Beyond the Click: A First Look at the Role of a Microblogging Platform in the Web Ecosystem

Jingtao Ding, Zhongjin Liu, Shaoran Xiao, Yang Chen, *Senior Member, IEEE*, Yong Li, *Senior Member, IEEE*, Depeng Jin, *Member, IEEE*, and Steve Uhlig, *Member, IEEE*

***Abstract*—As the largest microblogging service in China, Sina Weibo, is associated with a huge amount of traffic to and from other major websites. Thanks to a massive set of anonymized data from a major operator covering a large-scale metropolitan area, we extract the click-based transitions between Weibo and other websites. Such transitions have not been looked into in most studies that rely on the visible hyper-links among websites. Based on this unique dataset, we show that 38% of the outgoing transitions (leaving Weibo) go to video clip services. Comparatively, links related to news and long videos are clicked less but posted more. This indicates that photograph and text-sharing in microblog have been gradually replaced by video clip sharing. Also, ignoring video clip sharing, 53% of the transitions go to five of the largest Chinese websites, showing how closely connected Weibo has become in the Chinese Web ecosystem. We further look at users' click behavior in transitions, and find that through co-clustering there do exist a number of prevalent and distinct behavior patterns, suggesting that only seven co-clusters can capture the click behavior of 192 000 studied users. Overall, our work is the first to unveil the significant role of microblogging platforms in the Web ecosystem. Considering the massive user base of Weibo, our findings can aid the design of more personalized link recommendation in Weibo, thus generating more referral traffic, i.e., user visits, from Weibo to other major websites.**

***Index Terms*—User click behavior, microblogging service, Web ecosystem, external hyper-link.**

Manuscript received April 2, 2018; revised August 10, 2018; accepted October 15, 2018. Date of publication November 30, 2018; date of current version June 10, 2019. This work was supported in part by China Ministry of Education-CMCC Research Fund Project no. MCM20160104, the National Key Research and Development Program of China under grant 2017YFE0112300, the National Nature Science Foundation of China under 61861136003, 61621091 and 61673237, Beijing National Research Center for Information Science and Technology under 20031887521, and research fund of Tsinghua University–Tencent Joint Laboratory for Internet Innovation Technology. This work was also sponsored by National Natural Science Foundation of China (No. 61602122, No. 71731004), Natural Science Foundation of Shanghai (No. 16ZR1402200). The associate editor coordinating the review of this paper and approving it for publication was Z. Zhu. *(Corresponding author: Yong Li.)*

J. Ding, S. Xiao, Y. Li, and D. Jin are with the Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing 100084, China, and also with the Department of Electronic Engineering, Tsinghua University, Beijing 100084, China (e-mail: dingjt15@mails.tsinghua.edu.cn; xsran@163.com; liyong07@tsinghua.edu.cn; jindp@tsinghua.edu.cn).

Z. Liu is with the National Computer Network Emergency Response Technical Team/Coordination Center of China, Beijing 100029, China (e-mail: lzj@cert.org.cn).

Y. Chen is with the School of Computer Science, Fudan University, Shanghai 200433, China (e-mail: chenyang@fudan.edu.cn).

S. Uhlig is with the School of Electronic Engineering and Computer Science, Queen Mary University of London, London E1 4NS, U.K. (e-mail: steve@eecs.qmul.ac.uk).

Digital Object Identifier 10.1109/TNSM.2018.2877819

I. Introduction

SINA Weibo (weibo.com) is the largest microblogging service in China. According to the Alexa ranking in October 2016, it ranks 7^{th} in China and 17^{th} in the world among all websites (www.alexa.com). Similar to Twitter, the world's largest microblogging service, it allows users to post short texts and get news quickly. In 2016, Weibo maintains 282 million monthly active users and 126 million daily active users. Besides the traditional microblogging services, Weibo has evolved into a multimedia blogging platform with features that combine Twitter, Pinterest and Tumblr, offering users the ability to share more images and video. In a way that is similar to YouTube, Weibo encourages users to create and upload video content around areas including food, sport and entertainment. Obviously, it plays a significant role in the whole Web ecosystem, especially in China.

The hyper-links that are such a fundamental part of the Web define click-based "transitions" from one site to another [1], [2]. Given the massive user base of Weibo, understanding these transitions is fundamental. Two specific aspects about them are most relevant. First, by investigating these transitions between Weibo and other external websites, we can better understand its role in the Web ecosystem. These transitions generated by users' actual clicks accurately describe the relationship between Weibo and other websites, thus being more reliable. Second, as a representative Chinese social media, Weibo and its massive user base have great value in social traffic referral. If there are frequent transitions between Weibo and other top websites, both Weibo itself and these Internet entities can consider using Weibo as an important source of referral traffic. Since different users have their own preference on clicking external hyper-links, one specific application can be a more personalized and well-designed scheme of hyper-link recommendation that ranks links by users' potential preference, instead of choosing those currently retweeted most times in Weibo to recommend.

However, *capturing these click-based transitions is notoriously difficult.* To our knowledge, since the transition data is not publicly accessible, no previous work has investigated transitions between Weibo and other websites. Indeed, most

existing Weibo studies are based on crawled data only containing publicly viewable information, such as social graphs, tweets, and user profiles [3]–[6]. Unfortunately, these datasets do not contain actual clicks made by users in Weibo. Weibo itself surely has such data, while it would not publish these data due to commercial sensitivity. The situation is similar for other microblogging services like Twitter. Without users' actual click information, investigating the typical usage of a microblogging service is significantly limited [7]–[9]. Kwak *et al.* [8] analyse the trending topics and news-related information diffusion in Twitter, showing that one typical usage of Twitter is news media. However, retweeting tweets is only one of activities in Twitter. Twitter users are also attracted by massive hyper-links posted in tweets. Once observing users' clicks on these, we may obtain unexpected insight. Gabielkov *et al.* [9] address this aspect by showing that high shares of tweets containing hyper-links do not imply high numbers of clicks. However, their click statistics are provided by a third-party URL shortening service, where the number and type of these links are limited.

In this work, by collaborating with one of the largest Chinese mobile operators, we obtained an anonymous dataset containing over 190*k* subscribers' HTTP traffic collected by using a deep packet inspection system. Our dataset was collected between April 21st and April 26th, 2016 in *Shanghai*. As clicked links are recorded and Weibo does not rely on HTTPS, this dataset fully supports our study.

Our analysis is mainly conducted in two aspects, overall transition analysis and user characterization. For the first aspect, we look at the links that drive incoming and outgoing transitions to and from Weibo. After that, we further characterize the transitions into those popular domains, as well as Weibo's connection with other large Chinese websites. To the best of our knowledge, this is the first work to study such transition activities between a mainstream online social media and a large number of other websites. For the second aspect, we ask whether the click behavior in transitions is distinct for different users. To answer this, we conduct a co-clustering process which groups both users (who share similar click behavior) and clicked URLs (of like-minded users) simultaneously. We further analyse users' preference in different transitions and their crawled profile information (the number of *following*, *tweet*s and *follower*s) to characterize their behaviors.

Our extensive analysis and results draw the following important conclusions.

- With the unique click data, we observe different roles of Weibo in Web ecosystem. As a microblogging service, Weibo still serves as a news media, with news-related traffic occupying 35% and 10% in incoming and outgoing transitions, while news-related links are posted more (18%) but clicked less (11%). Comparatively, video clip links are clicked more, i.e., 38% of outgoing transitions go to video clip service, while photo and article links are less popular. This indicates that photo and text-sharing in microblog have been gradually replaced by video clip sharing.
- Owing to information filtering when reading tweets in Weibo, exponential truncated power-law tail is observed in the distribution of transitions into other domains, indicating the long tail phenomenon. Except video clip service, 53% of transitions go to five of the largest Chinese websites. With such frequent transitions between Weibo and other large websites, we assert that Weibo has become a close-connected part of Chinese Web ecosystem.
- Weibo users' click behavior in transitions is distinct, and seven user-click co-clusters can capture the behavior of 192*k* users with 1.03*m* outgoing transitions. Among them, nearly 50% prefer clicking video clip links in Weibo. Unlike users generating more clicks (preferring leisure or social activities), other users (preferring video clips or news) click less, while with higher predictability. Moreover, popular users in Weibo, i.e., with higher number of followers, tweets and followings, prefer leisure activity and social interaction compared to others who use Weibo as a news media. These findings can aid the design of more personalized link recommendation in Weibo.

