Discovering Social Photo Navigation Patterns

Luca Chiarandini
Universitat Pompeu Fabra
Web Research Group
Barcelona, Spain
chiarluc@yahoo-inc.com

Michele Trevisiol
Universitat Pompeu Fabra
Web Research Group
Barcelona, Spain
trevi@yahoo-inc.com

Alejandro Jaimes
Yahoo! Research
Barcelona, Spain
ajaimes@yahoo-inc.com

***Abstract*—In general, user browsing behavior has been examined within specific tasks (*e.g.*, search), or in the context of particular web sites or services (*e.g.*, in shopping sites). However, with the growth of social networks and the proliferation of many different *types* of web services (*e.g.*, news aggregators, blogs, forums, *etc.*), the web can be viewed as an ecosystem in which a user's actions in a particular web service may be influenced by the service she arrived *from* (*e.g.*, are users browsing patterns similar if they arrive at a website via search or via links in aggregators?). In particular, since photos in services like Flickr are used extensively throughout the web, it is common for visitors to the site to arrive via links in many different types of web sites. In this paper, we depart from the hypothesis that visitors to social sites such as Flickr behave differently depending on where they come from. For this purpose, we analyze a large sample of Flickr user logs to discover social photo navigation patterns. More specifically, we classify pages within Flickr into different categories (*e.g.*, "add a friend page", "single photo page," *etc.*), and by clustering sessions discover important differences in social photo navigation that manifest themselves depending on the type of site users visit before visiting Flickr. Our work examines photo navigation patterns in Flickr for the first time taking into account the referrer domain. Our analysis is useful in that it can contribute to a better understanding of how people use photo services like Flickr, and it can be used to inform the design of user modeling and recommendation algorithms, among others.**

***Index Terms*—Social navigation, behavioral modeling, session analysis**

I. Introduction

Insights into how users behave within a website or domain are extremely important in informing business decisions, in developing strategies to provide new functionalities, and in general for devising new algorithms that directly improve such services. For instance, having deep insights on what pages or sections are visited most and when can be used not just to create better user models, but also to improve the design of such pages and the overall "flow" of the website (*e.g.*, by highlighting certain sections on particular page layouts).

Flickr has become a rich resource for research in multimedia, in large part because its clear copyright policies and APIs have facilitated the gathering and analysis of Flickr data. While a lot is known about the data that resides in Flickr, however, there aren't that many insights into how people actually use Flickr, and in particular, on their social navigation patterns.

As the functionality of the web has become more complex, and sharing of content (*e.g.*, Flickr photos) is done in multiple ways (*e.g.*, by posting to social networks such as Facebook, or information networks such as Twitter; posting on blogs, in news articles, *etc.*), it has become increasingly more difficult to understand the dynamics of how users browse and look at (*i.e.*, "consume") photos once they arrive at Flickr from other sources. Although Flickr remains very popular, there are many similar services for social sharing and viewing of photos, thus the work we present here should provide insights that, although computed from Flickr data, should easily generalize.

In this paper, we perform an in-depth analysis of social navigation patterns on Flickr. In particular, we analyze a sample of user logs from approximately two months, by clustering sessions, and specifically considering the referrer domain (*i.e.*, the site or domain the user visited before arriving at Flickr).

Our work aims at addressing several questions, among which we include the following (i) are photo social navigation patterns different depending on the referral website? if there are differences, what kinds of differences are there? (ii) do similar types of websites (*e.g.*, "search" lead to similar behavior?) (iii) what types of pages (*e.g.*, within Flickr) are more popular depending on the referral website?, and (iv) does user behavior in terms of time spent vary depending on the referral website?

Although there is a reasonable amount of published work on Flickr, with a few exceptions detailed below, there is little knowledge on how users actually behave within the service and the relationship between such behavior and the referral pages. Our main contribution is thus providing insights into social photo navigation patterns. Such insights may be useful for understanding the dynamics of photo sharing sites, although the same type of analysis could be extended to other domains.

II. Related Work

Some authors have studied user navigation patterns in Flickr. Most notably, Lerman and Jones [1] studied how users find new images on Flickr, highlighting that people often navigate through photo streams of their contacts. They refer to such behavior as "social browsing" because users tend to browse the photos of their closest contacts. Such behavior has also been highlighted by other authors (*e.g.*, [2], [3]).

