Flickr.

weekdays and weekend, where the active time period is from 9 AM to 10 PM. However, on Instagram, the uploading peak on weekdays appears at around 6 PM, and on weekends, the uploading activity remains active over a longer period from 1 PM to 6 PM. (3) A further comparison of the temporal distributions of the two OSNs on weekdays showed that the peak uploading time on Instagram (around 6 PM, when people are getting off work) appears much earlier than that on Flickr (around 9 to 10 PM, before people go to bed). The above differences in the upload times of the two OSNs are largely related to the devices used by users and the movements of the users: compared with Flickr, Instagram is more mobile-oriented and thus influenced more by users' movement patterns. For example, the photo uploading peak on weekdays is primarily due to frequent phone usage while on the way home; on weekend afternoons, persons are more likely to do some shopping or get together with friends. Originally designed for desktop use, Flickr enjoys fewer mobile characteristics and has similar temporal distributions on weekdays and weekends.

In addition to the photo uploading time, the Flickr API also provides the photo capturing time. In Fig. 3, we illustrate the time intervals between the capture and upload timestamps on Flickr. We see that only 18.4% of the photos are uploaded within two hours of being captured, about one third of the photos are uploaded 2 $\sim$ 30 days after capturing, and one fifth of photos are uploaded more than 30 days after capturing. Combined with the photo-uploading distribution over a 24-h period, we observe that a significant percentage of Flickr photos are captured during the daytime and uploaded together at night. Moreover, users tend to share artistic photos such as photographic works on Flickr. This indicates that careful selection or even post-processing is necessary before uploading.

We further examined the distribution over a seven-day period for Flickr/Instagram photo uploading and Flickr photo capturing. Fig. 4 shows the respective distribution curves. We found that Flickr exhibits a steady photo-uploading temporal distribution, with no obvious difference between weekdays and weekends. It is interesting to note that the photo uploading on Instagram has a similar distribution as the photo capturing on Flickr; weekends show a higher level of activities and the peaks appear on Saturday. The above data observations demonstrate that the photo-uploading behavior is consistent with personsÆ photo-capturing behavior. Although the precise photo capture time is unavailable on Instagram, by combining the results from Figs. 3 and 4, we roughly conclude that Instagram enjoys a more obvious "capture & share" characteristic.

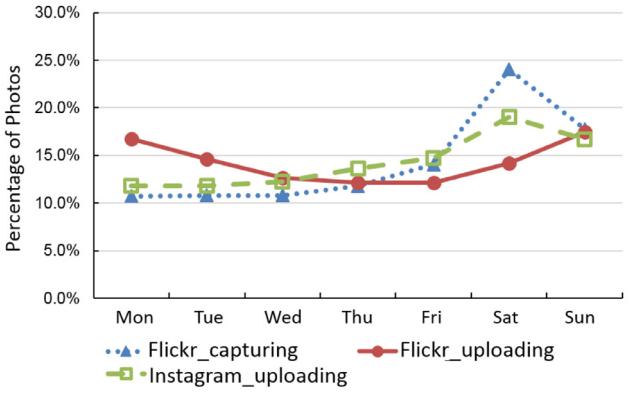


Fig. 4. Distribution of photo capturing/uploading time over a 7-day period.

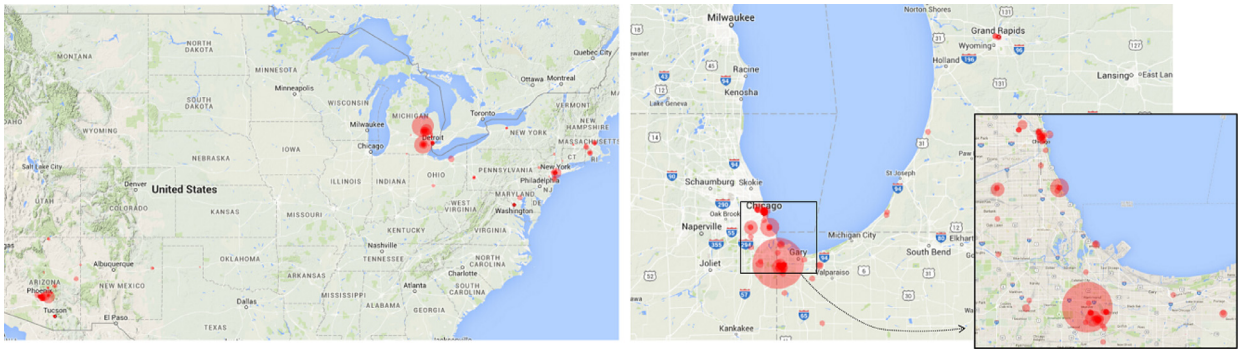


Fig. 5. The photo capturing location heat maps for the same sample user. **Left**: Flickr; **Right**: Instagram.

3.2. Photo capture location

In this subsection, we first examine the percentage of photos with available geographical information on the two OSNs. We calculated that of the 5,642,908 Flickr photos, 18.63% are associated with the geographical information, while on Instagram, 31.48% of the examined 5,387,879 photos contain publicly available geographical information. This may be because the Instagram client App is set to record and show the photo capturing locations by default unless the users selected special privacy settings. The fact that over half of the Instagram photos contains publicly available geographical information makes Instagram the best platform for analyzing users' movement patterns and developing location-based social media services.

We used a case study to distinguish between the photo capturing location patterns. For the same example user, Fig. 5 illustrates the heat map of his/her photo-capturing locations on Flickr and Instagram, respectively. The size of the red circle is proportional to the number of photos captured at the circle center. The results show that the user's shared photos on Flickr have a relatively scattered spatial distribution with several hubs, from Phoenix in the western United States, central cities of Detroit and Cleveland, to the eastern states of Washington and New York. By examining the photo capturing time, we found that photos with close capture locations also have a close capture time. This indicates that most of these Flickr photos may have been taken during business or vacation trips. However, this user's shared photos on Instagram are more concentrated, are mainly located around Chicago, and have a very small spatial overlap with the Flickr photos. From the enlarged illustration on Fig. 5(Right), we can see that the photos shared to Instagram are naturally clustered around two centers, i.e., downtown Chicago and the Hammond located in the southeast of Chicago. A possible interpretation of this observation is that the two centers correspond to the user's workplace and home, which indicates that this user is likely to share photos to record his/her daily life on Instagram.

To quantitatively compare the same user's photo-capturing location patterns, we introduced geographical information entropy to estimate the scattered degree of the photo-capturing locations. The geographical information entropy is calculated as follows. First, the world map is divided into many squares, with the length and width of each square being equivalent to 0.1 in longitude and latitude. Each resultant square spans an area of approximately 100 km^2. Then, the user-uploaded photos are assigned to different squares according to their geographical coordinates. Assuming the Flickr photos uploaded by user A are assigned to K squares, we formally define the geographical information entropy as follows:

$$\text{GeoEntropy}(A) = -\sum_{i=1}^{K} \frac{N_i}{N_A} \log \frac{N_i}{N_A} \tag{1}$$

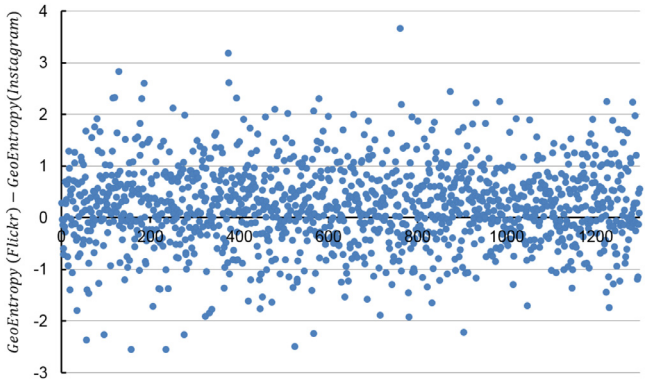


Fig. 6. The user scattered diagram for the difference between geographical information entropy between Flickr and Instagram.

