Volume 7, Issue 4
November 2014
ISSN: 1946-1836

JOURNAL OF INFORMATION SYSTEMS APPLIED RESEARCH

In this issue:

The **Journal of Information Systems Applied Research** (JISAR) is a double-blind peer-reviewed academic journal published by **EDSIG,** the Education Special Interest Group of AITP, the Association of Information Technology Professionals (Chicago, Illinois). Publishing frequency is currently quarterly. The first date of publication is December 1, 2008.

JISAR is published online (http://jisar.org) in connection with CONISAR, the Conference on Information Systems Applied Research, which is also double-blind peer reviewed. Our sister publication, the Proceedings of CONISAR, features all papers, panels, workshops, and presentations from the conference. (http://conisar.org)

The journal acceptance review process involves a minimum of three double-blind peer reviews, where both the reviewer is not aware of the identities of the authors and the authors are not aware of the identities of the reviewers. The initial reviews happen before the conference. At that point papers are divided into award papers (top 15%), other journal papers (top 30%), unsettled papers, and non-journal papers. The unsettled papers are subjected to a second round of blind peer review to establish whether they will be accepted to the journal or not. Those papers that are deemed of sufficient quality are accepted for publication in the JISAR journal. Currently the target acceptance rate for the journal is about 40%.

Questions should be addressed to the editor at editor@jisar.org or the publisher at publisher@jisar.org.

JOURNAL OF INFORMATION SYSTEMS APPLIED RESEARCH

Editors

JISAR Editorial Board

Similarity and Ties in Social Networks: a Study of the YouTube Social Network

Amir Afrasiabi Rad
a.afrasiabi@uOttawa,ca
University of Ottawa
Ottawa ON K1N 6N5, Canada

Morad Benyoucef
Benyoucef@Telfer.uOttawa.ca
Telfer School of Management
University of Ottawa
Ottawa, ON K1N 6N5, Canada

Abstract

Social networks and the propagation of content within social networks have received an extensive attention during the past few years. Social network content propagation is believed to depend on the similarity of users as well as on the existence of friends in the social network. Our former investigation of the YouTube social network showed that strangers (non-friends and non-followers) play a more important role in content propagation than friends. In this paper, we analyze user communities within the YouTube social network and apply various similarity measures on them. We investigate the degree of similarity in communities versus the entire social network. We found that communities are formed from similar users. At the same time, we found that there are no large similarity values between friends in YouTube communities.

Keywords: Social Network Analysis; Similarity; Social Ties; Influence; YouTube

1. INTRODUCTION

Social networking websites, such as MySpace, Facebook, Twitter, Flickr, Orkut, YouTube, etc. are becoming more and more popular. Statistics show that in the US, almost 90% of the teenage and young adult age group are social network users (Trusov, Bodapati, & Bucklin, 2010). The birth of Web 2.0 allowed Internet content users to become Internet content providers as well. Social networks, as Web 2.0 applications, contribute their share to this paradigm shift. Social network users upload more than 35 hours of videos to YouTube every minute (YouTube LLC., 2010); and they contribute to Facebook by generating more than 30 billion pieces of content when they spend over 23 billion minutes on Facebook every month (Facebook Inc., 2011). Also, a billion tweets every month (Twitter Inc., 2011) is another indicator of this paradigm shift. Hence social networks are turning into hubs of social activity. Along with their popularity as a new communication medium, social networks are regarded as tools for social presence and for building social identity (Rad, Amir, & Benyoucef, 2011). The interconnected nature of social networks is a building block for establishing social identity. This is because social identity has no meaning if it is not defined in the context of a society. Social identity is always accompanied with ideas, or user generated

content in the context of a social network. These ideas get propagated through interconnections between people in a social network, and they work as a way of further establishing social identity. Therefore, it is the interconnectivity of users in online social networks that allows user generated content, ideas, and influence to be easily propagated through the whole social network (Afrasiabi Rad & Benyoucef, 2012).

The wide use of social networks and their ability for propagating ideas attracted the attention of the marketing community which soon realized that content propagation along social links can lead to a huge community of users who can be used as viral advertisers. Moreover, the unique characteristics of social networks provide the opportunity to harness the collective opinions of the population in order to shape user behavior through adequate marketing campaigns while gaining insights into current and future market trends (Asur & Huberman, 2010; Bearden, Calcich, Netemeyer, & Teel, 1986; Leskovec, Adamic, & Huberman, 2007). There has been numerous studies on the different aspects, enablers, and contributing factors of viral advertisement on social networks (Bearden et al., 1986; Domingos & Richardson, 2001; Duan, Gu, & Whinston, 2008; Evans, 2009; Hu, Tian, Liu, Liang, & Gao, 2011; Kempe, Kleinberg, & Tardos, 2005; Kim & Srivastava, 2007; Stephen & Toubia, 2009; Van den Bulte & Joshi, 2007). However, there is little research dedicated to discovering why and how idea propagation occurs in the online world.