Several authors have analyzed sessions and browsing behavior for various purposes. Benevenuto *et al.* [4] showed a clickstream study over several social networks, proposing a clickstream model to characterize user behavior, while Jiang *et al.* [5] studied the Chinese social network Renren, creating latent interaction graphs as a different representation of interaction based on "profiles" of browsing events.

Huang *et al.* [6] take into account parallel browsing, *i.e.* when a user navigates using multiple browser windows at the same time. Other researchers model user behavior considering the content of the pages [7] and even use it for tag recommendation in Flickr [8].

Gamon and König [9], studied session logs collected from the Microsoft Live Toolbar. They grouped URLs into categories, for a manually defined list of websites, obtaining 5 categories. A somewhat related approach is proposed by Kumar and Tomkins [10], in which a URL taxonomy is generated by an automatic categorization. Other authors have focused on clustering or using Markov Chains (*e.g.*, Sharma *et al.* [11] and Vakali *et al.* [12]) to model user sessions.

The main difference between our work and previous work is that we take into account the referrer URL in order to model user behavior. Most work on session analysis on the web focuses on modeling behavior independently of where the user comes from when she visits a site. In addition, our work differs from Lerman and Jones' [1] in that we do not focus only on new images. More specifically, we take into account not just whose photos users view, but also consider categories of pages within the Flickr site, and given the referral information, explicitly analyze users' behaviors.

III. Dataset

In this section we describe the dataset used in our study.

A. Dataset Collection

We analyze user logs, which are plain text files that contain a line for each web server request. For this study we took a sample of the page views of more than 10 million anonymous users from approximately two months of Flickr user log data[1]. For each pageview, our dataset contains the following information fields: *a*) **User identifier:** an unique anonymous identifier of the user accessing Flickr. *b*) **Timestamp:** of the specific pageview; *c*) **Referrer URL:** the page the user arrived from; *d*) **Current URL:** the url of the pageview; *e*) **User-Agent**[2]**:** an identifier of the browsing application in use (*e.g.* Mozilla, Opera, *etc.*). This is useful to filter out web crawlers.

In the next section we describe the filters applied to the data in order to remove noise.

B. Pageview Filtering & Data Selection

In order to obtain a coherent dataset in terms of both timezone and activity, we focused on users who are located in the US by extracting the location of the IP address from the source of the HTTP request and filtering out non-US locations. We then removed traffic derived from Web crawlers by preserving only the entries whose User-Agent field contains a well-known browser identifier (*e.g.*, Firefox, Chrome, *etc.*). In spite of this filtering, there are cases in which the User-Agent field indicates that a legitimate browser was used, but the corresponding "users" have a very large number of pageviews. The frequency, however, suggests that such server requests could not have been made by humans, but instead were done automatically for malicious crawling. We therefore applied an additional filter by which we set a maximum threshold on the total number of pageviews per user. The threshold was set to remove a small percentage of the users (1% of the total amount). Applying the filtering steps described above resulted in a sample containing approximately 309 Million pageviews.

C. Pageview Layouts

In most websites, multiple URLs can map to exactly the same page "layout" (*e.g.*, on Flickr, the URL of a page that shows a single image contains a unique ID for the image, thus two URLs for two different images are different even though the page layout is the same). Since our interest is in modeling navigation patterns in Flickr, we must map all URLs that refer to the same layout, to a single page layout (*e.g.*, "single image page"). For this purpose, we define a hierarchical taxonomy of urls: the **pageLayout**. We manually created a set of regular expressions to classify the urls to obtain a total of 96 different pageLayouts. Example of layouts include the following: *display all user photos*, *search photos*, *browse group photos*, *add contacts*, *accept invitation to join Flickr*, *etc.*

D. Source URL Taxonomy

In order to analyze the referrer URLs (*i.e.*, the websites users arrive to Flickr pages from), we built a taxonomy for external urls (*i.e.*, whose domains are different from www.flickr.com). The first attempt of categorizing urls was based on the Open Directory Project[3] and the Yahoo! Directory[4]. However, by manually inspecting the results, we realized that the classification was too detailed and did not capture the aspects we are interested in. In fact, url categorization usually works by *topic* (*e.g.* travel, economy, food, *etc.*) whereas in our study we are interested on a categorization by *type* (*e.g.* blog, social networking site, search, *etc.*). We therefore opted to annotate them manually (*e.g.* search.google.com as *search*, *etc.*) and focused on defining 15 sourceCategories that we considered important. We created a set of regular expressions in order to identify about 210 different external url domains. Table III shows the most frequent sourceCategories.

