Selective Behavior in Online Social Networks

Chunjing Xiao*†, Ling Su†, Juan Bi*†, Yuxia Xue*, and Aleksandar Kuzmanovic†
*University of Electronic Science and Technology of China, Chengdu, China
†Northwestern University, Evanston, IL, USA
{chunjingxiao, yuxiaxue}@gmail.com; lingsu2012@u.northwestern.edu; {j-bi, akuzma}@northwestern.edu

***Abstract*—According to the classical communication theories, known as *Gatekeeping* and *Selective Exposure*, individuals tend to have selective behavior when they disseminate and receive information based on their psychological preferences. Selective behavior related to these two theories have been broadly studied separately. While, thanks to the advent of Online Social Networks (OSNs), larger-scale feedback and user information can be collected. In this paper, based on these data, We analyze the correlation among users' properties (such as age, gender, and cultural background) and analyze their selective behavior by tagging users as disseminators and/or audiences in YouTube, Flickr, and Twitter.**

We find that despite enormous amount of content available in OSNs, users have a comparatively small selective range and do exhibit selective behavior properties. In particular, they pay the most attention to the content published by disseminators that share similar properties, i.e., gender, age, and country. Nonetheless, we also find significant differences and commonalities among the three OSNs with respect to selective behavior. In particular, (i) the proportion and properties of disseminators, audiences, and dual-role users are quite different for the three networks; (ii) the global level of information spread in Flickr is almost two times than that in Twitter and YouTube is approximately the median one; (iii) For a given country, the global level of information spread is different for different OSNs. For a given OSN, it is different for different countries; (iv) despite ubiquitous presence of dual-role users in OSNs, most of such users are very active as either disseminators or audiences, but not both. Our findings are not only useful for understanding these two theories, but also have applications ranging from advertising and recommendation systems to developing predicting models.

I. Introduction

In a social network, information propagation is playing a vital role since it provides the basic way to people's communication and cooperation, as well as their perception of the world. At the same time, selective behavior during the information dissemination and reception impacts the way in which people understand the world. In 1920s, Lippmann found that the world that people experience through media, called pseudo-environment [1], is different from the real one. Selection is one of the crucial reasons for this phenomenon, which biases our understanding of the world. In 1940s, Kurt [2] found that people choose to spread information that fits their own values, and ignore other. This is called Gatekeeping theory. At the same time, other studies have shown that the similar result holds for audiences. In particular, they selectively accept information instead of accepting all the information [3], [4]. This is the principle of the Selective Exposure theory.

This project is supported in part by the National Science Foundation (NSF) via grant CNS-1064595.

Both two theories focus on studying selective behavior from aspects of disseminators and audiences respectively and have been broadly studied separately in traditional media [5], [6]. However, the advent of Online Social Networks (OSNs) makes it easy to collect the larger-scale feedback and user information, which provides a chance to explore the general correlation and interaction characters between disseminators and audiences.

In this paper, by tagging users as disseminators and/or audiences, we study the correlation between users' properties and their selective behavior in OSNs. Here, among many users' properties, we focus on the basic and generic aspects: *gender*, *age* and *cultural background*, which in our study correspond to the country that the user comes from. We explore OSNs that feature different primary information types, i.e., video, photo, and text. In particular, we select YouTube, Flicker and Twitter which represent these three information types, respectively. Based on large-scale data, we firstly conduct overall analysis about disseminators and audiences, and analyze the disseminators and audiences respectively. Then, we study the interaction between disseminators and audiences. Finally, we explore unique characters in OSNs, i.e., *dual-role* users, who are both disseminators and audiences.

In addition to better understanding the two classical theories in OSNs, some others reasons also motivate us to analyze users' selective behavior. (i) Understanding user's behavior (who influences whom) is a key factor to improve effects of information dissemination because it will enhance the focus of targeted audiences. This is particularly relevant for advertisement. Guha *et al.* [7] find that user profiles, such as location, gender and age affect the content of ads, i.e., ad networks will provide different ads contents depending on users' profile. Therefore understanding audiences' selective behavior relative to these three properties will be helpful for ad selection. (ii) The related video recommendation is the main source of views for the majority of the videos on YouTube [8]. Understanding users' preference can be used to design recommendation systems based on users' properties to recommend more focused contents. (iii) The comparison of three OSNs provides insight for the future OSNs development. For example, in terms of location, understanding how widely information spread for different types of mediums: video, photo and text and how different are the interaction levels among counties is valuable in devising caching mechanisms and exploring business; (iv) Finally, analyses of users' behavior will provide useful insight ranging from building models of user behavior to predicting audiences properties.

Our findings are the following. (i) By analyzing the correlation of disseminators and audiences, we find a striking homophily in terms of gender, location and age. For example, DE (German) male audiences pay about 60% attention to the videos uploaded by DE male disseminators in YouTube and Twitter. (ii) The types of mediums have hardly impact on homophily, but they place significant influence on the global level of information spread. In particular, the global level in Flikcr is almost two times than that in Twitter and YouTube is approximately the median one. (iii) Moreover, for a given country, the rank of global level of information spread in different OSNs is different. For example, US is more global than DE in YouTube, while the reverse is true in Twitter. (iv) Finally, the categories of contents uploaded by Dual-role users are high consistent with that they commented. And most of the dual-role users are very active as either disseminators or audiences, but not both.

