Tracking the dynamic variations in a social network formed through shared interests

Gerold Pedemonte and May Lim*
National Institute of Physics, University of the Philippines, Diliman, 1101 Quezon City
*Corresponding Author: may@nip.upd.edu.ph

Received: 3 December 2009; Revised: 12 December 2009; Accepted: 14 December 2009

ABSTRACT

We tracked the dynamics of a social network formed by a shared interest in movies. Users-, movie ratings-, and rental date-data from the Netflix Prize dataset were used to construct a series of date-filtered social networks, wherein viewers were linked when they rented the same movie and gave the same rating. We obtained a nearly constant high clustering coefficient (0.60 – 0.85), and a low average path length (1.4 – 2.3) indicating a static 'small-world' network despite the dynamic behavior of the borrowers.

Keywords: networks, time series analysis, complex systems, small-world network

INTRODUCTION

Social network analysis helps us in understanding the nature of interactions between individuals in a society. Studies on social networks provide a useful framework for analyzing social events such as opinion dynamics (Amblard and Deffuant 2004), failure analysis (Motter and Lai 2002) and information spread (Mossa et al. 2002). Social network dynamics is routinely described as a static network with dynamical aspects, or temporal changes, occurring on the network (Barrat, Barthélemy, and Vespignani 2008). In a social context, this is analogous to constructing a network of people with given connectivities, i.e. friends, enemies, acquaintances, etc. which are defined by the weights between the links and actions occurring within the boundaries of the said network. For example, a disease will more likely spread from a patient who has a higher number of interactions (or links) with others. As such, it is the dynamics of the spreading of the disease that is modeled over a static network framework (Pastor-Satorras and Vespignani 2001). Compared to static networks (Albert and Barabasi 2002), a longitudinal study provides a more realistic picture of real world networks which are mostly dynamic in nature (Kossinets and Watts 2006; Braha and Bar-Yam 2006).

We can think of a dynamic network as a time-series of static networks. These series of networks can be reconstructed as a single static network if we take the sampling time interval to be equal to the total observation time. At different times, nodes continuously change connections to other nodes or even connect and disconnect from the network. The main difficulty with such an analysis is the dearth of available data describing social networks with such high time resolution. This has been recently addressed with the release of the Netflix Prize dataset. The dataset was first made available as the data input to the $1M Netflix Prize (Netflix), a contest to beat the accuracy of the recommendation engine used by Netflix in advising its customers on what movies to rent.

In this paper, we tracked the changes in the overall structure of the shared-interest network as we varied the observation period and the degree of shared inclination.

MATERIALS AND METHODS

Social network of movie viewers

The Netflix Prize dataset (Netflix 2006) provides an account of movie rentals, with rental dates and corresponding user ratings culled from the information database of Netflix, an online movie rental service available only in the United States. Users select the movies they wish to rent online, and the physical disks are delivered to them by the US Postal Service. While no other information (e.g., geographical location) about the users was released, we note that the online aspect of the service presents an additional minimum requirement on the end-user (i.e., Internet access) which may link the users even more than just by a shared interest in movies. The entire dataset has over 2 million users and 17,770 movies for the period October 1998 - December 2005. Integer ratings (highest = 5, lowest = 1) are associated with each rental. We constrained our analysis to a subset of the data with 21,324 viewers and 6,582 movies covering the period January 2000 - December 2001.

Network Construction

In constructing the network of viewers, we assumed that a shared interest is reflected in the user ratings. A pair of viewers were connected if they watched the same movie and gave the same rating. We note that our definition for shared-interest evenly considers shared-like (rating = 5) or shared-dislike (rating = 1) for a particular movie. Since monthly networks were constructed based on the date of rating, and not the release date of the movie, this does not discount the effect of a surge in rental requests for new releases. Furthermore, ratings subnetworks were also constructed from each monthly network to see if the dynamics would change for shared-like and shared-dislike networks. Our network construction method creates a fully-connected cluster for a group of viewers who watched the same movie and gave it a common rating.