[1] All processing was anonymous and performed in aggregate. In addition, only profiles and photos marked as "public" were considered.

[2] http://en.wikipedia.org/wiki/User_agent

[3] Netscape (AOL), "Open directory", http://www.dmoz.org/, June 1998.

[4] Yahoo!, "Yahoo! directory", http://dir.yahoo.com/, March 1995.

IV. Session analysis

This section describes how we extracted sessions from the dataset, and also gives hints about the discriminate power of the *source URL* (*i.e.* the web site from which the user arrives when starting a new Flickr session).

A. Session Identification

Since users' behavior varies over time, we group pageviews into *sessions*. Traditionally, in each session, a user's behavior is assumed to be oriented towards a single goal. We split a user's activity into different sessions when either of these two conditions hold:

- **by time:** when the inactivity between two pageviews is longer than 25 minutes.
- **by external url:** when the user comes to Flickr from a different domain (*i.e.* a domain that is different from www.flickr.com) even if these visits are within the 25 minute threshold. For example, when a user enters and exits from Flickr via another domain (*e.g.*, mail account).

B. Session Characteristics

Table I shows some statistics computed over aggregate sessions in our sample dataset. The average number in the pageLayout shows the number of different types of pageLayouts present in the sessions. The values suggest that a large number of sessions, tend to consist of only a few page categories. It is important to note, however, that "complex" use of Flickr is not uncommon, and represented by sessions in which a maximum of 39 different page types are visited.

Total number of sessions	40'446'676
Total number of users	10'912'431
Avg. number of distinct pageLayout	1.83
Max. number of distinct pageLayout	39

TABLE I: Basic statistics about the sessions in the dataset.

C. Analysis of Types of Pages Visited

Table II shows the ten most visited pageLayouts in the dataset. We can see that there are a few pageLayouts that are visited most frequently: although we defined a total of 96 pageLayouts, users tend to navigate through a small subset of them, namely to explore *photos of users* and *groups*. This is compatible with the results of Table I that shows that users usually browse in just a few categories during one session. We will now move our focus to the *source URL*, which is the referrer of the session.

D. SourceCategory Analysis

One of our main assumptions is that there is a relationship between the source URL and the type of navigation behavior of the user.

In Table III we show the most frequent domain categories from which the user arrives to Flickr pages and some example urls. The histogram in Figure 1 shows the distribution of the source URL categories. The two most common sources are *search* and *social*. The presence of search is reasonable due to the contribution of image search and navigational queries.

pageLayout	**Occurrences**
Display all user photos Displays the photos of a user on a grid	26.71%
Browse user photos Displays full-page photo of a user and allows browsing to the next and previous photos	20.67%
Browse user album Displays full-page photo of an album of a user and allows browsing to the next and previous photos	14.12%
Display single photo Displays full-size photo	7.22%
Homepage Home page of Flickr	5.60%
View user albums Lists the album of a user	4.59%
Browse group photos Displays full-page photo of a group and allows browsing to the next and previous photos	2.63%
Search photos Photo search in Flickr	2.38%
Browse user fav. Displays full-page photo of the favorite photos of a user and allows browsing to the next and previous photos	2.09%
Group photos Displays the photos of a group on a grid	1.79%

TABLE II: Top ten most frequent pageLayouts in the dataset.

sourceCategory	**Occurrences**
search: search.yahoo.com, google.com, *etc.*	34.87%
social: facebook.com, tumblr.com, *etc.*	26.95%
mail: mail.yahoo.com, gmail.com, *etc.*	13.22%
aggregator: reddit.com, stumbleupon.com, *etc.*	7.76%
blog: blogspot.com, blogger.com, *etc.*	6.65%
photo: flickrhivemind.net, compfight.com, *etc.*	2.32%
microblog: twitter.com, *etc.*	2.26%
forum: discussion forums	2.00%
news: news.yahoo.com, cnn.com, *etc.*	1.67%
shop: ebay.com, *etc.*	0.85%

TABLE III: Top ten most frequent sourceCategories in the dataset.