II. Data and Methodology

A. Data Description

Here, we present the data collected from YouTube, Flickr, and Twitter. Since our goal is to explore users' selective behavior related to information propagation, we need to collect information about users, topics, and their response. The topics in YouTube, Flickr, and Twitter refer to videos, photos, and tweets, respectively. For YouTube and Flickr, comments, a simple and popular reaction, are selected as topics' response. For Twitter, we use retweets as the measure of response. All the data is collected from June to October 2011. Table I provides the details.

	YouTube	Flickr	Twitter
# all users	45.342M	11.785M	48.072M
# users with location	44.545M	1.658M	16.606M
# users with gender	44.723M	4.193M	9.397M
# users with age	41.613M	0	0
# video/photo/tweets	10.166M	8.199M	9.517M
# comments	792.473M	344.429M	12.122M

TABLE I
DATA DESCRIPTION

1) YouTube: We collect a list of popular YouTube videos by relying on 6 YouTube APIs associated with standard feeds: most discussed, most popular, most responded, most viewed, top favorites, and top rated. In order to obtain videos from different countries for different categories, we use the region-specific standard and category-specific standard feeds. In this way, we retrieve video feeds from 25 counties and 15 regular categories, as shown in [9]. The crawler runs every 6 hours for a month from June 10 2011 to July 10 2011. In addition, after collecting these feeds, we query the related videos from these feeds. In the end we collect over 10 million video IDs and their profiles.

We also download the comments corresponding to the videos. Because for each video only first 1,000 comments can be obtained via the YouTube API, we download the rest comments by screen-scraping the HTML pages with comments above 1,000. In addition to comments, we also collect users' profiles related to videos and comments using APIs. We find that more than 90% profiles include standard formatted information about gender, age, and location that contains two letter ISO country codes. For age, users younger than 10 and older than 60 take up about 1% of all the users; hence, we focus only on the age range between 10 and 60.

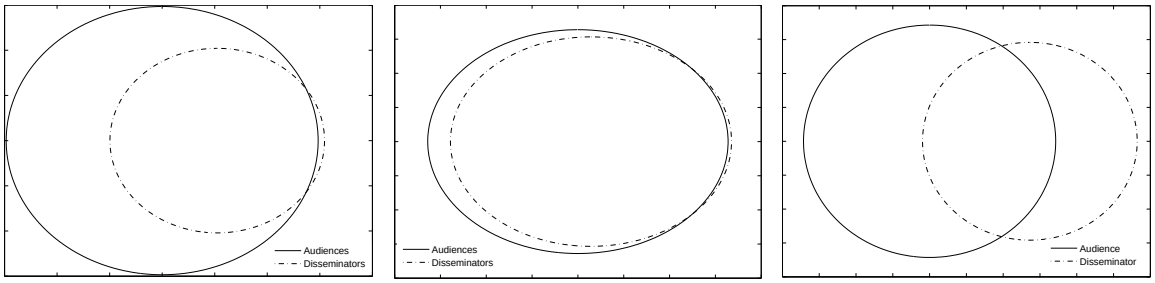
2) Flickr: We collect the data from Flickr using their APIs. First, to obtain a user list, we start with a randomly-selected Flickr user and recursively obtain all the friends along the breadth first search direction. After getting about 11M users (see Table I), we acquire their favorite photo lists, obtaining close to 110M photos. Because we are interested in popular photos, we select approximately 8M photos which have been added into favorite lists by more than 10 users. Finally, we download profiles and comments of the associated photos as well as corresponding users' profiles.

While the information is formatted data in Filckr, location is a self-reporting field. To identify the users' location, we aggregate location field and manually assign the two-letter country code to them. In this way, we obtain the location information for 14% of the 11M users.

3) Twitter: Based on a 42M users list [10], we re-retrieve their follower list and thus obtain a total of approximately 51M users. Of the 51M users, approximately 48M users have available profiles that we download, as shown in Table I. We collect 590M tweets during 31 days using Twitter streaming APIs which return roughly 10% of all public statuses. Focusing on tweets and retweets with location and gender information, we end up with approximately 9M tweets and 12M retweets.

There are two fields in users' profile that directly or indirectly provide location information: location and local Time Zone. Because we only need the country-level information, Time Zone can be mapped to unique country except for USA and Canada whose Time Zones are not uniquely identifiable (e.g., Central Time US & Canada). Thus, for USA and Canada, we further refine their location by exploring the location field. Specifically, we utilize two name lists including big cities names, state names, and abbreviation names for USA and Canada that we retrieve from wikipedia. Finally, we disambiguate a user's country if its location field can be matched to an entry from the corresponding lists. In the way we obtain 16M users' country code, which correspond to 34% of 48M users.

There is no information about gender in users' profiles. Nonetheless, user names can be used to detect gender. In particular, similar to [11], for English-speaking countries, i.e., USA, Canada, Great Britain, Australia and New Zealand, we first obtain the most popular 1,000 male and female names for babies born in the period 1880-2010 [12]. Then, we calculate the total frequency of each name as male or female. We remove dual-gender names if the ratio between low frequency as a gender and high one as the opposite gender is more than 10%. As a result, we remove 4,459 names while 56,347 female and 32,149 male names are utilized to infer gender. Similarly, for Japanese, German and French names, we also collect their female and male names from [13] to infer their gender. Finally, we end-up with about 9M users with gender information.



(a) YouTube (b) Flickr (c) Twitter

Fig. 1. Distribution of disseminators and audiences

B. Classifying Users' Interest Categories

Determining users' interest categories is essential for analyzing selective behavior of disseminators and audiences. Because there are no official categories for photos in Flickr, we apply a Naive Bayes classifier, i.e.,[14], to classify photos into six categories: People, Travel, Animal, Automobile Art and Plant which are created based on categories of YouTube and our observation. We select the available photos' tags and titles as photos' feature to classify.

