# Comparing user activities across social networks to determine similarity

Kiran Kumar, David Crandall, and Michael Conover

School of Informatics and Computer science
Indian University, Bloomington
{knkumar,djcran,midconov}@indiana.edu
http://soic.indiana.edu

**Abstract.** Online social networks have penetrated all aspects of human life. The question we try to answer is how different is one social medium from another and what autocorrelation is there among these different media. different This report explores techniques in online social network analysis using time based activity patterns across twitter and flickr. We explore extraction of interesting features from online feed of users across social networks. We also explore, time based user comparison techniques to determine similarity between users based on their activity patterns. In conclusion, we discuss implications on privacy and possible applications to spam detection.

**Keywords:** social networks, privacy, activity pattern, time series

## 1 Introduction

Online social networks have become pervasive to every aspect of our lives. There is a large amount of data that is generated everyday in social networks by users. This data is generated differently in each of the online social networks (OSN). We tend to see a flux of regular activity based on text in twitter, while the activity on Flickr tends to be in a large chunk in regular intervals. Knowledge of this underlying generation potential can help us better correlate data across the two networks. In the first part we analyze the organization of the data in both the social networks, and draw estimates of parameters we use to match users across the two sites. In the second part we analyze the jaccard and cosine similarity across a sample data set, and check for statistical significance of our hypothesis.

The users are generally identified across the two networks by their pseudonym and their place of residence [1]. This information provides a very good framework to establish a validation test for our hypothesis. We were able to collect users with similar names across both flickr and twitter and compare them for our validation case. The place of residence was unknown for many users, and hence we used the upload time of the photos, which was in UTC and the tweet time also in UTC ofr comparison. The information of residence and travel geographies and their dependence on the post time was mostly ignored for convenience.

## 2 Vector Analysis

The user activity stream is viewed as a time series signal. We cannot make good comparisons with the original time series because it is a continuous feature space. We convert this feature space to discrete bins to make correlations across the two networks. We look at the time at which the post was made and generate two features for comparison. First, we form time bins spread across the data. We generate time bins of 2 hours,4 hours and 24 hours to get signals of varying frequency. We then generate the frequency of posts for each user in each of these time bins. The flickr dataset does not have the upload time along a single timezone and the timezone information is hard to decipher across both the social media. We do not have a timezone for the photo taken, hence we look at the upload time for the photo.

We now form a vector with the post frequency for every time bin and compare the vectors with jaccard similarity and cosine similarity. We compute the jaccard index as follows, where $\boldsymbol{F_k}$ is the flickr vector and $\boldsymbol{T_k}$ is the twitter vector, comprising of time bins and frequency of posts per time bin. The jaccard index computes the similarity between the two vectors including only the time bins where there was user activity.

$$Jaccard(\boldsymbol{F_k}, \boldsymbol{T_k}) = \frac{\sum_{k=1}^{N} min(\boldsymbol{F_k}, \boldsymbol{T_k})}{\sum_{k=1}^{N} max(\boldsymbol{F_k}, \boldsymbol{T_k})}$$

We compute the cosine similarity between the $\boldsymbol{F_k}$ flickr vector and $\boldsymbol{T_k}$ twitter vector in a similar way. The cosine index computes across similarity across the two vectors using all the time bins and frequency of posts.

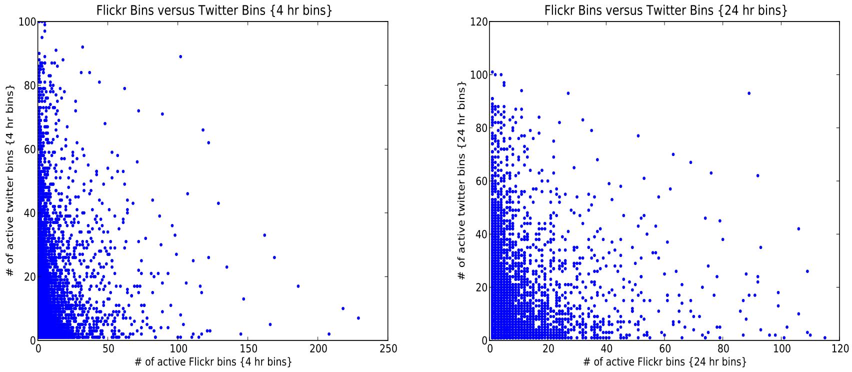
$$Cosine(\boldsymbol{F_k}, \boldsymbol{T_k}) = \frac{\boldsymbol{F_k}.\boldsymbol{T_k}}{\|\boldsymbol{F_k}\|.\|\boldsymbol{T_k}\|} = \frac{\sum_{k=1}^{N} F_k \times T_k}{\sqrt{\sum_{k=1}^{N} F_k^2}\sqrt{\sum_{k=1}^{N} T_k^2}}$$

From our experiments, the cosine similarity tends to perform better compared to the jaccard similarity in discerning the users across the two social media.

