http://pubsonline.informs.org/journal/mnsc

MANAGEMENT SCIENCE
Articles in Advance, pp. 1–25
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

The Similarity Network of Motion Pictures

Yanhao Max Wei[a]

[a] Marshall School of Business, University of Southern California, Los Angeles, California 90089
Contact: yanhaowe@usc.edu, https://orcid.org/0000-0001-6613-5993 (YMW)

Received: August 19, 2016
Revised: November 15, 2017; July 26, 2018
Accepted: October 30, 2018
Published Online in Articles in Advance: October 16, 2019

https://doi.org/10.1287/mnsc.2018.3261

Abstract. Ideas are connected. New ideas are often seen as creative combinations of previous ideas. I study these connections in the context of motion pictures. A network of 4,445 movies is constructed to indicate which movies are similar. I first examine the properties of the network using descriptive and regression analysis; then I develop a model of network formation for counterfactual analysis. It is found that most movies imitate and evolve around a "core" of the more successful movies. In addition, imitation is both conventional and atypical: a new movie usually follows a stream of similar movies yet simultaneously combines atypical elements from movies outside this stream. This atypicality, if well balanced, has a positive effect on the individual movie's box office. However, I find that, in the long run, atypical combination may lead to a worse collective box-office performance because of the way it changes the market structure.

History: Accepted by Matthew Shum, marketing.
Supplemental Material: Data files are available at https://doi.org/10.1287/mnsc.2018.3261.

Keywords: market structure • product similarity • networks • creativity • firm learning • movies

1. Introduction

Ideas are connected. Many models of creativity see new ideas as innovative combinations of existing ideas. This concept was discussed as early as in Schumpeter's theory of economic development, which defines the essence of entrepreneurship as "the carrying out of new combinations." Combinatorial creativity has been studied across fields, including psychology, information theory, economics, sociology, and organizational theory.[1] Earlier studies focused on understanding the implications of the combinative process. More recent studies have moved to the question of how to make this process more effective. For example, to produce the most fruitful papers in science, should a researcher stay in her comfort zone or reach out for unconventional ideas?

Although the concept of combinatorial creativity has generated great insights in various fields, it sees little application in studies of markets. Products, just like ideas and scientific papers, are not isolated but connected. The goal of this paper is to show how the application of this simple concept can uncover patterns and insights that have remained unknown to us with traditional market analyses. Specifically, I study a large network that represents the similarity pattern between motion pictures.

The network can be conceptualized using a standard diagram of market analysis. Given a market, a product is often represented as a point in the characteristic space (whose dimensions correspond to product characteristics). Two similar products are placed close to each other, and when a new product enters at a certain location, we can think of it as imitating the nearby existing products. A more compact yet still informative way to present these relations is a network; one uses nodes to represent products and adds a link between two nodes if they are similar. As I demonstrate throughout the paper, the network representation has several important advantages. First, from a data point of view, constructing a network is often more feasible than a full characteristic space. One example is the citation network: although it is hard to quantify all the characteristics of a paper, it is simple to identify the citation from one paper to another. Second, networks do not suffer from the curse of dimensionality, which is particularly useful when one deals with complex products that cannot be easily characterized using a few dimensions, such as movies. Third, there have been many recent exciting developments in network science that allow us to scale up the analysis to millions of product pairs. Most of these tools have yet to see applications in market analyses.[2]

I chose the movie industry as the empirical setting partly because of the seemingly conflicting nature of this industry. Although being one of the so-called creative industries, it seems to heavily rely on imitation (probably because of the great amount of uncertainty associated with investing in movies).[3] This makes the topic of combinatorial creativity particularly interesting and relevant. One may ask: is there evidence of imitation? To what extent do prior similar movies predict and reduce the uncertainty in the return on investment (ROI)? What types of combinations are more

likely to create box-office hits? How does imitation shape the landscape of the industry?[4]

Another reason for the choice of the movie industry is data availability. Thanks to high public interest, information on budget, cast, and box-office revenue has been kept track of for most movies. Apart from this information, my study also requires the data to identify which movies are similar. Although the most straightforward approach here is to measure the distances between movies in the characteristic space, many defining characteristics of a movie are difficult to quantify (e.g., narration, acting, camera work, music, special effects, ideology).[5] In addition, it is not simple to decide a priori which characteristics should be given relatively more weight in calculating the distance. An alternative approach here is using the revealed preferences of consumers, based on the idea that, if two movies are similar, they should get similar receptions from the same consumer. This indirect approach circumvents the difficulties of dealing with a full characteristic space. It has been widely used in recommender systems (most notably Amazon's) and has seen much development in computer science (Linden et al. 2003, Desrosiers and Karypis 2011). Following their practices, in Section 2 of the paper, I construct a similarity network among 4,445 movies from a data set of 20 million movie ratings by 138,000 individuals (known as the MovieLens data set). An illustration of a segment of the network is given in Figure 1.

Section 3 conducts model-free analyses on the network. I start with the macro-level structural properties. The network is found to have a clump or "core" of densely linked nodes, surrounded by a "periphery" of nodes that are directly or indirectly connected to the core. Interestingly, the movies in the core tend to have relatively high ROIs. This depicts an industry in which most of the products evolve around a set of successful stereotypes. Indeed, it is very hard to divide the network into "communities" that are clearly distinctive from each other even with the latest network algorithms.

An important question about combinatorial creation is whether the combination happens more out of randomness or follows certain patterns. To this end, I move analyses to a more microlevel and examine a class of local networks (subsets of the similarity network). For each movie j, the local network displays the similarity relations between the movies that j imitates. I find that such a local network usually features one cluster of connected nodes together with several isolated nodes. This indicates that a new movie usually follows an established stream (cluster) of movies yet simultaneously combines "atypical" elements from movies outside this stream. More importantly, this atypicality positively impacts a movie's ROI. Most interestingly, the impact exists only when the amount of atypicality is well balanced: too little or too much has no significant impact.

The impact of atypical combinations should go beyond individual movies; in the longer term, it can change the structure of the network. To examine this dynamic effect requires a model of network formation. I develop such a model in Section 4, which formulates both the births of candidate movies and the studios' decisions to accept/reject candidates based on the performances of past similar movies. The model is estimated and then used for counterfactual analysis (Sections 5 and 6). I find that a lower level of atypicality leads to a network of less stereotypical and more diverse movies. This diversity turns out to allow studios to achieve higher ROIs over time. So, in the long run and at the aggregate level, the impact of atypical combination can be negative. I compare my result to the relevant findings on citation networks and social networks.

Finally, in Section 7, I discuss how the network approach developed in this paper can be applied to other market contexts.

2. Data

2.1. Movie Characteristics

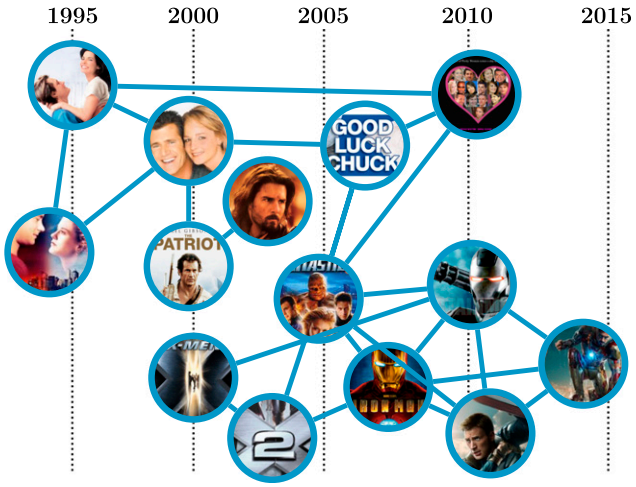
I collect from the Internet Movie Database (IMDb.com) information on a movie's title, language, region, genre, MPAA rating, production companies, release date, *production start date*, production budget, writers, directors, leading actors, and domestic box-office revenue. In the case in which the budget number is missing on IMDb, I try to collect it from Wikipedia.com. The box-office revenues are cross-checked with the numbers on Box Office Mojo (a box-office revenue tracking website).

I am able to collect complete data on 4,445 movies that were released in the United States from 1975 to 2014 (included) except for the production start date, which is missing for about one third of the movies. To overcome this problem, I use a nonparametric regression of the production time (i.e., the interval between the start date and release date) on production budget and impute the start date for each movie. On average, a movie takes slightly more than a year to produce. In addition, I exclude movies with a budget less than $1 million (in 2014 dollars); the mechanism behind the production and distribution of these "microbudget" movies is often different from that of the larger movies.

I focus my analyses on movies that started production in 1995–2012 (the release dates of these movies extend to 2014). This amounts to a sample of 3,079 movies or an average of 171 movies per year. This sample size per year is largely in line with previous studies (e.g., Einav 2007, 2010; and Goettler and Leslie 2005).[6] The data before 1995 is not discarded but used as the initial condition in some of my analyses.

To adjust for inflation over time, all budget and box-office numbers are normalized to be in 2014

Figure 1. Illustration of the Similarity Network



Notes. Movies are ordered from left to right by year of release. They are *Sleepless in Seattle* (1993), *While You Were Sleeping* (1995), *What Women Want* (2000), *The Patriot* (2000), *X Men* (2000), *X2* (2003), *The Last Samurai* (2003), *Fantastic 4* (2005), *Good Luck Chuck* (2007), *Iron Man* (2008), *Valentine's Day* (2010), *Iron Man 2* (2010), *Captain America* (2011), and *Iron Man 3* (2013).

dollars using the consumer price index. Yearly box-office ticket price, collected from The-Numbers.com, is used in some of the analyses to adjust for the fluctuation of ticket price over time.

2.2. The Similarity Network

As explained earlier, in this paper, I uncover the similarity relations between movies not by directly measuring their distances in the characteristic space, but using consumer-revealed preferences. The similarity network is constructed primarily from the MovieLens data. The data set includes 20 million movie ratings by 138,000 individuals (updated in October 2016) and is made available by the GroupLens laboratory at the University of Minnesota. To supplement the MovieLens data, I also make use of the movie recommendation data on IMDb and Amazon. Specifically, I anonymously scraped the recommendations under "People Who Liked This Also Liked" (IMDb) and "Consumers Who Watched This Also Watched" (Amazon).

At a conceptual level, the basic idea behind using individual ratings to construct a similarity network is that if two movies are similar, they should be rated similarly by the same person. This approach to measure similarity is at the core of many recommender systems, including those of IMDb and Amazon, and has seen much development in computer science (Linden et al. 2003, Desrosiers and Karypis 2011). In implementation, a similarity score between two items is often calculated as the correlation between the ratings given by the individuals who rated both items, that is, common raters. Two movies are considered similar if the correlation exceeds a predefined threshold.[7] In essence, the idea echoes a stream of works in marketing that uses consumer-panel data to uncover product positions in characteristic space (Chintagunta 1994, Elrod and Keane 1995, Goettler and Shachar 2001).[8]

In principle, either MovieLens or the IMDb/Amazon data can be used to construct the similarity network. The upside of the IMDb/Amazon data is that the websites are able to calculate similarity scores to high precision using their huge database. The downside, however, is that the exact details of their algorithms are unknown to the public. This is particularly problematic if websites have incentives to bias recommendations toward more recent or popular products. On the contrary, with the MovieLens data, I am able to directly calculate similarity scores, but the smaller data size means that, for some movie pairs, there are not enough common raters for a precise similarity score.