Following the steps of the classification procedure, we first conduct data cleaning. In particular, we remove the non-English words in the photos' tags and select manually 91 stop words which are not useful for classification. Example words are "aplusphoto" and "Nikon". Finally, 13% of photos with no more than a single tag are dropped.

Next, we select 4,000 photos randomly and classify them manually into six categories. Then, we divide these photos into the training set and testing set evenly. Both sets are given to the classifier to construct a classification model. We achieve the classification accuracy of 81%. Finally, based on 4,000 photos, the remaining photos are classified into different categories.

III. Disseminators and Audiences

In this section, we break the users into disseminators and audiences, and then present the overall user information about three OSNs. For YouTube and Flickr, users who upload videos or photos are regarded as disseminators. Users who author comments are regarded as audiences. Correspondingly for Twitter, users who initiate tweets being retweeted are disseminators while those who retweet them are audiences. Here, we first explore the ratio of disseminators and audiences in the three OSNs. Moreover, we explore *dual-role* users, which refers to users playing the roles of both disseminators and audiences.

A. Overall Analysis

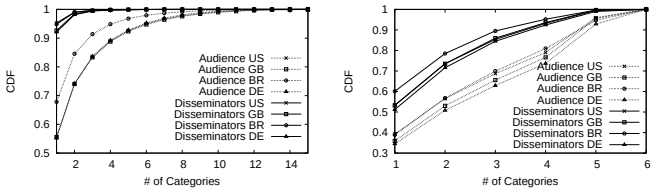
Figure 1 shows the relationships among disseminators, audiences as well as dual-role users for the three OSNs. In YouTube, the number of audiences overwhelms disseminators. Moreover, the audience group almost fully overlap disseminators. In Flickr, the disseminator and audiences group largely overlap, hence the percentage of dual-role users is the largest for the thee OSNs. On the contrary, the phenomenon of high proportion of dual-role users does not apply on Twitter. Indeed, while the number of disseminators and audiences is approximately the same, only half of disseminators are dual-role users.

The results shown in Figure 1 are certainly caused by the different working principles for each of the OSNs. In particular, while the number of disseminators is not small, the enormous YouTube audience is much larger. Hence, the proportion of disseminators relative to audiences is smaller. The motivation for most of the users, who register in Flickr, is sharing their photos, which can explain the numbers of disseminators, audiences and dual-role users are about the same. On the other hand, even though it is easiest to publish a topic in Twitter due to its text message form, there is a large number of topics and messages, most of which are ignored by users. This specific principle causes the number of audiences and disseminators to be about the same, yet the dual-role users just hold minority.

In the following part of data analysis, we will observe the characteristics of disseminators and audiences when they are attributed to different aggregations. In particular, we will analyze the disseminators and audiences sets respectively in Section III-B; the union of disseminators and audiences are our target set to explore the interaction between disseminators and audiences in Section IV; the intersection of two aggregations, users acted as dual-role, are studied in Section V.

B. Disseminators and Audiences Individually

In this section we analyze and compare the selective behavior of disseminators and audiences as separate groups. Compared to traditional media, disseminators in OSNs are more free to publish contents, and audiences have much more contents to select. Thus, we want to know if disseminators and audiences still focus on a narrow range of categories.

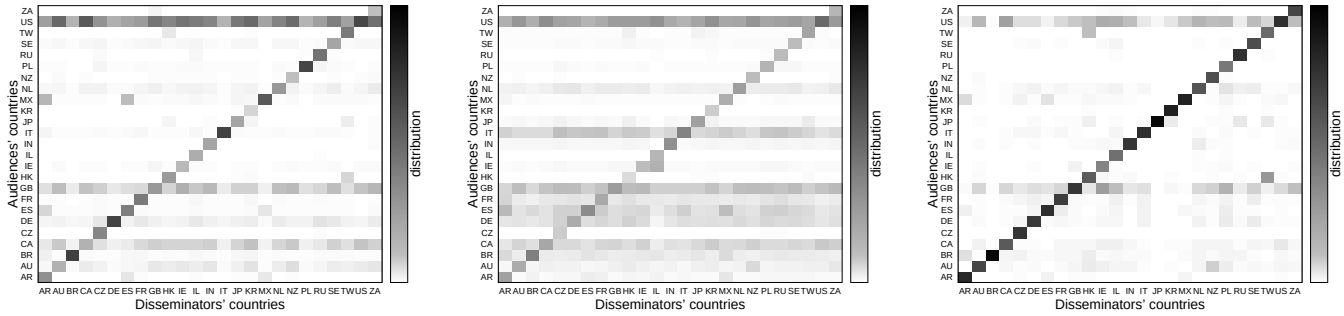


(a) YouTube (b) Flickr

Fig. 2. Selective ranges of different countries

First, we explore user selective behavior in light of location properties. We select four countries, US, GB, BR and DE, which are the top 4 counties with the most registered users for YouTube in our dataset.

Figure 2(a) and 2(b) show the results. The x-axis represents the categories number. The first insight from the figure is that audiences have a broader selection than disseminators.



