Competition for popularity in bipartite networks

Mariano Beguerisse Díaz,[1,2,a)] Mason A. Porter,[3,4,b)] and Jukka-Pekka Onnela[4,5,6,7,c)]
[1]*Centre for Integrative Systems Biology, Imperial College London, South Kensington Campus, London SW7 2AZ, United Kingdom*
[2]*Division of Biology, Imperial College London, South Kensington Campus, London SW7 2AZ, United Kingdom*
[3]*Oxford Centre for Industrial and Applied Mathematics, Mathematical Institute, University of Oxford, OX1 3LB, United Kingdom*
[4]*CABDyN Complexity Centre, University of Oxford, OX1 1HP, United Kingdom*
[5]*Department of Physics, University of Oxford, OX1 3PU, United Kingdom*
[6]*Department of Biomedical Engineering and Computational Science, Helsinki University of Technology, FIN-02015 HUT, Finland*
[7]*Harvard Kennedy School, Harvard University, Cambridge, Massachusetts 02138, USA*

(Received 25 June 2009; accepted 15 July 2010; published online 8 October 2010)

We present a dynamical model for rewiring and attachment in bipartite networks. Edges are placed between nodes that belong to catalogs that can either be fixed in size or growing in size. The model is motivated by an empirical study of data from the video rental service Netflix, which invites its users to give ratings to the videos available in its catalog. We find that the distribution of the number of ratings given by users and that of the number of ratings received by videos both follow a power law with an exponential cutoff. We also examine the activity patterns of Netflix users and find bursts of intense video-rating activity followed by long periods of inactivity. We derive ordinary differential equations to model the acquisition of edges by the nodes over time and obtain the corresponding time-dependent degree distributions. We then compare our results with the Netflix data and find good agreement. We conclude with a discussion of how catalog models can be used to study systems in which agents are forced to choose, rate, or prioritize their interactions from a large set of options. © *2010 American Institute of Physics*. [doi:10.1063/1.3475411]

***Human dynamics*, which is concerned with the characterization of human activity in time, has been the subject of intense and exciting research over the past few years.[1–3] In one typical problem setting, individuals are endowed with limited resources, and there are numerous activities, behaviors, and/or products that compete against each other for those resources. Although such situations admit a natural formulation using bipartite (two-mode) networks that connect individuals to activities, human dynamics has surprisingly seldom been studied from this perspective. In the present paper, we analyze bipartite networks constructed from a large data set of video ratings by the users of a video rental company over a period of six years. To analyze the time evolution of these networks, we introduce the concept of a *catalog network*, and we use this approach to explore the driving forces behind the video rating behavior of individuals. We believe that such a framework can be used to study many other phenomena in human dynamics that involve the allocation of and competition for scarce resources.**

I. INTRODUCTION

Numerous natural and man-made systems involve interactions between large numbers of entities. The structural configuration of interactions is typically rather complicated, so the study of such systems often benefits greatly from network representations.[4–6] A network is usually abstracted mathematically as a graph whose nodes represent the entities and whose edges represent the interactions between the entities.[7] In many cases, edges can be weighted or directed, and more complicated frameworks such as hypergraphs can also be employed. The number of edges connected to a node in an unweighted network is known as its *degree*, and the *degree distribution* of a network is given by the collection of numbers that give the fraction of nodes that have degree k (for all values of k).[5] In weighted networks, one considers the weight of an edge rather than simply whether or not it exists.

Because networked systems are not static, the past decade has witnessed a particular interest in models that attempt to address their growth and evolution.[6] Perhaps the best-known model of network growth was formulated by Barabási and Albert.[4,8] Similar models were also constructed decades earlier by Simon[9] and de Solla Price.[10] Barabási and Albert examined networks arising from diverse settings and found that their degree distributions often seemed to follow power laws, which are functions of the form $f(x) \sim x^{-\alpha}$ (with $\alpha > 0$). They proposed a growth mechanism, which they called *preferential attachment* (de Solla Price called it *cumulative advantage*) to try to explain their observations. One starts with a small seed network and—in the simplest form of the mechanism—iteratively adds individual nodes that

a)Electronic mail: m.beguerisse-diaz08@imperial.ac.uk.
b)Electronic mail: porterm@maths.ox.ac.uk.
c)Electronic mail: onnela@hcp.med.harvard.edu.

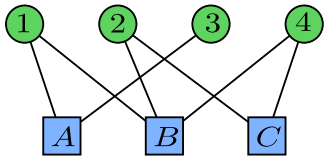

FIG. 1. (Color online) A bipartite network with nodes in the partite sets $\mathcal{U}=\{1,2,3,4\}$ and $\mathcal{V}=\{A,B,C\}$. Each edge connects a number to a letter.

each possesses exactly one edge. One connects each new node to an existing one chosen at random with a probability proportional to its degree. That is, the probability to choose node v_i with degree k_i is

$$P(v_i) = \frac{k_i}{\sum_{j=1}^{N} k_j},$$

where the total number of nodes N indicates the size of the network. Because nodes with higher degrees have correspondingly higher probabilities to receive new edges, the preferential attachment growth mechanism can result in a power-law degree distribution.[8,11]

Because of ideas like preferential attachment and the resulting insights on the origin of heavy-tailed degree distributions that one sees, e.g., in the World Wide Web or scientific collaboration networks, the study of networks has grown immensely during the past ten years.[5,6,12] However, most of this research has concentrated on one-mode (unipartite) networks, in which all of the nodes are of the same type. It is perhaps underappreciated that other graph structures are also very important in many applications.[13] Even the simplest generalization, known as a two-mode or *bipartite network*, has been studied much more sparingly than unipartite networks. Bipartite networks contain two categories (partite sets) of nodes: $\mathcal{U}=\{u_1,u_2,\ldots,u_U\}$ (with U members) and $\mathcal{V}=\{v_1,v_2,\ldots,v_V\}$ (with V members). As shown in Fig. 1, each (undirected) edge connects a node in $\mathcal{U}$ to one in $\mathcal{V}$.[7] Bipartite networks abound in applications: they can represent affiliation networks in which people are connected to organizations or committees,[14] ecological networks with links between cooperating species in an ecosystem,[15] and more.[16–20]

A bipartite network possesses a degree distribution for each of the two node types. We denote the adjacency matrix of a bipartite network by $\mathbf{G} \in \mathbb{R}^{U\times V}$. Each matrix element $\mathbf{G}_{ij}$ has a nonzero value if and only if an edge exists between nodes u_i and v_j. We denote the matrices that result from the two unipartite projections as $\mathbf{G}_{\mathcal{U}}=\mathbf{G}\mathbf{G}^T \in \mathbb{R}^{U\times U}$ and $\mathbf{G}_{\mathcal{V}}=\mathbf{G}^T\mathbf{G} \in \mathbb{R}^{V\times V}$. The degree of a node in each unipartite-projection network is the number of nodes of the same type with which the node shares at least one neighbor in the original bipartite network. The node strengths similarly incorporate connection strengths from the original bipartite network. (Recall that the "strength" of a node is the sum of the weights of the edges connected to it.) For example, in an unweighted affiliation network, the two projections encode the weighted connection strengths (the numbers of common affiliations) among individuals and the interlocks (the numbers of common people) among organizations.[14,21]

Many of the real-life systems that can be represented by bipartite networks are dynamic, as the existence and connectivity of both nodes and edges can change in time. For example, a person might retire or leave one organization to join another. One of the simplest types of changes is edge rewiring, in which one end of an edge is fixed to a node and the other end moves from one node to another (such as in the aforementioned change of affiliation). Because of the important insights they can offer, network rewiring models have received considerable attention.[18,19,22–26] They are closely related to abstract urn models from probability theory,[27–29] models of language competition,[30] and models of transmission of cultural artifacts.[31] More generally, they can help lead to a better understanding of any system in which the nature or existence of interactions among agents changes over time.[2]

The rest of our presentation is organized as follows. In Sec. II, we analyze a large data set of time-stamped video ratings from the video rental service Netflix that we model as a bipartite network of users and videos. In Sec. III, we examine the bursty behavior of individual users. In Sec. IV, we develop a catalog model of bipartite network growth and evolution. We then study the Netflix data using this model in Sec. V. Finally, we discuss our results and present directions for future research in Sec. VI.