Given these considerations, when I define a link in my network, I primarily rely on the correlation in the MovieLens data and simultaneously take into account (i) the precision of this correlation, (ii) whether there is a recommendation on Amazon or IMDb, and (iii) to what extent the recommendation reflects similarity versus other factors, such as the ages and popularities of the movie pair. Technically, this is handled by a latent factor model; interested readers are here referred to the appendix. Figure 1 illustrates the links between some example movies.

As a preliminary check on the constructed network, Table 1 displays a logit regression across movie pairs in which the dependent variable is a dummy indicating whether the pair is linked in the constructed network. All the coefficients have expected signs and are statistically significant. For example, having a common member in the crew, whether it is an actor, director, or writer, is significantly associated with linkages. On top

Table 1. Logit Regression Predicting Links

	Coefficient (standard error)
Intercept	−3.489 (0.011)
Sharing leading actor(s)	1.770 (0.034)
Sharing director	1.535 (0.057)
Sharing writer(s)	1.608 (0.057)
Significant overlap of genres	1.173 (0.001)
Same MPAA rating	0.437 (0.001)
Same production company	0.556 (0.001)
Difference in log budget	−0.797 (0.007)
Difference in release time (in year)	−0.212 (0.002)
Sequel/prequel	4.479 (0.287)
Pseudo-R^2	0.13
N	4.74×10^6

Notes. Each observation is a movie pair; the dependent variable is whether the pair is linked in my constructed network. Pseudo-R^2 equals one minus the ratio between fitted deviance and null deviance.

of this, being a sequel–prequel pair almost guarantees a link. Notice that the R^2 is fairly low. This is expected as much of the similarity is likely attributed to the characteristics not in the data, for example, plot, narration, ideology, and visual and sound effects.

Several variants of the constructed network are considered as robustness checks on the main results of the paper. One variant is constructed by explicitly incorporating the pair characteristics in Table 1. The details are given in the appendix.

3. Descriptive and Regression Analysis

This section presents some model-free results. These results are not only interesting in their own right, but also motivate some of the model specifications in Section 4. I start with patterns at the network level, then gradually "zoom" the analysis in toward individual movies.

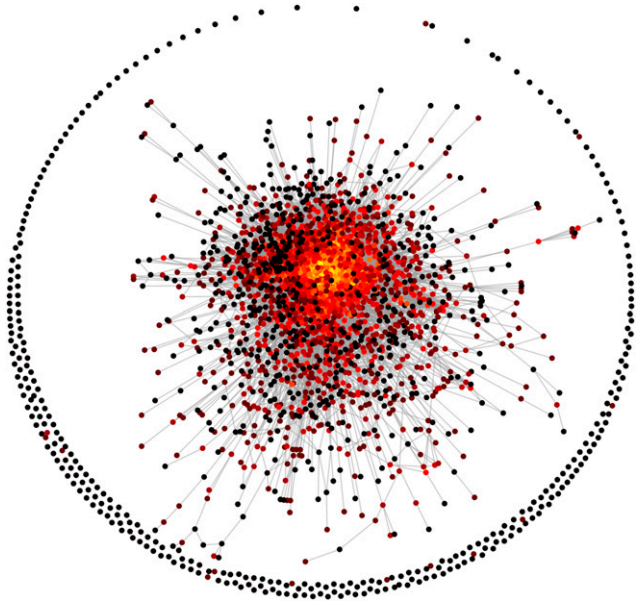
3.1. Core-Peripheral Structure and Imitation

Figure 2 visualizes the similarity network. The graph is plotted by force-directed placement, a popular method introduced by Fruchterman and Reingold (1991). In a nutshell, it attempts to place linked nodes close to each other and, in doing so, helps reveal the basic structure of the network. For example, if it were a citation network between scientific papers, the visualization would show distinctive clusters, each corresponding to a discipline or field. As we can see, this is clearly not the case for Figure 2; there is a single dense core in the center of the graph with most of the movies connected to the core, directly or indirectly.

To help understand what the core represents, I color the nodes by number of imitators; an imitator of a movie j is any of j's neighbors in the network that started production after j's release. A brighter color indicates a larger number of imitators. It is clear that the core consists of the movies that have been imitated more times.

An immediate question is why some movies have more imitators than the others. To answer this, Table 2 groups the movies by number of imitators and displays the ROI distribution within each group. There is a clear positive association between the ROI and the number of imitators. This suggests that one reason why a movie has had more imitators is because it was financially successful. Although intuitive, this provides direct empirical evidence of herding in the motion picture industry: most movies in the industry have been evolving around a core consisting of the higher-ROI types.

Figure 2. The Similarity Network and Its Core



Notes. Visualization of the network among the movies in/after 1995. The graph is drawn by force-directed placement (Fruchterman and Reingold 1991), which tries to place connected nodes closer to each other. The outer ring collects the singletons and small isolated groups of nodes. The coloring of a node indicates the number of imitators of that node. A brighter color indicates a larger number of imitators.

A potential issue with Table 2 is that the more recent movies tend to have an artificially smaller number of imitators because they are close to the end point of my data. To account for this, Table 3 restricts the attention to the movies in the earlier years. The result still holds.

3.2. Community Structure and Diversity

Movies have been historically classified into a number of genres: drama, comedy, romance, horror, war, etc. So it is reasonable to expect that the similarity network has a clear community structure, meaning that movies can be divided into groups, and the links are dense within groups but sparse between groups. However, as we have seen, Figure 2 suggests the opposite. At some level, this is not too surprising because many movies are cross-genre and studios like

Table 2. Log ROIs by Number of Imitators, 1995–2012

Number of imitators	Group count	Mean	First quantile	Third quantile
0	881	−0.59	−1.44	0.42
1–2	721	−0.53	−1.24	0.45
3–9	726	−0.1	−0.72	0.67
≥10	751	0.08	−0.41	0.61

Notes. Included are the movies in/after 1995. The cutoffs in the first column are chosen to make the four groups of as equal size as possible.

Table 3. Log ROIs by Number of Imitators, 1995–1999

Number of imitators	Group count	Mean	First quantile	Third quantile
0–1	192	−0.55	−1.3	0.37
1–4	122	−0.49	−1.21	0.46
5–19	141	−0.16	−0.72	0.51
≥20	152	−0.02	−0.55	0.54

Notes. Included are the movies in 1995–1999. The cutoffs in the first column are chosen to make the four groups of as equal size as possible.

to make diverse uses of elements from a successful movie regardless of to which genre it belongs. For example, there is a high similarity score between the action movie *Fantastic 4* (2005) and the romantic comedy movie *Good Luck Chuck* (2007), probably thanks to Jessica Alba who starred in both movies.

To see the exact extent to which the network has a community structure, one can use sophisticated network algorithms. Table 4 displays the result from modularity maximization, a widely used algorithm to detect communities in networks (Newman 2006, 2010). It is able to identify, for example, a group of family-type movies that are rarely R rated and often feature animations. It also identifies horror movies. However, overall, the division is coarse with the three largest communities making up more than half of the network. In addition, the boundaries between communities are very much blurred. This is seen in Figure 3, which colors the nodes by community membership.

To summarize, the network depicts a market characterized more by convergence toward a core of high-return movies than diversity of distinctive movie types.

3.3. Clustering and Similarity

An important property of similarity is transitivity; if movie j is similar to k, and k is similar to ℓ, it should often be the case that j is similar to ℓ. The property corresponds to a widely used statistic in networks called the clustering coefficient (Newman 2010), which measures the presence of triangles (i.e., a set of three nodes linked to each other). Roughly speaking, it is the probability of getting a triangle when we randomly pick three nodes from the network. The clustering coefficient of my similarity network is 0.228, which is substantial. To compare, most random network models have a clustering coefficient of zero; Facebook has a clustering coefficient around 0.3. Clustering is an important property to be captured later by the model in Section 4.

3.4. Atypical Combination

Now I move the analysis from macrolevel structures of the whole network to an important type of local networks. For any given movie j, there is a local network consisting of the movies that j imitates (i.e., j's prior similar movies). The local network is of particular interest because it helps answer the following question: does combinatorial creation usually happen out of randomness or follow a certain pattern?

Figure 4 provides some representative examples of these local networks. One can readily see an interesting pattern: many prior similar movies are directly or indirectly connected to each other, forming a cluster. Intuitively, the cluster represents a stream of similar movies on which j is based. For instance, if one looks at the prior similar movies of *Red Riding Hood* (2011), a mystery fantasy, one sees a cluster consisting of previous mystery fantasies with similar themes, such as *Season of the Witch* (2011) and *The Sorcerer's Apprentice* (2010).

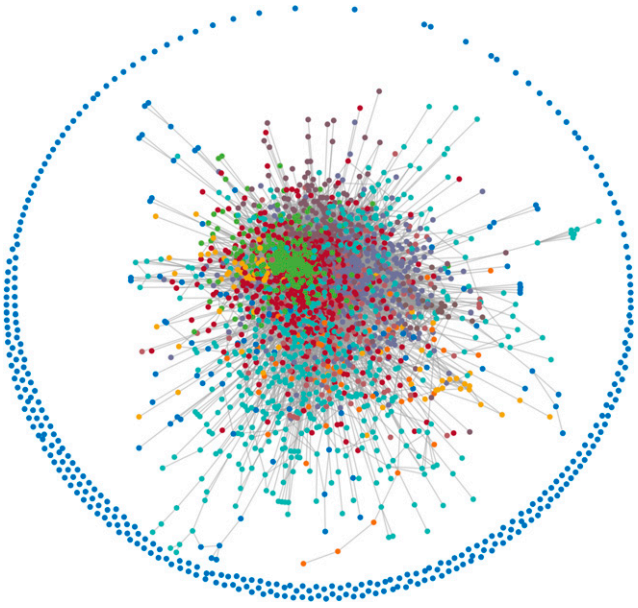
More interestingly, in some examples in Figure 4, we also see singletons disconnected from the cluster. These singleton nodes represent the instances in which movie j borrows elements from the movies

Table 4. Communities in the Network

Size	Major genres	Rated R, %	Median budget	Example
19	Drama, comedy, romance, crime	37	14.5	*Good Deeds* (2012)
50	Drama, comedy, romance, crime	64	21.3	*Small Time Crooks* (2000)
72	Drama, adventure, animation, family	14	51	The Harry Potter Series
265	Drama, action, thriller, crime	66	54.7	*U.S. Marshals* (1998)
274	Horror, thriller, mystery	68	26.8	*The Haunting* (1999)
361	Action, adventure, thriller, crime	27	78.9	X-Men/Wolverine Series
514	Drama, comedy	67	18	*The Campaign* (2012)
516	Comedy, drama	30	31.8	*Jack* (1996)
603	Comedy, drama, romance	13	35.6	*27 Dresses* (2008)

Notes. The communities with a size smaller than 10 are not displayed here. Budgets are normalized to be in millions of 2014 dollars.