In one of our earlier studies, and in an attempt to analyze propagation, its characteristics, and its contributing factors, we investigated the propagation of data in an open social network (i.e., YouTube) (Afrasiabi Rad & Benyoucef, 2012). We define an "open" social network as a social network where privacy settings allow for content posted by a user to be seen by all members of a social network. In other words, privacy settings do not restrict viewing, commenting on, or sharing content to only friends or followers (also called subscribers on certain social networks such as YouTube) of a user. Based on our definition, social networks such as YouTube, Twitter and Flickr fall into the category of open social networks. Our previous study (Afrasiabi Rad & Benyoucef, 2012) revealed that content propagation in online open social networks follows different patterns compared to what has been observed in offline social networks (i.e., pre-internet social networks) (Judea, 1986). Although the actions of individuals are usually open to a wide range of other users in both offline and online open social networks, interestingly, propagation in offline social networks is mostly affected by the number of ties (i.e., friends, coworkers, and family) and their networks, while our study revealed that in an online open social network, propagation is far more affected by individuals who are neither in the network of friends nor the network of followers of the content generator.

Other studies also revealed contradictory results. For instance, Crandall et al. (Crandall, Cosley, Huttenlocher, Kleinberg, & Suri, 2008) studied multiple online and offline social networks and discovered that an increase in similarity between online social network users boosts both the magnitude and speed of content propagation. On the other hand, and focusing merely on offline social networks, Feld (Feld, 1981) discovered that similarity is one of the major factors that define the strength of ties between members of a social network. Note that in this paper, we use "ties", "links", "connections" and "contacts" interchangeably to refer to friendship or following (also called subscribing to) relations between users in social networks, and that the focus here is mainly on friendship. A tie means the existence of a direct path between two social network users. It can be argued that since friends of a user have stronger ties with that user (assuming that friendship in online social networks has the same meaning as friendship in the offline world), and consequently a greater similarity, they should participate more in propagating the user's content, and consequently affect its propagation more than non-friends.

According to the literature, similarity is a boosting agent for content propagation, while our previous study (Afrasiabi Rad & Benyoucef, 2012) interestingly showed that strangers (non-friends, and non-followers) affected YouTube content propagation more than friends. Our objective here is to analyze communities (communities are formed by ties between users of a social network, and detected using random walks (Pons & Latapy, 2005)) within the YouTube social network to measure the similarity between members of those communities. For that we compute and analyze similarity metrics within the entire social network, and within its communities. This gives us a comparative tool for investigating similarity

values. We also evaluate the ratio of friendship over similarity with the goal of understanding if similar community members are in fact friends.

We focus on interest similarity since it is one of the most effective similarity measures contributing to the propagation of content or influence (Tang, Sun, Wang, & Yang, 2009). Although online social networks differ in their settings and content types, and probably follow different similarity patterns, a look at the work of Mislove et al. (Mislove, Marcon, Gummadi, Druschel, & Bhattacharjee, 2007) leads us to conclude that social networks that fall into the same category based on their privacy settings, user demographics, and applications, display similar information dissemination and similarity patterns. Considering that, we selected YouTube for our analysis as a good representative of online open social networks. We measure interest similarity between YouTube users based on the common topics they share with their friends, followers, and strangers in communities. We measure the similarity of connected and unconnected users in each community, and analyze the ratio of links between similar users versus dissimilar users. This will lead us to answer the question: "do similar users in communities befriend each other, and to what extent?"

Researchers in sociology, mathematics, and physics have proposed different similarity measures, and Social Network Analysis has adopted them to study similarity in social networks. In this paper we evaluate some of these similarity measures in a real social network setting and evaluate them based on the ratio of friendship between similar users.

The rest of the paper is organized as follows. The next section provides an overview of YouTube. Section 3 provides an introduction to the similarity measures used in our study. Section 4 is devoted to the results of our analysis. We continue on with a discussion in Section 5, and conclude the paper in Section 6.

2. Background

Similarity in social networks has been investigated from different angles. McPherson et al. categorized similarity into two categories: status homophily, and value homophily (McPherson, Smith-Lovin, & Cook, 2001). Status homophily can be regarded as structural similarity and value similarity is what we define in our research as interest similarity. According to McPherson et al. value homophily is derived from status homophily, hence it can be concluded that connected parties show similar interests and behavior.