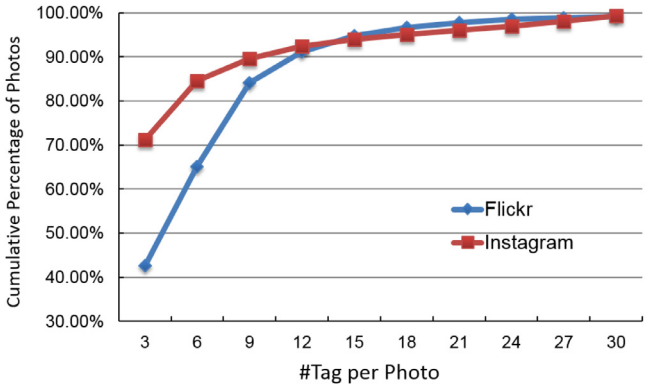


Fig. 7. The cumulative photo distribution regarding the number of tags.

where N_i, N_A indicate the number of photos assigned to the ith square and the total photos shared by user A on Flickr with geographical information, respectively. The *GeoEntropy* of user A on Instagram can be calculated in the same way. The larger the value of *GeoEntropy*, the more scattered will be the photo-capturing locations. We randomly selected 1299 overlapped users who have shared more than 50 photos with available geographical information on both Flickr and Instagram. Their *GeoEntropy* values on both Flickr and Instagram are calculated based on Eq. (1). In Fig. 6, we plot the scattered diagram of the *GeoEntropy* difference between the two OSNs: *GeoEntropy*(Flickr) - *GeoEntropy*(Instagram). According to the definition of *GeoEntropy*, if the difference is larger than 0, the user's photos on Flickr are more scattered than those on Instagram. Fig. 6 shows that the scatters representing the majority of users are located on the upper half quadrant. Of the 1299 examined users, 62.7% have a *GeoEntropy* difference larger than 0, and the average difference value is 0.215. This result is consistent with the case study illustrated in Fig. 5; compared with Flickr, the same individual has a more concentrated photo-capturing spatial distribution on Instagram.

3.3. Photo annotation

In this subsection, we study the differences and similarities on the photo annotations between Flickr and Instagram. On a macro level, we first calculated the number of associated tags per photo on each OSN, and we plot the cumulative photo distribution in Fig. 7. The results show that the overlapped users show different distributions on the two OSNs; about 40 and 70% of the Flickr and Instagram photos contain no more than three tags, respectively. The average number of tags per photo on Flickr is 5.45, and on Instagram it is 3.63. The lower tag number on Instagram is closely related to its mobile-oriented characteristic, i.e., the limited size of the screen/keyboard and the inconvenience associated with typewriting on-the-move.

On a micro level, we examined the respective number of tags that each overlapped user added to the uploaded photos on Flickr and Instagram. A total of 8189 overlapped users uploaded photos to both Flickr and Instagram associated with self-contributed tags. For each of the 8189 users, we calculated the average number of tags per photo on Flickr, denoted as *Ave_Tag*(*Flickr*), and on Instagram, denoted as *Ave_Tag*(*Instagram*). Fig. 8 illustrates the scattered diagram of the *Ave_Tag* difference values between Flickr and Instagram. The calculations show that 74.2% of the overlapped users are located on the upper half quadrant, with the average difference value as 1.39. This indicates that at the micro level, most of the users add more tags per photo on Flickr than on Instagram. We interpret this observation to mean that Flickr was originally designed as a tool for online photo management, and is thus more focused on encouraging users to add annotations for organization

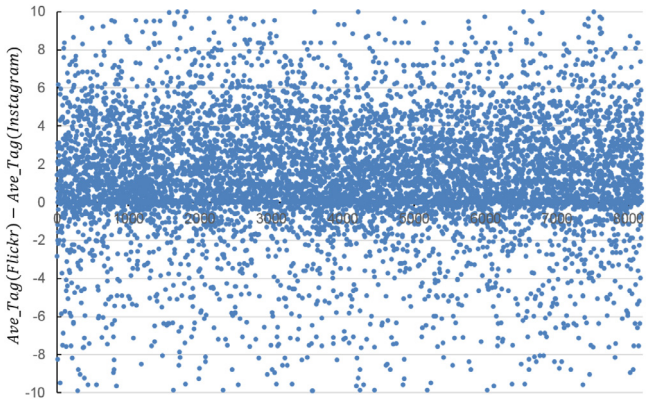


Fig. 8. The user scattered diagram for the difference of *Ave_Tag* between Flickr and Instagram.

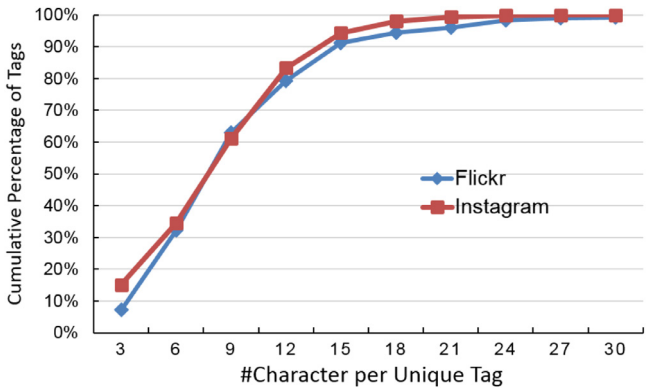


Fig. 9. Cumulative tag distribution regarding the number of characters.

and search purposes. However, Instagram is more focused on photo sharing, where the goal annotation is mainly for photo-content description in order to attract the attention of others.

In addition to the number of tags per photo, we also calculated the number of characters per unique tag. The cumulative tag distribution regarding the number of characters is shown in Fig. 9. Flickr and Instagram show a similar cumulative distribution; the percentage of unique tags with no more than six characters on Flickr and Instagram are 32.1 and 34.5%, respectively, and the percentage of tags with no more than nine characters are 62.9 and 61.2%, respectively. The average number of characters per tag on Flickr and Instagram is 9.0 and 7.3, respectively. This indicates that there is a common characteristic for photo-sharing websites in that users tend to use simple and short tag words to describe their shared photos. This tendency on Instagram is more obvious because of the device limitation and the usage environment.

