Social Media – are they underpinned by social or interest-based interactions?

Katarzyna Musial
King's College London

Nishanth Sastry
King's College London

ABSTRACT

On many social media and user–generated content sites, users can not only upload content but also create links with other users to follow their activities. It is interesting to ask whether the resulting user–user Followers' Network is based more on social ties, or shared interests in similar content. This paper reports our preliminary progress in answering this question using around five years of data from social video–sharing site vimeo.

Many links in the Followers' Network are between users who do not have any videos in common, which would imply the network is not interest–based, but rather has a social character. However, the Followers' Network also exhibits properties unlike other social networks, for instance, clustering co–efficient is low, links are frequently not reciprocated, and users form links across vast geographical distances. In addition, analysis of the relationship strength, calculated as the number of commonly liked videos, people who follow each other and share some "likes" have more video likes in common than the general population. We conclude by speculating on the reasons for these differences and proposals for further work.

1. INTRODUCTION

The shift from traditional online fora such as bulletin boards and discussion lists to social media has allowed users to explicitly declare fine-grained connections with individual users based on shared interests, rather than participate in a broadcast medium where every user is connected to every other user. In this context, it is interesting to ask whether the interest-based social network that forms is different from or similar to traditional online social networks, and whether users are efficiently creating links that captures their interest commonalities with other users.

This paper addresses the issue by using empirical traces of five years of user activity and link formation on social video-sharing site vimeo. The data is investigated using the concept of Multirelational Social Network in which the users are network nodes and two different relationships between these nodes can be defined: (i) direct one – when two users are connected if at least one of them follows the other one and (ii) a quasi-direct video–based relation that is present in the system between two users if they like at least one video in common.

Based on the concept of multi-relational social network, we first induce a Video-based Network (VN) between users who have liked the same videos. This is clearly a "purely" interest-based network. We then compare the structural properties of the induced network with the user-declared network, as well as with a merged Followers' Video-based Network (FVN) created from links present in both the user-declared and interest-based networks. The extracted relations are the basis to find out if people directly connected by the "follow" relationship, like the same videos, which would be an indicator of interest–based relations. On the other hand if people do not share their likes then it would suggest that these direct "follow" relations are social ones.

The three networks, FN, $VN$and FVN, are studied using different complex network metrics such as number of nodes and relations, clustering co-efficient and network density. We find evidence that users do not share many videos with their friends, indicating that links are not necessarily interest-based. However, we also find differences from purely social networks, with low clustering co-efficient and reciprocity values, and links with geographically distant users. We present this as a problem for further study to the community.

The rest of the paper is structured as follows. First, the related work in the area of social and interest–based networks is presented. In Section 3 presents preliminary theoretical background and describes our data. In Section 4 studies whether links are interest-based or social, using various complex network measures. Section 5 concludes.

2. RELATED WORK

Recently, a number of studies [5, 4, 3, 13] have looked at various macroscopic properties of YouTube videos and suggested ways to exploit these properties. However, these have been mainly studies of individual videos and popularities, whereas we examine the social network of declared user–user links. It should be noted that the social network in [4] is the video–video network where two videos are linked if they are deemed to be related by YouTube. [10] studies the declared social network of links on vimeo in the context of distinguishing popular "head" items from unpopular "tail items". To the best of our knowledge, this work is one of the first attempts to characterise multirelational social networks

SIMPLEX '12 April 17 2012, Lyon, France.

around video sharing by augmenting the declared social network with links inferred based on shared likes.

Social networks around content have been better studied in the context of photo sharing site Flickr [8, 11, 6]. Interestingly, [8] finds that links are quickly reciprocated, contrary to the network on vimeo. In [11] authors consider the space of tags defined or appropriated by users, and finds that cosine similarity between two users decreases with number of hops separating them. However, even with direct friends, similarity values are quite small (average 0.05; [11, Fig. 4]). Complementary to this, [6] shows that direct relations are rather social ones whereas quasi–direct connections are more interest–based and these two groups do not overlap with each other.