Figure 3. The Similarity Network and Its Communities



Note. The same network visualization as Figure 2 with the exception that the coloring tries to distinguish the different network communities (listed in Table 4).

that are atypical of the stream (or cluster) of movies in which j is grounded. An analogy in scientific research is when a scholar reaches outside the scholar's "comfort zone" to borrow from other disciplines. Take *Red Riding Hood* (2011) as the example again; although mostly a mystery fantasy, its twisted love story with the performance of Amanda Seyfried seems intended to copy the actress' success in the romance movie *Chloe* (2009).

Seeing each movie as a creative combination of previous movies, the cluster–singleton pattern allows us to define a measure of atypicality in this combination. For each movie j, let the atypicality be the number of singletons in j's local network divided by the size of the local network (equal to the number of j's prior similar movies). For example, the first local network in Figure 4 has an atypicality of 1/7. If j has only one prior similar movie, there is really no context for atypicality, so I define it to be zero. If j is novel, that is, having no prior similar movie, the atypicality is defined to be zero as well. These definitions are used in the subsequent analyses.

3.5. Returns and Risks

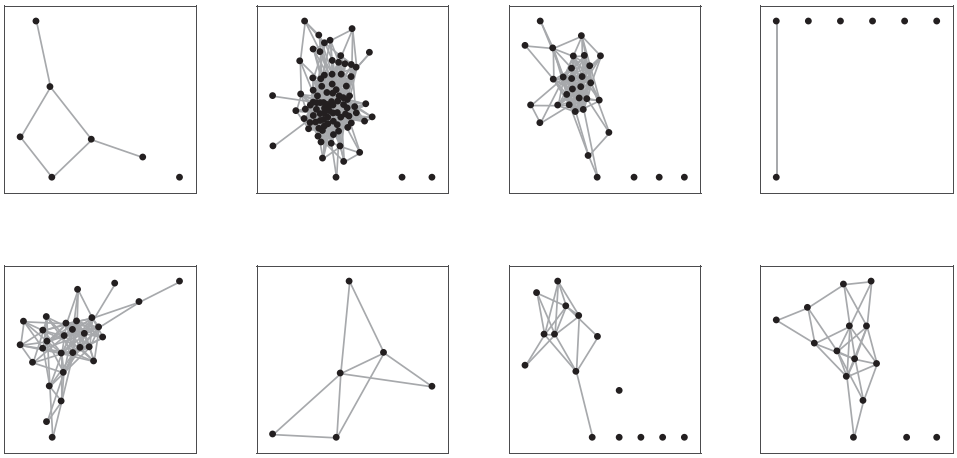
As the last part of the analysis in this section, I examine what determines individual movies' box-office performances, focusing on variables that are available at the time of green-light decisions (e.g., production budget, genre, casting, similarity with previous movies). To this end, Table 5 displays regressions of log ROI (the ratio between the domestic box-office revenue and budget).[9]

Column (1) regresses the log ROI on the observed movie characteristics (with detailed definitions of the regressors given in the table notes). Notice, in particular, that the R^2 of the regression is very low, only 0.088, which is actually not surprising given that movie success has been known to be notoriously difficult to predict (De Vany and Walls 1996, 1999; Squire, 2005).[10]

Column (2) is the same as column (1) except that it additionally controls for the performance of prequels. As expected, the performance of a prequel(s) is highly predictive of the performance of a sequel. The R^2 does not change much, mainly because not many movies are sequels (only about 6% in the data).

Column (3) is the main regression in the table. It is the same as column (2) except that it adds the information from the similarity network, which sets it apart from the regressions in previous movie studies. Specifically, for each movie j, I introduce a new regressor equal to the average log ROI of j's prior similar

Figure 4. Examples of Local Network Among Prior Similar Movies



Notes. Each plot takes one movie and then draws the links between that movie's prior similar movies. For example, the first plot is about a movie that has seven prior similar movies; what the plot shows is the similarity network among these seven movies.

Table 5. Regressions of Log ROI

	(1)	(2)	(3)	(4)	(5)	(6)
Year dummies	Yes	Yes	Yes	Yes	Yes	Yes
Genre dummies	Yes	Yes	Yes	Yes		Yes
Group dummies for number of prior similar movies			Yes	Yes	Yes	Yes
Log budget	0.076**	0.074**	0.014	0.014	0.036	−0.001
Seasonality	0.13**	0.12**	0.098**	0.099**	0.1**	0.1**
Rating as restricted	−0.22***	−0.21***	−0.13**	−0.13**		−0.11**
Star actor	−0.021	−0.018	0.091	0.093*		0.087
Star director	−0.012	−0.026	0.007	0.012		0.005
Star writer	0.11**	0.073	0.04	0.049		0.057
Dummy for being a sequel		−0.001	−0.05			
Mean log ROI of the prequels		0.36***	0.14**			
Log number of subsequent similar movies						0.22***
Atypicality groups:						
0.1–0.2 atypicality	0.15**	0.15**	0.11	0.11*	0.12*	0.23***
0.2–0.3	0.26***	0.26***	0.22***	0.23***	0.21**	0.32***
0.3–0.5	0.09	0.079	0.042	0.041	0.029	0.16***
0.5–0.7	−0.22*	−0.21*	−0.071	−0.069	−0.1	−0.003
≥0.7	−0.37***	−0.36***	−0.13	−0.13	−0.17	−0.051
Average log ROI of prior similar movies:						
1 prior similar movie			0.26***	0.26***	0.28***	0.26***
2–4 prior similar movies			0.65***	0.65***	0.67***	0.63***
5–10 prior similar movies			0.76***	0.77***	0.81***	0.74***
>10 prior similar movies			1.2***	1.3***	1.3***	1.3***
R^2, all observations	0.088	0.094	0.19	0.19	0.18	0.21
R^2, with ≥1 prior similar movies	0.097	0.11	0.23	0.23	0.22	0.25
N	3,079	3,079	3,079	3,079	3,079	3,079

Notes. Dependent variable is the log ROI. ROI is defined as the ratio between domestic box-office revenue and budget, both of which are normalized to be in 2014 dollars. The observations are the movies that started production in/after 1995. "Star actor" is a dummy for movies with at least one leading actor who had previously taken a leading role in a top 10% grossing movie. "Star director" and "star writer" are defined in the same way. "Seasonality" is a dummy for releases in June, July, August, and December. There are 17 genre dummies. The variable "group dummies for number of prior similar movies" uses the same grouping under "average log ROI of prior similar movies."

***$p < 0.01$; **$p < 0.05$; *$p < 0.10$.

movies. Moreover, I group movies by the number of prior similar movies and allow the regressor's coefficient to vary across groups. We see that the coefficient in each group is positive and significant; the coefficient is higher for the group with more prior similar movies. Importantly, the R^2 is significantly higher compared with columns (1) and (2).

The results imply that, (i) conceptually, we can think of the ROI of each prior similar movie as a signal of j's ROI; (ii) the informational value of the signal average increases with the number of the signals; and (iii) more importantly, aggregation of the signals is Bayesian-like, and the marginal value of each additional signal decreases with the number of signals. These implications are instructive for how I specify the model later in Section 4.

As to the prequel's ROI, its coefficient is still significant in column (3) but much smaller than that in column (2). This is expected because, as we have seen in Section 2, prequel/sequel relation is captured by the similarity network.

The coefficients for atypicality in column (3) display an exceptionally interesting pattern: neither a very small nor large amount of atypicality has a significant impact on ROI. It is around a modest value (0.2 ~ 0.3) at which the impact is positive and significant. This modest value represents a movie's balance between staying within its conventional domain and reaching out for atypical elements. This result turns out to nicely echo the finding in Uzzi et al. (2013) in a different context. They find that, in citation networks, the highest impact scientific papers are often grounded in a conventional field yet simultaneously reach outside for atypical knowledge from other fields.

Columns (4) and (5) are robust checks of the regression in column (3). More specifically, column (4) drops the control for prequel performance; the coefficient estimates and the R^2 stay almost unchanged. Column (5) further drops the genre dummies, MPAA rating, and the star dummies—a total of 21 variables; the decrease in R^2 is small, and the coefficients for the remaining variables stay roughly the same. This indicates that the similarity network is, relatively speaking, a powerful source for ROI prediction; once included, the additional predictive power from the observed movie characteristics is only marginal.

Column (6) extends the regression in column (3) by adding a regressor equal to the log number of imitators. (Recall that an imitator of j is any movie similar to j and started after j.) The coefficient is significantly

positive, indicating that the number of a movie's imitators "predicts" the ROI of this movie. The caveat, of course, is that the imitators come only after the ROI is realized, so the regression cannot really be seen as predictive. The purpose of column (6) is to provide correlational evidence for studio selection as in Tables 2 and 3 but in a regressional setting for which many factors can be controlled.

So far, I have focused on expected ROI but have not said anything about risks, particularly, how risks are related to imitation. In these regressions, risks are presented by the uncertainty in the residuals. Table 6 groups the movies by number of prior similar movies and shows, for each group, the distribution of the squared residuals from column (5) of Table 5 (using other columns gives similar results). The distribution shifts quickly toward zero as we move to the group with more prior similar movies. The result shows that the usually enormous risks involved with producing a movie can be greatly reduced by imitation.

To summarize the findings in Tables 5 and 6: (i) prior similar movies are a powerful source for the prediction of a movie's ROI; (ii) atypicality positively impacts a movie's ROI, but not when there is too little or too much of it; and (iii) imitation reduces the risks of movie investments.

4. Model

The purpose of this section is to detail a model for the formation of the similarity network. However, it is helpful to start with a brief discussion on how movies are made in practice. Typically, a movie starts with an idea by a person, usually a producer, a director, a screenwriter, or sometimes even an actor or actress. Whoever this person is, the writer is always the one that translates the idea into a script. The script lays out not only the flow of the story, but also the movement, actions, and expressions of the actors. The actual script writing takes three to six weeks. The cast is usually determined after the script is finished. In the meantime, an artist is often called upon to draw a storyboard, which is a sequence of comic book–like sketches that help illustrate the script. A producer typically presents the whole package (the script, casting, and storyboard) to a studio, which decides whether to finance the movie (i.e., the green-light decision; more details in Section 4.3). If the decision is yes, the actual shooting of the movie starts. Once completed, the films are sent to postproduction for editing, after which the movie becomes ready for marketing and release.