We further examined the addressed topics of the shared photos to Flickr and Instagram. We canonicalized each user's annotated photo tags into a text document representation. We employed a statistical topic model, the Latent Dirichlet Allocation (LDA), to analyze the resultant text corpus. LDA models documents and their vocabulary in the same space by clustering similar documents and words together based on their co-occurrence. Within the context of this work, each user's tag textual representation on either Flickr or Instagram corresponds to one document, and the unique tags that appear in either Flickr or Instagram photos are used to construct the vocabulary. After removing the stop words and infrequent tags, the final vocabulary consists of 1,300,000 unique tags. A total of 5675 overlapped users have annotated tags on both OSNs and constitute the topic modeling corpus (i.e., $5,675 \times 2 = 11{,}350$ documents). We tuned LDA parameters to achieve the lowest perplexity and found that $T = 20$ topics worked well. We selected the document-topic and topic-word hyperparameters as a relative small value of $\alpha = 1$ and $\beta = 0.05$, and each document (user) is encouraged to have a dominant document-topic distribution to facilitate the subsequent micro analysis. Then, we normalized and aggregated the derived user topical distributions in order to obtain the OSN distributions over the 20 discovered topics, which are illustrated in Fig. 10. It is shown that Flickr and Instagram enjoy quite similar topical distributions, yet with slightly different dominant ones. Table 2 lists the dominant topics for Flickr and Instagram represented by the five most probable tags. Flickr exhibits a high proportion of tags on the landscape related topics, while Instagram has a major focus on the food topic. These observations are generally consistent with those from temporal-spatial comparative studies.

Next, we analyze the micro difference of the addressed topics in the two OSNs. For each of the 5675 examined overlapped users, we calculated the cosine similarity of his/her topical distributions on Flickr and Instagram. Fig. 11 plots the

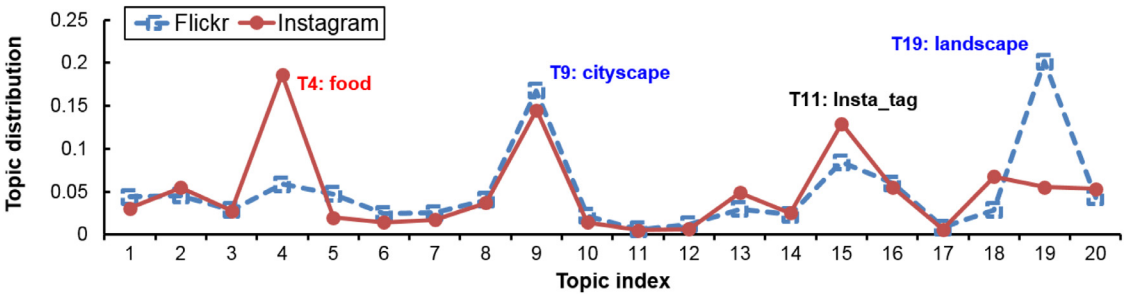


Fig. 10. Topic distributions based on photo annotation (some topics are manually labeled based on the most probable tags).

Table 2
Dominant topics for Flickr and Instagram.

OSN	Topic index	Most probable tags
Flickr	#9	park city street building lake
	#11	instagood instamood photooftheday picoftheday instadaily
	#19	photo beach sky nature sunset
Instagram	#4	food eat make cream dinner
	#9	park city street building lake
	#11	instagood instamood photooftheday picoftheday instadaily

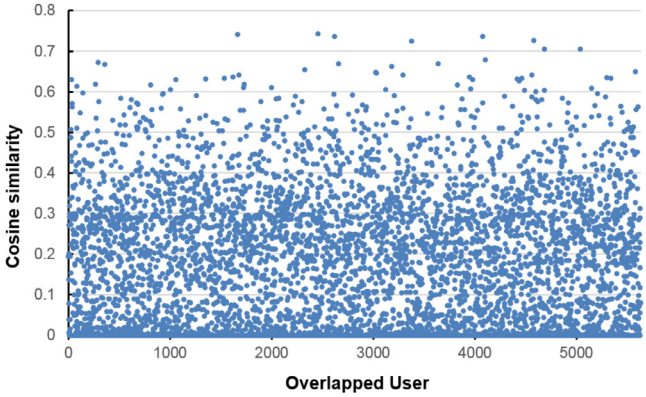


Fig. 11. The scattered diagram for the cosine similarity of overlapped users' topical distributions on Flickr and Instagram.

scattered diagram of the calculated cosine similarity scores. The figure shows that the topical distributions of the same overlapped user on Flickr and Instagram are significantly different from each other. The average similarity score is very low, i.e., only 0.165. This is very different from the consistent topic distributions observed on the macro analysis results. A possible explanation for this is that the individual's behavioral difference on the two OSNs is concealed by the aggregated behaviors at the macro level. From this comparison data, we observed that although Flickr and Instagram show similar topical distributions, on the micro level, an individual user may have a very different focus on the two OSNs. Note that only by comparing the same individual's coupled behaviors (i.e., overlapped user in this work) can we discover this type of difference.

Discussion. Next, we consider Table 2. In addition to *Topic 9*, which is cityscape-related, Flickr and Instagram share another dominant topic, i.e., *Topic 11*. We find that many probable tags of *Topic 11* are Instagram oriented. Instagram predefines several hashtags to assist users to annotate photos efficiently. For example, "Instagood" is a hashtag for expressing the feeling of like, "Instamood" is used to annotate photos that capture the essence of nature or beauty of everyday life. This helps to explain their prevalence on Instagram, but yet they are also popular on Flickr. Determining the reason for this requires us to consider a very interesting cross-OSN posting behavior.

Informal cross-OSN posting indicates the behavior when a user posts his/her document multiple times on different OSNs. The observation of Instagram hashtags as a dominant topic on Flickr is one manifestation of this behavior. In cross-OSN posting, there exists a source OSN (where the document originates) and a sink OSN (on which the document is shared). We conducted data analysis to investigate the source/sink role between Flickr and Instagram. The photos cross-posted from Instagram to Flickr are easily identified as having an automatically added mark of "uploaded:by = instagram." In this way, of the 5,642,908 Flickr photos, we identified 751,647 that were cross-posted from Instagram. To identify the photos cross-posted from Flickr to Instagram, we built a photo-sharing timeline for each overlapped user by chronologically organizing

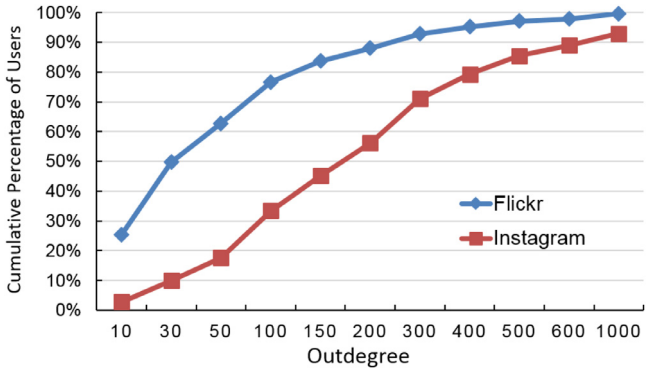


Fig. 12. The cumulative user distribution regarding outdegree.