II. NETFLIX VIDEO RATINGS

Netflix is an online video rental service that encourages its users to rate the videos they rent in order to improve their personalized recommendations. As part of the Netflix Prize competition,[32] in which the company challenged the public to improve their video recommendation algorithm, Netflix released a large, anonymized collection of user-assigned ratings of videos in its catalog. In this paper, we use the Netflix data to study human dynamics in the form of video ratings from a limited catalog. One can also examine the dynamics of the ratings themselves, which would complement a recent empirical study of video ratings that used data from the Internet Movie Database (IMDB).[33] The Netflix data consist of 100 480 507 ratings of 17 770 videos. The ratings, which were given by 480 189 Netflix users between October 1998 and December 2005, were sampled uniformly at random by Netflix from the set of users who had rated at least 20 videos.[34] Each entry in the data includes the video identification (ID), user ID, rating score (an integer from 1 to 5), and submission date. To illustrate some of the temporal dynamics in the data, we show in Fig. 2 the total number of ratings for each day in July and August 2003. The number of daily ratings exhibits a weekly pattern in which Mondays and Tuesdays have the highest activity and Saturdays and Sundays have the lowest. This reflects the weekly patterns in human work and leisure habits.

Figure 3 shows the total number of ratings from 2000 to the end of 2005. These ratings seem to grow exponentially, which we confirm by fitting the data to the function

$$r(t) = a_r(e^{b_r t} - 1) \tag{1}$$

using nonlinear least squares. We obtain the parameter values $a_r \approx 6.3656\times 10^5$ and $b_r \approx 0.0024$.

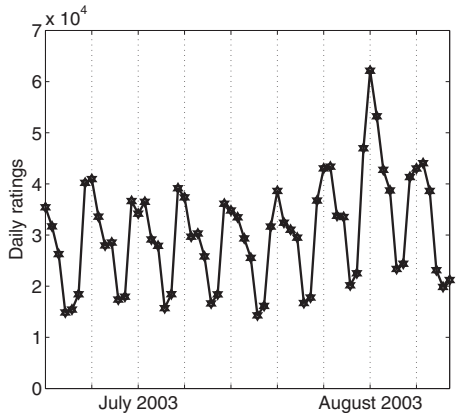

FIG. 2. Number of daily video ratings for each day in July and August 2003. The mean number of ratings per day over this period is 30 449. The dashed vertical lines indicate Tuesdays.

The number of users also grows exponentially, as shown in the top panel of Fig. 4. The dashed curve in the plot is a fit to

$$u(t) = a_u(e^{b_u t} - 1), \tag{2}$$

where we obtain $a_u \approx 1.0018 \times 10^4$ and $b_u \approx 0.0018$. We will need to take the exponential growth of the system into account when comparing data from dates that are far apart from each other.

In the bottom panel of Fig. 4, we show the number of videos from 2000 to 2005. The number of videos appears to grow roughly linearly as a function of time, but in fact it is better described by the relation

$$v(t) = a_v + b_v t^{c_v}, \tag{3}$$

where fitting yields $a_v \approx 2780.00$, $b_v \approx 0.6705$, and $c_v \approx 1.3097$.

A. Bipartite network formulation

The Netflix data can be represented as a bipartite network. The two types of nodes in this network are users and videos. We use $\mathcal{U}$ to denote the set of users and $\mathcal{V}$ to denote the set of videos. We ignore the rating values and consider

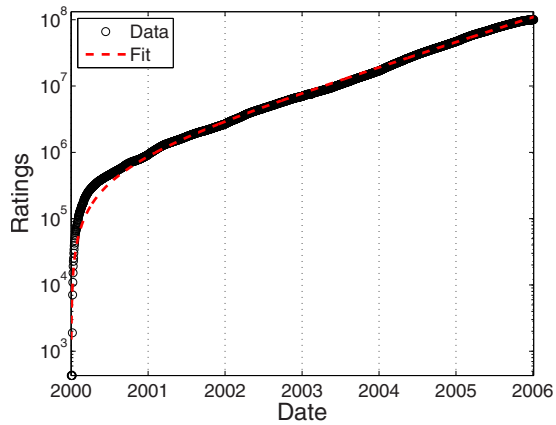

FIG. 3. (Color online) Number of video ratings in the Netflix data vs time from the beginning of 2000 to the end of 2005. Circles indicate data from Netflix and the dashed red curve is a fit to Eq. (1).

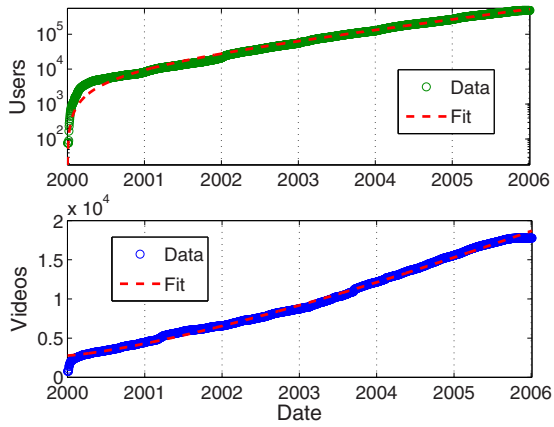

FIG. 4. (Color online) Number of users (top) and videos (bottom) in the Netflix data vs time from the beginning of 2000 to the end of 2005. Circles indicate data from Netflix and the dashed red curves are fits to Eqs. (2) and (3) for users and videos, respectively.

only the presence or absence of a rating event, which corresponds to an edge between a user and a video in the unweighted bipartite network. The large size and longitudinal nature of the data provide a valuable opportunity to study video rating in the context of human dynamics, as has been done previously with mobile telephone networks,[3,35] book sale rankings,[36] and electronic and postal mail usage patterns.[1,37]

B. Degree distributions

The bipartite video-rating network has one degree distribution for the user nodes and another one for the video nodes. Keeping in mind the observations in Fig. 2, we examine the cumulative degree distributions of individual days. The distributions have a similar functional form for each day in the data set. We fit them to a power law with an exponential cutoff,

$$F(k) \sim k^{-a} e^{-bk}, \tag{4}$$

using a modification of the method discussed by Clauset *et al.*[38] As an example, we show in Fig. 5 the cumulative degree distributions for one day. Table I gives the parameter values that we found in our fits of the data to Eq. (4). Despite the weekly pattern of the ratings shown in Fig. 2, we did not find any significant differences between the values of a and b for different days of the week. Hence, although the number of daily ratings does differ significantly among weekdays, such differences seem to not have much effect on the aggregate structure of the network.

The problem setting sheds some insight into the observed functional form of the degree distribution. Users select which videos to rate from a large set of possibilities and possess time limitations on the number of videos that they are able to watch and rate. As in any market, videos must compete against each other for users' attention. One can also anticipate that certain videos saturate their market, especially in the case of niche videos whose audience is necessarily small. Once the demand for a niche video has been met, it

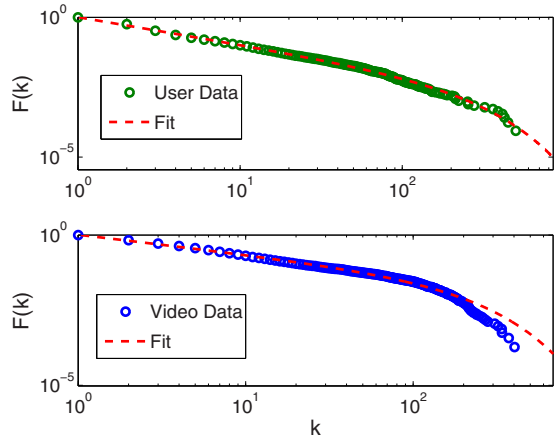

FIG. 5. (Color online) Cumulative degree distributions of user (top) and video (bottom) nodes for 26 August 2003 (a Tuesday). The dashed curves are the fits to Eq. (4) with parameters $a \approx 0.9828$, $b \approx 0.0057$ for the users and $a \approx 0.6622$, $b \approx 0.0070$ for the videos.

virtually ceases to receive further ratings. On the other hand, blockbusters might continue receiving numerous ratings for a long time.