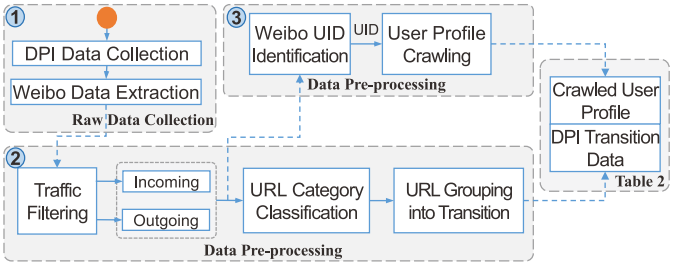


Fig. 1. Data collection and pre-processing.

This paper is structured as follows. We first detail the utilized dataset as well as data pre-processing. Then we conduct an overall analysis of transitions between Weibo and other websites. Next, we move forward to identify the key patterns of users' click behavior. After reviewing the related work, we discuss the implications, applications and limitations of this work. Finally, we summarize our work and discuss the future work.

II. Data Collection

In this section, we introduce the dataset we collected, as well as the pre-processing applied to it. The dataset contains two important parts: (1) transition data extracted from DPI traffic trace and (2) user-profile data crawled from Weibo website. The detailed method and process are presented in Fig. 1 and discussed now.

A. Raw Data Collection

The raw data was collected by one of the largest mobile operators in China using Deep Packet Inspection (DPI) appliances, between April 21*st* and 26*th* in 2016. Note that this week includes no holiday or special events. The coverage of data collection is the whole metropolitan area of *Shanghai*, including both urban and rural areas. The detailed information about users' Web visits are recorded in this dataset, including anonymized user IDs, timestamps for all HTTP requests and responses, and corresponding HTTP headers. As shown in

TABLE I
CATEGORIES FOR URL CLASSIFICATION

Category	Keywords	Category	Keywords
Advertisement	display-ad	**Entertainment**	star, show
BBS & Blog	forum, blog	**Finance**	finance, stock
Books	ebook, novel	**Life**	weather, cooking
Cloud	cloud drive	**Local**	local info.
E-commerce	online shopping	**Music**	online music
News	online news, news aggregator	**Video**	online video
Photo	picture, photo	**Video Clip**	video clip
SNS	social netw., q&a website	**Web Portal**	web portal
Software	App market	**Sports**	sports news, live streaming

Step 1 of Fig. 1, we extract the information related to Weibo by referring to several fields from the dataset, such as the destination URL field (*DestURL*), the "User-Agent" field (*UA*) and the "Referer" field (*Ref*). If *DestURL* or *Ref* matches the pattern ".*\.weibo\.cn/", or the App name in *UA* is "Weibo", this entry will be extracted to our dataset.

One limitation about the data collection is that DPI data captures no HTTPS traffic. Please see Section VI-C for further discussion.

B. Data Pre-Processing

1) DPI Transition Data: We then extract the transitions to and from Weibo in the obtained DPI dataset (Step 2).

Filtering Outgoing/Incoming Traffic: With the collected raw data, we further filter out the incoming or outgoing Weibo traffic. Specifically, we consider two different cases that users visit Weibo either through a browser or through a mobile App of Weibo. To determine whether an entry is related to "Web" or "App", we look at the UA field to check the App name. For traffic from an App, we consider all traffic with *DestURL* not matching the following pattern as outgoing: ".*\.weibo\.cn/". Note that our data collection cannot capture incoming Weibo traffic for the App for following two reasons. One is that most users manually launch Weibo App. The other is that the action of opening Weibo inside another App cannot be identified based on our captured data. As for traffic from a browser, the outgoing traffic is filtered out similarly as for the App. Oppositely, if the *Ref* field does not match the above pattern, then it will be considered as incoming traffic. From this step, we obtain 630*k* incoming URLs from 102*k* users, and 12.7*m* outgoing URLs from 192*k* users.[1]

Mapping URLs to Categories: After filtering outgoing and incoming Weibo traffic, we build a dictionary that maps the extracted URLs to their categories. Specifically, we look up the corresponding category on a per-domain basis for each URL. This dictionary combines category information provided by multiple navigation websites (as well as manual inspection), leading to 18 categories listed in Table I. Among these categories, Advertisement category refers to those URLs generated by clicking display-ads. BBS represents bulletin board system. The main difference between Video and Video Clip categories is that the latter mainly includes short videos. While a dictionary will be incomplete to classify all URLs, we find up to 95% unique visited URLs can be matched with our dictionary, and the unmatched URLs correspond to small websites with infrequent clicks. In this process, the category of traffic that goes into and out of Weibo depends on the domain in the *Ref* and *DestURL* fields respectively. After mapping, we keep the corresponding category of each URL, as well as the key, i.e., the matched domain.

Definition of Transition: In general, users reach or leave Weibo through hyper-links (URLs). One click of a URL corresponds to one transition. However, clicking one URL may generate multiple HTTP requests, potentially biasing the count of transitions in our data. Therefore, we merge multiple consecutive URLs into one transition when they satisfy the following: (a) they have the same *Ref*, (b) the same category are seen within a short ($< 60s$) time interval.[2] Finally we obtain the DPI transition data with each record as a transition or click, including user ID, time stamp, platform ("Web" or "App"), transition type (outgoing or incoming), corresponding category and matched domain. The basic statistics are summarized in Table II.

2) Crawled User Profile: To help with investigating click behavior in terms of transitions between Weibo and external sites, we crawl the users' corresponding Weibo profile information (number of *following*, *follower*s, and *tweet*s).

As shown in Step 3, we first extract the *uid* of a user's Weibo account in our transition data. We collect manual traces for which we know the ground truth—by passively monitoring our interactions with Weibo and extracting the patterns for matching *uid*. With these patterns, we extract 95,259 users' *uid*s in Weibo. Then, we implement a crawler which downloads the profile page of each of these Weibo users, with needed profile information of *following*, *follower*s and *tweet*s.

III. OVERALL TRANSITION ANALYSIS

Based on the obtained transition data, we present a macroscopic analysis on the traffic transitions through Weibo. First, we investigate the overall incoming ("In") and outgoing ("Out") transitions of Weibo. Then, we characterize transitions by comparing their clicked URLs with those URLs posted in users' tweets. Finally, we analyse transitions at domain-level.

[1]Since our data are collected by one operator with a uniform sampling of users and a large proportion of Weibo users does not click hyper-links and jump out of Weibo, about 192*k* users are identified.

[2]The smaller the time interval between two consecutive URLs, the more we expect they are related. However, we observed a local maximum at about 60s, indicating a good crossover point to split two transitions.

TABLE II
SUMMARY OF THE DPI TRANSITION DATA

Date Range	Type	# Users	# Transitions	Fields
2016/04/21	out	192k	1.03m	user ID, time, platform,
∼2016/04/26	in	102k	320k	type, category, domain

TABLE III
MAIN TRANSITION CATEGORIES

Dataset	Percentage	Category	Typical Domains
Web (In)	17%	News	*news.sina*, *toutiao*
	15%	Books	*book.ifeng*, *qidian*
	14%	Web Portal	*sina*, *3g.163*, *m.sohu*
	11%	Sports	*hupu*, *sports.sina*
	7%	Finance	*finance.sina*, *eastmoney*
App (Out)	37%	Video Clip	*miaopai*
	13%	Ad.	*pos.baidu*, *tanx*
	10%	News	*ifeng*, *toutiao*
	7%	Photo	*meitudata*, *photo.sina*
	5%	Video	*youku*, *v.qq*, *iqiyi*
Web (Out)	43%	Video Clip	*miaopai*
	16%	Photo	*meitudata*, *photo.sina*
	15%	Ad.	*pos.baidu*, *tanx*
	8%	SNS	*qzs.qq*, *weixin.qq*, *zhihu*
	6%	Music	*music.sina*, *music.163*

A. Incoming and Outgoing Transitions

We classify traffic transitions into three groups, namely incoming/outgoing transitions ("In/Out"), either from the Weibo App or a browser ("App/Web"). Table III summarizes top five categories that three different kinds of transitions fall into.