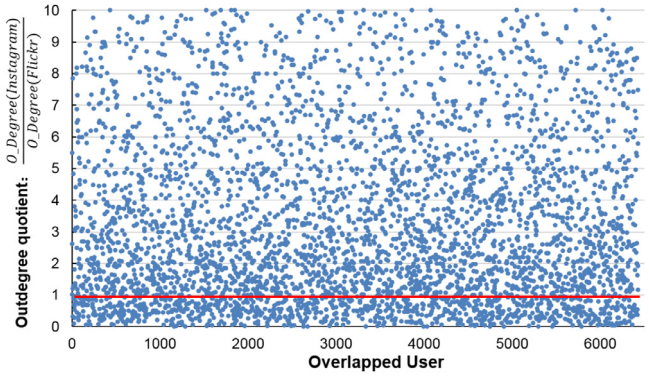


Fig. 13. The user-scattered diagram for the quotient of the outdegree of Instagram and Flickr.

his/her shared photos on Flickr and Instagram. The cross-posted photo from Flickr to Instagram is identified if: (1) the identical photo was shared by the overlapped user on both Flickr and Instagram, and (2) the sharing timestamp on Flickr is earlier than that on Instagram. Based on the results, no photos were identified in this way. A possible interpretation of this is that Instagram acknowledges original content because of its efficient mobile access, while Flickr serves more as a collection for beautiful and artistic photos.[4]

3.4. Social popularity

This subsection examines the social popularity of the overlapped users on Flickr and Instagram by analyzing the number of social friends and the social endorsements received by the photos.

The social link is directed on both Flickr and Instagram. The number of outward social links, i.e., the outdegree, reflects the social popularity of the users within the social network. We first examined the outdegree at the macro level. The distribution over 6434 overlapped users who have outdegree values larger than 0 on both OSNs is calculated and plotted in Fig. 12. Flickr and Instagram have very different outdegree distributions: the percentage of users with an outdegree of no more than ten is 27.6 and 3.1%, respectively, and the percentage of users with an outdegree of no more than 50 is 76.7 and 17.6%, respectively. The average outdegree is 84.2 and 302.3 on Flickr and Instagram, respectively. We can therefore conclude that the overlapped users have much denser social interactions with other users on Instagram than on Flickr.

On the micro level, we examined the outdegree difference of the two OSNs for each user. We calculated the quotient between the same user's Instagram outdegree and Flickr outdegree. The quotients of the 6434 overlapped users are obtained in the form of a scattered diagram and shown in Fig. 13. When the quotient equals 1, the overlapped user has the same outdegree on Instagram and Flickr. The larger the quotient, the larger will be the outdegree that the user has on Instagram. From Fig. 13, we can see most of the scatters located above the horizontal line $y = 1$. 84.9% of the 6434 users have larger outdegree on Instagram than on Flickr. This is consistent with the observation from Fig. 12.

Because users on both Flickr and Instagram interact with each other with photos, the level of social endorsements obtained by the photos also reflects the uploader's social popularity. We measure the level of social endorsements by the

[4] One of the future works here is to validate this assumption by conducting semantic and aesthetic analysis on the photos cross-posted from Instagram to Flickr.

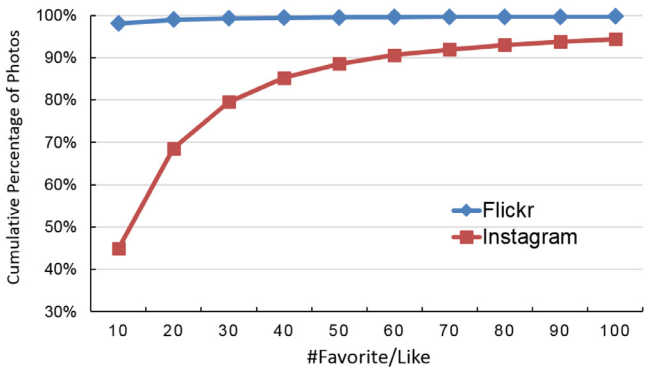


Fig. 14. The cumulative photo distribution regarding social endorsement.

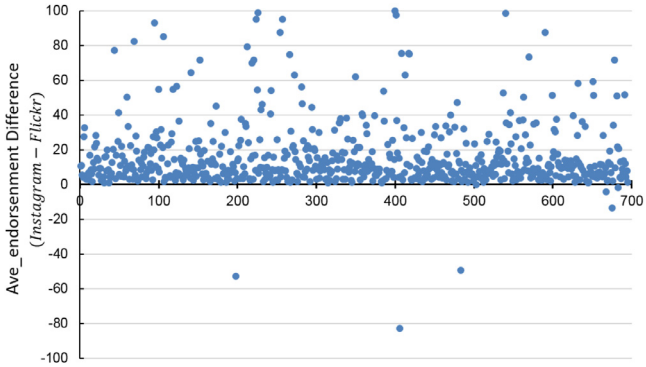


Fig. 15. The user scattered diagram for the difference in the average social endorsement per photo.

number of "Favorites" on Flickr and "Likes" on Instagram. On the macro level, Fig. 14 shows the cumulative photo distribution regarding the number of social endorsements. Flickr and Instagram show very different distributions: on Flickr, over 98% of the photos receive "Favorites" no more than ten times, while on Instagram over half of the photos receive "Likes" more than ten times. The average number of "Favorite/Likes" received per photo is 5.78 and 22.62 for Flickr and Instagram. Although the indications represented via "Favorite" and "Like" are not exactly the same, the large difference in quantity is adequate to demonstrate the superiority of Instagram in social interaction.

On the micro level, to examine the different levels of social endorsement received by each user between the two OSNs, we randomly selected 695 overlapped users who have shared more than 100 photos to both Flickr and Instagram. For each of the 695 users, we calculated his/her average number of "Favorites/Likes" received per photo. The difference values (*Instagram* – *Flickr*) are drawn on a scattered diagram and shown in Fig. 15. We can see that only six (0.86%) users have a difference value smaller than 0. Almost all the users receive more photo social endorsements on Instagram than on Flickr. From the above observations, it is recommended that users share their photos on Instagram if more attention and endorsements are expected.

Based on the comparative study of different users' characteristics on both a macro and micro level, we identified significant differences between Flickr and Instagram. This helps to interpret the multi-OSN participation phenomenon on the two photo sharing websites. The differences were that users have different purposes and exhibit different behavior patterns. This is consistent with the OSN focus. Flickr is photo-centric and focused more on the photo content, while Instagram is social interaction-centric with photos serving as the medium. The fact that Instagram caters to the trend of social interaction contributes greatly to it rapidly gaining popularity among young people and its globally explosive growth.

4. Statistical confidence discussion

In this section, we discuss the advantage of leveraging the overlapped users in comparative studies between OSNs. In most social media-oriented data analyses, examining the whole population is impossible as the number of involved users is very large. Most researchers therefore utilize a sampling-based data analysis for potentially useful observations. Because our data analysis in the previous section is also sample-based, we will investigate the advantage by analyzing the sampling statistical confidence.