C. Clustering coefficients

To investigate the local connectivity of nodes and examine the impact of highly connected nodes, we calculate bipartite clustering coefficients.[16,39] In bipartite networks, a clustering coefficient for a node can be calculated by counting the number of cycles of length 4 (i.e., the number of "squares") that include the node and dividing the result by the total possible number of squares that could include the node. As stated in Ref. 16, the possible (or underlying) number of squares is calculated by adding the potential links (including existing ones) between a particular node and the neighbors of its neighbors. In Fig. 6, we show how a square occurs in a bipartite network when two neighbors of a node have another neighbor in common. Bipartite networks cannot have triangles (three mutually connected nodes) because two nodes of the same type cannot be neighbors, so a square is the shortest possible cycle.

The definition of a clustering coefficient of node v_i in an unweighted bipartite network is[16]

$$C_4(v_i) = \frac{\Sigma_{h,j} q_{i_{jh}}}{\Sigma_{j,h}[(k_j - \eta_{i_{jh}}) + (k_h - \eta_{i_{jh}}) + q_{i_{jh}}]}, \qquad (5)$$

where $q_{i_{jh}}$ is the observed number of squares containing v_i and any two neighbors u_h and u_j. The degrees of the neighbors are k_h and k_j, respectively, and $\eta_{i_{jh}} = q_{i_{jh}} + 1$. The possible number of squares is calculated by adding the degrees of the nodes u_h and u_j minus the link that each shares with v_j if the three nodes are not part of a square (to avoid double counting). If the three nodes are part of a square, then the square represented by the deleted link must be added again, which yields $(k_j - \eta_{i_{jh}}) + (k_h - \eta_{i_{jh}}) + q_{i_{jh}}$ in the denominator of Eq. (5).

TABLE I. Fitting parameters of the daily video and user degree distributions from 2000 to 2005 for the power law with exponential cutoff in Eq. (4).

	a		b	
	Mean	Variance	Mean	Variance
Videos	0.6580	0.0200	0.0686	0.0100
Users	0.8381	0.0573	0.0116	0.0007

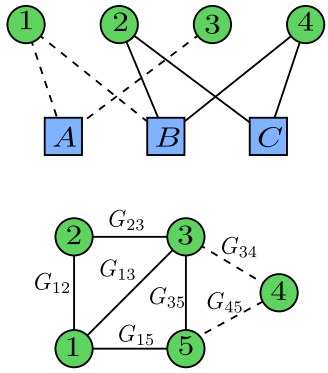

FIG. 6. (Color online) Examples of how to calculate clustering coefficients for bipartite (top) and unipartite (bottom) networks. In the bipartite network, solid lines indicate edges that form the square that includes node B, whose bipartite clustering coefficient calculated according to Eq. (5) is $C_4(B) = 1/5$. One obtains this result because there are five possible squares for this node ($\{1A2B, 1C2B, 1A4B, 1C4B, 2C4B\}$) but only one of them ($2C4B$) actually exists. In the unipartite network, the solid lines indicate edges that form the triangles that include node 1. If this were an unweighted network, for which $G_{ij} \in \{0,1\}$ for all i and j, then one would obtain an unweighted clustering coefficient of $C_3(1) = 2/3$. To calculate the value of its weighted clustering coefficient $\tilde{C}_3(1)$, we use Eq. (6).

In Fig. 7, we show the values of $C_4(v_i)$ and $C_4(u_i)$ versus k_i for the video and user nodes for a single day (Tuesday, 12 August 2003). In Table II, we show the mean values and variances of the bipartite clustering coefficient of all one-day snapshots of Netflix in 2003. Despite the weekday-dependent variation in the number of daily ratings, the values of the bipartite clustering coefficient do not vary significantly across weekdays. However, on the weekends, the values of $\langle C_4 \rangle$ for both the user nodes and the video nodes increase by almost 80%. For a network constructed from a single day's

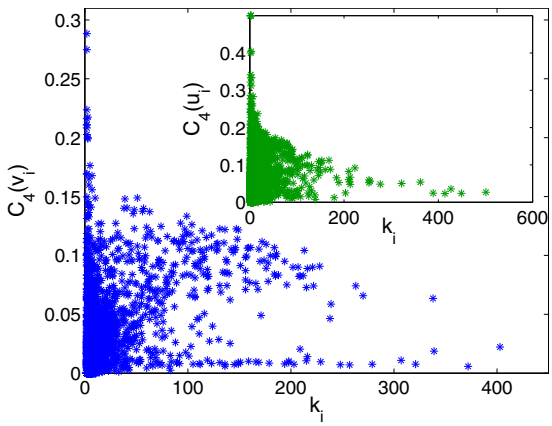

FIG. 7. (Color online) Bipartite clustering coefficients $C_4(v_i)$ for video (blue) and user nodes (inset/green) for 12 August 2003 (a Tuesday). The mean values for this day are $\langle C_4 \rangle = (1/V)\Sigma_{i=1}^{V} C_4(v_i) \approx 0.02606$ for the videos and $\langle C_4 \rangle = (1/U)\Sigma_{i=1}^{U} C_4(u_i) \approx 0.03144$ for the users.

TABLE II. Means and variances of $\langle C_4 \rangle$ (for the bipartite network) and $\langle \tilde{C}_3 \rangle$ (for the projections) of videos and users on single-day snapshots of 2003, calculated using Eqs. (5) and (6).

	$\langle C_4 \rangle$		$\langle \tilde{C}_3 \rangle$	
	Mean	Variance	Mean	Variance
Videos	0.0204	0.0007	0.0056	10^{-6}
Users	0.0209	0.0012	0.0044	10^{-6}

data, only about 2% of the possible squares typically exist; this is comparable to what would occur in a random network with the same degree distributions. To investigate whether the presence of blockbuster nodes (which have high degrees and increase considerably the number of possible squares) has any effect on the value of $\langle C_4 \rangle$, we calculated the clustering coefficient after removing the top ten most rated videos. We did not find any conclusive evidence that blockbusters drive the value of the clustering coefficient; some of them caused the value of $\langle C_4 \rangle$ to decrease and others caused it to increase.

One can also examine clustering coefficients in the weighted unipartite networks given by the projected adjacency matrices $\mathbf{G}_{\mathcal{U}}$ and $\mathbf{G}_{\mathcal{V}}$. We calculate the weighted clustering coefficient for each projection using the formula[40]

$$\tilde{C}_3(v_i) = \frac{2}{k_i(k_i-1)} \left[\frac{1}{G_V} \sum_{j,h} (G_{ij} G_{ih} G_{hj})^{1/3} \right], \tag{6}$$

where k_i is again the degree of node v_i, G_{ij} is the weight of the edge between v_i and v_j, and $G_V = \max(G_{ij})$ denotes the maximum edge weight in the network. The geometric mean $(G_{ij}G_{ih}G_{hj})^{1/3}$ of the edge weights gives the "intensity" of the (i,j,h)-triangle. When the network is unweighted, $(G_{ij}G_{ih}G_{hj})^{1/3}$ is 1 if and only if all edges in the (i,j,h)-triangle exist and 0 otherwise. This reduces Eq. (6) to the unweighted unipartite clustering coefficient

$$C_3(v_i) = \frac{2t_i}{k_i(k_i-1)}, \tag{7}$$

where t_i is the number of triangles that include node v_i. A node v_i with degree 1 has $C_3(v_i)=0$.

In Fig. 8, we show the $\tilde{C}_3(u_i)$ values for the user projection $\mathbf{G}_{\mathcal{U}}$ (with 10 228 nodes and 814 667 edges) from Tuesday, 4 August 2003. In Table II, we show the mean clustering-coefficient values for the projected user and video networks for all single-day snapshots of 2003. The values of $\langle \tilde{C}_3 \rangle$ did not vary much among weekdays, except for the videos' $\langle \tilde{C}_3 \rangle$, which almost doubled its value on the weekends from an average of 0.0045 between Monday and Friday to 0.0086 on Saturday and Sunday.