Network analysis

We analyzed the time evolution of the constructed monthly networks from January 2000 to December 2001, as well as the time evolution of each of the five user-ratings values. We described the overall structure or topology of a network using the degree distribution, the average clustering coefficient, and the average path length. The degree k_i of a node i is the number of direct connections to it. In a sparse network, a high degree is indicative of the importance of a node in maintaining network connectivity. The clustering coefficient C_i measures the connectedness of neighbors of node i relative to a fully-connected idealization (Eq. 1),

$$C_i = 2e_i / k_i(k_i - 1) \qquad (1)$$

where k_i is the degree and e_i is the number of links between the neighbors of i. The denominator in Eq. (1) represents the total number of possible connections $k_i(k_i - 1)/2$ of the k_i neighbors. If the average clustering coefficient $<C>$ over all nodes is close to unity, then most of the nodes are directly connected to one another. Lastly, the shortest path length measures the minimum number of hops to reach another node from a reference node. The average shortest path length is taken over all possible node pairs. When the average shortest path length is low, a network is tagged as a 'small-world' network because any two nodes are separated by just a few hops.

RESULTS AND DISCUSSION

Popularity dynamics

From January 2000 to December 2001, the population of viewers $n(t)$ actively renting movies (Fig. 1a) from Netflix generally increased with time. On the other hand, the popularity of a movie $p_m(t)$ defined as the fraction of viewers $n_m(t)$ / $n(t)$ renting said movie during the specified time interval, where the subscript m is the movie index, showed fluctuations in demand (Fig. 1b shows the top five movies). Though some movies are very popular, the range of $p_m(t)$ for all movies is 0.0-0.565 (mean = 0.001535, stdev = 0.009236). In the absence of the giant clusters of viewers linked by the blockbuster movies, the network linkages would suddenly become sparse. While sustaining a high $p_m(t)$ is difficult to achieve for a single movie, there is always a movie which we call a 'hub movie' that is able to fulfill said role (Fig. 1c), whether via the controlled movie release date or seasonal demand. On the average, the most popular movie is watched