However, Hinds et al. showed that, in a work environment (e.g., corporate social network), value homophily is a stronger indicator of tie formation than status homophily (Hinds, Carley, Krackhardt, & Wholey, 2000), which leads to the conclusion that McPherson et al.'s argument does not hold for every type of social network. However, research on social marketing reports that value homophily is an enabler of word-of-mouth distribution in online social networks (Anderson, 1998; Bernard J. Jansen, Mimi Zhang, Kate Sobel, & Abdur Chowdury, 2010; Hu et al., 2011). The importance of value homophily for online word-of-mouth distribution, hence for tie development, motivates us to investigate the relationship between value homophily and tie creation in online social networks, namely in YouTube as a representative of online social networks.

YouTube: an Open Social Network

YouTube, a subsidiary of Google, is the largest video sharing website containing about 43% of all videos found on the Internet (Flosi, 2010). Since its launch in 2005, the popularity of YouTube has consistently increased, and more web users, from various demographics, registered on this video sharing website to benefit from its contents and features. YouTube is not just an online repository for videos uploaded by users. YouTube also accounts for being a social network since it has a large number of registered users (aka channels) who can upload videos, follow (aka subscribe to) other channels, and be friends with other users (aka channels). Thus, many channels in YouTube have millions of friends and subscribers (YouTube LLC., 2010). YouTube, to fully qualify as a social network, provides facilities that enable communication and interaction between its members. YouTube satisfies this requirement by implementing a broad infrastructure that allows users to communicate with each other in many different ways which resulted in users commenting on nearly 50% of YouTube videos (YouTube LLC., 2010). YouTube's communication infrastructure includes the following features: Private messaging, Commenting on channels, Commenting on videos, Marking a video as favorite (favorite marking), Publishing video descriptions, Liking

or disliking a video description or a comment (rating), and Replying to a comment. In reality, users (channels) who subscribe to a channel will receive updates about the channel's activities on their news feed, and whenever they make a comment about or favorite-mark an activity, this act will appear in the news feed of their followers, and, in this way, the activities will propagate in the network.

YouTube also provides APIs that can be used by other web platforms interested in integrating YouTube services. By being integrated with many other web platforms, YouTube videos are not only displayed on a user's profile page, but they can be delivered directly to subscribers, and even the general public (online users) via email, Really Simple Syndication (RSS), and even in connection with other social networking platforms such as LinkedIn and Facebook. Videos can be also searched in search engines such as Google and Bing. These functionalities help YouTube videos to be propagated not only inside YouTube, but also on other platforms, which provides a unique advantage for word-of-mouth distribution, and is actually the reason for us choosing YouTube in our study.

YouTube provides the advantage of allowing two types of ties between channels: friendship, which creates a two-way relationship for channels, and subscription, which allows channels to get updates on any other channel while having a one-way relationship with those channels (Chakrabarti et al., 1999).

Another reason for choosing YouTube in our study is the fact that it allows (as of December 2010) for the existence of groups. By joining different groups, YouTube users could have access to a set of contents of their interests, all gathered in one location. Although Google has decided to revoke access to YouTube groups in December 2010, and has integrated it with Google+, our data, which was collected in 2007, shows a large participation of users in YouTube groups. Hence, we use YouTube group membership as an indicator of the interests of YouTube users. We argue that being members of the same group is indicative of the similarity of interests. In the next section, we explore different similarity measures used to evaluate similarity between users.

3. Similarity Measures and Functions

This section is devoted to a review of popular similarity measures used in social network analysis. According to Lin (Lin, 1998), similarity is a function of commonality and difference, in a way that if two objects are not exactly the same, their similarity depends positively on the amount of their common features, and will have negative relations with their differences.

Many similarity measures have been developed; each tied to an application or requiring a specific domain and design. Therefore, not all similarity measures are suitable to be applied on social networks to compute interest similarity. To measure the similarity of YouTube users, first, we selected a set of similarity measures that can be applied to interest similarity, and then we applied each measure (all of them discussed in this section) as a function of common group memberships of YouTube users. According to Baatarjav et al. , a group in a social network has specific characteristics that match the profiles of most of its members (Baatarjav, Phithakkitnukoon, & Dantu, 2008). Therefore, users who share a set of group memberships should have a similar profile. Note that analyzing similarity based only on group membership may not provide results as accurate as those that can be obtained by semantically analyzing, for instance, the content of users' postings, and considering the demographic information of users.