Incoming Transitions: Among the incoming transitions in browsers, we find that websites of news-related categories, i.e., News, Sports and Finance, contribute 35% of the transitions. To ease the reading, we focus on the frequently occurring domains in these categories. *news.sina*, *sports.sina*, and *finance.sina* all provide online news services. *Toutiao* is the most popular news content aggregator in China, which offers similar service as *Flipboard*. Tracing back the original clicked URLs, we find that these webpages often provide a share widget to Weibo, which indicates that news websites tend to use Weibo as a venue to spread their content. Similar to news-related categories, literature websites like *qidian* and *book.ifeng* also rely on Weibo to promote their content, such as novels. Besides, Web portals also provide an entrance point to Weibo, occupying 14% of the incoming transitions. Note that *sina*, *3g.163* and *m.sohu* belong to *Sina*, *Netease* and *Sohu* respectively, which are the three largest Chinese Web portals.

Outgoing Transitions: One striking similarity between the outgoing transitions on both the App and browser platforms, is the dominance of the category Video Clip. The most common domain is *miaopai*, a video clip sharing App like *Vine*. It dominates the video clip market in Weibo, through an aggressive strategy such as inviting celebrities to share Miaopai video clips in Weibo and thus attract many users. Besides video clip sharing, the category Advertisement also ranks high on both platforms, allowing Weibo to leverage its large user base. *tanx* and *pos.baidu* provide display-ad services like *googlesyndication*, a similar kind of service provided by *Google*. These ads come in multiple forms, such as banners and skyscrapers. Picture websites are also one category popular in outgoing transitions. One of them, *meitudata*, corresponds to the popular image editing software *Meitu*.

In terms of the difference between two types of platforms from which transitions are observed, 10% of outgoing transitions from the Weibo App go to news websites, while 8% of those from browsers go to SNS websites like *qzs.qq*, *weixin.qq* and *zhihu*. This difference suggests that Weibo is more of a news media for App users and a platform of online social interaction for browser-based users. Just to mention a few examples, *qzs.qq* refers to *Qzone*, a Chinese clone of *Myspace*, and *weixin.qq* links to shared articles in the *WeChat* application, a popular instant messaging service that also allows users to post images and share music. We observe that *zhihu*, a Chinese clone of *Quora*, is classified into the SNS category due to its social media feature.

Summary: As a microblogging service, Weibo still serves as a news media, with large amounts of news-related traffic in both incoming and outgoing transitions, covering about 35% and 10% respectively. By providing sharing widgets to Weibo, news websites promote their content and in return gain large amounts of traffic from Weibo. More importantly, we observe that 38% of outgoing transitions (413k) go to the *miaopai* video clip service. We investigate next whether this suggests a shift in the role of Weibo towards becoming an important video-sharing platform in China.

B. Clicked and Posted Links

Since an outgoing transition is generated by clicking on an external link, it tells us how Weibo drives traffic to other websites. Besides, in our transition data, the number of outgoing transitions is about three times that of incoming transitions (1.03m vs 320k). Therefore, we focus on outgoing transitions in further analysis.

Intuitively, links posted in users' tweets reflect the usage of Weibo, since most transitions out of Weibo, except for those related to display-ads, are generated by clicking these links. Given its fundamental importance of understanding Weibo's role, we investigate *how it has evolved over time*. Moreover, unlike most posted links, links corresponding to outgoing transitions are actually clicked by users. Therefore, we also investigate *the difference between these two types of links*. To address the above two aspects, we introduce another Weibo tweet dataset, containing Weibo users' posted links in 2012 and 2016 [10]. To be specific, it contains 770k posted links with 190k users in 2012 and 1.40m posted links with 54k users in 2016. To investigate the first aspect, we compare the posted links in 2012 and 2016. For the second, we analyse the differences between these posted links and the clicked links.

In Table IV(a), we list the top 5 URL categories in these three types of links. Since users generally do not post links related to display-ads in their tweets, we have removed the category Advertisement from our statistics. Comparing the links posted in 2012 and 2016, the most significant difference is that video-related categories, especially video clips, have increased and taken over the place of SNS. As for differences between posted links and clicked links in 2016, the SNS and

TABLE IV
POSTED LINKS VERSUS CLICKED LINKS

(a) top 5 categories

Posted Links, 2012		Posted Links, 2016		Clicked Links, 2016	
Percentage	**Category**	**Percentage**	**Category**	**Percentage**	**Category**
79%	SNS	44%	Video Clip	44%	Video Clip
5%	Video	18%	News	11%	News
5%	BBS&Blog	16%	Video	9%	Photo
2%	News	10%	SNS	5%	Video
2%	Music	3%	BBS&Blog	5%	E-commerce

(b) rank correlation analysis

Pair	**Kendall's τ**	**Spearman's ρ**
2012-2016	0.3725	0.4365
posted-clicked	0.2941	0.3519

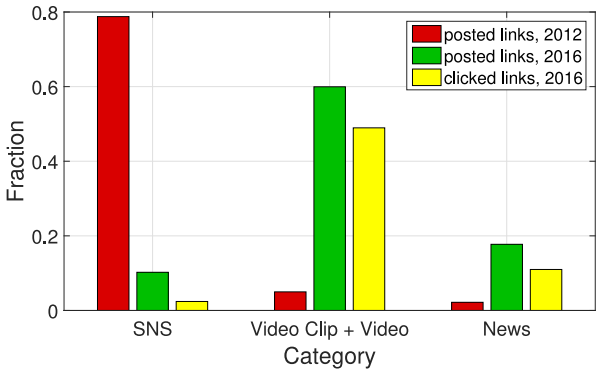


Fig. 2. Comparison in Specific URL Categories.

BBS&Blog categories are replaced by Photo and E-commerce categories in clicked links. Also, the ranks and proportions for the other three categories are not the same between posted and clicked links.

To highlight the differences observed above, in Fig. 2, we plot the fractions of the three aforementioned categories within each kind of links, i.e., posted links in 2012 (red), posted links in 2016 (green) and clicked links in 2016 (yellow). For the evolution of the Weibo usage, the SNS category, which takes up to 79% of the links in 2012, is replaced by video-related categories (Video Clip plus Video) in 2016, with about 60% of such links. This corresponds to our finding earlier in the paper that video-related services are becoming increasingly popular over time, taking over the place of once popular photo or article sharing services in SNS. More importantly, different from posted links, the fraction of clicked links for the category Video decreases from 16% to 5%. This happens while the fraction of the Video Clip category remains almost the same, indicating that video clip services are actually much more popular. This stems from the ease of watching and sharing video clips in Weibo, making it particularly attractive to users. Also, the fraction belonging to the News category in posted links (2016) is about twice as in clicked links (2016), i.e., 18% vs. 11%, indicating that news-related links are posted more and clicked less.

To quantitatively characterize the difference between these three types of links, we conduct a rank correlation analysis for the categories of the links, by computing the Kendall's τ and

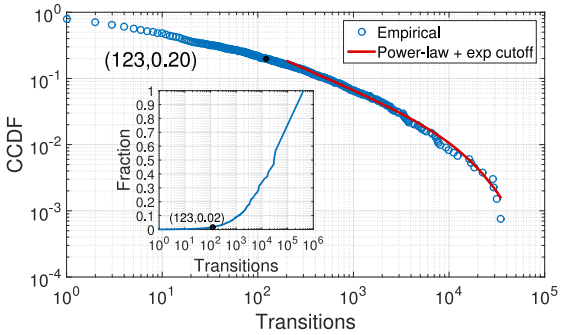


Fig. 3. The CCDF of transitions that go into each domain, with the inset showing the fraction.

Spearman's ρ [11]. Both two metrics are nonparametric measure of rank correlation (statistical dependence between the ranking of two variables). They assess how well the relationship between two variables can be described using a monotonic function, with a range from -1 to 1 and higher absolute value indicating higher correlation. The results are summarized in Table IV(b). Notice that all four values in Table IV(b) are lower than 0.5, indicating a low correlation among them. With the values for posted and clicked links lower than those for posted links in different years, we show there exists the larger difference between posted links and clicked links in Weibo.

In summary, confirming what we concluded in the previous part, our investigation of the temporal evolution of the Weibo usage shows that photo-sharing and text-sharing in Weibo are gradually replaced by video clip sharing. Different from the statistics of posted links, video clip links are clicked more while news links are clicked less.

C. Domains in Transitions

Besides URL categories, we also investigate transition characteristics at the domain-level by focusing on popular domains, i.e., domains attracting more traffic from Weibo. These popular domains can tell us what gets most clicks in Weibo, which is more reliable compared with those observed from link analysis.