Given the values of $\langle C_4 \rangle$ in Table II, it is unsurprising that the values of $\langle \tilde{C}_3 \rangle$ are also typically low. In the inset of Fig. 8, we show the values of the users' unweighted clustering coefficient C_3, which are naturally much higher. For example, about 4000 users have $C_3=1.0$, indicating that all potential triangles exist among these users. This differentiates one set of nodes from the rest. This feature, which we

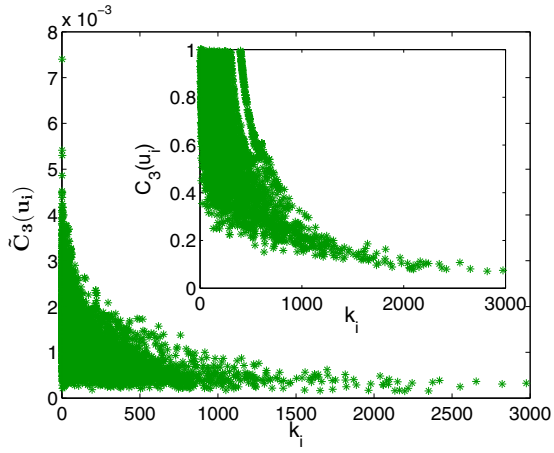

FIG. 8. (Color online) Weighted clustering coefficient $\tilde{C}_3(u_i)$ for nodes in the unipartite projection onto users for 4 August 2003. The x-axis represents node degrees and the y-axis represents $\tilde{C}_3(u_i)$. The mean values for this day are $\langle \tilde{C}_3 \rangle = (1/U)\Sigma_{i=1}^{U} \tilde{C}_3(u_i) \approx 0.0013$ for the projection onto users and $\tilde{C}_3 \approx 0.0086$ for the projection onto videos (not shown). The inset shows values of the unweighted clustering coefficient $C_3(u_i)$ from the same data.

observe often in the data, arises from the dominant video of the day. On 4 August 2003, this video (which is typically a blockbuster) was Daredevil, which had 396 ratings and created many edges in the user projection among the users who rated it. Removing Daredevil from the bipartite network also removes these deviant nodes. This feature is not apparent if one calculates only the weighted unipartite clustering coefficient $\tilde{C}_3$. Just as we did with $\langle C_4 \rangle$, and given the dramatic effect observed by removing Daredevil, we calculated $\langle \tilde{C}_3 \rangle$ and $\langle C_3 \rangle$ for the projected network of users after removing the ten most rated videos. We found that for each top-ten video that we removed, the value of $\langle C_3 \rangle$ increased by about 0.2%, and the value of $\langle \tilde{C}_3 \rangle$ increased by a somewhat larger percentage.

III. USER BURSTS

A close examination of the rating habits of individual users can also yield rich and informative insights. Recent research has shown that people tend to have bursts of e-mail and postal correspondence, in which they send and receive numerous messages within short periods of time, which are followed by long periods of inactivity.[1,37,41] We find similar features in the Netflix data, as about 70% of the users exhibit bursty behavior by rating several videos in one rating session after several days without activity. We illustrate this phenomenon in Fig. 9 by plotting the cumulative distribution of inter-event times between the ratings of one user over a period of almost five years. We fit this distribution to a power law $F(x) \sim x^{-\alpha}$ using the method discussed in Ref. 38 to determine the value of the exponent α. We can similarly provide estimates for possible power laws (with actual power laws over roughly two decades of data) among the other bursty users, although the value of α depends on the final degree (i.e., the total number of rated videos) of the user. For example, the mean exponent for bursty users with final degrees between 100 and 1000 is $\alpha \approx 2.54$, whereas it is

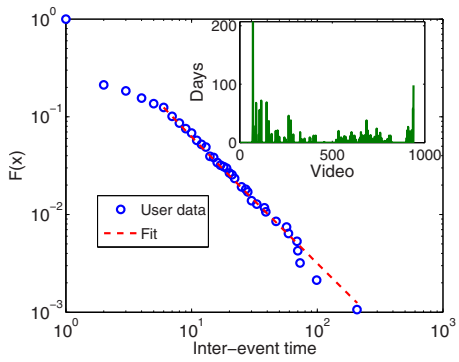

FIG. 9. (Color online) Cumulative distribution of the inter-event time between the ratings of one Netflix user. The user signed up on 4 April 2000 and has a degree of 940 based on ratings cast over a period of almost five years. The dashed curve indicates a fit to the function $F(x) \sim x^{-\alpha}$, which yields $\alpha \approx 2.27$ in this case. The inset shows the number of days that separate consecutive video ratings.

$\alpha \approx 3.17$ for those with final degrees of at least 4000. Additionally, there are several types of users among those who do not exhibit bursty dynamics. In particular, some users rated only a very small number of videos (which might be due to the sampling done by Netflix) and others exhibit seemingly unrealistic levels of rating activity. (For example, there are 47 users who signed up in January 2004 or later and who have rated more than 4000 videos each.)

IV. CATALOG NETWORKS

The above empirical investigation of the Netflix data motivates the development of an evolution model for bipartite *catalog networks*, which arise in a diverse set of applications. Such networks have two sets of nodes whose numbers can be fixed or dynamic, and edges are placed one at a time between previously unconnected nodes that are chosen according to predefined rules. One continues to add edges until a predefined final time has been reached or the system has become saturated, at which point every node in one partite set is connected to every node in the other partite set. The Netflix network can be studied using such a catalog network framework; it starts completely disconnected (nobody has rated any videos), and the users start choosing and rating videos from the catalog. Depending on the way the data set is sampled, the catalogs can be static (e.g., a one-day snapshot) or dynamic (e.g., the full data set). Catalog models of network evolution are closely related to the network rewiring problem studied by Plato and Evans[2,19] that features fixed sets of artifacts and individuals. Every individual has one affiliation (a connection) with an artifact and can reassign this connection to another node as the network evolves. In contrast, in a catalog network, any edge that has been placed between two nodes in the network is permanent. Consequently, catalog networks are suited to describing records of interactions that are assigned dynamically and then remain permanently in the system.

As before, $\mathcal{U}$ denotes the set of users and $\mathcal{V}$ denotes the set of videos. The size of $\mathcal{U}$ is $U(r)$ and the size of $\mathcal{V}$ is $V(r)$, where r denotes a discrete time that is indexed by the ratings. That is, we take every rating event as a time step, so when we discuss time in this context, we are referring to "rating time" and not physical time unless we indicate otherwise. The time-dependent catalog vectors, $D_{\mathcal{U}}$ and $D_{\mathcal{V}}$, have components given by the degrees of each node in the catalog:

$$D_{\mathcal{U}}(r) = \begin{bmatrix} k_{u_1}(r) \\ k_{u_2}(r) \\ \vdots \\ k_{u_U}(r) \end{bmatrix}, \quad D_{\mathcal{V}}(r) = \begin{bmatrix} k_{v_1}(r) \\ k_{v_2}(r) \\ \vdots \\ k_{v_V}(r) \end{bmatrix}. \tag{8}$$

These vectors have size $U(r)$ and $V(r)$, respectively. We denote by $N_{\mathcal{U}}(r,k)$ (with $k \in \{0,1,\ldots,V(r)\}$) and $N_{\mathcal{V}}(r,k)$ (with $k \in \{0,1,\ldots,U(r)\}$) the numbers of users and videos, respectively, that have degree k at rating time r. One can normalize $N_{\mathcal{U}}(r,k)$ to obtain the proportion of nodes with degree k; this is given by $\hat{N}_{\mathcal{U}}(r,k) = N_{\mathcal{U}}(r,k)/U(r)$. An analogous relation holds for $\hat{N}_{\mathcal{V}}(r,k)$.

Based on our intuition about the choosing and rating of videos, we add edges to the network using a combination of linear preferential attachment and uniform attachment. On one hand, one expects the choice of a user to be driven in part by the choices made by others, as popular videos are more likely to attract further viewings and hence ratings. On the other hand, one also expects an element of idiosyncrasy on the part of each user, allowing him or her to choose any video from the catalog regardless of the choices of others. This results in two time-dependent probabilities—one for users and one for videos—each of which consists of a convex combination of preferential and uniform attachment terms. More specifically, each time an edge is added to the network, we select a user and a video to be connected by this new edge. The video (user) node is chosen using uniform attachment with probability $1-q$ (respectively, $1-p$) and linear preferential attachment with probability q (respectively, p). The addition of an edge occurs during a single discrete (rating) time step, as is common in models of network evolution. Combining these ideas, a video node with degree k_i is chosen with probability

$$P_{\mathcal{V}}(r,k_i) = \frac{1-q}{V(r) - N_{\mathcal{V}}(r,U(r))} + \frac{qk_i}{\|D_{\mathcal{V}}(r)\|_1 - U(r)N_{\mathcal{V}}(r,U(r))}, \tag{9}$$

and a user node with degree h_i is chosen with probability

$$P_{\mathcal{U}}(r,h_i) = \frac{1-p}{U(r) - N_{\mathcal{U}}(r,V(r))} + \frac{ph_i}{\|D_{\mathcal{U}}(r)\|_1 - V(r)N_{\mathcal{U}}(r,V(r))}, \tag{10}$$

where the values of the parameters $p,q \in [0,1]$ are fixed, $\|D_{\mathcal{U}}(r)\|_1 = \Sigma_{i=1}^{U(r)} k_i(r)$, and $\|D_{\mathcal{V}}(r)\|_1 = \Sigma_{i=1}^{V(r)} h_i(r)$. The probabilities $P_{\mathcal{U}}(r,h_i)$ and $P_{\mathcal{V}}(r,k_i)$ change over time as the degrees of the nodes change when edges are added to the network.