While most photo websites retain proprietary rights on the retrieved results or do not have clear photo licensing policies, we can assume that Flickr is one of the main sources of Creative Commons-licensed material. Social network websites, such as Facebook, constitute very popular access points to Flickr since users are highly interested in photos shared by friends. We did not expect *mail* to have high importance, as usually the attachments are sent within the message itself and not as external links. As we will see in Section V-B, many session derive from invitations of friends to join Flickr. The fact that many sessions come from the *news* domain is indicative that the image is often considered as appealing or significant as

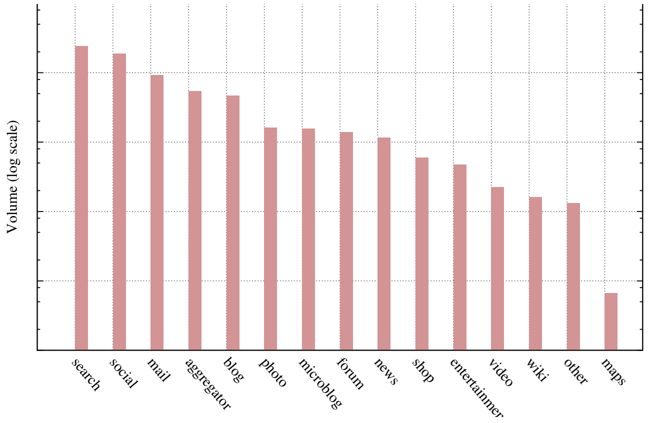


Fig. 1: Distribution of the 15 main categories for the external source urls.

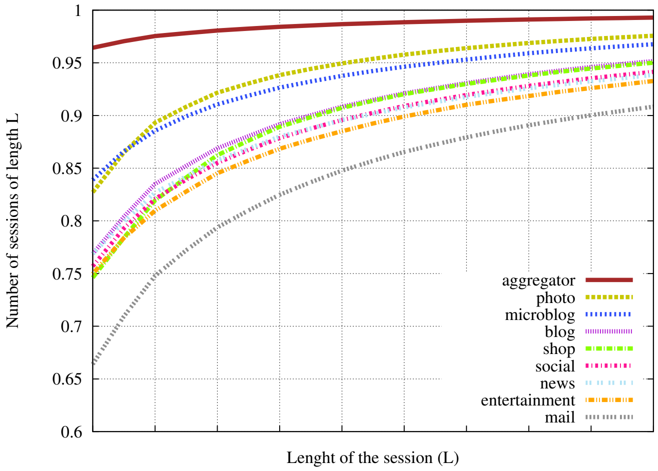


Fig. 2: Cumulative distribution of the nine main categories of source urls.

the actual text of the article.

The raw analysis of volume gives us first insights into how the initial context may affect navigation patterns. However, we understand this even better by observing the cumulative distribution of session lengths depicted in Figure 2. In the figure we represent only the 9 most frequent categories. The categories have a different behavior from one another. The lines that reach value 1 sooner correspond to the situation in which the user spends less time on Flickr on average. On the contrary, the ones which grow slowly show users with longer sessions on average. Based on this analysis, we see that the shortest sessions originate in aggregators. One example is www.reddit.com, in which the links to Flickr appear inside news posts. It may appear strange that the sessions deriving from *news* sites last longer. An explanation for this might be that the visual material in news sites (such as Yahoo! News) is curated by professional editors and photographers and often consists not only of a single photo, but rather of a collection of photos related to a particular event. For example, an article about the earth-quake in Japan is linked to a group or a set of photos all related to that topic. The user is therefore prone to see more than one picture.

Extreme behavior is observed in the *mail* category where the users spend the longest time interacting with Flickr. One possible explanation might be that only the "closest" contacts send e-mail, and thus a stronger bond exists between the sender and the receiver of the message. Moreover one could assume that users that share links via e-mail, may share entire sets or albums which contain many photos, leading to longer and more complex interactions with Flickr.