3. SETUP AND METHODOLOGY

In this section, we first lay down a taxonomy of user relations in social networks. (Sec. 3.1). We then discuss the dataset to which this framework is applied (Sec. 3.2). Finally we describe the application of the framework to the dataset (Sec. 3.3)

3.1 Types of relations

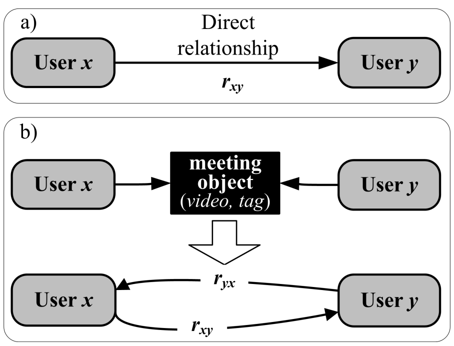

Figure 1: Types of relations in Multirelational Social Network: a) direct relationship, b) quasi–direct relationship

In all social media, in which people can share their content with other users and in the same time are able to maintain direct relationships, two different types of relations can be distinguished: (i) direct and (ii) quasi–direct ones (see Figure 1). The former ones result from adding directly somebody to the contact list or explicitly stating that activities of a given person are of interest for somebody, e.g. "following" other peoples' actions at the video sharing systems (Figure 1a). The quasi–direct relations exist when people perform some actions towards the same meeting object (Figure 1b). This object is an element that is shared within Web 2.0 application, e.g. photo, video, blog, entry at the forum. People can add objects to their "likes", comment them, tag them, etc. Although two persons can show some interest towards the same objects, it does not imply that they know each other. Thus, the social networks built based on these types of relations are perceived as interest–based networks. The important fact is also that the direct relations result in directed network whereas quasi–direct in undirected one.

More detailed description of different types of relations between users in the virtual world can be found in [9], where in addition to direct and quasi–direct connections, authors also discuss indirect relations. The last ones are relations created based on the fact that people have similar demographic profiles, e.g. they live in the same place.

In the latter part of this work both direct and quasi–direct types of relations are investigated in order to find out to what extent they are driven by common interests. The direct relation is built based on the fact that one "follows" activities of others and the quasi–direct based is derived from the information about what videos people like. It means that two persons are in the quasi–direct relation if they like at least one common video.

If more than one type of relation is extracted from the available data, then the extracted network is called a Multirelational Social Network, where on one set of nodes (users) more than one type of connection can be defined.

3.2 Dataset

Our dataset is obtained from video sharing site vimeo. Vimeo streams videos uploaded by its users. Other users can vote for videos by 'like'ing them, or leave comments. Users can also link to other users, creating a social network of users as well as follow other video creators by subscribing to their channels. In addition, users also form groups and can co–create videos. The vimeo website highlights the activities of contacts by allowing users to follow their friends' channels, subscriptions, likes etc.

Using public APIs made available by the site, all the channels and groups on the site were first crawled, obtaining information about users who are subscribers or members, and videos of the channels and groups. From here, additional video and user objects were obtained by snowball sampling, using links between users to obtain further users and videos.

The vimeo data comprises five years of activity, from Feb 16, 2005 (within 3 months of when vimeo was founded) up to Mar 27, 2010. Some statistics are summarised below. Note however that the numbers in derived networks (e.g. VN and SVN below) are different from the aggregate statistics. This is due to the procedure used to select data for the experiments (see Sec. 3.3).

Video statistics	
Number of videos	443,653
Likes	2,427,802
Social graph statistics	
Number of users	207,468
Directional links	718,457

3.3 Networks Extraction

First step is the data preparation stage in which networks of connections between users are created for further analysis. Based on the information gathered from the vimeo system, a multirelational network, consisting of two different one–layered networks, is extracted. The created one–layered networks are: (i) Followers' Network (FN) — which is a directed social network where relationship between two users exists if one user explicitly follows the activities of another user and (ii) Video Network (VN) — which is an undirected network connection created if two people like at least one common video. According to the description in Section 3.1 relations in FN are direct ones and those in VN are quasi–direct ones. Also an FVN network created by merging FN and VN is analysed. A connection between two nodes exists in this network if the two nodes are connected in both FN

and *VN*.