The denominators in Eqs. (9) and (10) contain the terms $N_{\mathcal{V}}(r,U(r))$ and $N_{\mathcal{U}}(r,V(r))$ because once a node of either type is fully connected, it is no longer eligible to receive any

new connections and is effectively no longer in the catalog until a new node of the other type arrives. When $r=0$, one obtains $\|D_v(0)\|_1=\|D_u(0)\|_1=0$ and $N_{\mathcal{V}}(0,U(r))=N_{\mathcal{U}}(0,V(r))=0$, which would result in division by zero. To overcome this problem, we follow the standard procedure employed in network growth models[4] by seeding the algorithm with an edge that connects two randomly chosen nodes (one from each of the partite sets). This is equivalent to shifting the rating-time variable and changing the initial conditions to $\|D_v(0)\|_1=\|D_u(0)\|_1=1$.

A. Rate equations

One can use rate equations (i.e., master equations) to investigate the dynamics of the degree distributions of a catalog network. This type of approach has been used successfully to study a large variety of other networks.[2,5,11,19,42,43] The analysis of the degree distribution for videos in the catalog model is identical to that for users, as only the constants and sizes of the catalogs are different. Accordingly, we present our results for the degree distributions of the videos; one obtains the results for user distributions by changing q to p, $V(r)$ to $U(r)$, and $P_{\mathcal{V}}(r,k)$ to $P_{\mathcal{U}}(r,k)$. For notational convenience, we also drop the subscripts in this subsection, so $N(r,k)$ denotes the number of nodes with degree k at time r. To construct the rate equations, one must consider how many nodes pass through $N(r,k)$ (i.e., turn into nodes of degree k and $k+1$) for $k\in\{0,1,2,\ldots,U(r)\}$. This yields

$$\frac{dN(r,0)}{dr}=v'(r)-P_{\mathcal{V}}(r,0)N(r,0),$$

$$\frac{dN(r,k)}{dr}=P_{\mathcal{V}}(r,k-1)N(r,k-1)-P_{\mathcal{V}}(r,k)N(r,k),\quad k>0 \tag{11}$$

where $v(r)$ is a continuous approximation of $V(r)$, and $v'(r)=\frac{dv}{dr}$. The initial conditions are

$$N(0,0)=V(0)-1,$$

$$N(0,1)=1, \tag{12}$$

$$N(0,k)=0,\quad k>1.$$

Equation (11) is a system of coupled nonlinear ordinary differential equations. The positive and negative terms account, respectively, for an increase and decrease in the number of nodes of a given degree as nodes receive new edges. The equation for $N(r,0)$ has $v'(r)$ as a positive term to indicate the entry of new nodes (with degree 0) to the network. The time-dependent probabilities $P_{\mathcal{V}}(r,k)$ are defined in Eq. (9). In the case of fixed catalogs, there is a maximum value of k, so the final equation in Eq. (11) takes a slightly different form (see below).

1. Fixed catalogs

We begin by analyzing the evolution of the network with fixed catalog sizes, so $U(r)=U$, $V(r)=V$, and $v'(r)=0$ for all r. Because a finite, fixed number of users and videos are available in the catalogs, the network can only evolve until time $r=UV$. At this point, the system becomes saturated (i.e., $N_{\mathcal{U}}(VU,V)=U$ and $N_{\mathcal{V}}(VU,U)=V$), and no additional edges can be added to the network. Note that Eq. (11) changes slightly for fixed catalogs. In particular, the last equation for nodes with degree U changes to

$$\frac{dN(r,U)}{dr}=P_{\mathcal{V}}(r,U-1)N(r,U-1), \tag{13}$$

which only has the positive term because nodes with degree U stay that way until the end of the process.

Additionally, although the degree distribution of a network generated using the catalog model with static node sets is time-dependent, the long-time asymptotic behavior is always the same:

$$\lim_{r\to UV}N(r,k)=\begin{cases}V & \text{if } k=U\\ 0 & \text{if } k<U,\end{cases}$$

which gives a de facto final condition to the system in Eqs. (11)–(13). Accordingly, we examine degree distributions for $r\le UV-1$.

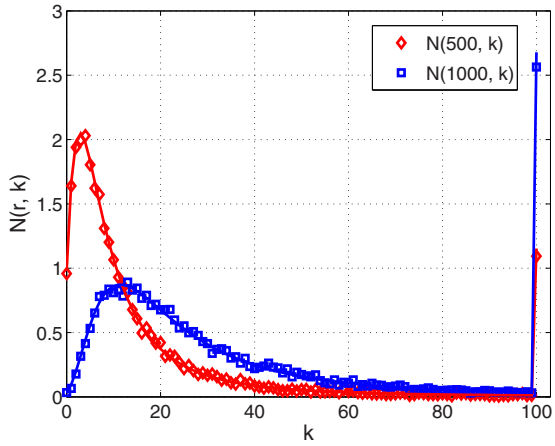

FIG. 10. (Color online) Degree distributions of video nodes averaged over 500 simulations of a fixed catalog network with $U=100$ users, $V=30$ videos, and $q=0.8$ at rating times $r=500$ (red diamonds) and $r=1000$ (blue squares). The solid curves are the solutions to the differential Eq. (11).

In Fig. 10, we show the degree distribution of the video nodes averaged over 500 simulations of a fixed catalog network (with $U=100$, $V=30$, and $q=0.8$) at different times during its evolution. As the discrete time r increases, the peaks of the functions travel toward higher values of k and decrease as if they were diffusing. We also observe a jump in $N(r,k)$ at $k=U$. This occurs because there are nodes in the network that become fully connected during the edge-assignment process (see Fig. 11). Interestingly, Johnson *et al.*[44] showed recently that the time-dependent degree distributions in some networks that undergo edge rewiring with preferential attachment follow nonlinear diffusion processes.

Figure 12 reveals how the user nodes achieve full connectivity between $r=0$ and $r=UV-1$. The image shows the "paths" that user nodes follow in the (r,k)-plane between (0, 0) and $(UV-1,V)$. For example, the nodes that follow a steep (high k for early r) trajectory are the ones that receive many links early on. Their degree grows mostly from preferential attachment in the edge-assignment mechanism, and

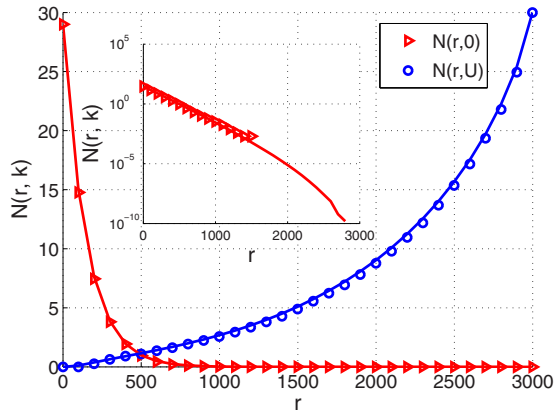

FIG. 11. (Color online) Numbers of video nodes $N(r,0)$ with degree 0 (red triangles) and $N(r,100)$ (blue circles) with degree 100 from 500 simulations of a fixed catalog network with $U=100$, $V=30$, and $q=0.8$. Inset: decrease of $N(t,0)$ on a semilogarithmic scale; the decrease appears to be exponential. The solid curves come from the solutions of Eq. (11).