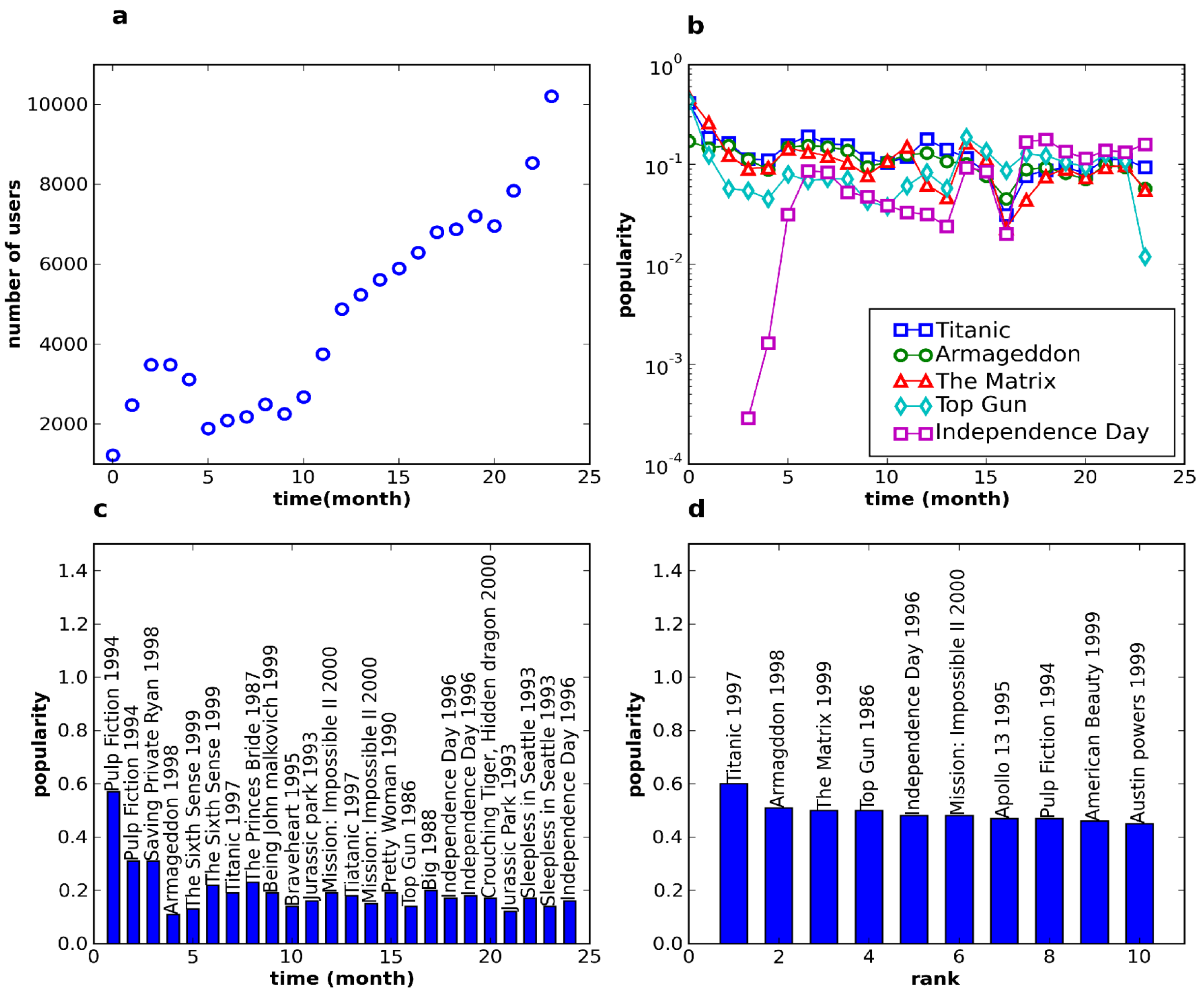

Figure 1. (a) Monthly population of users, (b) popularity of the five most watched movies, (c) monthly blockbuster movies (defined as hub movies); and (d) the top ten movies of 2000-2001.

by 10% of the active renters. Overall, the top ten movies were seen by half of the active renters during the entire observation period. Since the sequence of movies that the viewers will check out is not predictable, the corresponding fluctuation in Fig. 1b is unpredictable as well, sans the expected surge in popularity of new releases.

Popularity, however, is not equivalent to satisfaction which is reflected in the ratings given by the viewers. We constructed the five rating networks constructed for each monthly network from January 2000 to December 2001. The ratings frequency were: [1 (hated it): 169,490; 2 (didn't like it): 380,887; 3 (liked it): 849,534; 4 (really liked it): 838,414; 5 (loved it): 455,149]. As expected, borrowing dynamics was not random; viewers borrowed 3.9 times more of the movies they liked over those they didn't like. Since movies are available for rent at the earliest only months after the first release, reviews are already widely available by the time a viewer makes a decision to borrow. We propose that shared interest is most reflected in extreme opinions; a rating of 5 (or 1) would be more indicative of interest (or dislike) than an average rating of 3. Thus, aside from a monthly network description, we also looked into the disaggregated (by ratings) description of the monthly networks. Though beyond the scope of this work, we note that the rate of ratings bias