(a) Youtube (b) Flickr (c) Twitter

Fig. 3. Interaction between countries

Focusing on location results, we find that although there are subtle difference among different countries, their selective ranges are very narrow. For example, for all the countries in YouTube, there are more than 90% disseminators that published videos in a single category; on the other hand, there are less than 50% audiences that commented in more than one category. Ding *et al.* [15] also find there is a strong tendency for YouTube uploader to concentrate videos in a small number of categories. Figure 2(b) shows that users in Flickr have a relatively broader selective range relative to YouTube. Still, users focus on narrower categories. There are approximately 50% of disseminators and 40% of audiences who published pictures in one category. Finally, the selective ranges for gender and age, not shown here for space constraints, show similar trends to the location analysis.

IV. UNION OF DISSEMINATORS AND AUDIENCES

In traditional media, there is little interaction between disseminators and audiences. On the other hand, in OSNs, users can easily release any kind of information, and receive feedback, almost immediately. Consequently, each user has a lot of choices for publishing or receiving various contents and can select more freely contents according to their value. Under these circumstances, the key issue we want to explore in this section is the characteristics of interaction between disseminators and audiences with different attributes. In other words, we want to know if the audiences have similar attributes to disseminators who uploaded the content.

A. The Role of Location

Here, we explore the role of users' location and its impact on the choice of content that users select. In particular, we want to understand if people prefer to select domestic content rather than foreign one, e.g., because of the same cultural context. We collect information from more than 250 countries and regions. For space constraints, we select 25 countries, listed in [9], which come from continents with different cultural contexts. For country i and k, we calculate the ratio: $\frac{c_{ik}}{c_i}$, where c_{ik} is the number of comments posted by country k to videos of country i, and $c_i = \sum_{k=1}^{N} c_{ik}$ is the sum of all of the comments that the videos of country i received (N=25).

Figure 3 shows the ratio among different countries. The x-axis represents the countries of disseminators: i, i.e., the owners of videos, photos or tweets. The y-axis represents the countries of audiences: k, i.e., those who posted comments.

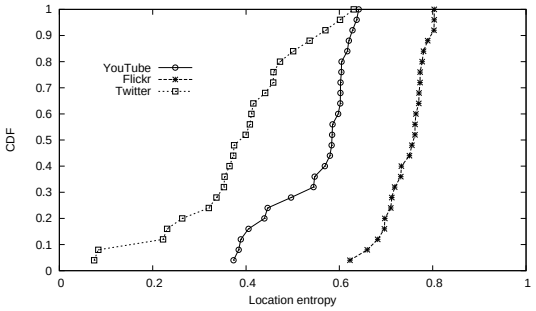


Fig. 4. CDF of location entropy

The darker the color is, the stronger the correlation is between the corresponding countries. Also, the sum of each of the columns in Figure 3 equals to 100%. The figures clearly show that the domestic factor is quite strong, i.e., the darker colors at the diagonals. This is particularly true for Twitter, which generally shows much more localized behavior, as we will further demonstrate later in the paper. The figures also show that countries such as USA (US), Great Britain (GB) and Canada (CA) always pay more attention to contents generated in other countries. This is likely due to the multi-cultural nature of the three countries.

1) Global Level of Information Spread: To study to what extent will the foreign audiences select contents uploaded by disseminators with a given country, i.e., how widely will the information spread, we introduce the *Location Entropy* L_i for country i as below:

$$L_i = -\frac{1}{\ln N} * \sum_{k=1}^{N} \frac{c_{ik}}{c_i} \ln \frac{c_{ik}}{c_i}$$

Where c_{ik} and c_i are defined above. And we still take 25 countries for example, so N is still 25. For country i, if all the comments of its videos only come from one country, its location entropy will be 0; while if the comments of its videos come from 25 countries uniformly, its location entropy will be 1. Hence, higher entropy denotes more global information propagation.

Figure 4 shows the distribution of the location entropy for three OSNs. It clearly shows that Flickr are the most global one, whose mean is almost 2 times than that of Twitter, and YouTube is approximately the median one among them. Indeed, pictures and videos have a more global nature and hence spread more widely. On the contrary, text messages on Twitter have a more localized spread. And, re-tweeting

text messages stays local due to language barriers. Besides, for each OSN, different countries have different global levels. Moreover, by comparing the values of the location entropy, we find it is not consistent for global levels of a country in three OSNs. For example, US is more global than DE in YouTube, while the reverse is true in Twitter.

2) Interaction Level among Locations: The above analysis shows that audiences always pay more attention to videos from their own countries, because of the same cultural background. By intuition, apart from their own countries, some countries have more similar culture than others. Correspondingly we ask whether interaction between countries with similar culture is more frequent than that with a different one. To this end, firstly we introduce the *Interaction Level*. Let a_{ik} be the degree of the attention paid by country i to country k, i.e., a_{ik} is the proportion between amount of comments posted by country i to videos of country k and all the comments posted by country i. The reciprocity between count i and k might be unbalanced, i.e., a_{ik} maybe not equal to a_{ki}. Thus we define the *interaction level* between country i and country k, I_{ik} as below:

$$I_{ik} = \sqrt{a_{ik}^2 + a_{ki}^2}$$

Secondly, according to Sundqvist *et al.* [16] and the geographic location, we put 23 of these 25 countries into four clusters: Cluster #1 comprises of 11 countries, mainly Anglo-Saxon cultures from Western Europe and North America: DE, ES, FR, GB, IE, IT, NL, NZ, AU, US and CA; Cluster #2 includes 3 countries from South America: AR, BR and MX; Cluster #3 consists of 3 Eastern Europe countries: RU, CZ and PL; Cluster #4 consists mainly of rapidly developed Asian countries: HK, TW, KR and JP.