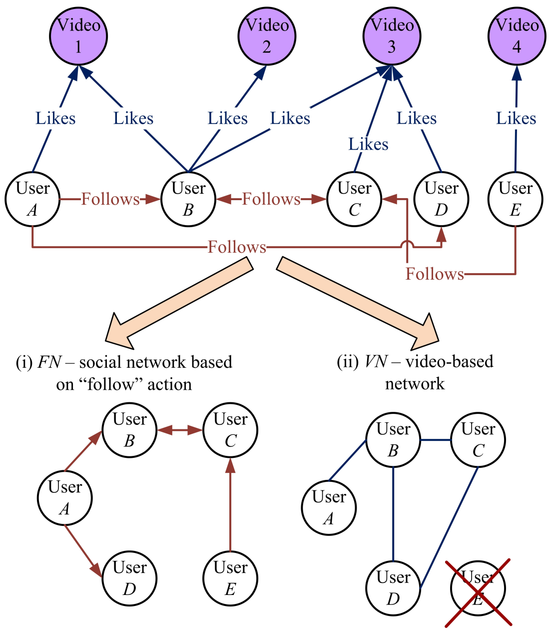

Figure 2: Followers' Network and Video–based Network extraction

We explain the above process with an example. We begin by identifying all people who follow others (users A, B, C, D, and E in Figure 2). This forms FN. For each user in FN, the videos that they like were extracted (videos 1, 2, 3, and 4 in Figure 2). If two people like the same video then the relationship between them is created in VN (e.g. User D and User C both like video 3). Note that if more than two people like the same video then in VN the relations are created between all users who like this video (e.g.users B, C, and D). The consequence of such approach is that each user in VN will have at least one relationship in the FN. In addition, not all users have to have connections in VN (e.g. User E who does not share any videos with selected users). This results from the fact that a person can follow another person's activities, but may not share any likes with that user. If a node is isolated in the VN network then the assumption is made that this user does not belong to the VN, and removed. We then create FVN by taking the intersection of links present both in $FN$and FVN. The general characteristics of the extracted networks are presented in Table 1.

4. INTEREST OR SOCIAL NETWORK?

Vimeo is primarily a content–driven site. Therefore, it is an open question whether the networks formed on vimeo resemble other social networks. In this section, we investigate the issue by examining attributes of the vimeo social network and comparing it to known properties of social networks.

The studies focus on investigating extracted networks in terms of different characteristics that can indicate whether relations between people in the video sharing system are interest– or social–based. The analysis of the results will reveal if people follow each other because they share common interest with them. The network metrics taken into account are: number of nodes and edges, network density, clustering coefficient, and reciprocity. In addition, the geographical position of users is considered to see if spatial position has any meaning in the virtual world.

4.1 Number of nodes and relations

We begin by simply studying the decreasing numbers of nodes in VN and FVN compared to the original network FN. Note that the number of nodes in FN is greater than in VN by more than 75,000 (Table 1). In other words, almost 40% of the users do not express any video "likes" and are present only in FN and not in VN. It may show that even if people follow others, they do not actively participate in the network activities and do not express their opinion about items on a public forum. In the merged FVN network there are 57,172 users. These means that roughly half (57,164) users from VN only like videos that none of their friends in FN like (as e.g. User E at the Figure 2). This is the first indicator that the fact that one user follows another does not imply that these users like the similar videos.

To sum up, in moving from FN to VN and then FVN, we lose almost 70% of the users because they either do not add any videos to their "likes" pool or have some videos in their "likes" but they do not share them with their friends from FN. Figure 3 shows the node-degree distribution for the remaining nodes follows a scale-free distribution. The fitting function that in the best way reflects the power–law trend in the node degree distribution for FVN is $y = 0.1818 \cdot x^{-1.54}$ with the $R^2 = 0.87$.