4.1. Error of sampling-based data analysis

In statistical theory, the quality of a sampler is related to two types of errors [1,25]:

$$\text{Total error} = \text{sampling error} + \text{nonsampling error}. \tag{2}$$

Of the two types of errors, sampling errors arise solely as a result of randomly selecting a sample instead of canvassing the whole population, and this will be effectively eliminated by increasing the number of samples. Nonsampling errors (also called systematic errors) are mainly due to adopting incorrect procedures during data collection or processing, which do not go away as the sample size increases. Nonsampling errors arise where randomization is not achieved, i.e., the sample obtained is not representative of the whole population to be analyzed.

In social media-oriented data analysis, it is difficult to achieve randomization during sampling [13]. The analyzed users are usually selected by setting a seed user and crawling the related users using the depth-first or breadth-first strategy. Moreover, because the sampling is based only on publicly available data, users who have made their data private are excluded. This leads to the inevitable non-coverage issue, which is an important source of nonsampling errors [6]. For example, on a specific OSN, if users who voluntarily make their data publicly available tend to have more friends than those who make their data private, the observations of the friend number analysis considering only public data are largely biased, and thus not valid. Simply increasing the sample size cannot help improve the statistical confidence.

This situation becomes worse when conducting comparative data analysis between OSNs. Assuming the goal is to compare the characteristic X (e.g., the average outdegree) between *OSN1* and *OSN2*, we associate the observed samples on two OSNs with random variables X^1 and X^2, respectively. Conducting a comparative analysis is equivalent to examining the difference between the two variables, i.e., $Y \triangleq X^1 - X^2$. Given the variance of X^1 and X^2, we calculated the variance of Y as[5]:

$$Var(Y) = Var(X^1) + Var(X^2) - 2Var(X^1)Var(X^2)Cov(X^1, X^2) \tag{3}$$

where $Cov(\cdot,\cdot)$ indicates the covariance between two variables. When the samples on *OSN1* and *OSN2* are independent, i.e., $Cov(X^1, X^2) = 0$, the sampling variance of the difference will be the sum of the respective sampling variances: $Var(Y^{independent}) = Var(X^1) + Var(X^2)$. For example, when comparing the social popularity of users on Flickr and Instagram, if a socially active user group is selected on Flickr and a socially inactive user group is selected on Instagram, the derived data observations will be significantly biased, and may generate contrasting conclusions.

4.2. Overlapped user-based comparative study

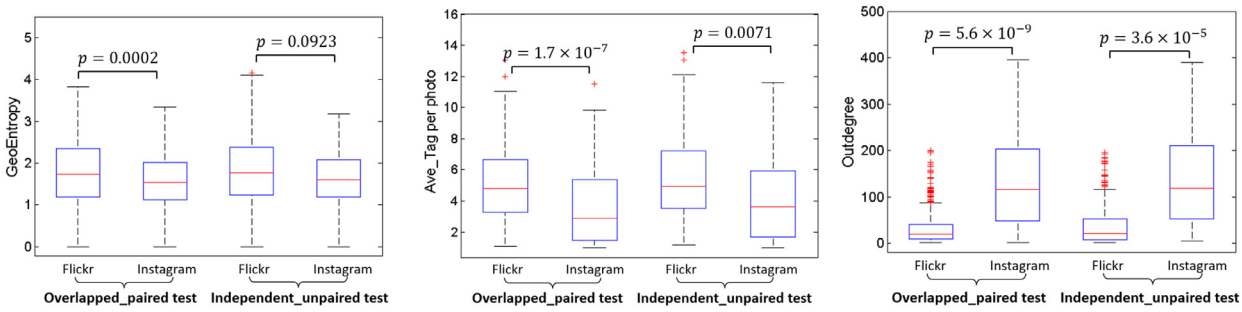
In this work, we conducted the multi-OSN comparative study based on overlapped users, i.e., selected samples on different OSNs are perfectly paired to the same group of subjects. We discuss how this contributes to the improved statistical confidence of sampling-based observations. First, reviewing Eq. (3), the random variables X^1 and X^2 that record certain characteristics of the same subject group are heavily correlated. Therefore, the covariance between the two variables $Cov(X^1, X^2) > 0$. We can easily have the following relation:

$$Var(Y^{independent}) > Var(Y^{overlapped}) \tag{4}$$

where $Y^{overlapped}$ indicates the variable recording the differences between the two OSNs for a specific characteristic of the overlapped users. The more correlated are X^1 and X^2, the more accurate will be the estimation that we can obtain of the difference variable $Y^{overlapped}$. This can be understood by re-examining the social popularity example mentioned in the previous subsection. We elaborate on this as follows. About.me users are generally highly motivated to perform social interactions and propagate themselves, making the overlapped users from About.me inevitably biased to represent the whole population. This can be clearly shown from the fact that the average outdegree calculated in the collected user set is much higher than that in literature [11,36]. However, with the explicit correspondence to the unique overlapped users, sampled users occupy the same position within the full population of the two OSNs, e.g., they will represent the socially active group on both Flickr and Instagram. In this case, although separate data analyses of each OSN are subject to bias, the observed differences on the overlapped users between the two OSNs are sufficient to guarantee a promising comparative study.

Further, we quantitatively examine the advantage of leveraging overlapped users in the comparative study. Because the focus of the comparative study is to highlight the differences between two sets of samples, it is appropriate to utilize the statistical significance test to examine the confidence of the data analysis. Specifically, the overlapped users' specific characteristic values on different OSNs can be viewed as the observed data samples in two trials. The goal of the statistical significance test is to determine whether the samples in the two trials are significantly different, i.e., whether the difference is significant enough to reject the null hypothesis. In the context of overlapped users, the tested samples are paired and subject to one special type of statistical significance test called the paired test [32]. Moreover, because it is reasonable to assume that the values of each characteristic in the *Flickr_Instagram* dataset follow the normal distribution, we utilize the paired t-test for our statistical significance analysis.

[5] The differences between OSNs can also be defined as the quotient of random variables, i.e., $Y \triangleq X^1/X^2$. In this case, although no closed formula exists for the variance of Y, an approximation can be made to obtain a similar conclusion with Eq. (3) [19].



(a) GeoEntropy (b) Average number of tags per photo (c) Outdegree

Fig. 16. Results of statistical significance test.

We investigated a subset of our collected overlapped users for the statistical significance test. Specifically, we constructed two scenarios for comparison:

- **Overlapped_paired test**: We randomly selected 300 overlapped users, and we examined the differences in their characteristic values on Flickr and Instagram using the paired *t*-test;
- **Independent_unpaired test**: We selected 300 users independently from Flickr and Instagram (i.e., without correspondence to the unique overlapped user), and we examined their characteristic values using the unpaired *t*-test [31].

We conducted a statistical significance test on the above two scenarios regarding three types of characteristics, *GeoEntropy*, *Ave_Tag* per photo, and *Outdegree*. The results are summarized in Fig. 16, where the *p*-value indicates the probability of observing an effect given that the null hypothesis is true.[6] The smaller the *p*-value, the more confidently will the null hypothesis be rejected. It is shown that although similar distributions (encoded by the boxplot) are observed, the *p*-value calculated from the *overlapped_paired test* is much smaller than that from the *independent_unpaired test*. This demonstrates that the differences observed from the overlapped users are more confident than those from the independent sample groups. Actually, by comparing the same overlapped user's characteristic values on Flickr and Instagram, we effectively utilize each subject as its own control. The statistical significance of the data analysis thus increases as the random between-subject variation has now been eliminated.