Jaccard and Dice's Similarity Coefficient

Jaccard and Dice's similarity coefficient measures are specific to measuring set similarity (Dice, 1945; Jaccard, 1901). They were first developed to measure similarities in ecological studies, but their nature of set operations made them applicable for measuring social similarity. They are computed by dividing the intersection of sets over their union. Jaccard and Dice's index can easily be converted to each other and provide monotonic asymmetric results. Therefore, in this paper, we only use Jaccard similarity coefficient for simplicity. Jaccard index is calculated using the following equation:

$$J(U_1, U_2) = \frac{|H_1 \cap H_2|}{|H_1 \cup H_2|} \qquad (1)$$

Where H_1 and H_2 are the group memberships of user U_1 and user U_2, respectively.

Russel and Rao Similarity

Russell and Rao similarity measure (RUSSELL & RAO, 1940) is close to Jaccard's similarity

coefficient. Russell and Rao measure the similarity of the common items compared to the whole vector including the attributes, here groups, that are absent from both vectors. In other words, the Russell and Rao similarity measure computes the common group memberships versus the whole set of unique groups in the system, and is calculated by:

$$R(U_1, U_2) = \frac{|H_1 \cap H_2|}{|H|} \qquad (2)$$

Where H represents the total number of group memberships.

Roger and Tanimoto Similarity
Roger and Tanimoto (Rogers & Tanimoto, 1960) devised a measure that is suitable for comparing the similarity of Boolean vectors. Their model gives double weight to disagreements. The Roger and Tanimoto index is calculated by:

$$T(U_1, U_2) = \frac{|H_1 \cap H_2| + |H_1^C \cap H_2^C|}{-3|H_1 \cap H_2| + 2(|H_1| + |H_2|) + |H_1^C \cap H_2^C|} \quad (3)$$

Where H_1^C represents the groups that do not have user U_1as their member.

Sokal and Sneath Similarity
Sokal and Sneath similarity measure (Sneath & Sokal, 1973) is comparable to Dice's measure and to Roger and Tanimoto measure. The only difference between Sokal and Sneath and Roger and Tanimoto similarity measures is in the heuristic constant components of the formulas, which produce almost similar results. Sokal and Sneath give double weight to matches instead of differences. Sokal and Sneath, however, founded their model on the Jaccard and Dice similarity measure by extending it to integrate dissimilarity of items into the calculation of similarity. It is calculated by:

$$S(U_1, U_2) = \frac{|H_1 \cap H_2| + |H_1^C \cap H_2^C|}{|H_1 \cap H_2| + |H_1| + |H_2| + 2|H_1^C \cap H_2^C|} \quad (4)$$

$L^1$and L^2 - Norms
With regard to sets, L^1-Norm, and L^2 -Norm (Gradshteyn, Ryzhik, Jeffrey, & Zwillinger, 2000) evaluate similarity to be the overlap between two groups divided by their sizes. L^2 –Norm compared to L^1-Norm decreases the level of effect that the sizes of individual sets have on the similarity measure. L^1 and L^2 –Norms are measured by:

$$L^1(U_1, U_2) = \frac{|H_1 \cap H\_2|}{|H\_1| \cdot |H\_2|} \qquad (5)$$

$$L^2(U_1, U_2) = \frac{|H_1 \cap H\_2|}{\sqrt{|H\_1| \cdot |H\_2|}} \qquad (6)$$

4. Interest Similarity and Ties in YouTube

According to Crandall et al. (Crandall et al., 2008), friends and followers in social networks are either similar to each other at the time the friendship (or follower) tie is made (aka selection process) or they grow in similarity over time after they become friends or followers through social influence. Also, rising similarity between two individuals is an indicator of current, and more specifically future, interactions between them (Crandall et al., 2008; Feld, 1981). Therefore, we argue that current activities of friends and followers of a user, who are presumed to have a certain degree of similarity, can be a predicator of that user's next activity. Hence, friends, also recognized as the most similar people by Crandall et al. (Crandall et al., 2008), should have the greatest effect on content propagation. But the question is: are friends the most similar people in their community? This section attempts to answer this question by analyzing data extracted from YouTube for similarity friendship ratios (the ratio yielding that what percentage of similar users in communities are friends). To do so, we utilize the similarity measures defined in Section 3 of this paper. Note that we cleaned the YouTube dataset to only keep friends in our evaluation and ignored all follower links in order to comply with the findings of Crandall et al. (Crandall et al., 2008) who only consider reciprocated links (here, YouTube friends).