Long Tail Phenomenon: We first analyse the distribution of popular domains in transitions, by plotting the complementary cumulative distribution function (CCDF) of transitions into each domain in Fig. 3. Our fitting shows that the tail of CCDF ($x \geq 200$) can be fitted with an exponential truncated power-law distribution $x^{\alpha}e^{-\lambda x}$. With α and λ set as -0.617 and 4.59×10^{-5}, the adjusted R^2, i.e., the coefficient of determination, equals to 0.9969. According to the main figure, 20% of the domains have more than 123 transitions. At the same time, the fraction of total transitions captured by these domains, which we plot in the inset, can be up to 98%, indicating a Pareto-like behavior.[3] Considering that each user only browses a limited number of tweets, a single user can only click a limited number of links related to these popular

[3]Applying Pareto's law in our context would mean that a majority of the transitions (say, for instance 80%) should come from a restricted, possibly very small, fraction of the domains. These domains are often referred to as *blockbusters*.

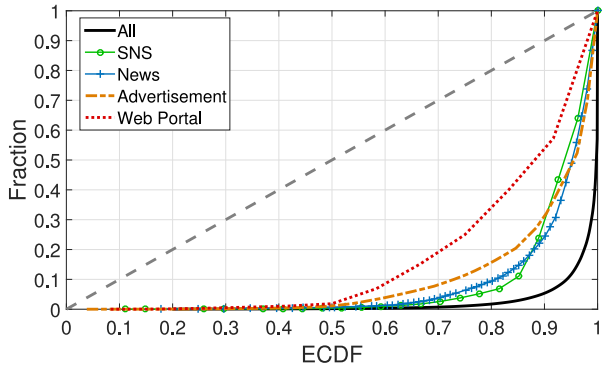
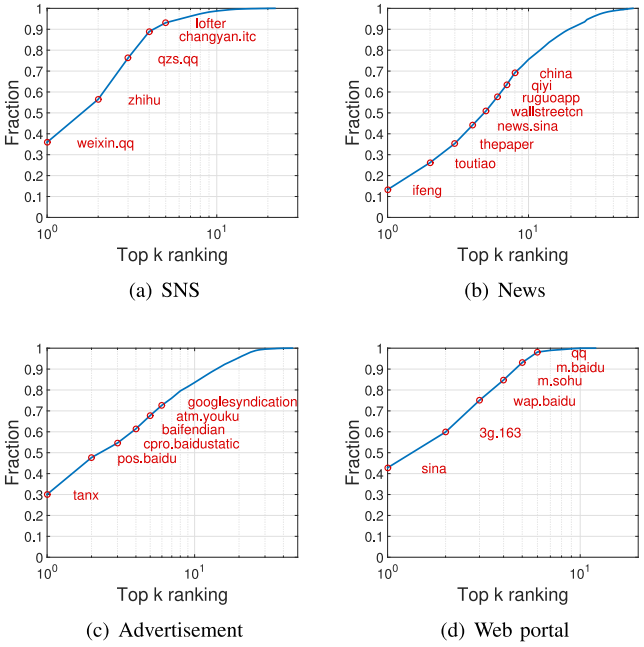


Fig. 4. Comparison of Pareto-like behaviors across categories. X-axis shows the empirical CDF of transitions that go into each domain and y-axis shows the corresponding cumulative fraction.



(a) SNS (b) News

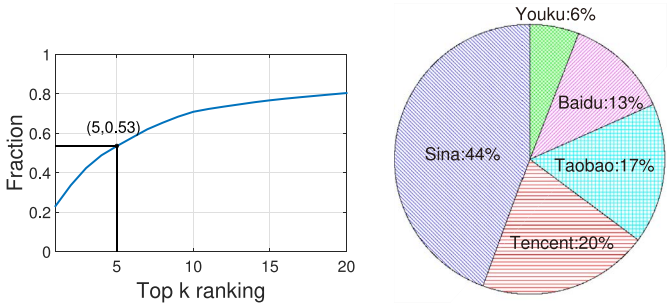
(c) Advertisement (d) Web portal

Fig. 5. Cumulative fraction of transitions for top k ranking domains.

domains, which causes the exponential cutoff of the tail. This corresponds to the concept of "information filtering" described in degree distribution of the world wide Web [12].

We plot in Fig. 4 the empirical CDFs (ECDF) of transitions versus the corresponding cumulative fraction for four different categories of domains, to compare different categories. Each curve corresponds to one category, while the grey dashed line is provided for reference, corresponding to the situation where each individual domain gets equal transitions. The distance between this line and each curve is longer when more transitions are captured by popular domains in this category. For example, when ECDF equals to 0.8, the captured transition fraction of unpopular domains (80%) is 0.07, 0.10, 0.16, and 0.40 for category SNS, News, Ad. and Web Portal, respectively. This indicates that Pareto's behavior in category SNS is the most significant.

We further analyse the highly popular domains in the above four categories. In Fig. 5, we plot top k ranking domains and compute the cumulative fraction of overall transitions



(a) overall (b) top 5

Fig. 6. Fraction of overall transitions for top k ranking websites (left) and individual fractions for top 5 websites (right).

they capture, for each category. First, *WeChat* (*weixin.qq*), *Qzone* (*qzs.qq*) and *Zhihu* (*zhihu*) are the three most popular domains in the SNS category. They contribute over 75% of the transitions in this category, indicating a very strong tail. In the News category, we find 8 very popular domains that contribute 70% of the transitions. Besides *Baidu* (*pos.baidu*) and *tanx*, display-ad services provided by *Youku* (*atm.youku*) and *Google* (*googlesyndication*) also appear in the category Advertisement. The top domain in the Web Portal category is *Sina* (*sina.com.cn*), followed by Web portal services provided by *Netease* (*3g.163*), *Baidu* (*wap.baidu*) and *Sohu* (*m.sohu*).

More importantly, among these popular domains in transitions, we observe that most of them belong to few Chinese Internet giants, such as *Sina*, *Baidu* and *Tencent*. To further investigate possible connections between Weibo and these websites, we aggregate the transitions whose destination domains belong to the same website.

Transitions to Top Websites: Since the video clip service accounts for 38% of transitions, we remove these transitions and show the fraction of remaining transitions across the ranked websites in Fig. 6(a). The top five websites account for 53% of the transitions,[4] whose own proportions are shown in Fig. 6(b).

These websites that account for most transitions also reflect their importance in the Web ecosystem. Their corresponding Alexa rankings in China are 8, 2, 3, 1 and 15. We further look into the details of these five websites, in terms of the most frequent sub-domains that appear in transitions. First, we investigate the sub-domains of *Sina*, which ranks highest in Fig. 6(b). We observe that *news.sina* and *photo.sina* occupy most transitions. As the parent company of Weibo, as well as being a large Web portal, it is expected that *Sina* manages to keep traffic inside its own system. Second, other websites act consistently with their main functions. For example, *Tencent* steers most traffic transitions to *WeChat* and *Qzone*, its two most popular social network services. As for *Baidu*, Weibo users mainly use its Web portal service (*wap.baidu,m.baidu*). Since all these five websites are Chinese Internet giants, large amounts of transitions from Weibo to these websites indicate

[4]The websites are Sina (Weibo's parent company and large Web portal), Tencent (largest Chinese social network provider), Taobao (largest e-commerce website in China), Baidu(largest search engine in China), and Youku (largest Chinese video provider). All ranked by Alexa.

that Weibo has managed to become an important player in the Chinese Web ecosystem, not a dedicated social networking player, therefore ensuring its future as a hub in the Chinese Internet market.

Standing as one of the largest social media, Weibo provides its users an entrance into other parts of Chinese Web that providing various kinds of services, like search, video play, online shopping and so on. Users are able to visit into and out of Weibo through the specific links posted in Weibo. If using proper method of link recommendation towards Weibo users, other websites can benefit from obtaining these users' visits from Weibo.

Summary of Overall Analysis: First, different from publicly accessible Weibo data that only contain users posted links, our analysis of user clicks shows that photo and text-sharing have been gradually replaced by video clip sharing in Weibo, where video clip links are clicked more while news links are clicked less. Second, with frequent transitions between Weibo and other five of the largest Chinese websites, Weibo has become a close-connected part of the Chinese Web ecosystem. More importantly, this implies Weibo's great value in social traffic referral.

IV. User Behavior in Transitions

Through clicks on external links in Weibo, a large number of referral traffic is driven to various external websites. To capture user behavior, we analyse actual clicked links in this section. We first conduct a co-clustering process on both Weibo users and their clicked URLs (*Key User Patterns in Transitions* Section). With the help of extracted user behavior patterns, we answer the following two questions.