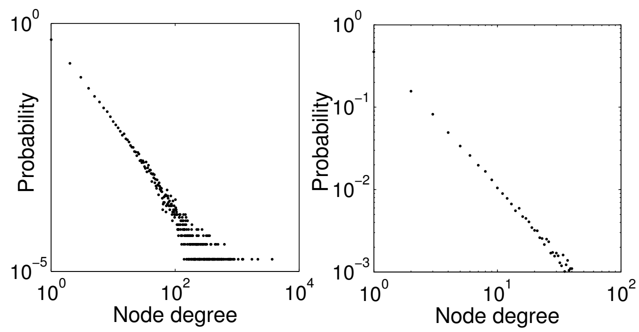

Figure 3: Node degree distribution of *FVN* for *Left*: all nodes. *Right*: 40 nodes with the highest degree.

4.2 Clustering coefficient

Next, we investigate the clustering coefficient (CC). Suppose a vertex v has neighbours $\mathcal{N}(v)$, with $|\mathcal{N}(v)| = k_v$. At most $k_v(k_v - 1)/2$ edges can exist between them (this occurs when v is part of a k_v-clique). The clustering coefficient [12] of the vertex, C_v, is defined as the fraction of these edges that actually exist. The clustering coefficient of the graph is defined as the average clustering coefficient of all the vertices in the graph. In friendship networks, C_v measures the extent to which friends of v are friends of each other, and hence, gives an estimation of the *cliquishness* of the graph. The local CC of a user quantifies how well its neighbours are connected. If they create a clique then the local CC equals 1. If there is no connections between an individual's friends then it CC equals 0.

Social networks are known to have high CC values. Here we investigate whether the social links between users who

Characteristic	*FN*	*VN*	*FVN*
Number of nodes	$190,097$	$114,336$	$57,172$
Number of relations	$650,639$	$493,359,394$	$221,414$
Network density $D = \frac{2 \cdot E}{V \cdot (V-1)}$ $E-$ no. of relations; $V-$ no. of nodes	$3.60 \cdot 10^{-5}$	0.075	$1.35 \cdot 10^{-4}$

Table 1: Main characteristics of the extracted networks

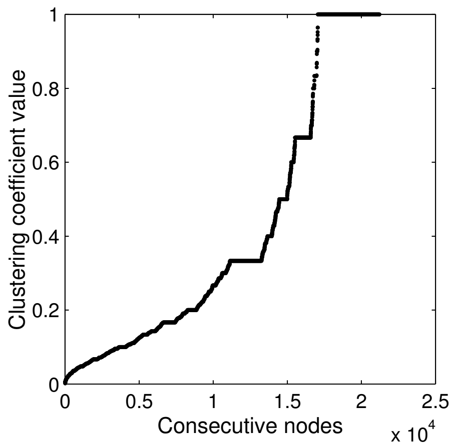

Figure 4: Clustering coefficient for nodes from FVN. Nodes are ordered according to the increasing value of clustering coefficient. Note that nodes for which CC equals to 0 are neglected in the plot.

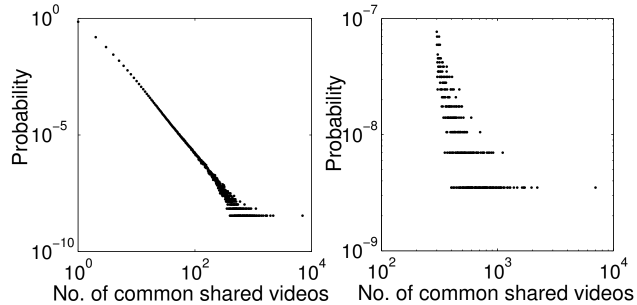

Figure 5: Distribution of the number of common "like" videos for pairs of users (percentage of links that have 1,2,...,n videos in common). *Left:* showing all links. *Right*: showing top links with more than 300 videos in common.

share some interest also exhibit high clustering. We find that the CC values in the undirected version of *FN* are low, with a mean of 0.14 and a median of 0. Next, to restrict attention to links with known shared interest, we look at the *FVN*, where links are created only between users who share a social link, and like at least one common video. Again, the CC is calculated for the undirected version of the *FVN*. Figure 4 shows the distribution of CC values. We find the CC values in the *FVN* are low as well, with mean value of 0.15 and median of 0.