Finally we compute the interaction level of the country *i* and *k*, I_{ik}, and put them into two disjoint groups: *Similar Culture* if *i* and *k* come from the same cluster and *Different Culture* if *i* and *k* come from the different clusters. Table II shows the average interaction level of two groups for three OSNs.

	YouTube	Flickr	Twitter
Similar Culture	0.0729	0.0881	0.0569
Different Culture	0.0239	0.0524	0.0238
P-Value	1.19E-12*	2.10E-06*	1.02E-08*

TABLE II
AVERAGE INTERACTION LEVEL BETWEEN SIMILAR/ DIFFERENT CULTURAL COUNTRIES

From this table, we can find the average of interaction level between countries with similar culture is much higher than that between countries with different culture. This indicates audiences trend to select the content from countries with similar culture, instead of countries with different culture. Moreover two-sample Kolmogorov-Smirnov (KS) test is conducted for these two groups and the p-values are shown in the last row of Table II. All the p-values are less than 0.001, which indicates the difference of the two groups is statistically significant.

B. The Role of Gender

Here, we explore the role of gender, i.e., whether audiences have a tendency to post comments to the content uploaded by disseminators with the same or different gender. Figure 5 shows the results for three OSNs. In the x-axis, for each of the OSNs we divide the content into two categories: one uploaded by females and the other by males. The y-axis represents the percent of comments posted by each gender. Necessarily, the sum of females' comments uploaded to each of the OSNs (e.g., YouTube) equals to 100%. Likewise, the sum of the males' comments uploaded to each of the OSNs equals to 100% as well.

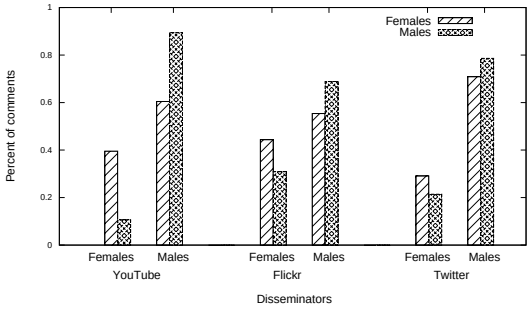


Fig. 5. Interaction between genders

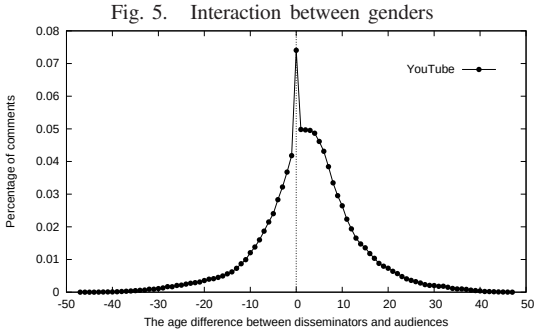


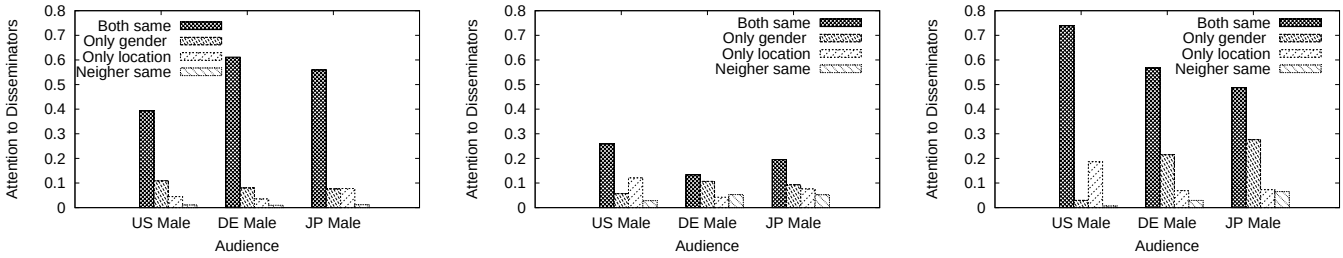
Fig. 6. Interaction between ages

Figure 5 shows two consistent selective behavior for the three OSNs. First, in all cases, the percent of comments is always larger for *males'* disseminators. This holds both for males' comments (always larger than 60%) and females' comments (always larger than 55%). This is because the number of male disseminators is larger than the number of female disseminators. The second insight is that there exists a clear *homophilic* selective gender behavior. In particular, in all cases, females tend to comment more the content uploaded by females, while males tend to comment more the content uploaded by males. This phenomenon is particularly emphasized for YouTube, where females are approximately 5 times more likely to comment the content uploaded by females than males are.

C. The Role of Age

Next, we explore the interaction between disseminators and audiences in the context of their age. Necessarily, we focus on YouTube for which we are able to extract the age information in more than 90% of cases, as shown in Table I. Figure 6 shows the results. The x-axis depicts the age difference between disseminators and audiences. The y-axis represents the percent of comments posted by the users of a given age difference.

Figure 6 shows a clear impact of the age factor, i.e., that the content uploaded by disseminators is more likely to be commented by audiences of similar age. In particular, the most striking example is the peak (of 7.4%) at the age difference



(a) YouTube (b) Flickr (c) Twitter

Fig. 7. Interaction between location and gender

of *zero*. This means it is most likely that the disseminators and audiences are of the same age. One final observation is that it is more likely that younger audiences comment on the content uploaded by older disseminators than it is for older audiences to comment on the content uploaded by younger disseminators. Indeed, Figure 6 shows that the age distribution is not symmetric. The Skewness value of this distribution is -0.054 (the negative/positive values for the Skewness indicate data are skewed left/right, and the Skewness value of a normal distribution is zero), and the average age difference is 2.66, which indicate the distribution is only a litter skewed left, and ages of disseminators are generally 2.66 higher than their audiences.