Which Types of Users Click More Links? In *User Preference in Transitions* Section, we provide a detailed analysis of users' preferences in outgoing transitions, including the total number of clicks, the predictability of clicked links, as well as temporal features.

Which Type of Users Are Popular Users in Weibo? We answer this question in the *Profile Analysis* Section, by looking at crawled users' profile information.

A. Key User Patterns in Transitions

Since both users and their clicking patterns are influenced by each other and are not independent, we need to capture the grouping among users induced by the clicking patterns, and the grouping among clicking patterns induced by users simultaneously. With the extracted 192*k* Weibo users and 1.03*m* outgoing transitions, we apply a hierarchical co-clustering algorithm [13]. The input data is a matrix with users as rows, and the URL categories of clicks as columns, with each entry representing the number of clicked links belonging to this category. We denote this matrix as D, with m rows (users) and n columns (categories). The goal of co-clustering is to group both users and URL categories at the same time. The final outputs are k user-click co-clusters, denoted as $C^{(u,l)} = \{c_1^{(u,l)}, \ldots, c_k^{(u,l)}\}$. Every co-cluster $c_j^{(u,l)}$ consists of users ($c_j^{(u)}$) and categories ($c_j^{(l)}$). The basic idea is to start with the entire input matrix as a single cluster, with all users and categories, then iteratively partition each cluster into two children clusters by using the SpectralGraph-2-Part co-clustering algorithm [14]. A metric called *cluster cohesiveness* is defined to evaluate how well the co-clusters are distinct from each other. For a co-cluster $c_j^{(u,l)}$, its *cluster cohesiveness*, $\gamma_j^{(u,l)}$, is the ratio of the sum of the weights of associations between users $c_j^{(u)}$ and categories $c_j^{(l)}$ to the sum of the weights of all the associations of users $c_j^{(u)}$, defined as

$$\gamma_j^{(u,l)} = \frac{\sum_{a \in c_j^{(u)}} \sum_{b \in c_j^{(l)}} D_{ab}}{\sum_{a \in c_j^{(u)}} \sum_{b \in C^{(l)}} D_{ab}}. \quad (1)$$

Once all children clusters cannot be partitioned anymore, i.e., the obtained *cluster cohesiveness* is lower than the given threshold, we obtain the final outputs $C^{(u,l)}$.

We obtain seven user-click co-clusters, listed in Table V. All *cluster cohesiveness* are higher than 0.78 except two, which supports the credibility of the obtained co-clustering results. A visualization of our seven co-clusters is provided in Fig. 7, where each label is one URL category and each edge represents a click made by a user. Users and categories in the same co-cluster are plotted in the same color. From the visual, we can observe that users prefer to click links of some certain categories. Therefore, there are several groups of users and

TABLE V
User-Click Co-Clusters

No.	Click Categories	Cohesiveness	User
C_1	Video Clip	0.8800	88609
C_2	Ad., E-commerce, BBS & Blog, Video	0.7249	39395
C_3	Photo, Music, SNS	0.7807	26728
C_4	Entert., News	0.8066	25699
C_5	Local, Software	0.8205	5913
C_6	Cloud, Web Portal	0.7265	4266
C_7	Life	0.8959	1698

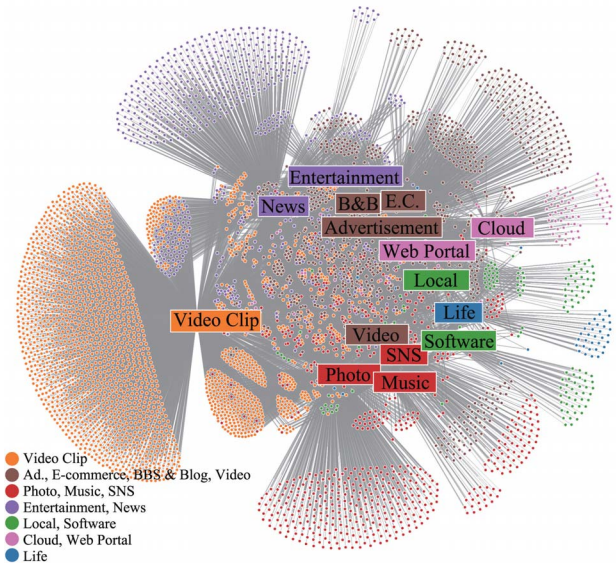


Fig. 7. Visualization of the user-click co-clusters.

categories with tight connections in this graph. This indicates that the co-clusters capture users' preference when clicking links in Weibo. We now look specifically into each co-cluster listed in Table V.

C_1 (Video Clip): Nearly half of the users are in this co-cluster. They choose to watch video clips, indicating a typical usage of Weibo.

C_2 (Leisure Activity): About 20% of users are in this co-cluster. They tend to visit external websites related to leisure activities, like watching videos, ads, reading blogs and shopping.

C_3 (Social Interaction): As a SNS website, Weibo provides a place for users to share liked images and music, as well as content from other SNS websites. We observe 14% of the users choosing to click on these links.

C_4 (News): As mentioned before, news media is one of the important roles played by Weibo. About 13% of users go to news-related websites from Weibo.

C_5 (App Download): Mobile Apps are promoted through Weibo, some of which have attracted quite a few users.

C_6 (Tool): A minority of users, about 2%, use Weibo as a tool when they need to obtain resources and information from cloud drives and Web portals.

C_7 (Dailylife): As the smallest co-cluster, the daily life category contains for example weather forecast websites.

Overall, there are several dominant and distinct behavior patterns from users' clicks in Weibo. With these extracted co-clusters, personalized recommendation schemes can be provided, which will recommend links to users across various co-clusters when they use Weibo. Compared with currently used schemes such as ranking links by their trends, we believe such recommendations can help generating more referral traffic to other websites.

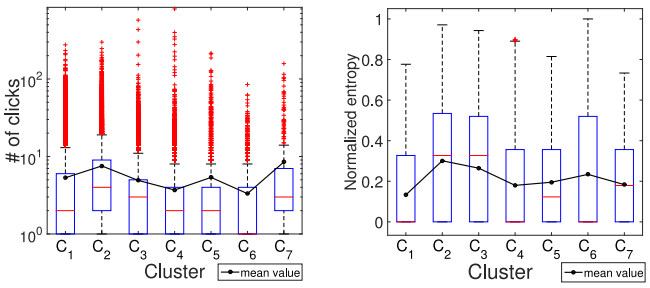
B. User Preference in Transitions

In order to understand which users click more external links in Weibo, we provide a detailed analysis of users' preference across two aspects.

Number and Predictability of Clicks: Intuitively, we need to count the sum of URLs clicked by each user. However, how predictable is the clicked links by a user? In other words, does the user always click links belonging to her own co-cluster? To answer this question, we investigate the entropy of clicked links at the category-level, which can be used to indicate predictability [15]. The entropy, H_{vj}, of user v in co-cluster $c_j^{(u,l)}$, is defined as

$$H_{vj} = -\sum_{i=1}^{k} \left\{ \frac{\sum_{b \in c_i^{(l)}} D_{vb}}{D_v} \right\} \times log_2 \left\{ \frac{\sum_{b \in c_i^{(l)}} D_{vb}}{D_v} \right\}, \tag{2}$$

where D_{vb} represents the clicks of b-category links by user v and D_v is the number of his total clicks. A low value of the entropy indicates a high predictability of user's clicks at the category-level. When v only clicks links in his own co-cluster $c_j^{(u,l)}$, H_{vj} reaches its minimum as 0, indicating that his clicks are highly predictable as the possible number of categories is 1. Otherwise, if v clicks equal number of links in each co-cluster, H_{vj} equals to its maximum as $log_2 k$ ($log_2 7$ in our case), while the possible number of clicked categories is k (7 in our case).



(a) Number of clicks (b) Normalized entropy

Fig. 8. Users' number and predictability of clicks in each co-cluster.