Before we proceed, it is important to comprehend that communities are different from groups, where communities are concepts that are generated based on existing links between social network members, and groups are a feature introduced on social networks to gather users with similar profiles into a single place.

Data Description
Before developing our analysis, the data must be cleaned and made ready for analysis. We have access to a large dataset of over 1.15 million YouTube users and their group memberships along with information about ties between them. This dataset was collected and formerly used in

an analysis by Mislove et al. (Mislove et al., 2007). The dataset covers more than 30 thousand groups and contains over 290 thousands recorded group memberships, so on average, every user in the dataset is a member of roughly four groups. Every user, on average, has more than four reciprocatory and non-reciprocatory ties with other users. The most connected user has over 28 thousand links, while the majority of users only have one link. Figure 1 shows the frequency distribution of ties per user in the YouTube social network.

TABLE 1. YouTube Statistics

Type of Data	Statistics
Users	1,157,827
Groups	30,087
Users That are member of at Least One Group	94,238
Users That are not Members of any Group	1,063,589
Links	4,945,382
Number of Group Memberships	293,360
# of Groups that a user with highest number of membership is subscribed in	1,035
# of memberships for a group that has highest number of memberships	7,591
# of Communities	139,142

The highest number of ties in the network belongs to a user with 28,644 connections while the second most connected user only has 11,239 connections. Interestingly, about 183 thousand users only have one connection, and more than 500 thousand are not connected at all. This shows the level of uneven distribution of inactivity and activity in the YouTube social network. As it is apparent in Figure 1, most users have less than 128 ties. The full statistics of the YouTube dataset used in this study can be found in TABLE 1.

A more detailed look at the statistics shows that about 8% of the users are members of groups, which accounts for about 10 memberships per group. From this point on, our analysis only considers users who are group members, and we simply discard from our analysis the users who did not use YouTube's group feature. The statistical data also illustrates that, on average, users have three common group memberships,

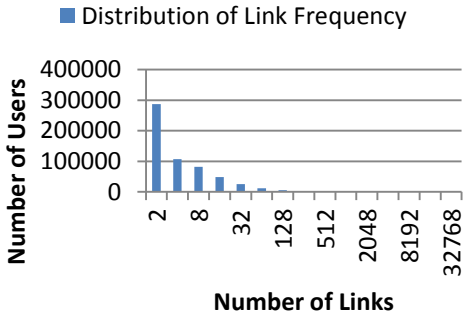

Figure 1. Frequency of ties per user

which shows a great potential for similarity between users.

As planned, we then extracted communities from the YouTube dataset. To do so, we relied on the random walk community detection technique described in (Pons & Latapy, 2005). The Random Walk community detection method discovers communities based on their structural similarity. It first estimates the distance of vertices, as a metric for estimating structural similarity, and assigns it to them as a weight. The next step is applying a hierarchical clustering model in order to identify clusters (communities). The algorithm works at the time complexity of $n^2\log(n)$, which is suitable for analyzing large graphs. We identified over 139 thousand communities with an average of 11 members per community, the largest community having 73 members.

Analysis of Similarities

As detailed earlier in this paper, we use common group memberships of users in the YouTube social network to measure the similarities between them. We argue that users who are members of the same set of groups are more likely to have similar interests, and that the similarity of interests increases as the number of common group memberships increases.

In order to perform this analysis, we implemented six programs, each of them responsible for performing one similarity measurement operation. The programs performed their analysis on a cleaned database of YouTube users that were previously clustered for communities using our RandomWalk clustering program developed using C++ and the iGraph (www. igraph. sourceforge. net) library.

To measure similarities, we selected six well-defined and generally accepted similarity measures as detailed in Section three of this paper. TABLE 2 describes the result of applying each technique on YouTube social network and its extracted communities.

TABLE 2 shows that for every similarity measure, the similarity of users within the communities is greater than the similarity within the entire social network. Being connected increases similarity, and therefore community members are more similar to each other than the rest of the network.

TABLE 2. Similarity Measures and the Result of Applying Them on the YouTube Social Network and its Communities

Metric	Social Network Average	Average Over Communities
Jaccard	0.14	0.31
Russel and Rao	0.90	0.91
L1	0.12	0.17
L2	0.26	0.34
Sokal and Sneath Similarity	0.50	0.54
Roger and Tanimoto Similarity	0.40	0.47

However, being a member of a community does not necessarily indicate friendship. A community is a collection of users who have transitive connections to each other. Therefore, there is a path between most community members. This also results in a high clustering coefficient for every node in the community. This means that a community is created from the collection of friends, friends of friends and so on. Based on our analysis, it is still not clear how much similarity induces friendship. To be able to answer this question, we selected users who have a more than average similarity with each other in their community, and examined if they are friends or not.