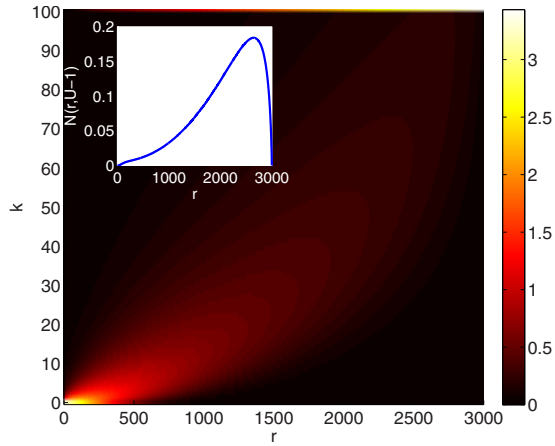

FIG. 13. (Color online) Numerical solution of $N(r,k)$ for video nodes from Eq. (11) with a fixed catalog and $q=0.8$, $V=30$, and $U=100$. [We again plot $\ln(N(r,k)+1)$.] The horizontal line at $k=100$ corresponds to the saturated nodes $N(r,U)$. The inset shows a plot of $N(r,U-1)$ for the same network.

they accordingly achieve full connectivity early in the process. The nodes that acquire edges more slowly initially begin to receive edges very fast as r approaches UV (because other nodes have already saturated), explaining the steep climb in the upper right corner of the figure.

The "final" condition that $N(UV-1,U)=V$ makes the system in Eq. (11) very stiff for high values of k and r. Figure 13 shows the path that the video nodes follow in the (r,k)-plane (i.e., the same information as in Fig. 12 but for video nodes) but for the numerical solutions of Eq. (11) instead of direct network simulations. In the inset of the figure, we show the profile of $N(r,U-1)$, which evinces the aforementioned stiffness. Because all nodes must be fully connected at $r=UV-1$, nodes with low degrees begin to receive many edges for high values of r. This causes $N(r,k)$ for high k to peak late in the process, and the nodes "travel" through values of k rather quickly, which explains the incredibly steep slope of $N(r,U-1)$ as r approaches $UV-1$.

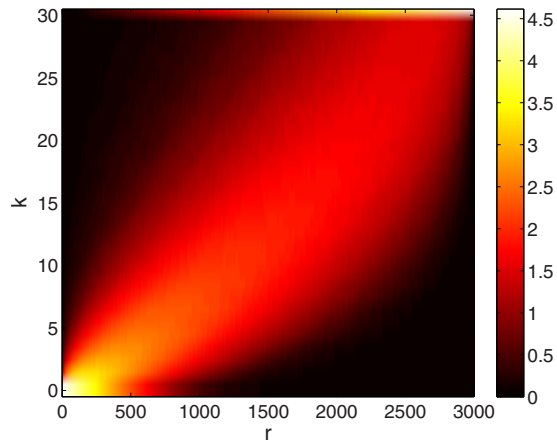

FIG. 12. (Color online) Mean of $N(r,k)$ for user nodes in 500 simulations of a fixed catalog network with $U=100$, $V=30$, and $p=0.5$. The axes are (rating) time r and degree k, and the color indicates the value of $\ln(N(r,k)+1)$. The horizontal line at the top of the image is the discontinuity (as seen with the video nodes in Fig. 11) that corresponds to the value of $N(r,V)$ and reflects the appearance of fully-connected user nodes.

The value of q affects the width of the region (light colored) in the (r,k) plane. For lower values of q (e.g., $q=0.3$), uniform random attachment dominates and the region of activity becomes narrower. The nodes attain edges at roughly the same pace. For larger values of q, the first nodes to receive edges become increasingly likely to continue receiving more nodes until they saturate, and the area of activity of the nodes becomes wider (see Fig. 13).

2. Growing catalogs

In the Sec. IV A 1, we described the dynamics of catalog networks when the sizes of the catalogs are fixed. While this provides a good baseline investigation, catalogs can grow in many applications—for example, Netflix gains both new subscribers and new videos almost every day. Accordingly, in this section we study the dynamics of Eq. (11) for growing. This entails that $v'(r)>0$.

The system no longer has an obligatory final time, and the saturation level of nodes is now time-dependent. For example, a user that has degree $V(r)$ is saturated temporarily until a new video "arrives"—i.e., until time $r+\Delta r$ so that $V(r+\Delta r)-V(r)>0$ and there is a new video to rate.

In Fig. 14, we show a numerical solution to Eq. (11) where $v(r)$ and $u(r)$ are linear functions of r. In this simulation, $q=0.8$. Instead of the horizontal line of fully connected nodes along $k=100$ in Fig. 13, the saturation of the nodes follows the growth of $U(r)$. In the inset of Fig. 14, we show the time profile of $N(r,0)$. Initially, it has what appears to be an exponential decrease before it starts to grow slowly as the catalog size increases, in contrast to what we observed in Fig. 11. The early rapid decay is caused by the dearth of high-degree nodes, so nodes with lower degrees receive edges. As r increases, the better-connected nodes receive more edges (because for $q=0.8$, the dominant mechanism is linear preferential attachment), and the population of nodes with fewer edges increases slowly. In Sec. V, we discuss how the Netflix data display some of these features.

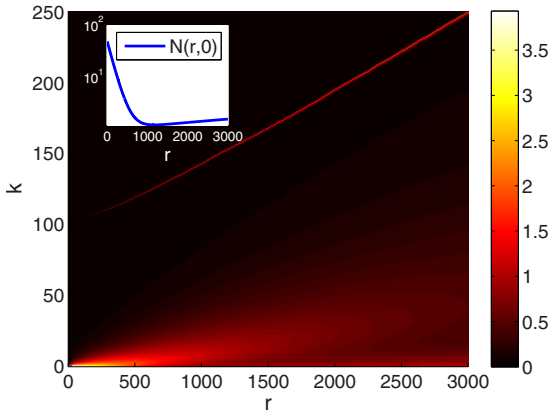

FIG. 14. (Color online) Numerical solution of $N(r,k)$ for video nodes from Eq. (11) with $q=0.8$, $v(r)=30+0.007r$, and $U=100+0.05r$. [We again plot $\ln(N(r,k)+1)$.] The increasing diagonal line indicates $U(r)$ and represents the temporarily saturated nodes. In the inset, we show a plot of $N(r,0)$ on a semilogarithmic scale. We observe a rapid initial decrease followed by a slower increase as the catalog grows.

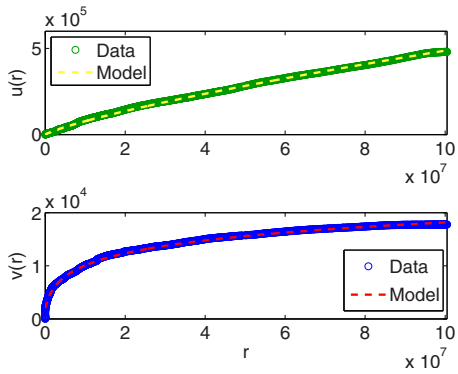

FIG. 15. (Color online) Users (top) and videos (bottom) as a function of ratings. We use circles to show the data from Netflix and dashed curves to show the predictions from Eqs. (15) and (16). We use the parameter values obtained in Sec. II.

V. NETFLIX AS A CATALOG NETWORK

We now investigate how well our catalog model captures the human dynamics revealed by the Netflix data. To do this, we sample the data set while keeping in mind the following considerations:

- Because of the way we have defined our catalog network growth model, we must consider the evolution of the Netflix data in rating time, in which every new rating (which adds an edge to the network) constitutes a time step.
- Although there might be a (physical) time difference between a node (either user or video) joining Netflix and the node receiving its first edge, this information is not included in the data. Many videos receive more than one rating on their first day, so their entry to the network is reflected by increases in the value of $N(r,k)$ for several values of k. We will have to take this into account when comparing our model to the data.

A. Growth and dynamics

To compare our results to the data, we express the growth of the numbers of videos and users as a function of rating time r. Solving for t in Eq. (1) gives

$$t=\frac{1}{b_r}\ln\left(\frac{r}{a_r}+1\right). \tag{14}$$

We substitute Eq. (14) into Eq. (2) to obtain an expression for the number of users as a function of the number of ratings:

$$u(r)=a_u\left[\left(\frac{r}{a_r}+1\right)^{b_u/b_r}-1\right]. \tag{15}$$

We follow the same procedure for the videos to obtain

$$v(r)=a_v+b_v\left\{\frac{1}{b_r}\ln\left(\frac{r}{a_r}+1\right)\right\}^{c_v}. \tag{16}$$

Because $v(r)$ and $u(r)$ are not always integers, we define $U(r)=\lfloor u(r)\rfloor$ and $V(r)=\lfloor v(r)\rfloor$ as the (nonnegative integer) numbers of user and video nodes, respectively.