D. The Role of Location and Gender

The above analyses reveal the selective behavior in terms of these properties individually. However, more intuitively, some behavior should be determined by multiple properties jointly. Therefore, here we address the condition for the combination of two properties: location and gender that all these three OSNs own. In particular, we try to answer whether more similar the properties of disseminators are with that of audiences, more likely their contents are selected by audiences.

For audiences with given country and gender, we divide all the comments posted by them into four groups: *both same*: comments posted to videos whose disseminators have same both location and gender; *only location*: comments posted to videos whose disseminators have only same location; *only gender*: comments posted to videos whose disseminators have only same gender; *neither same*: comments posted to videos whose disseminators have same neither location nor gender. Since there is only a country in former two groups while there are a lot of countries in the last two groups, we further distribute comments to each country in last two groups and select the maximum to represent the value of their groups.

For space constraints, three countries' male audiences are selected for three OSNs, as shown in Figure 7. The y-axis represents the degree of attention paid by audiences with given country and gender to videos whose disseminators own different properties, i.e., the fraction of comments of given groups and all the comments. The first observation is that values of *both same* are always higher than that of other three groups. This is particularly true in YouTube and Flickr whose values are always larger than 40%. Besides, values of *neither same* are lowest one among four groups except DE Male in Flickr, where its value is only negligible 0.01 higher than that of *only location*. These mean, as we expected, that

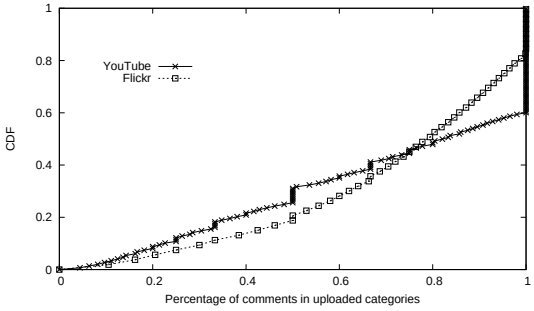


Fig. 8. Distribution of comments

audiences would like to pay more attention to videos published by disseminators with the same location and gender, instead of that with different ones. Second, the summary value of *both same* and *only location* indicates the global level of information spread, i.e., the lower the value is, the more global the information spread is. Therefore we can reach consistent conclusions with Figure 4 that Twitter is the most local one while Filckr is the most global one, and a country has different global levels of information spread for different OSNs.

V. Intersection of Disseminators and Audiences

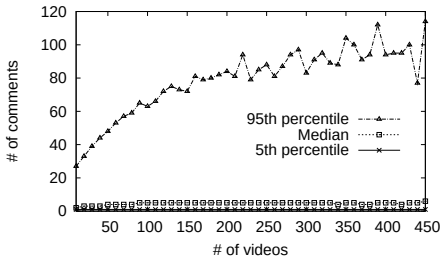
In addition to the high level of interaction relative to the traditional media, a distinct feature common for OSNs is that there exists a community of users who are both disseminators and audiences. We call them *dual-role* users. For each such user, the behavior at different roles, i.e., disseminators and audiences, is still based on the same psychology. Our goal in this section is to extract selective behavior of dual-role users.

A. Impact of Role

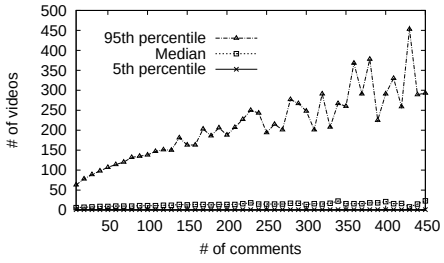
Here, we want to understand how similar (or not) are selective range when dual-role users act as disseminators and audiences. A potential similarity can explain the relationship between users' selective behavior and roles they playing.

First, we quantify the number of users whose uploaded categories are consistent with commented categories, where uploaded categories refer to the categories list in which users have uploaded contents, and correspondingly commented categories are the categories list in which users have posted comments. We calculate there are 86% dual-role users whose uploaded categories completely belong to their commented categories in YouTube and 92% in Flickr.

We take a further step in studying these users (86% in YouTube and 92% in Flickr). Figure 8 shows the CDF of dual-role users as a function depicted on x-axis, which is the ratio of number of comments in uploaded categories vs. all commented number. The figure shows that there are about