V. Clustering of Sessions

In this section we describe the clustering of sessions and analyze the clusters' general characteristics in terms of the pageLayouts that constitute the clusters, and in terms of browsing behavior depending on the referral domain categories (*i.e.*, the type of domain that users visit before arriving at Flickr).

We model each session s as a vector $v = (v_1, v_2, ..., v_P)$ where each v_i counts the number of views of pageLayout i in session s. *Cosine similarity* is used to compare vectors since it is not affected by the absolute number of pageviews but only by the relative distribution across the pageLayouts. We applied the Canopy algorithm [13] on the vectors followed by K-Means clustering to extract clusters of sessions to obtain a total of 62 clusters.

As discussed earlier, our hypothesis is that user browsing patterns are different depending on the source website. First of all, however, we are interested in examining session clusters without taking into account how users arrived at Flickr. We do this by examining the clusters that appear in the same percentage across all sourceCategories $srcCat$. More specifically, we compute the entropy distribution for each cluster c across $p(c|srcCat)$ with Equation 1:

$$\sum_{srcCat} [p(c|srcCat) \log_2 p(c|srcCat)] \qquad (1)$$

We then sort the clusters in ascending order and select the top 7 clusters. In order to understand the characteristics, we draw the heat-map (Figure 3) of $p(c|srcCat)$.

A. Patterns in Session Clusters

In this sub-section we refer to Figure 3, which displays 7 clusters (VL0, VL-1, ...) and the pageLayouts that constitute them. As the figure shows, `VL-0`, `VL-37` and `VL-51` contain a large number of *Display all user photos* and *Browse user photos* page views, which indicate a typical pattern of mainly browsing through the photos of a user or users. Cluster `VL-1`, on the other hand, contains more cases of users that import and add new contacts (*Add contact* row in Figure 3). A very clear case of browsing photo albums is cluster `VL-11`, where we can observe a large value in the *Browse user album* row. A similar behavior is in cluster `VL-59` where the sessions are more balanced between browsing a specific album (*Browse user album*) and seeing the list of albums (*View user albums*),

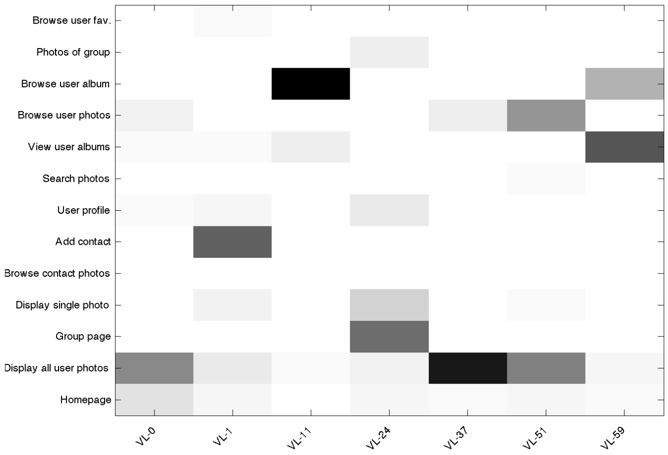


Fig. 3: Heat-map of pageLayout for the most frequent clusters. The clusters distribution is normalized by column, darker squares indicate a higher presence of the relative pageLayout views (row) in the current cluster (column).

maybe to explore a different one. Group-oriented navigation is specific of `VL-24`, due to the presence of *Group page* and *Photos of group*. In this case users switch between the main page of the group and its photos.

Although these clusters are useful to understand how users interact with Flickr, we would like to explore the peculiarities of the sourceCategories. We therefore manually inspected the clusters and selected the ones that show interesting patterns.

B. Browsing from Different Sources

As stated earlier, many clusters illustrate a very specific browsing behavior. We manually picked a few of them to show how well they describe some navigation patterns in relation with the sourceCategories.

Figure 4a shows the distribution of such clusters across sourceCategories whereas Figure 4b shows the distribution of the same clusters across pageLayouts. Due to the large amount of sessions originated from search engines, the *search* sourceCategory appears in most of the clusters. Despite this, there are still some clusters in which this is not the case.