In Fig. 15, we show the numbers of users and videos versus the number of ratings in the network. Observe that the predictions from Eqs. (15) and (16) agree very well with the data.

Figure 16 shows the time-dependent degree distribution of videos in the Netflix data set for the year 2000 (from 1 January through 30 December). The sample in the plot consists of 365 measurements (one for each day) of r and $N(r,k)$. The highest degree in this sample is 4794; this is well below the theoretical maximum of 9289 that one obtains using the expression for $u(r)$ in Eq. (15), so the network is not experiencing node saturation. We can rewrite the probability that a video node receives an edge as

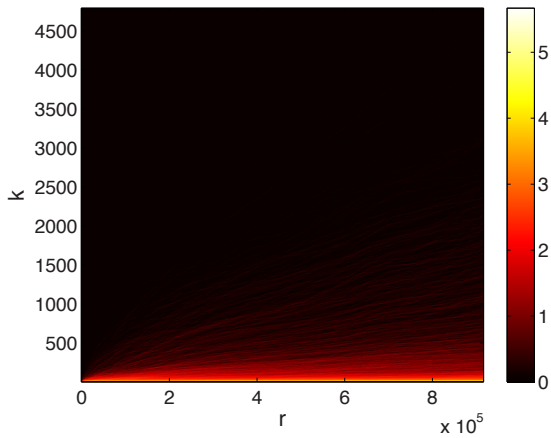

FIG. 16. (Color online) Video degree distribution $N_{\text{data}}(r,k)$ in the Netflix data set in 2000. [We again plot $\ln(N_{\text{data}}(r,k)+1)$.] We show data for videos with degrees ranging from 1 to 4794.

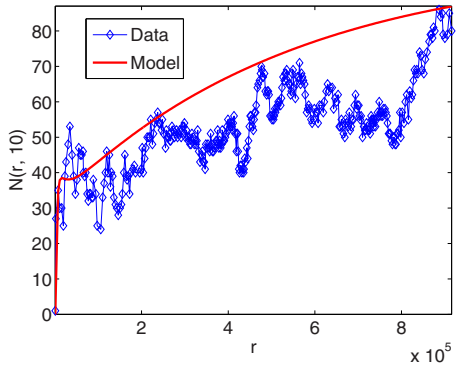

FIG. 17. (Color online) Values of $N(r,10)$ (videos with degree 10) obtained by solving Eq. (17) using $q=0.9795$ (red curve) and the data from Netflix that we report in Fig. 16 (blue dots).

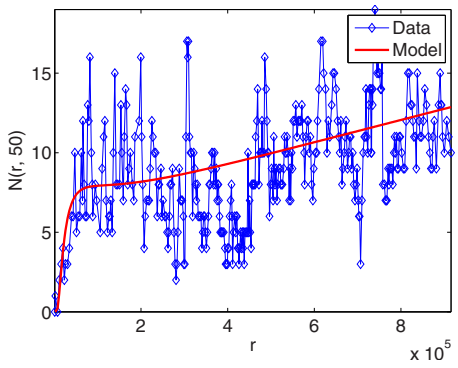

FIG. 18. (Color online) Values of $N(r,50)$ (videos with degree 50) obtained by solving Eq. (17) using $q=0.9795$ (red curve) and the data from Netflix that we report in Fig. 16 (blue dots).

$$P_{\mathcal{V}}(r,k_i)=\frac{1-q}{V(r)}+\frac{qk_i}{\|D_{\mathcal{V}}(r)\|_1}.$$

The rate equation for the evolution of the degree distribution is

$$\frac{dN(r,1)}{dr}=\delta_1 v'(r)-P_{\mathcal{V}}(r,1)N(r,1),$$

$$\frac{dN(r,k)}{dr}=\delta_k v'(r)+P_{\mathcal{V}}(r,k-1)N(r,k-1)-P_{\mathcal{V}}(r,k)N(r,k),\quad k>1. \tag{17}$$

The initial conditions are $N(0,1)=v(0)$ and $N(0,k)=0$ for $k>1$. As noted earlier, the lowest degree a node can have in the data is 1, and the entry degree of the nodes can have any value of k. We denote by δ_k the proportion of new nodes whose entry degree is k, and we note that $\Sigma_k\delta_k=1$. We investigated how many ratings the videos receive on the day they entered the system and found that over 97% of the new nodes receive three or fewer ratings. We found that $\delta_1\approx 0.8$, $\delta_2\approx 0.15$, and $\delta_3\approx 0.05$.

To see how well our model describes the Netflix video data in the year 2000, we define $N_k(q)$ as the 4794×365 matrix obtained solving the system in Eq. (17) and define N_{data} as the matrix of the same size that contains the values of $N(r,k)$ from the data sample. These two matrices contain the values of $N(r,k)$ from the sample and from the equations for all values of k and r. The matrices are of the given size because we sample the degree distribution once per day, and the maximum degree that we observe is 4794. We define the error function

$$E(q)=\|N_k(q)-N_{\text{data}}\|, \tag{18}$$

where $\|\cdot\|$ is the Euclidean matrix norm. To find the optimum value q^*, we minimize $E(q)$ using the Nelder–Mead derivative-free simplex method.[45] We found that the value of q that minimizes Eq. (18) is $q^*\approx 0.9795$, so about 98% of the decisions to rate a video by users are guided by its popularity (i.e., preferential attachment).

In Figs. 17 and 18, we compare the values of $N(r,k)$ that we obtained in our model to those in the data. Despite the noise in the data, our model is able to reproduce the temporal dynamics of $N(r,k)$.

In Fig. 19, we show the approximation of our model to the cumulative degree distribution of the videos on the last day of the sample (i.e., for all values of k and $r=915\,628$, the number of ratings at the end of year 2000), which agrees very well with the data. Although $q^*\approx 0.9795$ suggests that the way the users choose to rate videos is dominated by the popularity of the films, we should stress that the model that we have developed is a very simple one. There are many other processes influencing the decisions of the users, including different external (to the user) factors, such as advertisements, press, and the underlying social network in which the users are embedded.

VI. CONCLUSIONS

We have analyzed a large network of video ratings given by the users of the Netflix video rental service. We studied the system using a bipartite network of videos and users and employed this perspective to reveal interesting features in the dynamics of video rating, such as weekly patterns in video ratings and bursts of activity followed by long idle periods. We calculated clustering coefficients for one-day snapshots, concluding that their low values arise from the presence of high-degree nodes (i.e., videos with a large number of ratings

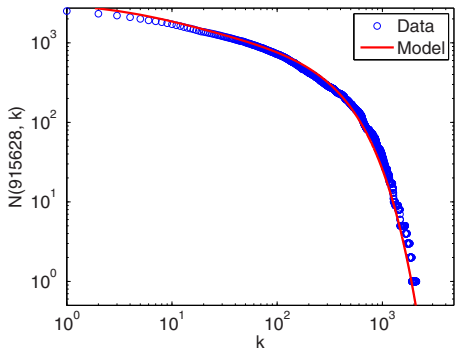

FIG. 19. (Color online) Cumulative degree distribution of video nodes on the last day (915 628 ratings) of the sample from year 2000. We obtained the distribution by solving Eq. (17) using $q=0.9795$ (red curve) and directly from the data (blue dots).

and users who rate many videos). We also showed that the degree distributions of both the user and video nodes resemble power laws with exponential cutoffs.