The above discussion fixes the sample size and examines the significance levels that can be attained in the two scenarios. We can also set an expected significance level and conduct the comparisons by estimating the required minimum sample size in the respective scenario. In statistical theory, the relation between the sample size and the statistical confidence is approximated in the following equations [14]:

for independent sample groups,

$$n = \frac{\sigma^2 (t_{n(m+1)-2,\alpha/2} + t_{n(m+1)-2,\beta})^2}{\delta^2} \tag{5}$$

for paired sample groups,

$$n = \frac{\sigma^2 (t_{n-1,\alpha/2} + t_{n-1,\beta})^2}{\delta^2} \tag{6}$$

where n is the minimum sample size, m is the ratio of control to subjects, σ is the standard derivation of the characteristic variable, δ indicates the mean of the difference, $t_{a,b}$ is the *student t* score with a degrees of freedom (DoF) and probability b, α represents the desired level of significance (typically $\alpha = 0.05$), and β represents the desired statistical power (typically $\beta = 0.2$, $power = 1 - \beta = 0.8$).[7] In the above two equations, because n appears on both sides, it is not easy to directly estimate the sample size by fixing the other arguments. For ease of comparison, additional simple equations can be utilized to approximate the calculation of the sample size as follows [21]:

for independent sample groups,

$$n = \frac{2\sigma^2 (Z_{\alpha/2} + Z_\beta)^2}{\delta^2} \tag{7}$$

for paired sample groups,

$$n = \frac{\sigma^2 (Z_{\alpha/2} + Z_\beta)^2}{\delta^2} \tag{8}$$

[6] Formally, a null hypothesis is rejected if the calculated *p*-value is less than the significance or α level.

[7] The α-level is related to the *p*-value that was discussed above, and it measures the probability of rejecting a true null hypothesis (called a type I error). β measures the probability of failing to reject a false null hypothesis (called a type II error). Therefore, in this discussion, we examine the advantage of utilizing an overlapped user to improve the statistical confidence by considering both type I and type II errors.

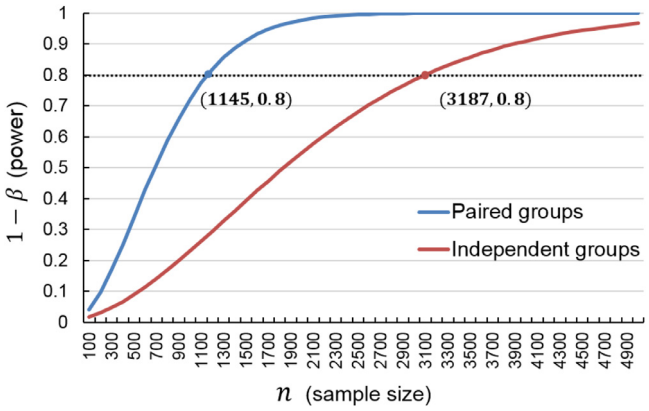


Fig. 17. The power of the statistical test regarding the sample size.

where Z_b indicates the Z score with probability b. We can see that the approximated number of samples required for independent groups is two times that of the paired groups.[8]

We utilized Matlab toolbox to examine the relation between the power of the statistical test and the sample size, which solves Eqs. (5) and (6) using root-finding methods [18]. Using the users' *GeoEntropy* values in our collected dataset as an example, the examined results are plotted in Fig. 17. The results show that to reach a power level of 80%, the number of samples for paired groups and independent groups are 1,145 and 3,187, respectively. This demonstrates that the paired statistical analysis is more powerful with fewer samples to prove a given difference between the study groups; this further validates the advantage of utilizing the overlapped users for the comparative study.

5. Photo-sharing behavior-based user modeling

In addition to the advantage of improving the statistical confidence, the overlapped user-based comparative study also yields many inspiring observations, especially from the micro-level data analysis. For example, in the photo-capturing location analysis, we observed that the same individual's shared photos show very different spatial distributions on Flickr and Instagram. The shared photos on Instagram are spatially focused and distributed, possibly around working and living places, while the shared photos on Flickr are relatively scattered and mostly captured during vacation and business trips. These observations directly provide useful instructions for many applications, such as photo-sharing behavior-based user modeling and personalized services. Two examples are: (1) behavior data on Instagram deserve more attention when modeling location-related user attributes; and (2) behavior data on Flickr are expected to contribute more to user interest modeling where more descriptive information is available. To validate the derived data observations in actual applications corresponding to the above two examples, in this section, we utilized the photo-sharing behaviors to address two specific user modeling tasks of living-city estimation and occupation inference.

5.1. Living-city estimation

Data preparation We utilized a different user set for user modeling. Because most Flickr and Instagram user profiles do not have detailed living location information, we turn to another OSN, LinkedIn, to retrieve the ground-truth for living-city estimation. Specifically, we selected 1500 About.me users who share their accounts on Flickr, Instagram, and LinkedIn. Of the 1500 overlapped users, 783 users were retained to construct the dataset for living-city estimation that: (1) have available living location information (at the city-level) on LinkedIn; (2) shared more than 50 photos with available geographical information on both Flickr and Instagram.

Methodology We considered two simple approaches to estimate the users' living city from the shared photos' geographical information. The first approach is directly *majority voting*-based [38]. For each examined user, we assigned each of the shared photos to a city by issuing the photo's GPS coordinates to Yahoo's geo-coding API, *Yahoo! PlaceFinder*.[9] The user's living city is then estimated as the most popular one among the shared photos using majority voting. The second approach is a *clustering*-based [22] approach. For each examined user, we utilized agglomerative clustering, which starts by treating each photo's GPS point as its own cluster. Cluster merging continues until all the cluster centers are at least 100 km apart. The resultant cluster with the most shared photos is considered as the living area. The living city is obtained by issuing the GPS coordinates of the cluster center into *Yahoo! PlaceFinder*.

[8] Note that many approximations are made here to obtain the above equations, e.g., the two independent groups are required to have the same σ. The actual ratio between the sample sizes needed in the two scenarios are subject to change in different tasks.

[9] https://developer.yahoo.com/boss/placefinder/.

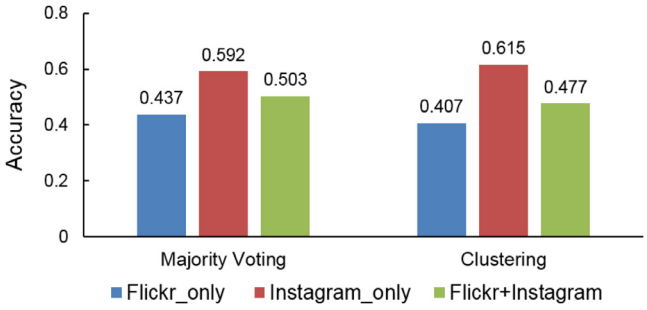


Fig. 18. Accuracy in living-city estimation.