Looking deeper, we find that the low CC in *FVN* arises in spite of 7.29% of users with a clustering coefficient of 1, because many others have a CC of 0. The 7,29% of users with a CC of 1 create cliques, but their size is small; 83% of users have degree 8 or less. Only 37.12% (21,222) of users within *FVN* have a local clustering coefficient greater than 0. The median CC for the group of people whose CC is greater than 0 is 0.33. This indicates that in the network exist groups of highly connected nodes. However, the fact that most of the users have a zero–valued clustering coefficient means that the clusters that exist are not the dominant characteristic feature of the analysed network.

To summarise, neither the original *FN* network of declared user-user links, nor the subset of those links sharing some videos in common (the *FVN*), exhibit a high clustering coefficient, unlike other purely social networks.

4.3 Network density and strength of links

The network density is the ratio of the number of existing within the network relationships divided by the number of all possible relations in this network. We use this metric to study whether users share videos with their declared friends on *FN*, an indication that the link is interest–based.

In the case where users are linked with every user who likes a video they like, *FN* would become identical to *VN*. In the data, however, *VN* is over 2,000 times denser than *FN*[1].

However, the density of the network is not necessarily a good measure of how efficiently users form links with others who have similar interests. The *number* of "video likes" shared is an indication of the *strength* of tie. People who share a specific number of commonly liked videos in the *VN* is presented in Figure 5. Interestingly, over 70% of links in *VN* are between users sharing only one video "like". Only 0.02% of these one-shared-video links survive in the *FVN*, i.e., when we focus on the subset of *VN* links which are also explicitly declared by users in *FN*. The mean numbers of shared video likes in *VN* and *FVN* are 1.16 and 9.45 respectively, which is indicative of stronger ties between users when they explicitly form following/followee links.

Thus, from studying the strength of ties, we find that declared links between users implies a greater similarity between them. However, not all links that could be created are created–the *FN*network could be 2000 times more denser if only video likes were used to infer links between users.

4.4 Reciprocity and similarity

Links on social networks such as Facebook are only formed if both parties agree to establish the connection. Thus, the

[1]This partly results from the fact that if n people share one video then $n \cdot (n-1)/2$ relations are created in the *VN*. However, in the obtained dataset only 0.7% of the videos are shared by more than 100 people. Thus, we do not expect many large cliques even with the $O(n^2)$ blowup.

social links are reciprocal. Even in networks which are directed, but socially motivated, it has been found that reciprocity is high [7]. Indeed, in many social settings, users may reciprocate a link out of politeness.

In *FN*, users can follow others without being followed back, i.e., it is a directed network. We find that activity is highly directional; with only about 20% of users initiating links. Fig. 6 shows the fraction of each user's links which are reciprocated, as a cumulative distribution. It can be seen that reciprocity ratios are fairly uniformly distributed (in comparison with the uniform random distribution over [0,1]), with a bias for links *not* being reciprocated (the CDF of actual data is above the uniformly random). This is another indicator that *FN* is unlike other purely social networks

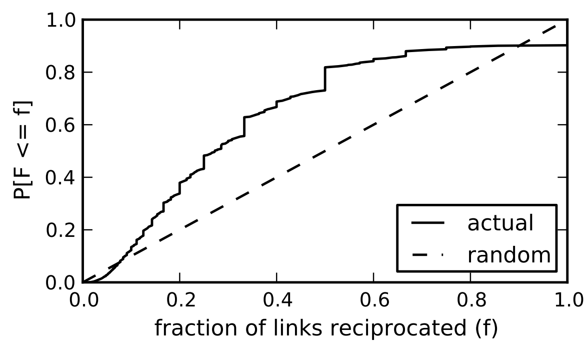

Figure 6: CDF of per-user fraction of links reciprocated (line marked "actual"). Line marked "Random" shows the CDF of the uniformly random distribution over [0,1] for comparison.

This naturally raises the question of whether bi–directional (reciprocated) links or uni–directional (unreciprocated) are more important. Unidirectional links can be interpreted as a "Follow" with one user seeking exposure to the activities of another user. Bidirectional links could arise either because the users are related to each other socially (rather than merely in the context of the videos), or because both users seek exposure to the activities of the other.