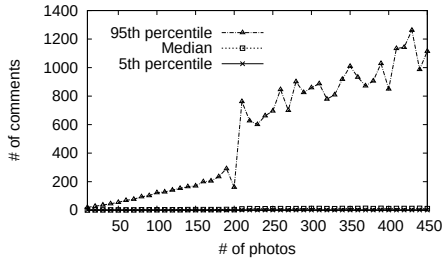


(a) # videos VS # comments

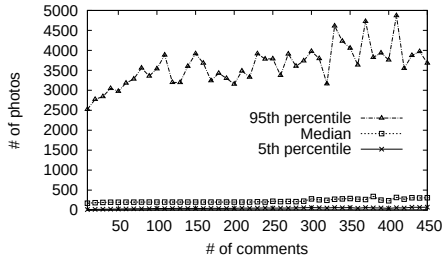


(b) # comments VS # videos

Fig. 9. YouTube

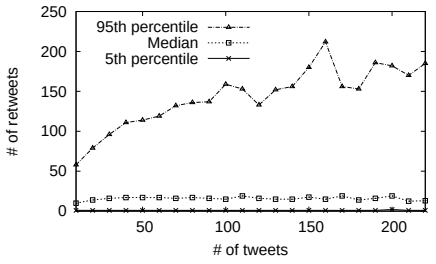


(a) # photos VS # comments

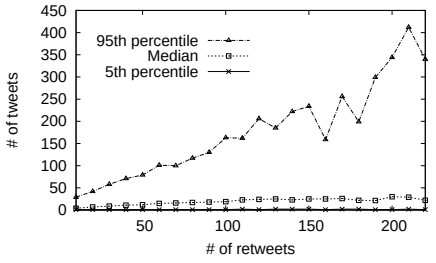


(b) # comments VS # photos

Fig. 10. Flickr



(a) # tweets VS # retweets



(b) # retweets VS # tweets

Fig. 11. Twitter

60% of dual-role users in YouTube and 80% in Flickr whose number of comments in uploaded categories is smaller than that in commented categories, and the rest have the same uploaded and commented categories. These are the points in the figure that correspond to the values in y-axis for $x = 0.5$. The reminder of users, i.e., only 30% in YouTube and 20% in Flickr, have less comments in uploaded categories than that in non-uploaded categories. Nonetheless, in the context of dual-role users, these results show highly similar selective behavior when dual-role users act at different roles. This means that user selective behavior is related to user attributes rather than the particular role that they are playing.

B. Role Preference

To understand the activity level of dual-role users at different roles, i.e., disseminators and audiences, we compare the amount of contents that they published and the amount of comments.

Figure 9(a) shows the number of videos uploaded by the dual-role users versus the number of comments of the same users for YouTube. The x-axis denotes the number of videos that dual-role users have published. The y-axis denotes the number of comments posted by users with the given number of uploaded videos. To comprehensively represent the distribution of comments, we show the 95th percentile, the median and the 5th percentile value of the comments, respectively.

According to Figure 9(a), as the number of published videos grows, the number of comments in 95th percentile is increasing continuously. However, the median number of comments stays steady and small. This indicates that as the activity levels of dual-role users who act as disseminators rises, a few keep correspondingly increasing the activity level as audiences. Indeed, the activity levels of the majority of such users do not grow on the audiences side, hence the median comments value remains low. Thus, when dual-role users have a high disseminating activity level, their audience activity level does not grow. One final point here is that the fluctuations in the tail of the 95th percentile arises due to a small fraction of users that upload higher number of videos. Still, the overall trend is obvious.

Next, we explore a complementary question, i.e., when dual-role users have a high activity level in acting as audiences, do they have comparatively high activity level in acting as disseminators? Figure 9(b) shows the results. The x-axis and y-axis are reversed relative to Figure 9(a). In particular, the x-axis denotes the number of comments posted by dual-role users, while the y-axis shows the number of videos uploaded by users who posted the given number of comments. Figure 9(b) shows that there are a few dual-role users publishing more videos as the number of their comments grows. Indeed, the median value of uploaded videos remains low. This suggests that when dual-role users have a high activity level in acting as audiences, they do not have a correspondingly high activity level in acting as disseminators.

Figures 10 and 11 show the results for Flickr and Twitter, respectively. The results show similar trends as for YouTube. For Flickr, Figure 10(a) shows a sudden change for the 95th percentile curve at x = 200 photos/user. This is because there exists a 200 photo limit for the number of photos that a regular user can publish. To publish more than 200 photos, users should become professional users and pay a fee to Flickr. The figure demonstrates that the top 5 percent of professional users have a higher audience activity level, i.e., post more comments, relative to the regular users and the remaining 95% of professional users. Indeed, the median number of comments is small and constant. A similar reason can explain that the 95th percentile curve for the number of published photos in Figure 10(b) stays above 2,500. This is because there is a small number of professional users who publish a lot of photos independently of the number of comments they make.

Moreover, we measure the strength of a linear relationship between the amount of contents and comments (i.e., correlation coefficient). These coefficients (0.01, 0.04 and 0.22 for YouTube, Flickr and Twitter respectively) are close to zero, which suggests there are no strongly linear correlations

between them. These also provide support for the above interpretation.

The above analysis shows that there are subtle differences for the three OSNs. Nonetheless, we conclude that although the dual-role phenomenon is ubiquitous, most of dual-role users show a role preference. They are very active as either disseminators or audiences, but not both.

VI. Related Work

The most closely-related work to ours is the thread of work in the area of information propagation and users' content selection behavior [17], [18], [19]. In particular, Singla and Richardson [17] study similarity in querying information among the users who chat with each other. The more time they spend talking, the more similar they are in selecting keyword searches. Hofman *et al.* [18] divide users into four categories: celebrities, bloggers, representatives of media outlets and other formal organizations. They find homophily within categories, i.e., that celebrities listen to celebrities, bloggers listen to bloggers etc. Wang *et al.* [19] explore a selective process of information spreading in email communication. They find that social and organizational context significantly impact the spread of information.

While similar to the above work in that it studies correlation between users and the content they access, our work differs from the above in the following: (i) We show that selective behavior is ubiquitous at the very large scales, i.e., it applies to millions of users in OSNs. (ii) Next, contrary to previous work, we demonstrate that selective behavior applies to the most generic categories, such as age, gender, and country-level location. (iii) Further, in addition to text, i.e., Twitter [18], instant messaging [17] and email [19], we show that selective behavior applies to different mediums, such as video, photo, text, and the combination of the three. (iv) Finally, we systematically study disseminators, audiences, and the dual-role users, at scale.