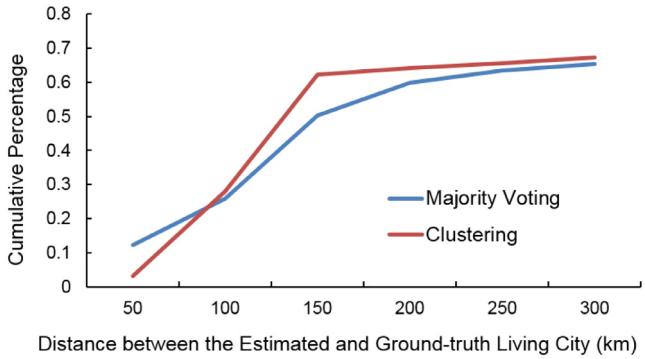


Fig. 19. Cumulative distribution regarding the distance between the estimated and ground-truth city center.

Experimental results To differentiate between the contribution of users' photo-sharing behaviors and user modeling tasks on the two OSNs, we considered the following three experimental settings:

- *Flickr_only*: utilizing only the overlapped user's shared photos on Flickr for majority voting or clustering;
- *Instagram_only*: utilizing only the shared photos on Instagram;
- *Flickr+Instagram*: considering the shared photos both on Flickr and Instagram.

Fig. 18 shows the accuracy of living-city estimation based on both approaches. We find that both approaches produce consistent results in that we obtained the greatest accuracy when we utilized only the shared photos on Instagram. It appears that the combination of Flickr and Instagram photos did not negatively affect the accuracy, and this may be because the combination of Flickr photos deviates the photo geographical distribution from the ground-truth city. To better understand the results, for each incorrect case in the *Instagram_only* setting, we examined the distance between the estimated and ground-truth living-city center. Fig. 19 shows the cumulative distribution of these distances. The results show that over 50% of the distances are under 150 km. Combining this result with the correctly estimated ones, both simple approaches can estimate the user living city within 150 km with an accuracy of 80%, and this was done by utilizing only the geographical information obtained from shared photos on Instagram.

5.2. Occupation inference

Data preparation Similarly, in the living-city estimation task, we used LinkedIn to obtain the ground-truth occupation information for the examined users. Among the 1500 overlapped users, we selected 1046 users that: (1) have *industry* information on LinkedIn; (2) have shared more than 20 photos to both Flickr and Instagram associated with self-contributed tags. By examining the available industry of the 1046 users, we identified four types of popular industry attributes, such as *Marketing and Advertising*, *Research*, *Human Resources*, and *Design*. Finally, we used 868 users with the above four types of industry information to construct the dataset for occupation inference.

Methodology Occupation inference is essentially a multi-class classification problem, i.e., classifying users into one of the following classes: *Marketing and Advertising*, *Research*, *Human Resources*, or *Design*. To address this problem, we also consider two simple supervised solutions. The first is *topic*-based, which involves directly using the derived latent distribution from topic modeling in Section 3.C as the user feature. We trained the multi-class SVM classifier on the topic-based feature and utilized it for occupation inference. The second solution is *fusion*-based. For each user, we extracted the tag unigram feature and used it to train a unigram-based SVM classifier. The topic-based and unigram-based classifiers are then fused by stacked-SVM [39].

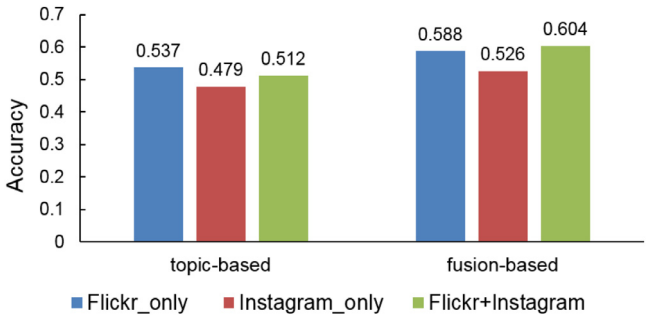


Fig. 20. Accuracy in occupation inference.

Experimental results Within the 868 selected overlapped users, half of them are used for training and the other half for testing. For this task, we considered the same experimental settings. Fig. 20 shows the accuracy of occupation inference based on the two introduced solutions. The figure shows that the photo annotations on Flickr are more descriptive and informative, enabling us to understand user interests and infer user occupation. As opposed to the results obtained from living-city estimation, a combination of Flickr and Instagram photo annotations achieves a comparable performance with a single OSN. Moreover, the fusion of different classifiers contributes to realizing a significant performance improvement for all three settings.

Note that with respect to the task of living-city estimation and occupation inference, recent works have proposed complex approaches with more features such as temporal patterns and photo visual content [15,27]. However, in this article, the goal of user modeling is not to realize state-of-the-art performance, but to differentiate the contribution of user behavior data on respective OSNs. We believe that simple approaches are more effective in assessing the potential of data on a specific OSN towards the user-modeling tasks.

6. Conclusions

We exploited the overlapped users to conduct extensive comparative data analysis between Flickr and Instagram, with regards to the characteristics of spatial-temporal distribution, photo annotation, and social attributes. We validated the confidence of overlapped user-based comparative data observations using both statistical theory and user-modeling applications. It is interesting to note that the derived observations also shed some light on the design of the two OSNs: (1) Flickr is proposed to improve user participation by encouraging more social-related features in user-user as well as user-photo interaction; (2) To make up for the limited semantic information, it will be helpful if Instagram adds functions such as tag recommendation or auto-annotation, and improves its support on search and content organization; (3) Publicly available behaviors on photo-sharing websites potentially reflect users' attributes such as home location and occupation. Privacy-preserving photo sharing will receive more and more attention.

In the future, we aim to focus on three directions: (1) investigate the interplay between characteristics, e.g., # tag per photo versus its social popularity, photo topics versus cross-OSN posting behavior, (2) conduct more theoretical analysis into the advantage of overlapped user-based comparative studies and test the conclusions in other OSNs, and (3) apply the derived observations in more applications, e.g., examine overlapped users' social popularity in the physical world.

References

[1] A. Achar, P. Sastry, Statistical significance of episodes with general partial orders, Inf. Sci. 296 (2015) 175–200.
[2] S. Ahern, D. Eckles, N. Good, S. King, M. Naaman, R. Nair, Over-exposed?: privacy patterns and considerations in online and mobile photo sharing, in: Proceedings of the 2007 Conference on Human Factors in Computing Systems, 2007, pp. 357–366.
[3] Y. Ahn, S. Han, H. Kwak, S.B. Moon, H. Jeong, Analysis of topological characteristics of huge online social networking services, in: Proceedings of the 16th International Conference on World Wide Web, 2007, pp. 835–844.
[4] T. Althoff, D. Borth, J. Hees, A. Dengel, Analysis and forecasting of trending topics in online media streams, in: Proceedings of the 21st ACM international conference on Multimedia, 2013, pp. 907–916.
[5] M. Ames, M. Naaman, Why we tag: motivations for annotation in mobile and online media, in: Proceedings of the 2007 Conference on Human Factors in Computing Systems, 2007, pp. 971–980.
[6] P.P. Biemer, L.E. Lyberg, Introduction to Survey Quality, 335, John Wiley & Sons, 2003.
[7] F. Buccafurri, V.D. Foti, G. Lax, A. Nocera, D. Ursino, Bridge analysis in a social internetworking scenario, Inf. Sci. 224 (2013) 1–18.
[8] F. Buccafurri, G. Lax, S. Nicolazzo, A. Nocera, A model to support design and development of multiple-social-network applications, Inf. Sci. 331 (2016) 99–119.
[9] F. Buccafurri, G. Lax, A. Nocera, D. Ursino, Moving from social networks to social internetworking scenarios: the crawling perspective, Inf. Sci. 256 (2014) 126–137.
[10] F. Buccafurri, G. Lax, A. Nocera, D. Ursino, Discovering missing me edges across social networks, Inf. Sci. 319 (2015) 18–37.
[11] M. Cha, A. Mislove, K.P. Gummadi, A measurement-driven analysis of information propagation in the flickr social network, in: Proceedings of the 18th International Conference on World Wide Web, 2009, pp. 721–730.
[12] M. De Choudhury, H. Sundaram, Why do we converse on social media?: an analysis of intrinsic and extrinsic network factors, in: Proceedings of the 3rd ACM SIGMM international workshop on Social media, 2011, pp. 53–58.
[13] S. del Rio, V. Lopez, J.M. Benitez, F. Herrera, On the use of mapreduce for imbalanced big data using random forest, Inf. Sci. 285 (2014) 112–137.