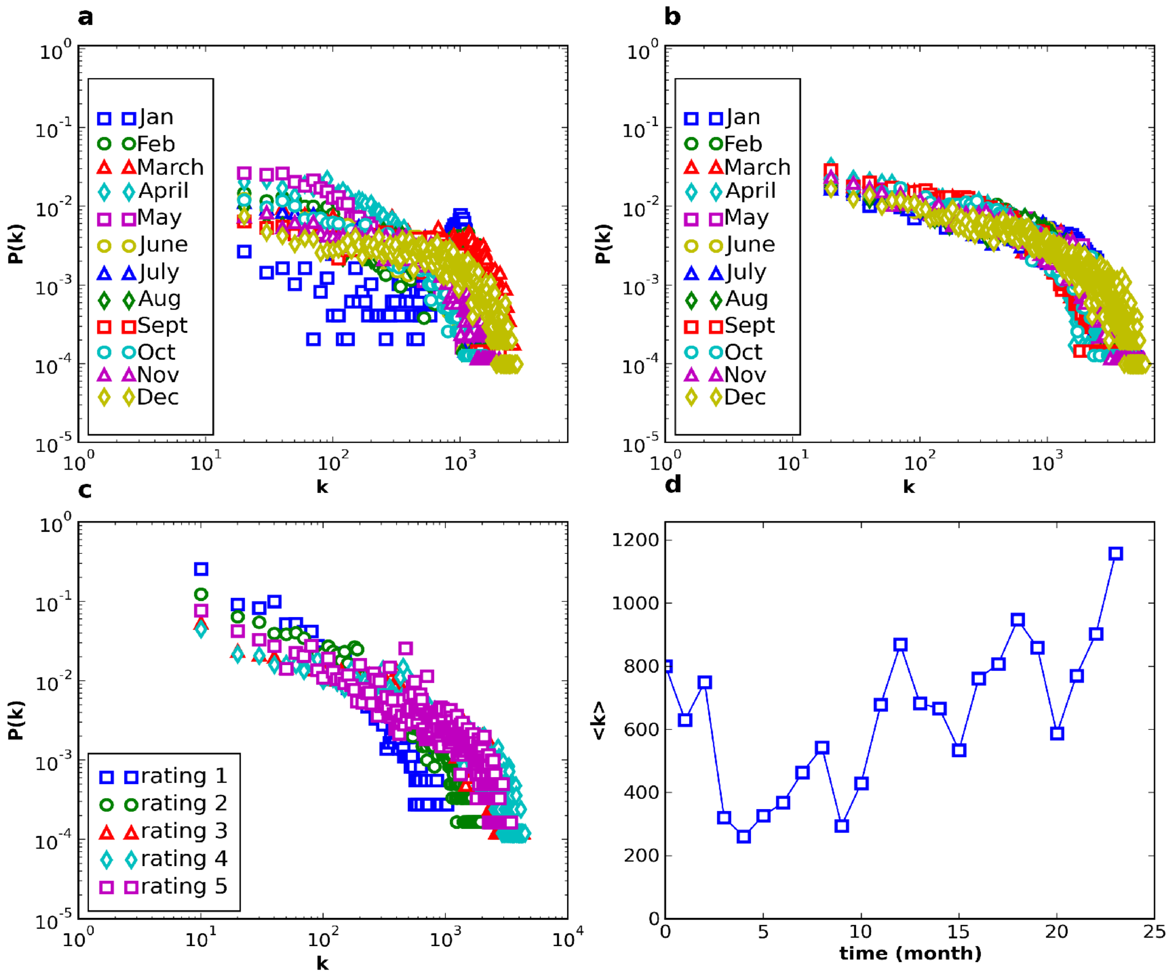

Figure 2. Degree distribution of monthly aggregated ratings network for year a) 2000 and b) 2001. (c) Degree distribution of rating networks for the month of December 2001. (d) Mean degree of the monthly aggregated rating networks of 2000-2001.

accumulation may hint on the rate of information dissemination.

Degree distribution

From the degree distribution, we gain a picture of the current direct connectivity of viewers who share the same interest. There are several possibilities through which a viewer would have a high degree: a) by watching a single hub movie, taking a value of k equal to the number of other viewers of that hub movie, b) by watching several movies m, each with its own unique following f such that $k \sim mf$, or a combination of both. Figures 2a and 2b illustrate a decreasing frequency (on a logarithmic scale) of viewers with increasing k. Majority of the viewers are able to watch only a few hub movies in each month. The rate at which $P(k)$ drops off gives us a picture of the relative homogeneity of the network. A sharper drop-off reflects a less diverse set of viewer interest. That is, there is a high concentration of viewers watching the hub movies (relatively low k). The range of k reflects not just the increasing population, but also the monthly dynamics and diversity of interests. Months which are known to have more holidays (and, consequently, downtime to allow movie viewing) have a larger range for k. By taking the average degree distribution $<k>$ on a monthly basis (Fig. 2d) and comparing it with the viewer population (Fig. 1a), we note that $<k>$ is correlated with the population size, but is not the only determinant of its value.