However, being a member of a community does not necessarily indicate friendship. A community is a collection of users who have transitive connections to each other. Therefore, there is a path between most community members. This also results in a high clustering coefficient for every node in the community. This means that a community is created from the collection of friends, friends of friends and so on. Based on our analysis, it is still not clear how much similarity induces friendship. To be able to answer this question, we selected users who have a more than average similarity with each other in their community, and examined if they are friends or not.

The result of our analysis shows that there is not a high correlation between similarity and friendship in communities (see TABLE). In other words, most similar users are not necessarily friends even in small communities within the social network. Note that being in the same community means either a direct friendship or the existence of a short path with many mutual friends between two users. The friendship similarity ratio in small communities of connected people is not large (a range of 11% to maximum 38%). Most similar users in communities are not friends with each other. In our former study (Afrasiabi Rad & Benyoucef, 2012), we observed that content propagation in social network communities is done mostly by non-friends or non-followers. Also, as argued in the literature, content propagation happens where there is a high similarity between the propagator and propagatee. Therefore, it can be deduced that it is possible for indirect friends to be more similar that direct friends. Thus, a comparison of the results presented in TABLE 2 and and TABLE 3 suggests that the higher average of similarity in communities might be the result of high similarity between indirect friends rather than similarity between friends.

TABLE 3. Similarity Friendship Ratios in Social Network and Communities

Metric	Similarity Friendship Ratio in Communities
Jaccard	0.12
Russel and Rao	0.38
L1	0.32
L2	0.11
Sokal and Sneath Similarity	0.12
Roger and Tanimoto Similarity	0.21

5. Discussion

Our analysis shows that every similarity measurement method consistently yielded some degree of similarity between users in communities. Based on the proposition in (Feld, 1981), the higher similarity within communities was expected to be higher than the average similarity in the whole social network. This was confirmed by our results. However, the subsequent analysis that resulted in relatively low friendship similarity ratios in the communities was unexpected. Feld (Feld, 1981) proposes similarity as a determining factor in social ties in offline social networks. Nevertheless, the situation can be different in online social networks. Offline social networks are known to be free of fake friends and spammers which is certainly not the case for online social networks (Manago, Taylor, & Greenfield, 2012). The problem starts to grow when we realize that fake friends have on average six times more friends than legitimate users (i.e., users whose friends are real) (Manago et al., 2012). Therefore, unless we have a mechanism to separate fake friends from real friends, the results cannot show the true ratio. Nonetheless, the friendship similarity ratio is so low that the general finding of low similarity between friends stands even if fake friends are removed from the network. The only difference would be a slight increase in the ratio.

Based on the research done by Feld (Feld, 1981), it is expected that, in offline social networks, similar people be friends with each other. Our study on YouTube found that this is not necessarily the case for online social networks. However, considering Feld's study, we expect that friends should have higher similarity. Therefore, similarity measures that result in a higher ratio between friendship and similarity provide more accurate results in the case of online social network.

By looking at the results presented in TABLE , the similarity measures that resulted in higher values of friendship similarity ratios in communities are Russel and Rao and L1 similarities. We have a second category including Jaccard, L2, and Sokal and Sneath Similarity, with relatively similar results. Comparing these results with the values presented in TABLE 2, we see that even though the similarity values resulting from different techniques vary, the techniques can be categorized into two major categories with regards to their approximate accuracy. A conclusion about which category provides better results will depend on more research to be conducted on the correlation between friendship and similarity in online social networks. In which case, a higher correlation will play in favor of the first category of measurement techniques, and a lower correlation will favor the second category.

6. Conclusion

In this paper we analyzed the YouTube social network with regards to the ties that exist between users and their common group memberships (which we used as an indicator of similarity of interests), to assess the relation between friendship and the similarity of interest inside communities of users within a social network. We found that the similarity between users increases if they are friends, but this increase does not define similarity as a determining factor in friendship.

Considering that, and also the fact that content propagation in online social network communities is done mostly by non-friends, and knowing that similarity is a driver for content propagation, we can conclude that, within communities, indirect friends are more similar to each other than direct friends (as they participate more in content propagation). The second possibility is that the YouTube communities are formed from users that have little similarity whether friends or non-friends. The deterministic conclusion on the findings discussed above needs more exploration on the similarities between indirect friends, which is one the paths for our future study.