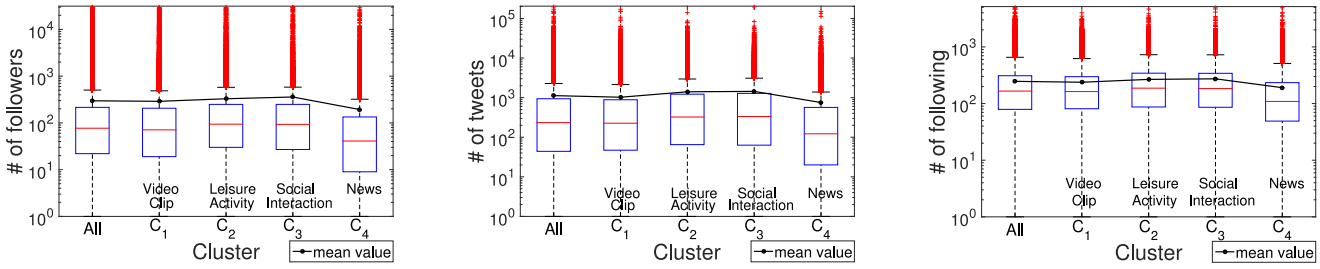
TABLE VI
TEMPORAL FEATURES OF CLICK RATIO

	Co-clusters						
Period	C_1	C_2	C_3	C_4	C_5	C_6	C_7
night	0.056	0.036	0.050	0.049	0.067	0.045	0.047
morning	0.233	0.300	0.290	0.306	0.274	0.299	0.244
afternoon	0.377	0.404	0.369	0.342	0.376	0.367	0.360
evening	0.334	0.260	0.291	0.303	0.283	0.289	0.349

The users' number of clicks and normalized entropy by possible maximum are displayed in Fig. 8, using boxplots[5] and curves of mean values. For the number of clicks, with mean values higher than medians and many outliers, we observe a significant right-skewed characteristic in all co-clusters. Among these, users in C_2 (leisure activity), C_3 (social interaction) and C_7 (dailylife) generate more clicks. Unlike news or video clips that may only be clicked once, links related to online shopping or photos can be clicked multiple times by the same user, thus generating more clicks. However, as shown in Fig. 8(b), clicks of users in C_1 (video clip) and C_4 (news) are more predictable, while those for C_2 and C_3 are much less. Though some users click less, their clicks are much more predictable, i.e., less links outside the considered co-cluster, indicating the consistent click behavior of these users.

Temporal Features: Investigating temporal features of users' clicks helps to understand the users' click behavior, and further develop time-dependent schemes for traffic referral. We divide one day into four periods, night (0-6), morning (6-12), afternoon (12-18) and evening (18-24). For each co-cluster, we compute the ratios of clicks during the four periods. The results are summarized in Table VI, where click ratios of each column sum up to 1. Overall, we observe that all co-clusters have the lowest and highest ratio in the night and afternoon, respectively, which corresponds to the users' lifestyle. As for the other two periods, click ratios of C_4 in the morning and C_7 in the evening are labelled using blue color because they are not only the highest in the corresponding period, but also

[5]The box shows the limits of the middle half of the data; the middle line represents the median; whiskers are drawn to the nearest value not beyond a standard deviation from the quartiles; outliers are drawn individually.



(a) Follower (b) Tweet (c) Following

Fig. 9. Users' crawled profile information in each co-cluster.

the closest to the highest ratio of their own co-cluster (in the afternoon). For C_4, the relative difference between click ratios in the morning and afternoon is 0.106, with those of other co-clusters higher than 0.184. As for C_7, similarly, the relative difference is 0.032, with others all higher than 0.112. Therefore, compared to others, users in C_4 (news) generate more clicks in the morning and users in C_7 (dailylife) click more in the evening. The underlying reason is that users tend to read news in the morning and care about the weather in the evening.

Summary: People use Weibo for leisure, social interaction and daily life click more. Unlike them, users preferring news or video clips click less, while with more predictable and consistent click behavior. In terms of time dynamics, users tend to click news-related links and weather-related links in the morning and evening, respectively.

C. User Profile Analysis

To better understand different clusters of users' roles in Weibo, we investigate their profile information (*follower*s, *tweet*s and *following*). Since the first four co-clusters (C_1 to C_4) capture over 90% of users and represent typical click behaviors, we only focus on them. The boxplots and mean values of the profile information for users in these four co-clusters (as well as all users) are shown in Fig. 9. Previous work [8] has reported a positive correlation between the number of tweets, followers and following in microblogging services. This explains why all three metrics exhibit the same characteristic among different co-clusters, as shown in Fig. 9. Since nearly 50% of the users belongs to C_1 (video clip), the statistics of C_1 are almost the same as the overall ones. Comparatively, users in C_2 (leisure activity) and C_3 (social interaction) have higher numbers of *follower*s, *tweet*s and *following*, while it is opposite for C_4 (news). This indicates that the former are more like popular users in Weibo. The underlying reason is that users preferring leisure and social activities share their feelings by tweets and interact more. Differently, users using Weibo as a news media just fetch news and rarely interact with other users. In summary, popular users generally use Weibo in leisure activity and social interaction, compared to others who barely use Weibo as a news media.

Summary of User Characterization: Through a co-clustering process, we show the click behavior of 192k studied users can be grouped into only seven typical patterns. These patterns of clicking links, along with users' preference and profile information, can help developing personalized link recommendation to provide social traffic referral for other major websites.

V. Related Work

We first review the existing works on analyzing the microblogging service. Since we analyze the transitions in Weibo based on click-through data, we also discuss the related works on utilizing users' click-through data as well as intra-website clickstream to improve Web services, followed by a few works that are related with inter-website transitions.

Microblogging Services: As the largest microblogging service in China, Weibo has received much attention recently. One hot topic is to investigate users' reactions after disasters or specific social events [16], [17]. Reference [16] observed a propensity to spread certain types of information (i.e., action-related messages) in Weibo after disasters. Reference [17] found out those URL-containing tweets are less likely to be reposted for specific social events. As not all microblogging platforms are equal, Gao *et al.* [18] have been focused on a comparative study between Twitter and Weibo, the two representative microblogging services. They reported a much more intensive usage of hashtags and URLs in Twitter, mainly related to both the lower power distance and the higher degree of individualism of Twitter users. Another popular topics include the microblog's ability to spread information [19]–[21], content value or credibility [22]–[24], and topic features [25], [26]. However, focusing on user behavior analysis in the Weibo, these works cannot reveal the relationship between Weibo and other Web services. Different from them, in this paper, we focus on the traffic transitions between Weibo and other websites, and observe that a majority of users click on videos and news, which helps discovering the role of microblogging platforms in the Web ecosystem.

Clickthrough on Web: User clickthrough data can be used to introduce more accurate description of user behvior and thus improve Web services, especially in Web search engines [27]. Joachims *et al.* [28] concluded that clicks are informative but biased when serving as the implicit feedback. Later, Agichtein *et al.* [29] incorporated it into the ranking process and improved the performance. A deep structured semantic model based on clickthrough data was also proposed in [30]. Besides search engines, click-through data have also been used in online advertising. Reference [31] demonstrated that the features of ads, terms, and advertisers can learn a model that

accurately predicts the clickthough rate for new ads. Recently, a more sophisticated design of handling multi-field features using adaptive weights was proposed [32]. Attention mechanism was also introduced to further improve the performance of clickthrough rate forecasting [33]. Different from above works directly using clickthrough data to improve Web services, we analyze Weibo users' clicks on those URL links from both the overall perspective and individual perspective. And our findings have the potential to further improve Weibo services.

Intra-Website Clickstream: Intra-website clickstream data have been widely used to model user behavior in Web services [34], [35]. Reference [34] modeled browsing patterns (sequence of clicks) via Markov chain theory to predict users' propensity to buy within a session. Similarly, students' clickstream data were used to improve MOOC performance prediction [36]. Recently, the e-commerce click streams irevealing usage patterns and shopping behaviors have been analyzed [37]. Clustering techniques has been utilized to capture users' behavior, either with large scale ground-truth data [38]–[40], or without ground-truth through unsupervised schemes [41]. In this paper, having captured Weibo users' click events, instead of characterizing users' browsing patterns inside Weibo, we focus on their click behavior in transitions to other websites and extract seven patterns rely on clustering process, which helps to develop more personalized schemes to provide social traffic referral for other websites.

Inter-Website Transition: Few works have studied inter-website traffic transitions [42], [43]. Qiu *et al.* [42] found that search engines help users reach 20% more sites by presenting them in search results, which may be otherwise unreachable. Meiss *et al.* [43] proposed a topical model of Web navigation, which served as a good benchmark for ranking and crawling algorithms. One close related work [9] has investigated both users' clicks and shares on news-related links in Twitter, where the clicks dynamics were observed to be long-lived. Different from them, we focus on the relationship between Weibo and other websites, instead of the impact of microblogs in Weibo. Therefore, our work relies on users' specific click events to analyse the inter-website traffic transition with the Weibo as the centre.