With that said, movie-making is a complex process. My goal here is not to model the details of movie making per se, but to find a reasonable abstraction that captures the process of combinatorial creativity in the empirical context of motion pictures. First, I model the arrival of movie ideas (or candidates), together with the connections between these ideas, with a stochastic network process. Once a movie candidate arrives, a studio makes a go/no-go decision. There is learning on the studio side based on past releases.[11] If the decision is yes, the movie goes into production, and later, consumers decide how much box-office revenue it realizes. In what follows, I start with the consumer side.

4.1. Consumer Demand

The main purpose of the demand-side model is to provide a way to compute the distribution of box-office revenue. This distribution gives the expected return and risks associated with a movie candidate, which are required later in the model of studio go/no-go decisions. Because the focus of the paper is not the box-office demand per se, I keep the demand-side model as simple as possible for the sake of tractability. Let me start with the following specification of consumer i's utility from movie j at the time of j's release:

$$u_{ij} = U\left(x_j;\beta\right) + \xi_j + \varepsilon_{ij}. \tag{1}$$

In this expression, vector x_j collects movie j's observed characteristics. In principle, x_j can include any aspect of j that is observed in the data up to j's release, such as budget size, atypicality, production start date, and the release date. The second term ξ_j captures the average consumer taste over the characteristics that are not included in x_j. I call ξ_j the "latent quality" of j. The last term ε_{ij} is the consumer's idiosyncratic utility.

Strictly speaking, the utility terms in (1) should have time subscripts as consumer's taste may change

Table 6. Residual Size from ROI Regression

Number of prior similar movies	Group count	Mean	First quantile	Third quantile
0–1	828	2.42	0.24	3.1
2–6	801	1.87	0.13	2.2
7–21	691	0.95	0.05	0.84
>21	759	0.5	0.05	0.62

Notes. The means and quantiles are calculated within each group with respect to the squared residuals from column (5) in Table 5. The cutoffs in the first column are chosen so that the four groups are roughly of equal size.

over time. However, given it is understood that u_{ij} refers specifically to the utility at the time of j's release, omitting time subscripts should not raise confusion.[12]

Assume that individual i chooses between going to a movie theater to watch j and an "outside option," for which the utility is specified as $u_{i0} = \varepsilon_{i0}$. Then, with ε_{ij} and ε_{i0} following type I extreme value distribution, the market share of j is given by $1/(1 + e^{-U(x_j,\beta)-\xi_j})$. The associated box-office revenue is

$$\pi_j = m_{r_j} / \left(1 + e^{-U(x_j,\beta)-\xi_j}\right), \tag{2}$$

where r_j is the release date of j and m_t is the product between the average theater ticket price and the number of moviegoers in the United States at time t. By the statistics published by MPAA, about two thirds of the population go to the cinema at least once a year. I regard this subpopulation as the moviegoers.[13]

The demand model does not take into account several previously studied factors that could affect box-office performance, including screening, advertising, competition, and word of mouth (e.g., Elberse and Eliashberg 2003; Ainslie et al. 2005; Hennig-Thurau et al. 2006; Liu 2006; Einav 2007, 2010; and Chintagunta et al. 2010; most papers have focused on one factor at a time). Note that these factors typically enter the picture after the production of a movie. I abstract away from them to keep the whole model tractable and keep the paper's focus on the green-light decisions, when the quality of a movie seems most important (Hennig-Thurau et al. 2006). This is not to say that postproduction factors are not taken into consideration at green-light decisions (e.g., a studio might decide not to pursue a movie to avoid competition at release with another studio). It is, therefore, wise to make sure that a simple model, such as (2), still captures demand (at least at the level of aggregate box office for each movie) reasonably well compared with more complex models. As we see later with the estimation results (Section 5.4), this is indeed the case.

4.2. Movie Candidates

Movie candidates arrive at a Poisson rate η. Fix a candidate movie j that arrives at time t. The candidate is characterized by (i) the observed characteristics, x_j; (ii) the latent quality, ξ_j; and (iii) its location in the network, that is, what prior movies j imitates. All three aspects are generated by a stochastic process or, more precisely, an evolving network model.

Before getting into the modeling details, it is instructive to have a short discussion on evolving network models in general and how they fit my setting. Given the nature of networks, it is not surprising that they have frequently appeared in models of creativity, particularly as a way to keep track of the connections between ideas. In information theory, Price (1965, 1976) proposed a model for the formation of citation networks in which he treats every scientific paper as a creative combination of previous papers. His pioneering work was rediscovered much later by Barabasi and Albert (1999), which became one of the most influential papers in network science. A theme of the Barabasi–Albert model, as well as many later models based on it, is the so-called attachment process in which new nodes arrive over time and "attach" (i.e., link) to existing nodes. The attachment process is a general way to model network growth. As in Price (1976), many specifications of the attachment process were proposed to describe how ideas or projects are created and accumulated (e.g., Kleinberg et al. 1999, Kumar et al. 2000). My model largely follows this line of literature. There is, however, an important difference. In almost all variants of the attachment process, every new node becomes a part of the network. In my setting, a new node is only a candidate, which may be rejected by studios. What stays in the observed network consists of only accepted candidates.

The aforementioned works are all outside economics. A sizable collection of works on network formation has also accumulated in the economic literature. A distinctive feature of the economic literature is that it examines strategic network formation (using game-theoretic tools) in contrast to the more statistical modeling in the computer science and physics literature (Jackson 2008). Strategic modeling is particularly relevant to social networks (as opposed to networks of "things"). This is not to say that statistical modeling of social networks is absent in economics; notable examples include Jackson and Rogers (2007) and Bramoulle et al. (2012), which explore the idea that people often form new ties by "meeting friends of friends." As we see, my model is closely related to this idea.

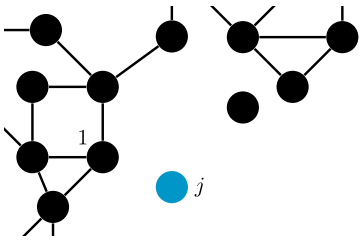
4.2.1. Location in Network. The location of j in the network is denoted as a binary vector y_j, where $y_{k,j} = 1$ indicates that j is similar to an existing movie k, and $y_{k,j} = 0$ indicates otherwise. My goal here is to specify the distribution $\Pr(y_j|\mathcal{S}_t)$ from which y_j is drawn, where $\mathcal{S}_t$ is the state of the model at time t. To formally express $\mathcal{S}_t$ requires the introduction of several notations. For each movie k in the data, let a_k denote its arrival time. The set of existing movies (either already released or still in production) at time t is written as $\{k : a_k < t\}$. The similarity network among this set of movies is denoted as Y_t with $Y_{k\ell,t} \in \{0, 1\}$ indicating whether k and ℓ are linked. Then $\mathcal{S}_t = \{\{x_k : a_k < t\}, \{\xi_k : a_k < t\}, Y_t\}$.

In its simplest form, one may specify that $\Pr(y_{k,j}|\mathcal{S}_t)$ is independent and uniform across k. This would be in the same spirit as the Erdos–Renyi model, the most

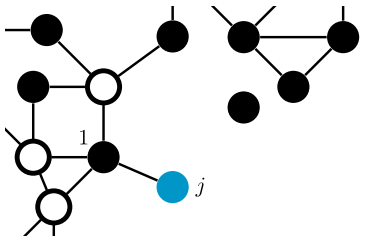
basic network formation model that says links are established in an independent and identically distributed manner. However, it is a very strong abstraction and does not give rise to some important properties that we saw in the similarity network, for example, clustering, and atypicality (see Section 3). To this end, several papers in the network literature become particularly relevant here. Kleinberg et al. (1999) and Kumar et al. (2000) specify that, after a new node arrives, it first attaches to an existing node and then forms links with the existing node's immediate neighbors. In some sense, the new node "copies" the neighbors of an existing node, and this behavior is common in citation networks: a researcher often finds additional papers through the bibliography of a paper that the researcher reads. Holme and Kim (2002) apply the same idea and show it gives rise to a wide range of clustering in networks. Of course, one can imagine that the new node can follow neighbors' neighbors so that it forms links in the vicinity (not just immediate neighbors) of the first existing node to which it attaches. This would nicely capture how "typical" (as opposed to atypical) combinations happen in idea creation. In addition to copying immediate neighbors, Kumar et al. (2000) also introduced a step in which links are formed randomly with any existing node. This echoes the concept of atypical combinations.

Figure 5 provides a graphical description of my specification of $\Pr(y_j|\mathcal{S}_t)$ without going into technical details. The links are formed in a sequential manner. Intuitively, the imitation process picks one existing movie to start with and then spreads out from that movie. This typically determines the stream (or cluster) of movies in which j is grounded. After this, the imitation may pick several movies elsewhere in the network, creating atypical combinations.

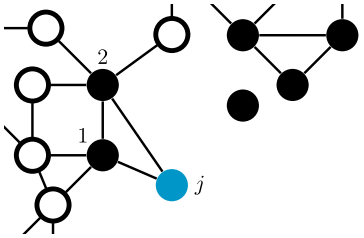
Figure 5. Illustration of the Linking Process



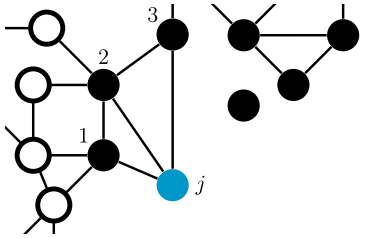
(i) Before the imitation starts. The candidate is represented by node j.



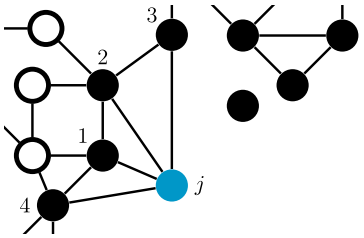
(ii) The candidate imitates movie 1, which makes movie 1's neighbors (marked as hollow nodes) open to imitation.



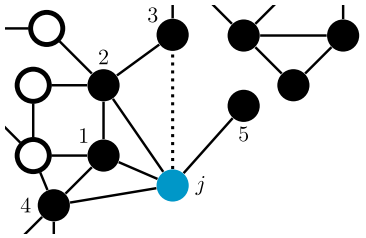
(iii) The candidate imitates movie 2, which makes movie 2's neighbors (marked as hollow nodes) also open to imitation.



(iv) The candidate imitates movie 3, which makes movie 3's neighbors (outside the plot) open to imitation.



(v) The candidate imitates movie 4, which makes movie 4's neighbors (outside the plot) open to imitation.



(vi) The link with movie 3 is re-wired to movie 5, creating an atypical combination.