Furthermore, we examined several similarity measures to find the most suitable ones for processing online social network data. We found that similarity measures can be categorized into two categories based on their accuracy, which is measured by the friendship ratio. The results yielded by the Russel and Rao as well as L1 similarity measures led to higher friendship similarity ratio, and Jaccard, L2, and Sokal and Sneath Similarity fell in the second category. More research is needed to determine which category provides better results for online social networks.

Our analysis can be developed further to extract better facts from a social network like YouTube. One of the limitations of this research is the lack

of comprehensive data on the YouTube network. We only used a sample of YouTube, where users are group members, and we ignored users who are not members of a group. This resulted in a large YouTube user base. Therefore, a higher group membership rate would have improved the results.

In our future work, we plan to investigate the validity of our findings on different types of social networks, such as photo sharing (Flickr), friendship (Orkut), professional (LiveJournal), and so on. Furthermore, we will try to detect fake friendships and remove them from our analysis to obtain more accurate results.

7. References

Afrasiabi Rad, A., & Benyoucef, M. (2012). Measuring Propagation in Online Social Networks: The case of YouTube. *Journal of Information Systems Applied Research*, *5*(1), 26.

Anderson, E. W. (1998). Customer Satisfaction and Word of Mouth. *Journal of Service Research*, *1*(1), 5–17. doi:10.1177/109467059800100102

Asur, S., & Huberman, B. A. (2010). Predicting the Future with Social Media. *SSRN eLibrary*. Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1579522

Baatarjav, E.-A., Phithakkitnukoon, S., & Dantu, R. (2008). Group Recommendation System for Facebook. In R. Meersman, Z. Tari, & P. Herrero (Eds.), *On the Move to Meaningful Internet Systems: OTM 2008 Workshops*, Lecture Notes in Computer Science (pp. 211–219). Springer Berlin Heidelberg. Retrieved from http://link.springer.com/chapter/10.1007/978-3-540-88875-8_41

Bearden, W. O., Calcich, S. E., Netemeyer, R., & Teel, J. E. (1986). An Exploratory Investigation Of Consumer Innovativeness And Interpersonal Influences. *Advances in Consumer Research*, *13*(1), 77–82.

Bernard J. Jansen, Mimi Zhang, Kate Sobel, & Abdur Chowdury. (2010, March 6). Twitter Power: Tweets as Electronic Word of Mouth. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.155.3321

Chakrabarti, S., Dom, B. E., Kumar, S. R., Raghavan, P., Rajagopalan, S., Tomkins, A., Gibson, D., et al. (1999). Mining the Web's link structure. *Computer*, *32*(8), 60 –67. doi:10.1109/2.781636

Crandall, D., Cosley, D., Huttenlocher, D., Kleinberg, J., & Suri, S. (2008). Feedback effects between similarity and social influence in online communities. *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '08 (pp. 160–168). New York, NY, USA: ACM. doi:10.1145/1401890.1401914

Dice, L. R. (1945). Measures of the Amount of Ecologic Association Between Species. *Ecology*, *26*(3), 297–302. doi:10.2307/1932409

Domingos, P., & Richardson, M. (2001). Mining the network value of customers. *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '01 (pp. 57–66). New York, NY, USA: ACM. doi:10.1145/502512.502525

Duan, W., Gu, B., & Whinston, A. B. (2008). Do online reviews matter? - An empirical investigation of panel data. *Decis. Support Syst.*, *45*(4), 1007–1016.

Evans, D. C. (2009, January 15). *Beyond Influencers: Social Network Properties and Viral Marketing*. Psychster Inc. Retrieved from http://www.slideshare.net/idealistdave/beyond-influencers-social-network-properties-and-viral-marketing-presentation

Facebook Inc. (2011). Facebook Statistics. Retrieved July 9, 2011, from http://www.facebook.com/press/info.php?statistics

Feld, S. (1981). The Focused Organization of Social Ties. *American Journal of Sociology*, *86*(5), 1015–1036.

Flosi, S. (2010, June). comScore Releases May 2010 U.S. Online Video Rankings - comScore, Inc. *ComScore Inc. - Press release*. Retrieved May 4, 2012, from http://www.comscore.com/Press_Events/Pre

ss_Releases/2010/6/comScore_Releases_May_2010_U.S._Online_Video_Rankings

Gradshteyn, I. S., Ryzhik, I. M., Jeffrey, A., & Zwillinger, D. (2000). *Table of Integrals, Series, and Products, Sixth Edition* (6th ed.). Academic Press.