To test which of these holds, we measure whether users are more similar to users with whom they share unidirectional links, or bidirectional links. Users are first represented as 0–1 vectors in high–dimensional video–space. Each video represents a unique dimension in the vector. A user has a 1 value in a dimension if the user likes the video. Similarity is measured using cosine similarity. Given two user vectors $\mathbf{u_1}$ and $\mathbf{u_2}$, their similarity is given by

$$\cos\theta = \frac{\mathbf{u_1} \cdot \mathbf{u_2}}{\|\mathbf{u_1}\|\,\|\mathbf{u_2}\|}.$$

For each user, we capture the ratio of cumulative similarity scores across bi–directional links to the cumulative similarities across uni–directional links. Fig. 7 shows the distribution of relative weights placed by individual users on bi– and uni–directional links. This indicates that across the population, there seems to be an even spread, with roughly half of users having more similarity with their bi–directional partners than their uni–directional partners, the other half, vice versa.

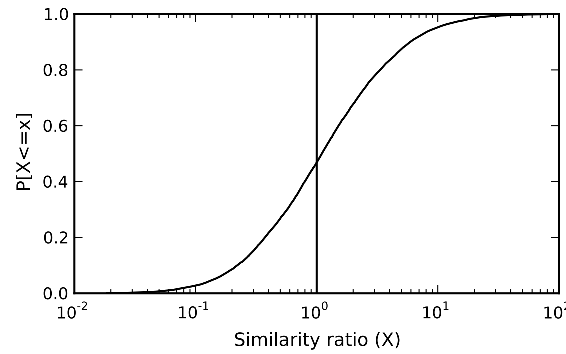

Figure 7: CDF of similarity ratios. Similarity ratio captures the ratio of the cumulative similarities of a user with their bidirectional friends to the cumulative similarities over unidirectional friends.

To summarise, links on *FN* are frequently not reciprocated. This is unlike other social networks. However, it is not clear whether this can be taken as an indicator of the interest-based nature of *FN*: some users are more similar to (like the same videos as) their bidirectional friends, whereas others are more similar to users with whom they have unreciprocated follows relationships.

4.5 Geographic locality

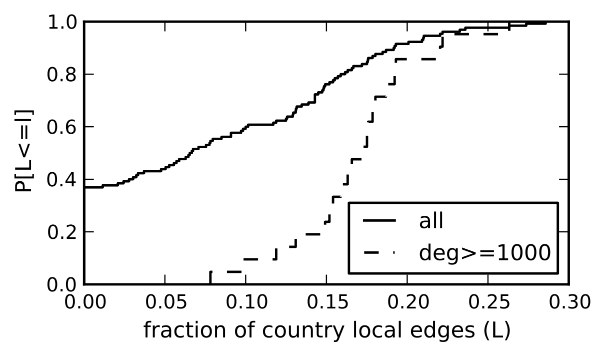

Figure 8: Fraction of country-local links per user (CDF).

Recent results suggest that in social networks, close links are preferred over long–distance links in online social networks [1]. To measure whether this holds in the *FN* on vimeo, users were first mapped to a country by extracting address information from the user's profile, and matching it to a latitude–longitude co-ordinate using Yahoo! Placefinder API[2].

Fig. 8 shows that contrary to what is seen in "purely" social networks such as Facebook where the probability of forming a contact decreases inversely with distance [1], users in vimeo have a strong tendency to form links across different countries. One can even say that this is expected given the fact that some videos might have universal and global

[2] `http://developer.yahoo.com/geo/placefinder/`

appeal (for example Hollywood movies). Nearly 40% of users have no friends from their own country (line marked "all"). High–degree users (line marked "deg >= 1000") have more country–local links, but this is still typically limited to at most a third of links. Thus, it appears that vimeo users are seeking to form links that are across geography.

This formation of user-user links across geographies appears to be consistent with Cairncross [2], who predicted that the ease of the Internet to connect users will bring a "Death of Distance". However, since this is the *opposite* of what is observed in social networks such as Facebook, which are believed to reflect "offline" or real-world friendships, we conjecture that user links in vimeo may not always represent offline friendships.