Formally, let binary vectors $y_j^1, y_j^2, \ldots, y_j^m, \ldots$ indicate the nodes that have been linked to j in each step of the sequential process. Exactly one link is formed in each step so that $\sum_k y_{k,j}^m = m$ for all $m = 1, 2, \ldots$ The last binary vector in the sequence is taken as the realization of y_j. Before the first step, there is a probability ι that the process terminates right away. In this case, j is completely novel. With probability $1-\iota$, a link is formed between j and one of the existing nodes, realizing y_j^1. The probability that the link is with k is specified as $\Pr(y_{k,j}^1 = 1|\mathcal{S}_t) \propto \exp(\theta|t-a_k|)$, where parameter $\theta < 0$ allows a higher probability of imitating the more recent movies. In step $m \geq 2$, the probability that a link forms between j and some node k is given by

$$\Pr\left(y_{k,j}^m = 1|y_j^{m-1}, \mathcal{S}_t\right) \propto \exp\left[\theta|t-a_k| + \omega \log\left(\sum_\ell Y_{k\ell,t} \cdot y_{\ell,j}^{m-1}\right)\right].$$

In this expression, $\sum_\ell Y_{k\ell,t} \cdot y_{\ell,j}^{m-1}$ equals the number of neighbors of k that have been imitated by j in the previous steps. In other words, if j has imitated many movies that are similar to k, then there is a high probability that j imitates k too.

As long as $\omega > 0$, this process tends to find a cluster around the first linked node in the sequence. It is reasonable to assume that the size of this cluster is larger if there are many movies similar to this first linked node. Hence, I specify that the process stops at step $m^* = \left(\sum_k y_{k,j}^1 \sum_\ell Y_{k\ell,t}\right)^\nu$. The term in the parentheses is simply the number of the first linked node's neighbors in Y_t. Parameter ν is to be estimated.[14]

So far, the process has not introduced any atypical link. To do so, I allow additional steps after m^*. In each additional step, with probability γ, a previously formed link is rewired to a random node k with the probability proportional to $\exp(\theta|t-a_k|)$; with probability $1-\gamma$, the process does nothing but terminate. For example, if the process terminates at step m^*+1, then the atypicality of j is zero. It should be clear that parameter γ here calibrates the amount of atypicality.

Robustness checks can be made by specifying variants of the linking process here. One such variant is considered in the appendix, in which atypical links are created in parallel with the typical links instead of being rewired in the end. The results of the paper hold without significant changes.

4.2.2. Observed Characteristics. Given y_j, next I specify the distribution $\Pr(x_j|y_j, \mathcal{S}_t)$ from which x_j is drawn. Intuitively, x_j should be correlated with x_k if j imitates k. For example, if j imitates a group of big-budget movies, then it should likely have a big budget too.

In principle, one can include in x_j any movie characteristic observed in the data. However, as shown by the earlier analysis, characteristics such as genre, MPAA rating, and star power of the crew add very little predictive power after the performance of prior similar movies is accounted for. In addition, as we see, the inclusion of these detailed characteristics is not necessary for the intuition behind the counterfactuals to carry through. Therefore, in an effort to keep the model tractable, I drop these characteristics from the supply side of the model. The remaining part of x_j includes the production budget and variables related to the release date.

I draw the budget, denoted as b_j, from a truncated normal distribution. Because the largest budget in the data are slightly below \$350 million (in 2014 dollars) and I exclude movies with less than a \$1 million budget, the truncation interval is set to be $[1, 350]$. Results are not sensitive to the exact choice of the upper bound. The mean of the normal distribution (before truncation) is set equal to the average budget of j's prior similar movies. If j is novel, then the mean is drawn from an exponential distribution with a mean parameter μ. The variance-to-mean ratio of the normal distribution is denoted as χ.

The release date r_j equals the current time t plus the time needed for producing j. The production time is specified as a function of the budget size b_j, which is estimated "off-line" using the data on movies' production start dates (see Section 2.1 for more details). Admittedly, this is a simplification and abstracts away from studios' postproduction strategies for timing the release (Einav 2010). However, it does provide a reasonable way to estimate r_j at the time of green-light decisions.

4.2.3. Latent Quality. Finally, I specify the distribution $\Pr(\xi_j|x_j, y_j, \mathcal{S}_t)$ from which the latent quality of j is drawn. Clearly, it should allow ξ_j to be correlated with ξ_k if j and k are similar (i.e., $y_{k,j} = 1$). In addition, it should allow the correlation to decrease with the age of k at the time of t. Intuitively, this is because ξ measures the consumer tastes *at the time of* the movie's release, and consumer tastes may vary over time; what consumers chose in the longer past is probably less indicative of what their preferences are today or will be tomorrow.

These considerations, together with the reduced-form results (Table 5), prompt me to specify the following distribution for ξ_j:

$$\Pr\left(\xi_j|x_j, y_j, \mathcal{S}_t\right) \sim \mathcal{N}\left(\frac{\lambda \sum_{y_{k,j}=1} \phi^{|r_k - r_j|} \xi_k}{1 + \lambda \sum_{y_{k,j}=1} \phi^{|r_k - r_j|}}, \frac{\sigma^2}{1 + \lambda \sum_{y_{k,j}=1} \phi^{|r_k - r_j|}}\right), \tag{3}$$

The mean of the distribution equals a weighted average over j's prior similar movies, which creates

correlations between the latent qualities of similar movies. Parameter $\lambda > 0$ and parameter $\phi < 1$ calibrate the correlations: a higher λ increases the level of correlation between any two similar movies; a smaller ϕ decreases the correlation between movies whose release dates are far apart from each other (which corresponds to a more rapid change in consumer tastes).

4.3. Investment Decision

Next, I describe the part of the model that determines which candidates get funded for production and which do not. This gives rise to a selection mechanism with which more potential movie candidates are more likely to get produced. The selection is consistent with what we saw earlier in the data (Section 3): movies with higher ROIs have more imitators.

To conceptualize the model, it is useful to have a brief look at how investment decisions are typically made in practice. A senior studio executive once described the green-light process as follows: "We bring together all studio department heads. [The production cost] is our most reliable estimate, and that thus forms the basis for our launch decision. . . . In the end . . . someone in the meeting has to put his or her reputation on the line and say yes."[15] To say yes (or no), it is essential to have a forecast, either in a quantitative or qualitative way, of the box-office revenue. The basis for this forecast often goes to the performances of past movies. When I spoke to industry executives, I was told that "the box office is the most important and the most difficult [to predict]. All you can use is the historical data; you look into the historical performance of the actors and other things. . . . You also need to make sure that these things mix well." In general, not only studios but also producers (who bring their projects to studios) frequently make references to past movies. Pitching a movie idea by comparing it to past similar releases is quite common.

I stylize the investment decision in the following way. Consider a candidate j arriving at t. The studio makes a forecast of j's box-office revenue, π_j. This forecast is a distribution $\Pr(\pi_j|\mathcal{F}_t)$, where $\mathcal{F}_t$ is the information set of the studios. Importantly, $\mathcal{F}_t$ includes the box-office performance of each movie that has been released up to t (but not the movies still in production). Formally, $\mathcal{F}_t = \{\{x_k : a_k < t\}, \{\pi_k : r_k < t\}, \{x_j, y_j\}, Y_t\}$. As a technical note, observe that each ξ_k with $r_k < t$ is known under $\mathcal{F}_t$ because it can be backed out from π_k using (2).

Suppose for a moment that I have a way to compute the studio's forecast $\Pr(\pi_j|\mathcal{F}_t)$; then I can calculate a risk-free equivalence for π_j denoted by $\widetilde{\pi}_j$:

$$V\left(\widetilde{\pi}_j\right) = \mathbf{E}\left[V\left(\delta^{r_j - t}\pi_j; \alpha\right)|\mathcal{F}_t\right].$$

In this equation, $V(\cdot;\alpha)$ is a utility function parameterized by α. Parameter α calibrates the concavity of V or $\alpha = -V''/V'$. It is known as the coefficient of constant absolute risk aversion. Parameter δ is a discounting factor. The studio puts j into production iff $\widetilde{\pi}_j > b_j \cdot e^{-\rho\zeta_j}$, where b_j is j's budget and ζ_j is an econometrician-unobserved independent shock. I specify ζ_j to follow the standard type I extreme value distribution. Notice that the probability of accepting j increases with $\widetilde{\pi}_j$ but decreases with b_j.

Back to the problem of computing of $\Pr(\pi_j|\mathcal{F}_t)$, recall that π_j is given by (2). So the problem basically reduces to computing $\Pr(\xi_j|\mathcal{F}_t)$.[16] In the simplest case in which all of j's prior similar movies happen to have been released by t, the distribution $\Pr(\xi_j|\mathcal{F}_t)$ is directly given by (3). In the more general case in which one or more of j's prior similar movies are still in production at t, computing $\Pr(\xi_j|\mathcal{F}_t)$ is more involved. One needs to estimate the ξ's of those in-production movies. I give details in the appendix.

5. Estimation

There are two main challenges in estimating the model. First, I do not observe the candidate movies that were rejected. In other words, I have a selected sample. As in most econometric models with selected samples (e.g., truncated Tobit, Heckit), consistent estimation relies on a model of how selection happens to make up for the missing data. Second, the similarity network is endogenous. So the estimation needs to not only account for the correlation pattern implied by the network but also the endogeneity of this pattern itself. This differs from the standard spatial econometrics, which takes the correlation pattern as exogeneous (Bradlow et al. 2005, LeSage 2008). I give the details on how I specifically deal with these challenges, after which I present the estimates.

5.1. Demand Side

Taking log on both sides of the box-office Equation (2) produces the regression equation for the demand side:

$$\log\left(\pi_j\right) - \log\left(m_{r_j} - \pi_j\right) = U\left(x_j, \beta\right) + \xi_j. \qquad (4)$$

The ordinary least squares (OLS) condition, $\mathbf{E}(\xi_j|x_j) = 0$, does not hold here. This is because the data only includes the movies that have been accepted. As in any estimation with selected samples, the key here is to control for the factors underlying the selection. In my model, each movie j is selected based on $\mathcal{F}_{a_j}$, studios' information set at time a_j. So $\mathbf{E}(\xi_j|\mathcal{F}_{a_j})$ can be used to control the selection. Although this gives me consistent estimates, it is not easily implementable as $\mathbf{E}(\xi_j|\mathcal{F}_{a_j})$ generally is not easy to compute (see Section 4.3). One solution is to use an information set that is slightly larger than $\mathcal{F}_{a_j}$: $\{\{\xi_k : a_k < a_j, r_k \geq a_j\}, \mathcal{F}_{a_j}\}$ or, equivalently, $\{x_j, y_j, \mathcal{S}_{a_j}\}$. First, because this information

set is larger than $\mathcal{F}_{a_j}$, it also is able to control for the selection. Second, $\Pr(\xi_j|x_j, y_j, \mathcal{S}_{a_j})$ has a relatively simple expression given in (3). To construct moments, I define $\epsilon_{1,j} \equiv \xi_j - \mathbf{E}(\xi_j|x_j, y_j, \mathcal{S}_{a_j})$, and by the law of iterated expectation,

$$\mathbf{E}\left(\epsilon_{1,j}|x_j, y_j, \mathcal{S}_{a_j}\right) = 0.$$

Moment conditions are constructed by interacting $\epsilon_{1,j}$ with functions of $\{x_j, y_j, \mathcal{S}_{a_j}\}$. To identify β, I interact $\epsilon_{1,j}$ with x_j. To identify λ, I interact $\epsilon_{1,j}$ with the average latent quality of j's prior similar movies. To identify ϕ, I interact $\epsilon_{1,j}$ with the average latent quality of a subset j's prior similar movies whose release dates are close to r_j.