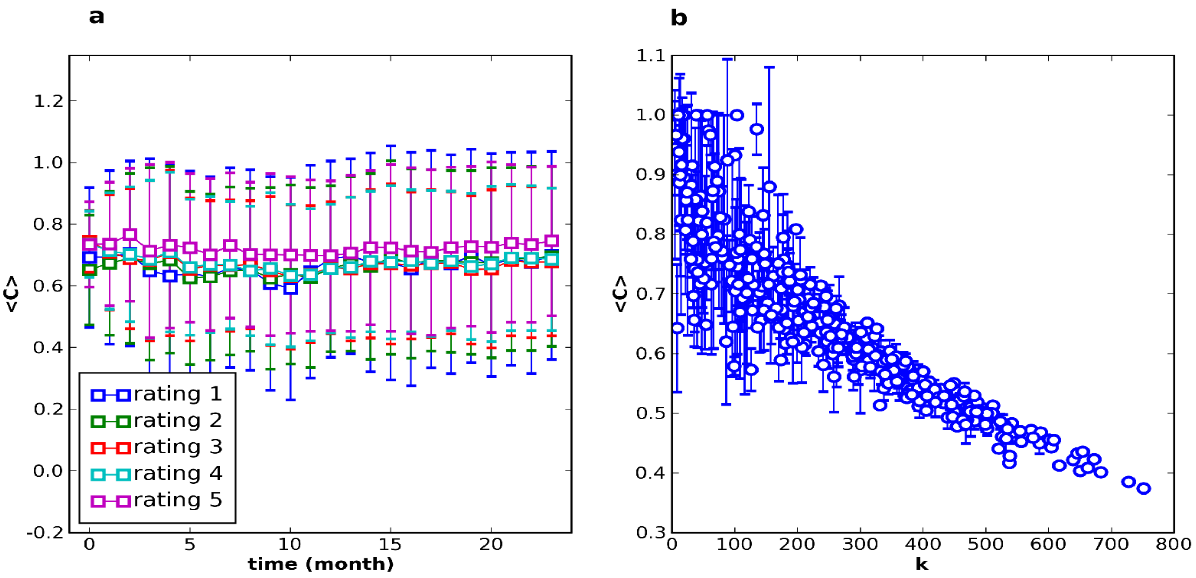

Figure 3. (a) Clustering coefficient for each rating network per month from 2000-2001. (b) Clustering spectrum of rating network 5 for the month of January 2000.

Clustering coefficient and the average shortest path

Within the uncertainty bounds, the average clustering coefficient <*C*> remains constant with time and ratings (Fig. 3a). Since seeing a common movie fully connects all the viewers, the value of <*C*> would be close to unity if all the viewers watched at least one popular movie each month. That <*C*> is within (0.6, 0.8) implies that a significant number of viewers saw non-hub movies, thus pulling down <*C*>. But since the movies being watched change with time, the nearly constant value of <*C*> for the monthly viewers networks imply that the underlying movie hierarchy of popularity (blockbusters to unknown) remain unchanged in time. Furthermore, the prevailing dynamics is robust to the addition of new viewers. The clustering spectrum *C*(*k*) shows the average clustering coefficient of nodes with the same degree *k* (Fig. 3b). In a network where there is a lack of correlations among nodes, the value of *C*(*k*) would remain constant with *k* (Vazquez, Pastor-Satorras, and Vespignani 2002). That *C*(*k*) decreases with *k* implies a correlation. The large variance for small *k* implies a large variation in the structure of viewers with small *k*, some belong solely to hub movie networks (*C* = 1.0) while others watch rarely-viewed movies along with the hub movie. On the other hand, the drop in <*C*> coupled with a small variance at large *k* implies the existence of a few viewers borrowing an expanded list of titles that make them the links between many small clusters. Such borrowers represent either multiple viewers borrowing under the same account, or viewers with such a wide range of interest that would make them poor predictors for a recommendation engine.

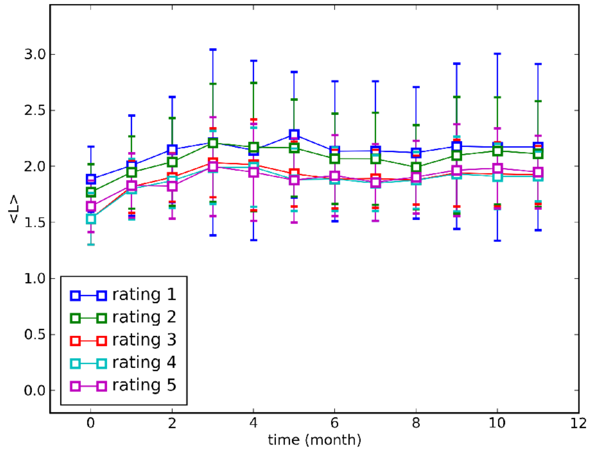

Figure 4. Mean shortest path per month for all rating networks for year 2000.