Cluster `VL-24` shows a large contribution of *news* and the distribution of pageLayouts for that cluster (first column of Figure 4b) is biased towards browsing of groups (*Group page*). This suggests that news editors embed sets of images into the article page. Moreover, photos of the same event are likely to be organized in the same group in Flickr. Cluster `VL-58`, slightly more evenly spread across all sourceCategories, is similar to `VL-24` but favors browsing through the photos of a group (*Photos of group*) on the home page of the group (*Group page*). Sessions in cluster `VL-25` are mainly originated from *aggregators* and are aimed at checking the recent activity on the Flickr website (*i.e.* recently added photos, albums, *etc.*). Indeed, aggregators are used by the user to get an overview on recent events in external websites, including Flickr.

Cluster `VL-29`, mainly originated from *search*, explores the list of favorite photos of a user (*Browse user fav.*). Cluster `VL-33` shows *mail* as a principal source and is composed of pageLayouts related to social actions: *a*) *Manage friends* is the set of all pages related to adding, editing or removing information about contacts in Flickr; *b*) *Add friend* is the page in which the user is asked for confirmation when adding a contact. Manual inspection of the sessions suggests that the traffic in this cluster mainly derives from accepted invitation mails sent to mail contacts[5].

Cluster `VL-42` contains sessions coming from both *news* and *aggregators* in which users visualize the tag cloud of photo tags used by another user. This visualization gives an aggregated vision of the content posted by her. Cluster `VL-9` in Figure 4b contains mainly *search* pageLayouts. It is not surprising that Fig 4a shows us that those sessions originate from search engines. One assumption is that in this case users are migrating the search task to Flickr in order to take advantage of the image search features, as for instance filtering photos by CreativeCommons (CC) license or tags.

VI. Discussion

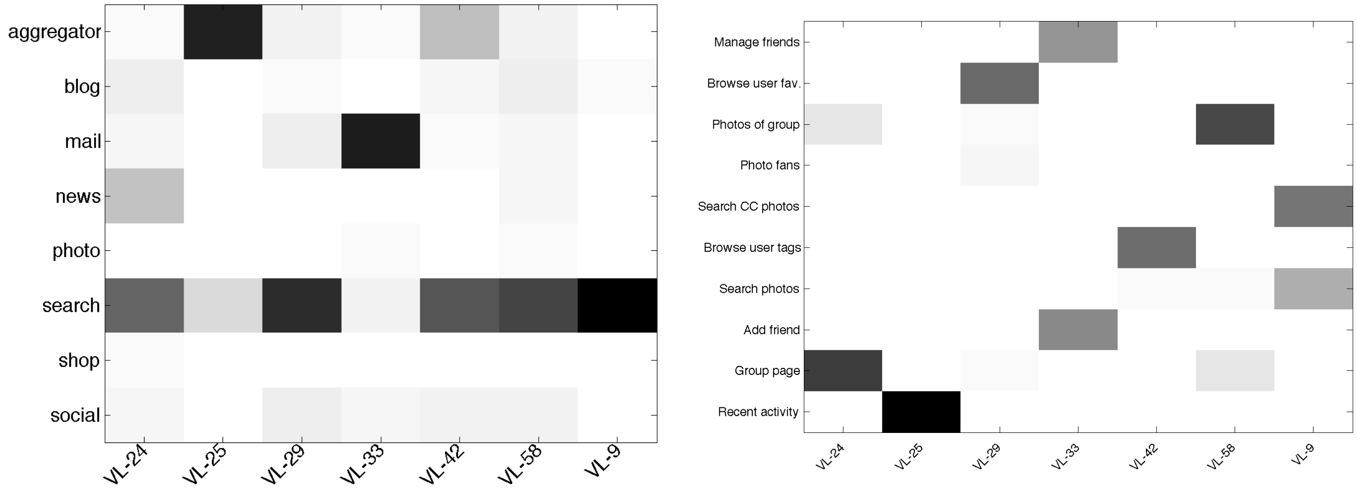
Our analysis shows that users arrive at Flickr from a variety of source domains (e.g, search, social, mail, aggregators, etc.) in varying degrees, and that the overall length of the sessions varies depending on the type of source domain (*e.g.*, users that arrive at Flickr from mail domains tend to spend more time than those arriving from any other sources). While the distribution of visits from different types of sources gives us interesting insights on the web as it is today (*e.g.*, social sites have a prominent place), it is possible to make some observations on the behavior in terms of session length (*e.g.*, users that click on mail links may be receiving photos from close social contacts, which might explain longer sessions). At the same time, we found that clear session clusters can be observed in the data (*e.g.*, some sessions are very focused on viewing photos of users, while others focus on viewing photos in groups), and that some of the behavior can be intuitively explained (*e.g.*, sessions that originate in mail domains have a stronger focus on managing and adding friends).