Parameter σ can be identified in a similar way but using the second-order conditional moment of ξ_j: $\mathbf{E}(\xi_j^2|x_j, y_j, \mathcal{S}_{a_j})$. Let $\epsilon_{2,j} \equiv \xi_j^2 - \mathbf{E}(\xi_j^2|x_j, y_j, \mathcal{S}_{a_j})$. Two more moment conditions are constructed by interacting $\epsilon_{2,j}$ with a constant term and the log number of j's prior similar movies.

As in the analysis of any dependent time series, a part of the data should be put aside as the initial condition for estimation. I use the data between 1975 and 1995 as this initial condition. More precisely, the generalized method of moments (GMM) only average across the movies that started production in or after 1995; however, data before 1995 is always included in $\mathcal{S}_{a_j}$ when computing $\Pr(\xi_j|x_j, y_j, \mathcal{S}_{a_j})$.

5.2. Supply Side

Now I describe the estimation of the parameters for candidate arrivals and studio decisions. This is done by the method of moments, which matches the model-implied distribution to the observed distribution in the data. Again, I need to account for the selection issue. When there is selection, one can no longer view the data as drawn from the population distribution. Instead, one generally relies on a model that captures the selection mechanism to derive the distribution *conditional* on selection. The parameters are then estimated by matching this conditional distribution with the observed distribution. In my context, this means that the moments to be matched should be what the model predicts for the accepted movies, not all the candidate movies.

Specifically, I index the movies in the data by arrival date so that j is the first movie that arrives after $j-1$. Let H_j collect some characteristics about movie j at time a_j, such as (i) the budget b_j, (ii) the latent quality ξ_j, and (iii) the number of movies that j imitates $\sum_k y_{k,j}$. Formally, H_j is any function of $\{x_j, y_j, \xi_j, \mathcal{S}_{a_j}\}$. Now, define

$$h_j \equiv H_j - \mathbf{E}\left(H_j|x_{j-1}, y_{j-1}, \xi_{j-1}, \mathcal{S}_{a_{j-1}}\right) \tag{5}$$

as the discrepancy between the realized H_j and the model-predicted H_j. Notice, in particular, that the previous conditional expectation is what the model predicts for the next accepted movie after $j-1$, not the next arrival after $j-1$.

By the law of iterated expectation, $\mathbf{E}(h_j) = 0$. Intuitively, what this moment condition says is that, under the right parameters, the prediction errors of the model should be mean-zero. Because the conditional expectation in (5) does not have a closed-form expression, $\mathbf{E}(h_j)$ needs to be evaluated through simulation. I follow the standard procedure of the method of simulated moments. Basically, I search for the parameter values that bring $\frac{1}{n-k+1}\sum_{j=k}^{n} h_j$ closest to zero, where k is the first movie produced in 1995. Again, the data in 1975–1994 is used as the initial condition.

One major challenge in estimation with a selected sample lies in choosing the moments to ensure that the parameters are identified. Here, the choices of the moments are specified by the entries of H_j. I give the intuitions on identification and specify H_j accordingly.

5.3. Identification

I focus on the parameters for which the identification is less obvious. The first parameter is the arrival rate of candidates, η. It is not directly observed in the data, and the identification actually relies on a normalization made in the model. More specifically, because the data only contains the accepted movie candidates, from an observational point of view, the effect of an increase in the arrival rate η can be exactly offset by a proportional decrease in the acceptance probability for *every* candidate movie. So to identify η requires fixing the acceptance probability for certain type of candidates. This has been done in Section 4.3 for the candidates with $\widetilde{\pi}_j = b_j$, whose acceptance probability is fixed at $\Pr(\zeta_j > 0) = 1 - e^{-1}$. Of course, the identification does not come for free; the value of η is ad hoc to the normalization. In particular, the estimate for η cannot be interpreted literally as the empirical rate at which studios receive scripts.[17]

The coefficient of risk aversion, α, can be identified from the average budget size in the data. Intuitively, the risks associated with a movie increase with its budget size, so a higher α should result in a lower acceptance probability for the bigger-budget candidates. In addition, α can also be identified from the joint distribution between movie budget and imitativeness (as measured by the number of prior similar movies). The intuition is that a smaller budget and a higher imitativeness are both ways to lower risks, so the level of α affects the extent to which the two substitute each other in studios' decisions.

The scale of the decision shocks, ρ, is identified from the average level of ROI in the data. Intuitively, ρ calibrates the extent to which the investment decision is based on the candidate's expected ROI versus the shocks. When ρ is smaller, the decisions are driven more by the candidate's expected ROI, meaning that the accepted movies will have higher ROIs (or latent qualities).

The other parameters whose identifications are not obvious pertain to the link formation process. The first of these parameters is ω. A larger ω gives a stronger tendency for the imitated nodes to be neighbors of each other. So ω can be identified by the degree of clustering in the network (see Section 3). The second parameter is γ. A larger γ implies more atypical links, so it can be identified by the average atypicality of the movies in the data. Finally, parameter ι is identified by the percentage of movies in the data that are novel.

With this said, I include the following entries in H_j: (1) the time elapsed since last movie production, that is, $a_j - a_{j-1}$; (2) a dummy indicating whether j is novel; (3) the log number of j's prior similar movies; (4) average log budget of j's prior similar movies; (5) average age of j's prior similar movies at time a_j; (6) log budget b_j; (7) the absolute difference between b_j and the average budget of j's prior similar movies; (8) log budget b_j times a dummy indicating whether j is novel; (9) log number of triangles created by j in Y_{a_j}; and (10) the latent quality ξ_j.

5.4. Parameter Estimates

Table 7 displays the estimates for the demand-side parameters. Column (1) displays the estimates from an OLS regression of the revenue Equation (4), column (2) displays the full GMM estimates; and column (3) displays the GMM estimates in which some observed characteristics are dropped from x_j.

The results here conform to those in the reduced-form analysis (Table 5), so I shorten the discussion to focus on several important points. First, λ is significant, which indicates a positive correlation between the ξ's (latent qualities) of similar movies. Second, ϕ is significantly less than one, which indicates that the correlation between the ξ's of two movies decreases with the gap between their release dates. This reflects time-varying consumer tastes. Third, a balanced level of atypicality has a significant and positive effect on box-office demand. In view of reduced-form results, the variable *balanced atypicality* here is defined as the negative of the distance between the movie's atypicality level and 0.25. Forth, the main source for the prediction of ROI is the similarity network; the added prediction power from the observed characteristics, such as genre, rating, and the quality of the crew, is only marginal.

The overall demand model fit can be compared with previous works, which mostly consider post-production factors. Einav (2007) accounts for competition in theaters, and the R^2 for movie demand is 0.39; Ainslie et al. (2005) additionally allows stronger competition within the same genre, and R^2 is 0.46 and rises to 0.61 if the number of screens is also used as an explanatory variable (R^2 for ROI, unfortunately, was almost never reported). Generally, such comparison can be made only roughly because of differences in data sample and modeling choice, but it suggests that my model captures the demand reasonably well even when compared with more complex models that account for postproduction factors. Table 8 displays

Table 7. Model Parameter Estimates, Demand Side

	(1)	(2)	(3)
Genre dummies	Yes	Yes	
Log budget	1.09 (0.03)	1.02 (0.04)	1.05 (0.03)
Trend	−0.0217 (0.005)	−0.0132 (0.007)	−0.0085 (0.008)
Seasonality	0.142 (0.05)	0.142 (0.05)	0.166 (0.04)
Rating as restricted	−0.22 (0.06)	−0.122 (0.06)	
Star actor	−0.0153 (0.06)	0.0155 (0.06)	
Star director	0.0243 (0.07)	0.0519 (0.06)	
Star writer	0.138 (0.07)	0.112 (0.06)	
Balanced atypicality	0.767 (0.13)	0.404 (0.14)	0.472 (0.13)
Parameters for latent quality:			
Similarity weight (λ)		0.249 (0.05)	0.274 (0.06)
Discounting factor (ϕ)		0.914 (0.05)	0.911 (0.04)
Standard deviation (σ)		1.71 (0.04)	1.74 (0.04)
R^2 (log box office)	0.528	0.572	0.564
R^2 (log ROI)	0.077	0.163	0.147

Notes. Column (1) displays the OLS estimates of Equation (4). Columns (2) and (3) display the GMM estimates. The numbers in parentheses are standard errors. The standard errors in columns (1) and (2) are computed by asymptotic formulas; the standard errors in column (3) are bootstrapped (see appendix). The regressor "balanced atypicality" is defined as the negative absolute difference between the movie's atypicality and 0.25.

the estimates for the supply-side parameters. First shown is the yearly arrival rate of the movie candidates. As discussed, the identification of η relies on a normalization of the acceptance probability in the model, so η should not be literally interpreted as an estimate of the number of proposals that studios actually receive in a year. However, conditional on this normalization, the estimate of η implies that around 59% of the candidates got rejected.

Next shown in Table 8 are the parameters for the imitation process (or the link-formation process). All the estimates have expected signs. In particular, θ is estimated to be significantly negative, implying that there is a tendency to imitate the more recent movies; parameter ω is significantly positive, consistent with the substantial clustering observed in the network; the estimate of parameter γ is in line with the level of atypical combinations that we saw in the descriptive analysis (Figure 4).

Last shown in Table 8 are the parameters for the studio's investment decisions. The coefficient of the risk aversion, α, is estimated to be statistically significant. It is also economically significant, which is shown in Figure 6. The top graph takes the estimate for α and plots a candidate's risk-adjusted ROI, $\widetilde{\pi}_j/b_j$, as a function of the uncertainty in ξ_j. The bottom graph plots the same function except that it lets $\alpha \to 0$. The clear differences between the two graphs tell us that risk aversion plays a substantial role in studios' investment decisions.

An interesting observation about Figure 6 is that, for a small-budget candidate, the risk-adjusted ROI *increases* with the level of uncertainty regardless of whether α is at the estimated level or set toward zero. This appears to go against the usual intuition about risk aversion. However, one needs to remember that there is a lower bound on how much a movie can lose: the worst scenario is not earning a single penny in the box office and losing the entire budget. So, if a candidate requires only a small budget, there is little to lose. In this case, the main effect of a higher level of uncertainty is increasing the potential gain of the investment rather than increasing the risks. As a result, interestingly, novelty actually makes small-budget candidates more attractive of an investment.