Motivated by the structural and dynamical features that we observed in the Netflix data, we formulated a mechanism of network evolution in the form of "catalog networks" for bipartite systems. Such networks are initially empty (aside from a seed), and edges are created between two types of nodes based on some predefined rules. New nodes can also be added to the network during the wiring process. In our model, we considered a combination of uniform random attachment and linear preferential attachment. We derived a set of coupled ordinary differential equations that describe the time evolution of the degree distributions of such catalog networks. Presupposing this mechanism and employing the Netflix data, we found that users seem to choose videos according to preferential attachment about 98% of the time and uniform attachment about 2% of the time. This suggests that the number of ratings for a given video is driven almost completely by its popularity (preferential attachment) and only in very small measure by the intrinsic preferences of users. While interesting, the extreme dominance of a preferential-attachment mechanism might be due in part to the simplicity of our model and the absence of information about the underlying social network of the users, which can have considerable influence over the video choices. Additionally, our model does not incorporate external influences such as media coverage and promotion campaigns that can certainly affect the popularity of videos. One can refine our model by considering more sophisticated attachment mechanisms that incorporate the actual scores of the video ratings (not just their existence), the age of the videos, user social networks (see Refs. 46 and 47 for recent interesting study), interactions among users, media presence of videos, and more. Our simple catalog model thereby serves as a good starting point for an abundance of interesting generalizations.

The Netflix data, which are both large and publicly available, provide an excellent vehicle to study many of the features that have been observed in network representations of systems in which agents exercise preferences or choices, such as citation, collaboration, and social networks.[4,10,36,37,48,49] In this paper, we formulated a catalog model to gain insights into the human dynamics of video rating. In our view, catalog models are suitable in many other contexts, including the study of certain electoral systems (such as preferential voting),[50] professional sports drafts,[51] and retail shopping. To achieve insights in such a diverse array of settings, the catalog model presented herein can be generalized in numerous interesting ways to incorporate external agents, underlying networks or cliques of individuals, and more.

ACKNOWLEDGMENTS

We thank M. Barahona, R. Desikan, T. Evans, P. Ingram, N. Jones, R. Lambiotte, S. Lanning, D. Lazer, P. Mucha, D. Plato, S. Saavedra, D. Smith, and J. Stark for useful comments and suggestions. We also acknowledge Netflix Inc. for providing the data, which were released publicly as part of their prize competition. This work was done in part as a dissertation for the MSc in Mathematical Modeling and Scientific Computing at the University of Oxford. M.B.D. was supported by a Chevening Scholarship and a BBSRC-Microsoft Dorothy Hodgkin Postgraduate Award. M.A.P. acknowledges a research award (Grant No. 220020177) from the James S. McDonnell Foundation. J.-P.O. is supported by the Fulbright Program.

[1] A.-L. Barabási, Nature (London) **435**, 207 (2005).
[2] T. S. Evans and A. D. K. Plato, Networks Heterog. Media **3**, 221 (2008).
[3] J.-P. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, D. Lazer, K. Kaski, J. Kertész, and A.-L. Barabási, Proc. Natl. Acad. Sci. U.S.A. **104**, 7332 (2007).
[4] R. Albert and A.-L. Barabási, Rev. Mod. Phys. **74**, 47 (2002).
[5] M. E. J. Newman, SIAM Rev. **45**, 167 (2003).
[6] G. Caldarelli, *Scale-Free Networks* (Oxford University Press, Oxford, UK, 2007).
[7] *Handbook of Graph Theory*, edited by J. L. Gross and J. Yellen (CRC, Boca Raton, FL, 2004).
[8] A.-L. Barabási and R. Albert, Science **286**, 509 (1999).
[9] H. A. Simon, Biometrika **42**, 425 (1955).
[10] D. J. de Solla Price, Science **149**, 510 (1965).
[11] P. L. Krapivsky, S. Redner, and F. Leyvraz, Phys. Rev. Lett. **85**, 4629 (2000).
[12] M. E. J. Newman, Proc. Natl. Acad. Sci. U.S.A. **98**, 404 (2001).
[13] M. Latapy, C. Magnien, and N. D. Vecchio, Soc. Networks **30**, 31 (2008).
[14] M. A. Porter, P. J. Mucha, M. E. J. Newman, and C. M. Warmbrand, Proc. Natl. Acad. Sci. U.S.A. **102**, 7057 (2005).
[15] S. Saavedra, F. Reed-Tsochas, and B. Uzzi, Nature (London) **457**, 463 (2009).
[16] P. Zhang, J. Wang, X. Li, M. Li, Z. Di, and Y. Fan, Physica A **387**, 6869 (2008).
[17] J.-L. Guillaume and M. Latapy, Physica A Stat. Theor. Phys. **371**, 795 (2006).
[18] J. Ohkubo, K. Tanaka, and T. Horiguchi, Phys. Rev. E **72**, 036120 (2005).
[19] T. S. Evans and A. D. K. Plato, Phys. Rev. E **75**, 056101 (2007).
[20] S. Saavedra, S. Powers, T. McCotter, M. A. Porter, and P. J. Mucha, Physica A **389**, 1131 (2010).
[21] G. F. Davis, M. Yoo, and W. E. Baker, Strat. Org. **1**, 301 (2003).
[22] D. J. Watts and S. H. Strogatz, Nature (London) **393**, 440 (1998).
[23] S. N. Dorogovtsev, J. F. F. Mendes, and A. N. Samukhin, Nucl. Phys. B **666**, 396 (2003).
[24] K. Park, Y.-C. Lai, and N. Ye, Phys. Rev. E **72**, 026131 (2005).
[25] H. Fan, Z. Wang, T. Ohnishi, H. Saito, and K. Aihara, Phys. Rev. E **78**, 026103 (2008).
[26] J. Lindquist, J. Ma, P. van den Driessche, and F. H. Willeboordse, Physica D **238**, 370 (2009).
[27] G. Polya, Ann. Inst. Henri Poincare **1**, 117 (1931).
[28] W. Feller, *An Introduction to Probability Theory and Its Applications* (Wiley, New York, NY, 1957), Vol. 1.
[29] C. Godrèche and J. M. Luck, J. Phys.: Condens. Matter **14**, 1601 (2002).
[30] D. Stauffer, X. Castelló, V. M. Eguíluz, and M. S. Miguel, Physica A **374**, 835 (2007).
[31] R. A. Bentley and S. J. Shennan, Am. Antiq. **68**, 459 (2003).
[32] Netflix Prize, http://www.netflixprize.com.
[33] J. Lorenz, Eur. Phys. J. B **71**, 251 (2009).
[34] J. Bennett and S. Lanning, Proceedings of KDD Cup and Workshop 2007, 2007.
[35] M. C. González, C. A. Hidalgo, and A.-L. Barabási, Nature (London) **453**, 779 (2008).
[36] D. Sornette, F. Deschâtres, T. Gilbert, and Y. Ageon, Phys. Rev. Lett. **93**, 228701 (2004).
[37] J. G. Oliveira and A.-L. Barabási, Nature (London) **437**, 1251 (2005).
[38] A. Clauset, C. R. Shalizi, and M. E. J. Newman, SIAM Rev. **51**, 661 (2009).
[39] P. G. Lind, M. C. González, and H. J. Herrmann, Phys. Rev. E **72**, 056127 (2005).
[40] J.-P. Onnela, J. Saramäki, J. Kertész, and K. Kaski, Phys. Rev. E **71**, 065103 (2005).
[41] A.-L. Barabási, *Bursts: The Hidden Pattern Behind Everything We Do* (Dutton, New York, 2010).

[42]R. Pastor-Satorras and A. Vespignani, Phys. Rev. Lett. **86**, 3200 (2001).
[43]A. Barrat and R. Pastor-Satorras, Phys. Rev. E **71**, 036127 (2005).
[44]S. Johnson, J. J. Torres, and J. Marro, Phys. Rev. E **79**, 050104 (2009).
[45]J. C. Lagarias, J. A. Reeds, M. H. Wright, and P. E. Wright, SIAM J. Optim. **9**, 112 (1998).
[46]S. Asur and B. A. Huberman, "Predicting the future with social media," e-print arXiv:cond-mat/1003.5699, 2010.
[47]J. Ratkiewicz, F. Menczer, S. Fortunato, A. Flammini, and A. Vespignani, "Characterizing and modeling the dynamics of online popularity," e-print arXiv:cond-mat/1005.2704v1, 2010.
[48]M. J. Salganik, P. S. Dodds, and D. J. Watts, Science **311**, 854 (2006).
[49]R. Lambiotte and M. Ausloos, Phys. Rev. E **72**, 066107 (2005).
[50]australianpolitics.com, History and features of the Australian electoral system, 2008, Consulted on 10 February 2009.
[51]A. Summers, T. Swartz, and R. Lockhart, *Statistical Thinking in Sports* (Chapman and Hall, London/CRC, Boca Raton, FL, 2007), Chap. 15, pp. 249–262.