## 3 Dataset

The dataset consisted of the feed from flickr and twitter over a 3 month period. We had 7000 users intersecting with the same pseudonym, and sampled the data over all these users for computing the jaccard and cosine. The dataset had a lot of blank bins owing to user inactivity which is explained better from the figures below.

As can be seen from Figure 1, the number of active bins taking into account 4 and 24 hour periods are varying by a small factor of 20 bins. This shows that there is a small signal drop as we look at larger time bins. The active time bins for a user hovers around the 40-50 mark. Seen over a 3 month period, this is a very small number. Also, the density is high below the 20 bins mark, and then diminishes rapidly, indicating the subset are not prolific users of the social media. This adds additional difficulty in discerning the user signals, since we have more bins that are empty.

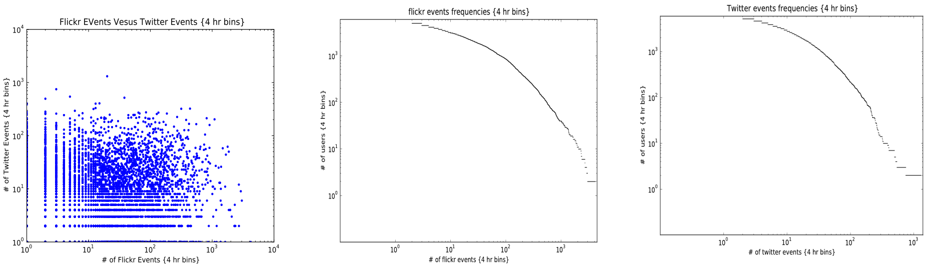


(a) 4 hour time bins

(b) 24 hour time bins

Fig. 1: The scatter of the flickr and twitter time bins

Figure 2, shows a scatter of frequency of posts among the flickr and twitter users. Most of the information is contained between 10 and 1000 posts. Users with a post frequency less than 10 are polarized to only one of the social medium. The time bins and the frequency of posts shows us that we need to look at prolific users. We determine prolific users as more than 10 posts and more than 5 bins.



(b) Flickr

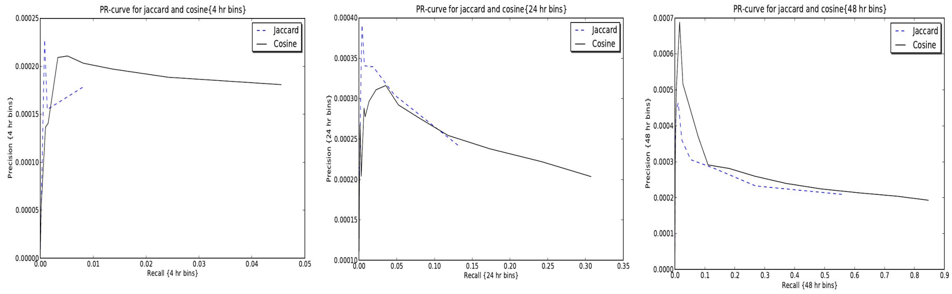
(c) Twitter

Fig. 2: The scatter of the flickr and twitter post frequency

If we consider users with less than 5 bins or 10 posts, it leads to a misclassification, since there is not much signal information. We can also see there is a direct correlation in activity of flickr and twitter, both have users with similar ranges of post frequency, which can be seen in the histograms in figure 2. This shows that there should be some similarity in the pattern of usage across the two social media. We are trying to leverage this similarity in pattern across twitter and flickr to determine the user across the two networks.

## 4 Results

When we applied the constraints on the number of bins and post frequency to be greater than 5 and 10 respectively, there were only around 1100 user left in each case of 4, 24 and 48 hour time bins. This gives us a baseline around 0.08 with a $+/-1$ deviation for the 4 hour ad 48 hour bins. As we can see in figure 3, the pr curve has a peak at 0.0002 for 4, 0.0003 for 24 ad 0.0007 for 48 hour bins respectively. Although this precision is low, this shows that we have better chance than the baseline case of identifying similar users between the two sites.



(a) 4 hour time bins (b) 24 hour time bins (c) 48 hour time bins

Fig. 3: PR Curve for different time bins

## 5 Future work

To obtain more signal information we can compute a tf-idf score on the tags in flickr versus the tweets across each bin and average the score. There is also expected correlation between this score and the similarity of users based on initial observations. This should help us discern users who have similar activity across the two social media. This work also can be used to compare users posting spam information across sites and discern them. This should be more prevalent since we have the posts at the same time or with a small shift in the timeline. The tf-idf score in these cases also would be very high.

There are definite implications to privacy and security in this area. It is definitely interesting to see if the tf-idf score adds more signal information and what the peak precision is for this problem.

# Bibliography

[1] Laura Garton, Caroline Haythornthwaite, and Barry Wellman. Studying online social networks. *Journal of Computer-Mediated Communication*, 3(1):0–0, 1997.