Finally, given all the parameter estimates, there is the question of how well the full model reproduces patterns in the data. In the appendix, I test the model on several key data distributions. The fit is exceptional considering the level of abstraction of the model.

6. Counterfactual

The preceding analyses have shown that, at the level of each individual movie, a balanced level of atypicality has a positive effect on ROI (Tables 5 and 7). However, this does not mean that, at the collective level or in the longer run, the effect on ROI is definitely still positive. Over time, any small change in the combinatorial process (e.g., an increase in atypicality) will gradually change the similarity structure of the market, which, in return, affects the combinatorial process. Such dynamics have important implications for the profitability of the industry and consumer welfare; however, it is difficult to account for them in a model-free analysis. This prompts me to conduct the counterfactuals in this section. A more comprehensive understanding of the effects of atypical combination is of interest not only in the context of movies but also for other types of creative works.

6.1. Simulation

Here I give a brief description of the technical aspects of the counterfactual analysis. Readers not interested in the technical details can safely skip to the results.

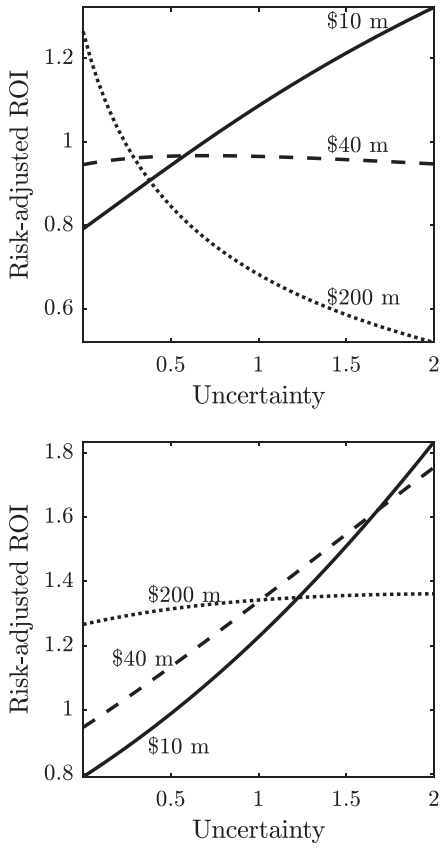
In my model, the average level of atypicality for the movie candidates is set by parameter γ. Notice that γ does not change the innovation rate (which is set by parameter ι) or the number of movies that a

Table 8. Model Parameter Estimates, Supply Side

Parameters	Estimates	Standard error
Number of yearly arrivals (η)	407.4	(19.3)
Similarity structure:		
Innovation probability (ι)	0.189	(0.014)
Propensity to imitate older movies (θ)	−0.153	(0.004)
Propensity to imitate (ν)	0.887	(0.014)
Propensity to cluster (ω)	1.568	(0.067)
Probability of atypical combination (γ)	0.762	(0.008)
Budget distribution:		
Mean budget for novel candidates (μ)	75.8	(9.8)
Dispersion parameter (χ)	6.32	(0.19)
Go/no-go decision:		
Risk aversion (α)	0.0163	(0.003)
Scale parameter of shocks (ρ)	0.299	(0.034)

Note. The standard errors are bootstrapped; see the appendix for more details.

Figure 6. Risk-adjusted ROI as a Function of Movie Uncertainty and Budget Size



Notes. Both plots display the risk-adjusted ROI, $\widetilde{\pi}_j/b_j$, of a hypothetical movie j as a function of the budget size b_j and variance in ξ_j. The top plot uses the estimated parameter values, and the bottom plot sets $\alpha \to 0$ so that firms are risk neutral.

candidate imitates (which is set by parameter ν). Rather, it specifically calibrates the balance between staying in a conventional zone versus reaching outside for atypical elements.

I simulate the model under different values of γ. The simulation starts with an empty set of movies and gradually grows the similarity network. Because the goal here is about the long-term implications, I focus on the steady states. For the model to have a steady state, I set the demand trend (a coefficient in β) to zero and fix the market condition m_t constant at the 2014 level. A "burn-in" period of the simulated data is discarded. The data after the burn-in period is used to visualize the network structure and compute various steady-state statistics, such as average budget size and average ROI.

6.2. Results

Figure 7 displays the networks formed under four different levels of γ. Recall that γ is the probability of creating atypical links; a higher γ implies a higher average level of atypicality. The four networks are drawn with the force-directed method that places linked nodes close to each other (in the exact same way that Figure 2 visualizes the network in the real data).

The network for $\gamma = 0.05$ is given in the top left plot. In this case, there is very little atypicality: each movie candidate tries to stay within its own conventional zone. Consequently, the network is divided into many groups with very sparse connections between them. These groups represent a diverse range of distinctive movie types. Conceptually, an analogy can be made with citation networks: if there is little cross-disciplinary work, science would break into many specialized fields.

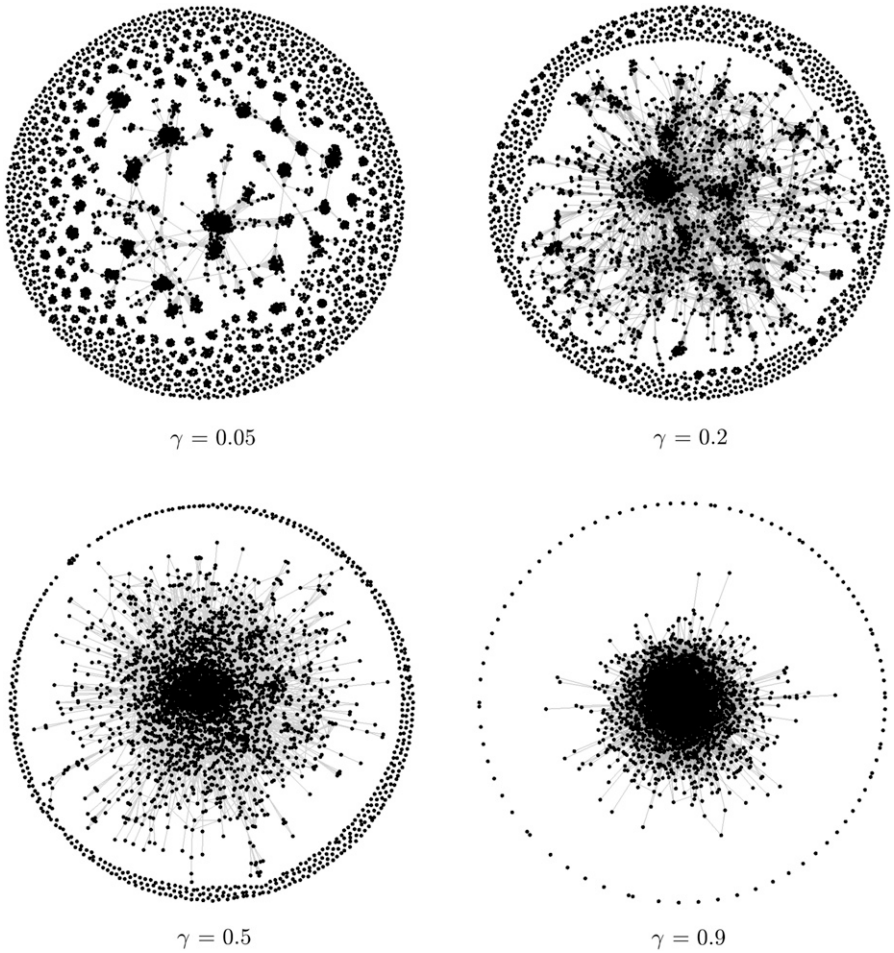
As γ increases, the pattern of isolation slowly transforms to a pattern of integration. At $\gamma = 0.2$, the network still exhibits diverse groups of movies but not as isolated as those under $\gamma = 0.05$. At $\gamma = 0.5$, the network approaches the one observed in the real data, in which a core is visually present (Figure 2). In this case of $\gamma = 0.9$, imitation is really not specifically focused on any stream of movies but becomes rather general. As a result, there develops a great deal of commonality among movies, which manifests itself as a larger and tighter core.

Figure 7 indicates an interesting disparity in the meaning of diversity between the individual and collective levels. For any individual movie, atypical combination sets it apart from its conventional realm, which implies individual diversity. However, for the movie population in the long run, atypical combination actually leads to commonality.

Given that different values of γ lead to different similarity structures of the market, the next question is how these different structures affect the ROI. This is answered by Table 9. The first column displays, under different values of γ, the average risk-adjusted ROI of the movies that are accepted. It is seen that the average ROI is downright decreasing in γ. Note that Table 9 assumes the parameter values when atypicality has an individual-level demand effect (i.e., the variable *balanced atypicality* in Table 7). To see whether the individual-level effect is driving the result, Table 10 recomputes the counterfactuals with this effect removed. The negative relation between the ROI and γ still holds. Notice that compared with Table 9, the ROI is slightly lower at moderate γ but higher at large γ; this is consistent with the optimal point of atypicality at the individual level (around 0.25).

The result points out another disparity between the individual and collective level. For any individual movie, atypicality can have a positive effect on its box office. However, for the whole industry in the long run, profitability is actually negatively related to any degree of atypical combination. The reason for this

Figure 7. Network Structures Under Different γ's



Notes. Each network is taken from the model simulation at its steady state for a period of 18 years (the same length as the data sample). The networks are drawn by force-directed placement (Fruchterman and Reingold 1991). Parameter γ calibrates the average level of atypicality of movie candidates.

negative relation lies in the similarity structure: as shown in Figure 7, the movie population is more diverse under a smaller γ. As a result, the movie candidates are more diverse too; in particular, the distribution of the risk-adjusted ROIs in the candidate pool is more dispersed (as evidenced by the third columns of Tables 9 and 10, which display the standard deviation of the risk-adjusted ROIs among all the candidates). The more dispersed candidate distribution is an advantage in the presence of studio selection because the better end of that distribution enjoys higher acceptance rates. The wider the candidate distribution is, the better the accepted movies are on average. In addition, better movie releases also foster better future candidates, resulting in a virtuous circle.

Table 9. Risk-Adjusted ROI Under Different γ's

Probability of atypical combination (γ)	Average of the accepted (%)	Average of the rejected (%)	Standard deviation of all (%)	Average budget of accepted	Yearly productions
0	8.78 (0.10)	−25.6 (0.03)	31.8 (0.07)	33.2 (0.1)	202 (0.4)
0.05	7.82 (0.09)	−25.8 (0.05)	31.3 (0.06)	33.4 (0.1)	199 (0.3)
0.2	6.09 (0.10)	−26.1 (0.05)	30.3 (0.06)	35 (0.1)	194 (0.4)
0.5	3.12 (0.07)	−27.1 (0.04)	28.9 (0.04)	36.5 (0.1)	184 (0.3)
0.9	−2.92 (0.08)	−30.3 (0.06)	26.3 (0.04)	37.2 (0.1)	160 (0.4)

Notes. The statistics in the first three columns are with respect to the risk-adjusted ROIs (expressed in percentage return). Parameter γ varies across rows; the other parameters are set at their estimates.