Many similar observations can be made based on the figures presented in this paper. Sessions that originate from search sites, for instance, cluster around the Flickr search functionality, suggesting that the user's main intent is indeed finding images of some sort. It's important to keep in mind, however, that such observations constitute hypotheses that need to be further examined.

VII. Conclusions and Future Work

In this paper, we analyzed a sample from two months of Flickr user data. Our analysis was performed on user logs. We classified pages within Flickr into specific categories, and analyzed how the behavior of users in viewing such page categories changes depending on the referral domain (*i.e.*,

[5] We do not examine mail contents, so this hypothesis cannot be verified, and is based solely on the aggregate views of the "add friend" page



(a) Heat-map of source urls

(b) Heat-map of pageLayouts

Fig. 4: Heat-map of the most interesting clusters. Darker squares indicate higher values for the presence of sessions with that category (row) in the relative cluster (column).

the page they come from). Our analysis shows that there are important differences users' social photo navigation patterns, and that these are largely affected by the referral domain. Future work includes deeper analysis of user actions within each of the pages, as well as content analysis and meta-data analysis to gain insights into how the content itself affects the navigation patterns. Based on our findings, we will create several recommendation algorithms that take advantage of the clustering results.

VIII. Acknowledgements

This research is partially supported by European Community's Seventh Framework Programme FP7/2007-2013 under the ARCOMEM and Social Media projects, and by the Spanish Centre for the Development of Industrial Technology under the CENIT program, project CEN-20101037 (www.cenitsocialmedia.es), "Social Media."

References

[1] K. Lerman and L. Jones, "Social browsing on flickr," *Arxiv preprint cs/0612047*, pp. 1–4, 2006.

[2] L. K., "Social browsing & information filtering in social media," *CoRR*, vol. abs/0710.5697, 2007.

[3] M. Valafar, R. Rejaie, and W. Willinger, "Beyond friendship graphs: a study of user interactions in Flickr," in *WOSN*. ACM, 2009, pp. 25–30.

[4] F. Benevenuto, T. Rodrigues, M. Cha, and V. Almeida, "Characterizing user behavior in online social networks," in *ACM SIGCOMM*. ACM, 2009, pp. 49–62.

[5] J. Jiang, C. Wilson, X. Wang, P. Huang, W. Sha, Y. Dai, and B. Zhao, "Understanding latent interactions in online social networks," in *Proceedings of the 10th annual conference on Internet measurement*. ACM, 2010, pp. 369–382.

[6] J. Huang and R. White, "Parallel browsing behavior on the web," *HT*, p. 13, 2010.

[7] L. Yu-Ru, H. Sundaram, M. De Choudhury, and A. Kelliher, "Temporal patterns in social media streams: Theme discovery and evolution using joint analysis of content and context," in *ICME*, 2009, pp. 1456–1459.

[8] T. Takashita, T. Itokawa, T. Kitasuka, and M. Aritsugi, "Tag recommendation for flickr using web browsing behavior," in *ICCSA*, ser. Lecture Notes in Computer Science, vol. 6017. Springer, 2010, pp. 412–421.

[9] M. Gamon and A. König, "Navigation Patterns from and to Social Media," in *ICWSM*, 2009, pp. 203–206.

[10] R. Kumar and A. Tomkins, "A characterization of online browsing behavior," *WWW*, p. 561, 2010.

[11] A. Jyoti, A. Goel, and P. Gulati, "A Novel Approach for clustering web user sessions using RST," in *ACT*. IEEE, Dec. 2009, pp. 657–659.

[12] A. Vakali, J. Pokorny, and T. Dalamagas, "An overview of web data clustering practices," in *EDBT Workshops*, no. I. Springer, 2005, pp. 500–501.

[13] A. McCallum, K. Nigam, and L. H. Ungar, "Efficient clustering of high-dimensional data sets with application to reference matching," *SIGKDD*, pp. 169–178, 2000.