VI. Discussion

A. Implications

By observing Weibo users' clicks on URL links, we investigate the role of Weibo as a microblogging service in the Web ecosystem. There are two existing works that are closely related to ours. One is a quantitative study on the topological characteristics and information diffusion of Twitter [8]. The other is a large scale study of social clicks on those news-related links that are propagated in Twitter [9].

Both two works have reported an important use of Twitter as a news media. Reference [8] extracted the trending topics and found that the majority (over 85%) of them were headline news or persistent news in nature. Reference [9] further pointed out that the secondary news resources, which were not promoted through headlines and less popular, generated more clicks. Based on these findings, we analyze users' actual clicks on all types of links in Weibo, which are not limited to news-related type. Similarly, we also observe a news media flavor in Weibo, with news-related traffic occupying 35% and 10% in incoming and outgoing transitions. However, news-related links are posted more (18%) but clicked less (11%). Comparatively, video clip links are clicked more, i.e., 38% of outgoing transitions go to video clip service, indicating the increasing popularity of video clips in Weibo. Moreover, instead of focusing on the information diffusion inside Weibo, we are able to analyze relationship between Weibo and other websites based on these click-based transitions. As 53% of traffic (video clips traffic are not considered) go to five of the largest Chinese websites, we conclude that Weibo has become a close-connected part in Chinese Web ecosystems.

In terms of predicting the actual influence of content in microblogging services, [8] observed a gap between influence inferred from the number of followers and that from the popularity of one's tweets. Reference [9] also demonstrated the difficulty of predicting whether a user will click a link in Twitter. Based on a co-clustering process, we extract seven patterns from user-link interaction history, which can serve as the user profile information that help predicting users' preference on various types of links in Weibo.

B. Applications

An important application of our findings is to improve the personalized recommender systems for Weibo hyper-links. Current tweet recommendation schemes in Twitter or Weibo generally leverage both user and content information to design the personalized models. For example, a well-known collaborative model [44] considered three major elements on Twitter: tweet topic level factors, user social relation factors and other explicit features. However, the sparsity of user interaction data remains as a challenging problem. Since a user only interacts with (clicks) a few links, it is hard to train the model with sparse data, especially for those long-tail, i.e., inactive, users.

From the co-clustering results of user-link click history, we extract seven typical user click patterns, each corresponding to a specific user-link group. For those users belonging to the same group, we can assume that they share the same interest. Then, for each user, besides a few number of actually clicked links, those clicked by other group users can also be leveraged to learn user preference. An intuitive way of integrating it into recommendation model is to augment the training data. More specifically, the interactions between users and their group users' clicked link are assigned with an intermediate label between 0 (not clicked) and 1 (clicked). Therefore, with more information of user preference, the improved recommender system can provide much more personalized and accurate link candidates for Weibo users.

C. Limitations

There are two limitations on data collection.

One limitation is about the missing HTTPS traffic. According to our observation, Weibo does not use HTTPS to encrypt its traffic. As for other external websites, appearing

in incoming or outgoing traffic of Weibo, we have checked the traffic of top Chinese websites [45] in our data and only 14% actually use HTTPS. Moreover, few websites choose to encrypt the whole traffic, indicating the number of websites in our captured traffic data does not shrink. For example, *Baidu* only encrypts its traffic for its search engine service, and URLs rely on HTTP when entering an online shop in *Taobao*.[6] Thus, the impact of missing HTTPS traffic is limited.

Another limitation is that our study focuses on Weibo users in Shanghai. Note that the coverage of data collection is the whole metropolitan area of *Shanghai*, which includes both urban and rural areas. Though without the sample of users across the whole country, it can still serve as a good lens to observe the role of microblogging services in the Web ecosystem.

VII. Conclusion

Based on transition data and crawled user profiles, we investigate the traffic transitions between Weibo and other external websites. We conduct a deep analysis of both links and the corresponding domains seen in these transitions. We find that photo-sharing and text-sharing in Weibo are gradually replaced by video clip sharing. From the frequent transitions between Weibo and other large Chinese websites, we observe that Weibo has now become a close-connected part of the Chinese Web ecosystem. In terms of Weibo users' distinct click behavior, our co-clustering process extracts seven patterns that capture the behavior of 192*k* studied users. Knowing different users' preference on clicking links in Weibo, we are able to develop more personalized schemes to provide social traffic referral for other websites.

References

[1] M. Henzinger, "Hyperlink analysis on the World Wide Web," in *Proc. ACM HYPERTEXT*, 2005.

[2] J. M. B. Cavalcanti and D. Robertson, "Synthesis of Web sites from high level descriptions," in *Web Engineering*. Heidelberg, Germany: Springer, 2001, pp. 190–203.

[3] Z. Guo, Z. Li, and H. Tu, "Sina microblog: An information-driven online social network," in *Proc. IEEE CW*, 2011, pp. 160–167.

[4] L. Liu and K. Jia, "Detecting spam in Chinese microblogs—A study on Sina Weibo," in *Proc. IEEE CIS*, 2012, pp. 578–581.

[5] L. Chen, C. Zhang, and C. Wilson, "Tweeting under pressure: Analyzing trending topics and evolving word choice on Sina Weibo," in *Proc. ACM COSN*, 2013, pp. 89–100.

[6] Q. Yu, W. Weng, K. Zhang, K. Lei, and K. Xu, "Hot topic analysis and content mining in social media," in *Proc. IEEE IPCCC*, 2014, pp. 1–8.

[7] A. Java, X. Song, T. Finin, and B. Tseng, "Why we Twitter: Understanding microblogging usage and communities," in *Proc. ACM WebKDD/SNA-KDD*, 2007, pp. 56–65.

[8] H. Kwak, C. Lee, H. Park, and S. Moon, "What is Twitter, a social network or a news media?" in *Proc. WWW*, 2010, pp. 591–600.

[9] M. Gabielkov, A. Ramachandran, A. Chaintreau, and A. Legout, "Social clicks: What and who gets read on Twitter?" in *Proc. ACM SIGMETRICS/IFIP Perform.*, 2016, pp. 179–192.

[10] T. Wang *et al.*, "The power of comments: Fostering social interactions in microblog networks," *Front. Comput. Sci.*, vol. 10, no. 5, pp. 889–907, 2016.

[11] J. L. Myers, A. Well, and R. F. Lorch, *Research Design and Statistical Analysis*. New York, NY, USA: Routledge, 2010.

[6] http://shop.m.taobao.com/*

[12] S. Mossa, M. Barthélémy, H. E. Stanley, and L. A. N. Amaral, "Truncation of power law behavior in 'scale-free' network models due to information filtering," *Phys. Rev. Lett.*, vol. 88, no. 13, 2002, Art. no. 138701.

[13] R. Keralapura, A. Nucci, Z.-L. Zhang, and L. Gao, "Profiling users in a 3G network using hourglass co-clustering," in *Proc. ACM MobiCom*, 2010, pp. 341–352.

[14] I. S. Dhillon, "Co-clustering documents and words using bipartite spectral graph partitioning," in *Proc. ACM SIGKDD*, 2001, pp. 269–274.

[15] J. Riihijarvi, M. Wellens, and P. Mahonen, "Measuring complexity and predictability in networks with multiscale entropy analysis," in *Proc. IEEE INFOCOM*, 2009, pp. 1107–1115.

[16] Y. Qu, C. Huang, P. Zhang, and J. Zhang, "Microblogging after a major disaster in China: A case study of the 2010 Yushu earthquake," in *Proc. ACM CSCW*, 2011, pp. 25–34.

[17] W. Guan *et al.*, "Analyzing user behavior of the micro-blogging website Sina Weibo during hot social events," *Physica A Stat. Mech. Appl.*, vol. 395, pp. 340–351, Feb. 2014.

[18] Q. Gao, F. Abel, G.-J. Houben, and Y. Yu, "A comparative study of users' microblogging behavior on Sina Weibo and Twitter," in *Proc. Int. Conf. User Model. Adapt. Personalization*, 2012, pp. 88–101.

[19] Q. Liao and L. Shi, "She gets a sports car from our donation: Rumor transmission in a Chinese microblogging community," in *Proc. ACM CSCW*, 2013, pp. 587–598.