To an extent, our work relates to the information propagation topic, which received wide attention on OSNs. Scellato *et al.* [20] study the relationship between popularity and locality of YouTube videos. Cha *et al.* [21] explore how widely and quickly does information propagate in Flickr. Cha *et al.* [22] also study how information disseminates via Flickr social links. Information diffusion has also been studied in the blogs domain, e.g., [23], [24]. In addition, models have been developed to characterize the information flow dynamics and trace the paths of diffusion and influence [25], [26]. Contrary to all the above work, which focuses on the information process or flow, the main focus of our work is on the *end users*, i.e., disseminators and audiences, and their properties.

VII. Conclusions

In this paper we explored selective behavior for disseminators and audiences in YouTube, Flickr, and Twitter. We have seen there are striking homophily for age, gender, and geographic properties. At the same time, global levels of information spread are explored. Contrary to traditional media, we find that dual-role users are ubiquitous in OSNs. And their selective behavior is strongly tied to user attributes, not a particular role that users are playing. Nonetheless, most of such users are very active as either disseminators or audiences, but not both.

Future work include a number of directions. One is to study the impact of other contexts, such as education, hobby, occupation etc, to users' selective behavior. Besides, predictive model of video popularity can take user's properties into account to refine the result. Finally, recommendation systems based on users' properties can be developed to provide more targeted contents.

References

[1] W. Lippmann, *Public Opinion*. New York: Harcourt, Brace and Co., 1922.

[2] K. Lewin, "Frontiers in Group Dynamics: II. Channels of Group Life; Social Planning and Action Research," *Human Relations*, vol. 1, no. 2, pp. 143–153, 1947.

[3] H. G. Paul Felix Lazarsfeld, Bernard Berelson, *The people's choice; how the voter makes up his mind in a presidential campaign*. New York: Columbia Univ. Press, 1948.

[4] J. T. Klapper, *The effects of mass communication*. New York: Free Press, 1960.

[5] P. Fischer, E. Jonas, D. Frey, and S. Schulz-Hardt, "Selective exposure to information: the impact of information limits," *European Journal of Social Psychology*, vol. 35, no. 4, pp. 469–492, 2005.

[6] A. Hoskins and B. O'Loughlin, "Remediating jihad for western news audiences: the renewal of gatekeeping?" *Journalism: theory, practice and criticism*, vol. 12, no. 2, pp. 199–216, 2011.

[7] S. Guha, B. Cheng, and P. Francis, "Challenges in Measuring Online Advertising Systems," in *IMC 2010*.

[8] R. Zhou, S. Khemmarat, and L. Gao, "The impact of youtube recommendation system on video views," in *IMC 2010*.

[9] "Reference guide: Data api protocol," http://code.google.com/apis/youtube/2.0/reference.html.

[10] H. Kwak, C. Lee, H. Park, and S. Moon, "What is twitter, a social network or a news media?" in *WWW 2010*.

[11] A. Mislove, S. Lehmann, Y.-Y. Ahn, J.-P. Onnela, and J. N. Rosenquist, "Understanding the Demographics of Twitter Users," in *ICWSM 2011*.

[12] "Popular baby names," http://www.ssa.gov/oact/babynames/.

[13] "Names from around the world," http://www.20000-names.com/.

[14] J. Han and M. Kamber, *Data mining: concepts and techniques*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2000.

[15] Y. Ding, Y. Du, Y. Hu, Z. Liu, L. Wang, K. Ross, and A. Ghose, "Broadcast yourself: understanding youtube uploaders," in *IMC 2011*.

[16] S. Sundqvist, L. Frank, and K. Puumalainen, "The effects of country characteristics, cultural similarity and adoption timing on the diffusion of wireless communications," *Journal of Business Research*, vol. 58, no. 1, pp. 107–110, 2005.

[17] P. Singla and M. Richardson, "Yes, there is a correlation: - from social networks to personal behavior on the web," in *WWW 2008*.

[18] S. Wu, J. M. Hofman, W. A. Mason, and D. J. Watts, "Who says what to whom on twitter," in *WWW 2011*.

[19] D. Wang, Z. Wen, H. Tong, C.-Y. Lin, C. Song, and A.-L. Barabási, "Information spreading in context," in *WWW 2011*.

[20] A. Brodersen, S. Scellato, and M. Wattenhofer, "Youtube around the world: Geographic popularity of videos," in *WWW 2012*.

[21] M. Cha, A. Mislove, and K. P. Gummadi, "A measurement-driven analysis of information propagation in the flickr social network," in *WWW 2009*.

[22] M. Cha, A. Mislove, B. Adams, and K. P. Gummadi, "Characterizing social cascades in flickr," in *WOSN 2008*.

[23] E. Adar and L. A. Adamic, "Tracking information epidemics in blogspace," in *WI 2005*.

[24] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins, "Information diffusion through blogspace," in *WWW 2004*.

[25] J. L. Iribarren and E. Moro, "Impact of human activity patterns on the dynamics of information diffusion," *Physical Review Letters*, vol. 103, no. 3, p. 038702, 2009.

[26] M. Gomez Rodriguez, J. Leskovec, and A. Krause, "Inferring networks of diffusion and influence," in *KDD 2010*.