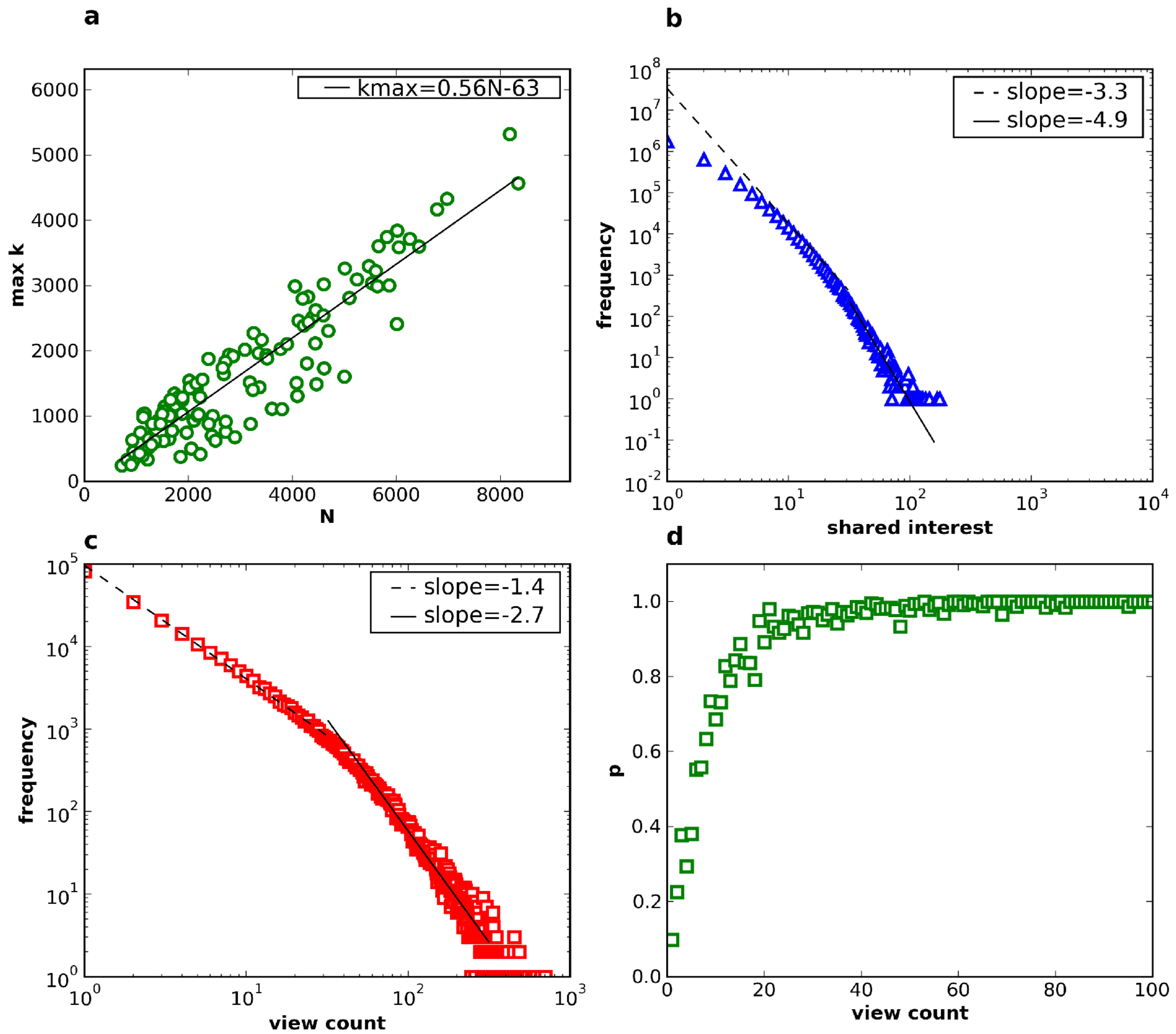

Figure 5. (a) Estimating the maximum degree k_{max} = 0.56N – 63 (R^2 = 0.85) as a function of the number of nodes N, (b) weight distribution of the edges, (c) movie viewers distribution, and (d) probability that a movie is connected to a blockbuster movie as a function of view count (p ≈ 1 for more than 100 views).

The constant trend in the average shortest path length (Fig. 4) confirms the constancy of the underlying hierarchical movie structure; there will always be movies (not necessarily the same movie) which persistently draw the majority of viewers at any given time. Thus, nodes are reachable in small hops, leading to the 'small-world' effect in movie viewership. It is telling, though, that viewers whorate a movie highly have a closer affinity to each other (slightly shorter <*L*>) than those who commonly hate a movie (rating = 1). That <*L*> is, within the error bounds, constant confirms the notion that negative ratings are just as good as positive ratings in determining shared interest.