[14] W.D. Dupont, W.D. Plummer, Power and sample size calculations: a review and computer program, Controlled Clin. Trials 11 (2) (1990) 116–128.
[15] Q. Fang, J. Sang, C. Xu, M.S. Hossain, Relational user attribute inference in social media, Multimedia, IEEE Trans. 17 (7) (2015) 10–31.
[16] M. Gomez Rodriguez, J. Leskovec, B. Schölkopf, Structure and dynamics of information pathways in online media, in: Proceedings of the sixth ACM international conference on Web search and data mining, 2013, pp. 23–32.
[17] L. Guo, E. Tan, S. Chen, X. Zhang, Y.E. Zhao, Analyzing patterns of user content generation in online social networks, in: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, 2009, pp. 369–378.
[18] E. Hansen, M. Patrick, A family of root finding methods, Numerische Mathematik 27 (3) (1976) 257–269.
[19] D.V. Hinkley, On the ratio of two correlated normal random variables, Biometrika 56 (3) (1969) 635–639.
[20] M. Kim, D. Newth, P. Christen, Trends of news diffusion in social media based on crowd phenomena, in: Proceedings of the companion publication of the 23rd international conference on World wide web companion, 2014, pp. 753–758.
[21] J. Kotrlik, C. Higgins, Organizational research: Determining appropriate sample size in survey research appropriate sample size in survey research, Inf. Technol., Learn., Perform. J. 19 (1) (2001) 43.
[22] J. Krumm, D. Rouhana, Placer: semantic place labels from diary data, in: Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing, 2013, pp. 163–172.
[23] K. Lerman, R. Ghosh, Information contagion: an empirical study of the spread of news on digg and twitter social networks, in: Proceedings of 4th International Conference on Weblogs and Social Media, 2010, pp. 90–97.
[24] J. Leskovec, M. McGlohon, C. Faloutsos, N.S. Glance, M. Hurst, Patterns of cascading behavior in large blog graphs, in: Proceedings of the 2007 SIAM International Conference on Data Mining, 2007, pp. 551–556.
[25] V.M. Lesser, W.D. Kalsbeek, Non-sampling error in surveys (1992).
[26] Z. Li, J. Liu, J. Tang, H. Lu, Robust structured subspace learning for data representation, IEEE Trans.Pattern Anal. Mach. Intell. 37 (10) (2015) 2085–2098.
[27] Z. Li, J. Tang, Unsupervised feature selection via nonnegative spectral analysis and redundancy control, IEEE Trans.Image Process. 24 (12) (2015) 5343–5355.
[28] Z. Li, J. Tang, Weakly supervised deep metric learning for community-contributed image retrieval, IEEE Trans. Multimedia 17 (11) (2015) 1989–1999.
[29] M. Lipczak, M. Trevisiol, A. Jaimes, Analyzing favorite behavior in flickr, in: Advances in Multimedia Modeling, 19th International Conference, 2013, pp. 535–545.
[30] M. Magnani, L. Rossi, The ml-model for multi-layer social networks, in: International Conference on Advances in Social Networks Analysis and Mining, 2011, pp. 5–12.
[31] M. Marusteri, V. Bacarea, Comparing groups for statistical differences: How to choose the right statistical test? Biochemia Medica 20 (1) (2010) 15–32.
[32] J.H. McDonald, Handbook of Biological Statistics, 2, Sparky House Publishing Baltimore, MD, 2009.
[33] A. Mislove, M. Marcon, P.K. Gummadi, P. Druschel, B. Bhattacharjee, Measurement and analysis of online social networks, in: Proceedings of the 7th ACM SIGCOMM Internet Measurement Conference, 2007, pp. 29–42.
[34] R.A. Negoescu, D. Gatica-Perez, Analyzing flickr groups, in: Proceedings of the 7th ACM International Conference on Image and Video Retrieval, 2008, pp. 417–426.
[35] R.-A. Negoescu, D. Gatica-Perez, Modeling flickr communities through probabilistic topic-based analysis, Multimedia, IEEE Trans. 12 (5) (2010) 399–416.
[36] O. Nov, M. Naaman, C. Ye, What drives content tagging: the case of photos on flickr, in: Proceedings of the SIGCHI conference on Human factors in computing systems, 2008, pp. 1097–1100.
[37] O. Nov, M. Naaman, C. Ye, Analysis of participation in an online photo-sharing community: a multidimensional perspective, JASIST 61 (3) (2010) 555–566.
[38] T. Pontes, M. Vasconcelos, J. Almeida, P. Kumaraguru, V. Almeida, We know where you live: privacy characterization of foursquare behavior, in: Proceedings of the 2012 ACM Conference on Ubiquitous Computing, 2012, pp. 898–905.
[39] D. Rao, D. Yarowsky, A. Shreevats, M. Gupta, Classifying latent user attributes in twitter, in: Proceedings of the 2nd international workshop on Search and mining user-generated contents, 2010, pp. 37–44.
[40] J. Tang, Z. Li, M. Wang, R. Zhao, Neighborhood discriminant hashing for large-scale image retrieval, IEEE Trans.Image Process. 24 (9) (2015) 2827–2840.
[41] C. Wang, X. Guan, T. Qin, T. Yang, Modeling heterogeneous and correlated human dynamics of online activities with double pareto distributions, Inf. Sci. 330 (2016) 186–198.
[42] P. Wang, W. He, J. Zhao, A tale of three social networks: User activity comparisons across facebook, twitter, and foursquare, Internet Comput., IEEE 18 (2) (2014) 10–15.
[43] M. Yan, J. Sang, C. Xu, Unified youtube video recommendation via cross-network collaboration, in: Proceedings of the 5th ACM on International Conference on Multimedia Retrieval, 2015, pp. 19–26.
[44] M. Yan, J. Sang, C. Xu, M.S. Hossain, Youtube video promotion by cross-network association: @britney to advertise gangnam style, IEEE Trans. Multimedia 17 (8) (2015) 1248–1261.