[20] Z. Jin, J. Cao, Y.-G. Jiang, and Y. Zhang, "News credibility evaluation on microblog with a hierarchical propagation model," in *Proc. IEEE ICDM*, 2014, pp. 230–239.

[21] K. Wu, S. Yang, and K. Q. Zhu, "False rumors detection on Sina Weibo by propagation structures," in *Proc. IEEE ICDE*, 2015, pp. 651–662.

[22] P. André, M. S. Bernstein, and K. Luther, "Who gives a tweet?: Evaluating microblog content value," in *Proc. ACM CSCW*, 2012, pp. 471–474.

[23] P. Bao, H.-W. Shen, J. Huang, and X.-Q. Cheng, "Popularity prediction in microblogging network: A case study on Sina Weibo," in *Proc. WWW*, 2013, pp. 177–178.

[24] M. R. Morris, S. Counts, A. Roseway, A. Hoff, and J. Schwarz, "Tweeting is believing?: Understanding microblog credibility perceptions," in *Proc. ACM CSCW*, 2012, pp. 441–450.

[25] D. A. Shamma, L. Kennedy, and E. F. Churchill, "Peaks and persistence: Modeling the shape of microblog conversations," in *Proc. ACM CSCW*, 2011, pp. 355–358.

[26] P. Bhattacharya *et al.*, "Deep Twitter diving: Exploring topical groups in microblogs at scale," in *Proc. ACM CSCW*, 2014, pp. 197–210.

[27] G.-R. Xue *et al.*, "Optimizing Web search using Web click-through data," in *Proc. ACM CIKM*, 2004, pp. 118–126.

[28] T. Joachims, L. Granka, B. Pan, H. Hembrooke, and G. Gay, "Accurately interpreting clickthrough data as implicit feedback," in *Proc. ACM SIGIR*, 2005, pp. 154–161.

[29] E. Agichtein, E. Brill, and S. Dumais, "Improving Web search ranking by incorporating user behavior information," in *Proc. ACM SIGIR*, 2006, pp. 19–26.

[30] P.-S. Huang *et al.*, "Learning deep structured semantic models for Web search using clickthrough data," in *Proc. ACM CIKM*, 2013, pp. 2333–2338.

[31] M. Richardson, E. Dominowska, and R. Ragno, "Predicting clicks: Estimating the click-through rate for new ADS," in *Proc. WWW*, 2007, pp. 521–530.

[32] J. Pan *et al.*, "Field-weighted factorization machines for click-through rate prediction in display advertising," in *Proc. WWW*, 2018, pp. 1349–1357.

[33] H. Gao, D. Kong, M. Lu, X. Bai, and J. Yang, "Attention convolutional neural network for advertiser-level click-through rate forecasting," in *Proc. WWW*, 2018, pp. 1855–1864.

[34] C. Lakshminarayan, R. Kosuru, and M. Hsu, "Modeling complex click-stream data by stochastic models: Theory and methods," in *Proc. WWW Companion USEWOD*, 2016, pp. 879–884.

[35] C. G. Brinton and M. Chiang, "MOOC performance prediction via click-stream data and social learning networks," in *Proc. IEEE INFOCOM*, 2015, pp. 2299–2307.

[36] S. D. Bernhard, C. K. Leung, V. J. Reimer, and J. Westlake, "Clickstream prediction using sequential stream mining techniques with Markov chains," in *Proc. ACM IDEAS*, 2016, pp. 24–33.

[37] J. Ge, Y. Tian, L. Liu, R. Lan, and X. Zhang, "Understanding e-commerce systems under massive flash crowd: Measurement, analysis, and implications," *IEEE Trans. Services Comput.*, to be published.

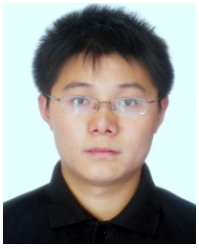
[38] Ş. Gündüz and M. T. Özsu, "A Web page prediction model based on click-stream tree representation of user behavior," in *Proc. ACM SIGKDD*, 2003, pp. 535–540.
[39] Q. Su and L. Chen, "A method for discovering clusters of e-commerce interest patterns using click-stream data," *Electron. Commerce Res. Appl.*, vol. 14, no. 1, pp. 1–13, 2015.
[40] G. Wang *et al.*, "You are how you click: Clickstream analysis for Sybil detection," in *Proc. USENIX Security*, 2013, pp. 241–256.
[41] G. Wang, X. Zhang, S. Tang, H. Zheng, and B. Y. Zhao, "Unsupervised clickstream clustering for user behavior analysis," in *Proc. ACM SIGCHI*, 2016, pp. 225–236.
[42] F. Qiu, Z. Liu, and J. Cho, "Analysis of user Web traffic with a focus on search activities," in *Proc. Int. Workshop Web Databases (WebDB)*, 2005, pp. 103–108.
[43] M. R. Meiss, B. Gonçalves, J. J. Ramasco, A. Flammini, and F. Menczer, "Agents, bookmarks and clicks: A topical model of Web navigation," in *Proc. ACM Hypertext*, 2010, pp. 229–234.
[44] K. Chen *et al.*, "Collaborative personalized tweet recommendation," in *Proc. ACM SIGIR*, 2012, pp. 661–670.
[45] Amazon. *Alexa's Digital Marketing Tools*. Accessed: Oct. 2016. [Online]. Available: http://www.alexa.com/topsites/countries/CN

Jingtao Ding received the B.S. degree in electronic engineering from Tsinghua University, Beijing, China, in 2015, where he is currently pursuing the Ph.D. degree with the Department of Electronic Engineering. His research interests include mobile computing, mobile data mining, and user behavior modeling.

Zhongjin Liu received the B.S. degree from the Beijing Institute of Technology in 2009 and the Ph.D. degree from Tsinghua University in 2014. He is currently with the National Computer Network Emergency Response Technical Team/Coordination Center of China. His research interests include Internet of Things, computer security, and user behavior modeling.

Shaoran Xiao received the B.S. degree from Beihang University in 2015 and the M.S. degree from Tsinghua University in 2018. He is currently with Face++ Cognitive Services. His research interests include Internet of Things and user behavior modeling.

Yang Chen (M'07–SM'15) received the B.S. and Ph.D. degrees from the Department of Electronic Engineering, Tsinghua University in 2004 and 2009, respectively. He was a Post-Doctoral Associate with the Department of Computer Science, Duke University, USA, where he served as a senior personnel with the NSF MobilityFirst Project. He is an Associate Professor with the School of Computer Science, Fudan University, China, where he has led the Mobile Systems and Networking Group since 2014. From 2009 to 2011, he was a Research Associate and the Deputy Head of the Computer Networks Group, Institute of Computer Science, University of Goettingen, Germany. In 2007, he was a Visiting Student at Stanford University, and from 2006 to 2008 with Microsoft Research Asia. His research interests include online social networks, Internet architecture, and mobile computing. He is serving as an Editorial Board Member for the *Transactions on Emerging Telecommunications Technologies* and an Associate Editor of IEEE ACCESS.



Yong Li (M'09–SM'16) received the B.S. degree in electronics and information engineering from the Huazhong University of Science and Technology, Wuhan, China, in 2007 and the Ph.D. degree in electronic engineering from Tsinghua University, Beijing, China, in 2012, where he is currently a Faculty Member with the Department of Electronic Engineering.

He has over 5200 citations. Among them, ten are ESI Highly Cited Papers in Computer Science, and four receive conference Best Paper (run-up) Awards. He was a recipient of the IEEE 2016 ComSoc Asia–Pacific Outstanding Young Researchers and Young Talent Program of China Association for Science and Technology. He is on the editorial board of two IEEE journals. He has served as the general chair, the TPC chair, and a TPC member for several international workshops and conferences.

Depeng Jin (M'09) received the B.S. and Ph.D. degrees in electronics engineering from Tsinghua University, Beijing, China, in 1995 and 1999, respectively, where he is currently an Associate Professor and the Vice Chair of the Department of Electronic Engineering. His research fields include telecommunications, high-speed networks, ASIC design, and future Internet architecture. He was a recipient of the National Scientific and Technological Innovation Prize (Second Class) in 2002.

Steve Uhlig received the Ph.D. degree in applied sciences from the University of Louvain in 2004. He is currently a Professor of networks with the Queen Mary University of London. His research interests are focused on the large-scale behavior of the Internet, Internet measurements, software-defined networking, and content delivery.