Extremal behavior, weights and small-world effect

Using the five ratings network per month for the 24-month period (total of 120 data points), we determined the maximum number of connections a particular monthly rating network can have given its number of nodes. This allows us to estimate the largest possible connection with the recruitment of more viewers. The maximum value of *k* grows linearly with the number of nodes *N;* the coefficient 0.56 implies that the viewer with the widest interest is able to share the interest of 56% of the entire network. From the perspective of advertising, such a viewer would be easy to target. Such a linear response is a signature of random networks, implying that there is no change in the prevailing dynamics of the network; the topological structure is largely the same even with more viewers. The 'small world' effect is a characteristic of random networks, yet the clustering spectrum in Fig. 3b points to an underlying correlation. We calculated a proxy for the “shared interest” between all pairs of viewers by taking the number of common movies for which each pair gave exactly the same ratings for the entire two-year dataset, and plotted the histogram in Fig. 5b. The histogram plotted on a log-log scale, spanning two orders of magnitude on the “shared interest” axis, shows a steep (and monotonic) decline in the number of viewer pairs with a “shared interest”; there is no characteristic value for “shared interest” which reflects a truly heterogenous population. Finally, we looked into the variety of movies present in the Netflix database and showed the histogram of monthly movie viewership in Fig. 5c. Most movies in the Netflix database are not known to many (low view count) and a few are considered blockbusters (which we operationally defined as the movie having the top view count in each month, Fig. 1c). Figure 5d shows the likelihood that a viewer of a movie with the corresponding view count will also see a blockbuster movie. The probability p is described by a sigmoidal function, p = tanh(0.08 view count), which is indicative of a sharp (threshold) response. Though a movie may be viewed a mere 23 times (roughly 3% relative popularity with respect to the top draw), 19 out of 20 of those viewers (p = 0.95) would have also seen a blockbuster movie, illustrating the role of hub (or blockbuster) movies in linking diverse viewer interests.

CONCLUSION

The diversity of interests of the movie-watching population effectively creates a social network with a nearly constant high clustering coefficient (0.60 – 0.85), and a low average path length (1.4 – 2.3) when analyzed on a monthly basis. Even if the movies that link the viewers change rapidly from month to month, the underlying network structure remains unchanged and promotes a view that a hidden shared interest persists between viewers. The detection of this community structure would aid in targeted advertisements and messages, and can be effective in minimizing the cost of information dissemination.

ACKNOWLEDGEMENTS

This work received financial support from the University of the Philippines Office of the Vice Chancellor for Research and Development through Project No. 080805PNSE.

REFERENCES

Albert, R., Barabasi, A.L. 2002. Statistical mechanics of complex networks. *Reviews of Modern Physics* 74(1): 47. doi:10.1103/ RevModPhys.74.47.

Amblard, F., Deffuant, G. 2004. The role of network topology on extremism propagation with the relative agreement opinion dynamics. *Physica A: Statistical Mechanics and its Applications* 343:725-738. doi:10.1016/j.physa.2004.06.102.

Barrat, A., Barthélemy, M., Vespignani, A. 2008. Dynamical Processes on Complex Networks. 1st ed. Cambridge University Press.

Braha, D., Bar-Yam, Y. 2006. From centrality to temporary fame: Dynamic centrality in complex networks. *Complexity* 12(2): 59-63.

Kossinets, G., Watts, D. 2006. Empirical Analysis of an Evolving Social Network. *Science* 311(5757): 88-90. doi:10.1126/science.1116869.

Mossa, S., Barthelemy, M., Stanley, H.E., Nunes Amaral, L.A. 2002. Truncation of Power Law Behavior in "Scale-Free" Network Models due to Information Filtering. *Physical Review Letters* 88(13): 138701. doi:10.1103/PhysRevLett.88.138701.

Motter, A.E., Lai, Y.C. 2002. Cascade-based attacks on complex networks. *Physical Review E* 66(6): 065102. doi:10.1103/PhysRevE.66.065102.

Netflix. Netflix Prize: Home. http://netflixprize.com/.

———. 2006. UCI Machine Learning Repository: Netflix Prize Data Set. http://archive.ics.uci.edu/ml/datasets/Netflix+Prize.

Pastor-Satorras, R., Vespignani, A. 2001. Epidemic Spreading in Scale-Free Networks. *Physical Review Letters* 86(14): 3200. doi:10.1103/ PhysRevLett.86.3200.

Vazquez, A., Pastor-Satorras, R., Vespignani, A. 2002. Large-scale topological and dynamical properties of the Internet. *Physical Review E* 65(6): 066130. doi:10.1103/PhysRevE.65.066130.