5. DISCUSSION AND CONCLUSIONS

In this work, we analysed several characteristics of networks of users formed on the video–sharing site vimeo to examine whether the network was based more on commonality of interest or on social links. In particular, we used the notion of Multirelational Social Networks to study both the declared networks of links between users, termed as the Followers' Network (FN), as well as networks inferred from users with common video "likes", termed as the Video-based Network (VN). We also studied an FVN network arising from merging FN and VN, by taking only edges which exist in both networks.

The analysis of number of nodes/relations shows that people who have relations in FN are not very keen on sharing their "likes" with their friends (only 30.08% of all FN users do so). Further, if they share videos, in 47% of relations, they have only one video in common. This suggests that the relations in FN network could be driven by social contacts rather than common interests.

However, we also found characteristics which do not match up with known features of social networks. For example clustering coefficient, unlike in other social networks, is low for both the FN and FVN. Also the analysis tie strengths, expressed as the number of commonly liked videos, shows that people in FVN on average share nine times more videos than those in VN. This is clearly visible when we merge FN and VN. In the resulting FVN network only 0.02% of connections that has one common like in VN survived. In addition, the FN network is highly directional; links are frequently not reciprocated. Further, users tend to form geographically distant links; which stands in direct contrast with purely social networks such as Facebook.

We present this to the community as a problem to be studied further. The interesting question is: what drives users to actively follow other users and become "friends", if they do not "like" many of the same videos? Also, to what extent is the relatively low number of shared likes between linked users caused by the large numbers of items available to consume? Some directions for further research include studying tags where they are available, to see if friends are more similar in "tag space" rather than in content spaces such as the video space of Sec. 4.4; studying channel subscriptions to see if "friends" like the same kinds of video uploaders, etc.

6. REFERENCES

[1] Backstrom, L., Sun, E., and Marlow, C. Find me if you can: improving geographical prediction with social and spatial proximity. In *Proc. 19th World Wide Web Conference* (2010).

[2] Cairncross, F. *The death of distance: How the communications revolution is changing our lives.* Harvard Business Press, 2001.

[3] Cha, M., et al. I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system. In *Proc. of the 7th ACM SIGCOMM conference on Internet measurement* (2007).

[4] Cheng, X., Dale, C., and Liu, J. Statistics and social network of youtube videos. In *IWQoS 2008* (2008).

[5] Gill, P., Arlitt, M., Li, Z., and Mahanti, A. Youtube traffic characterization: a view from the edge. In *Proceedings of the 7th ACM SIGCOMM IMC* (2007).

[6] Kazienko, P., Musial, K., and Kajdanowicz, T. Multidimensional social network and its application to the social recommender system. *IEEE Transactions on Systems, Man and Cybernetics - Part A: Systems and Humans 41*, 4 (2011), 746–759.

[7] Mislove, A., et al. Measurement and analysis of online social networks. In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement* (2007), ACM, pp. 29–42.

[8] Mislove, A., et al. Growth of the flickr social network. In *Proceedings of the first workshop on Online social networks* (2008).

[9] Musial, K., and Kazienko, P. Social networks on the internet. *World Wide Web Journal DOI: 10.1007/s11280-011-0155-z* (2012).

[10] Sastry, N. How to tell head from tail in user-generated content corpora. In *Proceedings of the 6th International AAAI Conference on Weblogs and Social Media* (2012).

[11] Schifanella, R., Barrat, A., Cattuto, C., Markines, B., and Menczer, F. Folks in folksonomies: social link prediction from shared metadata. In *Proc. WSDM* (2010).

[12] Watts, D. J., and Strogatz, S. H. Collective dynamics of 'small-world' networks. *Nature 393*, 6684 (1998), 440–442.

[13] Zink, M., Suh, K., Gu, Y., and Kurose, J. Characteristics of YouTube network traffic at a campus network-Measurements, models, and implications. *Computer Networks 53*, 4 (2009), 501–514.