Hinds, P. J., Carley, K. M., Krackhardt, D., & Wholey, D. (2000). Choosing Work Group Members: Balancing Similarity, Competence, and Familiarity. *Organizational Behavior and Human Decision Processes*, *81*(2), 226–251. doi:10.1006/obhd.1999.2875

Hu, N., Tian, G., Liu, L., Liang, B., & Gao, Y. (2011). Do Links Matter? An Investigation of the Impact of Consumer Feedback, Recommendation Networks, and Price Bundling on Sales. *Engineering Management, IEEE Transactions on*, *PP*(99), 1–12. doi:10.1109/TEM.2010.2064318

Jaccard, P. (1901). Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bulletin del la Société Vaudoise des Sciences Naturelles*, *37*, 547–579.

Judea, P. (1986). Fusion, propagation, and structuring in belief networks. *Artificial Intelligence*, *29*(3), 241–288. doi:10.1016/0004-3702(86)90072-X

Kempe, D., Kleinberg, J., & Tardos, É. (2005). Influential Nodes in a Diffusion Model for Social Networks. *Automata, Languages and Programming* (pp. 1127–1138). Retrieved from http://dx.doi.org/10.1007/11523468_91

Kim, Y. A., & Srivastava, J. (2007). Impact of social influence in e-commerce decision making. *Proceedings of the ninth international conference on Electronic commerce* (pp. 293–302). Minneapolis, MN, USA: ACM. doi:10.1145/1282100.1282157

Leskovec, J., Adamic, L. A., & Huberman, B. A. (2007). The dynamics of viral marketing. *ACM Trans. Web*, *1*(1), 5+. doi:10.1145/1232722.1232727

Lin, D. (1998). An information-theoretic definition of similarity. *Proc. 15th International Conf. on Machine Learning* (pp. 296–304). Morgan Kaufmann, San Francisco, CA. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.55.1832

Manago, A. M., Taylor, T., & Greenfield, P. M. (2012). Me and my 400 friends: the anatomy of college students' Facebook networks, their communication patterns, and well-being. *Developmental Psychology*, *48*(2), 369–380. doi:10.1037/a0026338

McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology*, *27*, 415–444. doi:10.2307/2678628

Mislove, A., Marcon, M., Gummadi, K. P., Druschel, P., & Bhattacharjee, B. (2007). Measurement and analysis of online social networks. *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement* (pp. 29–42). San Diego, California, USA: ACM. doi:10.1145/1298306.1298311

Pons, P., & Latapy, M. (2005). Computing Communities in Large Networks Using Random Walks. In pInar Yolum, T. Güngör, F. Gürgen, & C. Özturan (Eds.), *Computer and Information Sciences - ISCIS 2005*, Lecture Notes in Computer Science (pp. 284–293). Springer Berlin Heidelberg. Retrieved from http://link.springer.com/chapter/10.1007/11569596_31

Rad, A., Amir, & Benyoucef, M. (2011). A Model for Understanding Social Commerce. *Journal of Information Systems Applied Research*, *4*(2), 63.

Rogers, D. J., & Tanimoto, T. T. (1960). A Computer Program for Classifying Plants. *Science*, *132*(3434), 1115–1118. doi:10.1126/science.132.3434.1115

RUSSELL, P. F., & RAO, T. R. (1940). On Habitat and Association of Species of Anopheline Larvae in South-Eastern Madras. *Journal of the Malaria Institute of India*, *3*(1), 153–178 .

Sneath, P. H. A., & Sokal, R. R. (1973). Numerical taxonomy. The principles and practice of numerical classification. Retrieved from

http://www.cabdirect.org/abstracts/19730310919.html

Stephen, A. T., & Toubia, O. (2009). Deriving Value from Social Commerce Networks. *SSRN eLibrary*. Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1150995

Tang, J., Sun, J., Wang, C., & Yang, Z. (2009). Social influence analysis in large-scale networks. *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '09 (pp. 807–816). New York, NY, USA: ACM. doi:10.1145/1557019.1557108

Trusov, M., Bodapati, A. V., & Bucklin, R. E. (2010). Determining Influential Users in Internet Social Networks. *Journal of Marketing Research*, *XLVII*(August), 643–658. doi:10.2139/ssrn.1479689

Twitter Inc. (2011, March 14). Twitter Blog: #numbers. *Twitter Blog*. Blog. Retrieved March 8, 2012, from http://blog.twitter.com/2011/03/numbers.html

Van den Bulte, C., & Joshi, Y. V. (2007). New Product Diffusion with Influentials and Imitators. *MARKETING SCIENCE*, *26*(3), 400–421. doi:10.1287/mksc.1060.0224

YouTube LLC. (2010). YouTube - Press Statistics. Retrieved July 9, 2011, from http://www.youtube.com/t/press_statistics