Table 10. Risk-Adjusted ROI Under Different γ's, Demand Effect of Atypicality Removed

Probability of atypical combination (γ)	Average of the accepted (%)	Average of the rejected (%)	Standard deviation of all (%)	Average budget of accepted	Yearly productions
0	8.78 (0.10)	−25.6 (0.03)	31.8 (0.07)	33.2 (0.1)	202 (0.4)
0.05	7.66 (0.08)	−25.8 (0.05)	31.2 (0.04)	33.8 (0.1)	199 (0.4)
0.2	5.93 (0.08)	−26.2 (0.05)	30.2 (0.05)	34.7 (0.1)	193 (0.4)
0.5	3.42 (0.07)	−27 (0.05)	29 (0.04)	35.6 (0.1)	185 (0.3)
0.9	−0.15 (0.07)	−27.8 (0.05)	26.9 (0.04)	36.3 (0.1)	174 (0.4)

Note. This table is the same as Table 9 except that the atypicality of a movie no longer enters consumer utility.

6.3. Discussions

The counterfactuals reveal a mechanism through which atypical combination exerts a negative effect over time. In the context of motion pictures, this negative force overcomes the positive effect of atypicality on individual movies. Whether this is also the case in the contexts of other creative works depends on the relative sizes of the two effects. On citation networks, Uzzi et al. (2013) found a very large positive effect of atypicality on the scientific impact of individual papers. Given this, some level of atypical combination (which leads to cross-disciplinary research) is likely beneficial overall. However, an over-emphasis on a cross-disciplinary approach has double drawbacks: it raises little impact for individual papers and introduces a lot of commonality across papers, suppressing the development of specialized fields.

The counterfactuals also produce a seemingly contradictory result on the diversity at the individual versus collective level. The result actually can be related to social networks. Scholars have been familiar with opposing views on the cultural impact of globalization. The process of globalization brings together individuals with very different backgrounds and lifestyles. Although some people see it creating new cultural mixes, some people see it causing a homogenized society that mostly evolves around one dominating culture (Kraidy 2005). This bears quite a resemblance, at least at a conceptual level, with the counterfactual finding.

7. Concluding Remarks

Although the analyses in this paper have been tailored toward the motion picture industry, the modeling approach and most concepts (e.g., the core-peripheral structure, clustering, communities, atypicality) can be applied to other complex market structures. For example, there are millions of mobile apps listed on Google Play or Apple's App Store, many of which have similar functions. Do apps follow a common trend or diverge into distinctive groups? Is atypicality rewarded by a higher chance of success? To what extent should apps differentiate versus imitate each other? Is it more important for developers to innovate or refine an existing idea? What are some appropriate policies for the platforms to promote app quality as well as diversity? Similar questions can be asked about the hundreds of thousands of venture projects on crowdfunding platforms. An appealing aspect of crowdfunding platforms is that the similarity between two projects can be directly measured using text-mining techniques. More generally, because network tools are designed to tackle complex relational patterns across a large number of individual objects, in the era of big data and the long tail (Anderson 2006), it seems more appropriate than ever to deploy these tools in the study of markets.

Appendix

A.1. Construction of the Similarity Network

This part of the appendix details the construction of the similarity network, an outline of which was given in Section 2.2. The network is constructed using two data sets: the primary data set is the individual ratings provided by MovieLens (www.grouplens.org), and the other data set includes the movie suggestions on IMDb and Amazon Instant Video (scraped anonymously by me in October 2015).

The similarity score between two items is commonly calculated as the correlation or a correlation-like measure between the individual ratings for the two items (Desrosiers and Karypis 2011). Table A.1 provides summary statistics on the number of common raters (i.e., the individuals who rated both movies in a pair) in the MovieLens data. Given that my sample contains $n = 4,445$ movies, there are $n(n-1)/2 \simeq 9.88$ million movie pairs. About one third of the pairs have more than 100 common raters. For these pairs, a fairly precise similarity measure can be calculated directly from the MovieLens data. About one fourth of the pairs have fewer than 10 common raters. This is when the IMDb/Amazon data are most helpful.

To integrate the information in the two data sets, I treat the similarity as a latent factor; both the MovieLens ratings and the IMDb/Amazon recommendations are manifests of the latent similarities. This latent factor model also allows me to correct the potential biases in IMDb/Amazon recommendations

Table A.1. Common Raters in MovieLens Data

	All pairs	Pairs in/after 1995
$k \geq 100$	32.7%	31.1%
$k \geq 10$	75.4%	76.6%
$k \geq 10$ or both appear on Amazon	92.4%	94.7%
$k \geq 10$ or both appear on IMDb	100%	100%
Total count	9.877e6	4.739e6

Notes. k is the number of the individuals in the MovieLens data who have rated both movies in the pair. "Pairs in/after 1995" refer to the pairs in which the starting dates of both movies are in or after 1995.

introduced by factors such as the popularities and ages of the movies. I lay out the technical details as follows. However, it is important to note that my goal here is to provide a sensible way to integrate the two data sets, not a statistical model in the strict sense.

Fix a pair of movies. Let $\hat{c}$ denote the correlation between the ratings for the two movies in the MovieLens data. This correlation is calculated based on a sample of k common raters. Let the population correlation be c. For inferences, it is usually easier to work with the transformations $\hat{s} \equiv \operatorname{atanh}(\hat{c})$ and $s \equiv \operatorname{atanh}(c)$ (the atanh function is a strictly increasing mapping from $(-1,1)$ to $\mathbb{R}$). I regard s as the measure of similarity between the two movies. By Fisher approximation,

$$\hat{s} \sim \mathcal{N}\left(s, \frac{1}{k-3}\right). \tag{A.1}$$

Intuitively, the larger k is, the more precisely $\hat{s}$ measures s. Without further information, $\hat{s}$ would be the best estimate of s. However, $\hat{s}$ is not the only information I have about s: the recommendations on IMDb and Amazon also depend on s. Let $g \in \{0,1\}$ denote whether there is a recommendation between the two movies in the pair on either IMDb or Amazon. It is reasonable to expect that the chance for $g = 1$ increases with s, so I assume a probit model:

$$\Pr(g = 1|s) = \Phi(\tau s + \psi' w), \tag{A.2}$$

where ψ and τ are parameters to be estimated and w is a vector of additional observed factors that may affect g, such as the ages and popularities of the two movies. Here, the most straightforward way to estimate the parameters is the maximum likelihood (MLE). The likelihood of jointly observing $\hat{s}$ and g is given by

$$\Pr(\hat{s}, g) = \int_{-\infty}^{+\infty} \Pr(g|s) \cdot \Pr(\hat{s}|s) \cdot \Pr(s) ds,$$

where $\Pr(g|s)$ is given by (A.2) and $\Pr(\hat{s}|s)$ is given by (A.1). $\Pr(s)$ is a prior, which I specify to be a normal distribution whose mean and standard deviation are to be estimated as parameters. The objective function for the MLE sums the log of $\Pr(\hat{s}, g)$ over the $\frac{n(n-1)}{2}$ movie pairs. It is useful to mention that the preceding integral has a closed-form expression, which helps speed up the computation of MLE. However, the expression is quite lengthy, so I omit it here.

After the parameter values are estimated, I can compute my best estimate of s given both $\hat{s}$ and g:

$$\begin{aligned}\mathbf{E}(s|\hat{s}, g) &= \int_{-\infty}^{+\infty} s \Pr(s|\hat{s}, g) ds \\ &= \frac{1}{\Pr(\hat{s}, g)} \int_{-\infty}^{+\infty} s \Pr(\hat{s}, g|s) \cdot \Pr(s) ds \\ &= \frac{1}{\Pr(\hat{s}, g)} \int_{-\infty}^{+\infty} s \Pr(g|s) \cdot \Pr(\hat{s}|s) \cdot \Pr(s) ds.\end{aligned}$$

It can be shown that $\mathbf{E}(s|\hat{s}, g)$ is increasing with both $\hat{s}$ and g, which is intuitive.

Table A.2 displays the parameter estimates from MLE. Coefficient τ is significant and large, indicating that similarity is a driving force behind the movie recommendations made on IMDb/Amazon. The coefficient for the pair's average age is negative; the coefficient for the pair's popularity is positive. The signs are expected because the websites can increase traffic/sales by directing consumers to the more recent and popular items.

Given the parameter values, I compute the value of $\mathbf{E}(s|\hat{s}, g)$ for each movie pair. I define that a pair is similar if $\mathbf{E}(s|\hat{s}, g) > 0.45$, which is approximately the 3% right-tail cutoff of the prior, $\Pr(s)$. This also gives me a similarity network whose density is roughly at the same level as the network in which the link is

Table A.2. Parameter Estimates for Construction of the Similarity Network

	Point estimate	Standard error
Prior mean	0.2405	(5.87e-5)
Prior standard deviation	0.1140	(5.13e-5)
τ	5.414	(0.027)
ψ, constant	–5.429	(0.015)
ψ, pair's average age	–0.06444	(0.0015)
ψ, pair's average popularity	0.2513	(0.0012)
N	9.88×10^6	

Notes. The pair's average popularity is proxied by the log number of common raters. An alternative proxy is the log number of raters for one movie plus the log number of raters for the other movie; this does not change the coefficient estimates significantly.

defined directly by whether there is an IMDb or Amazon recommendation between the two movies. The main results of the paper are not sensitive to the cutoff choice (see Section A.5).

A.2. Studio's Belief

The goal of this part of the appendix is to derive the general expression for $\Pr(\xi_j|\mathcal{F}_t)$, the studio's belief of the latent quality of a candidate movie. Because the exercises in the paper require repeated computation of this belief, I also discuss ways to speed up the computation.

Fix a time point t. Use $R = \{k : r_k < t\}$ to denote the set of released movies and $Q = \{k : a_k \le t, r_k \ge t\}$ to denote the set of the yet-to-be-released movies (i.e., still in production). Notice that if there is a candidate movie j that arrives at t, then Q also includes that candidate at t. Use ξ_R to denote $\{\xi_k, k \in R\}$ and similarly for ξ_Q. Recall that $\mathcal{F}_t$ includes ξ_R but not ξ_Q. Also notice that $(\mathcal{F}_t, \xi_Q)$ contains $\mathcal{S}_t$. The probability of reaching a particular $(\mathcal{F}_t, \xi_Q)$ can be written as

$$\Pr(\mathcal{F}_t, \xi_Q) = \Psi(\mathcal{F}_t) \cdot \prod_{k\in Q\cup R} \Pr(\xi_k|x_k, y_k, \mathcal{S}_{a_k}).$$

In this expression, $\Pr(\xi_k|x_k, y_k, \mathcal{S}_{a_k})$ is the generating probability of ξ_k as specified in (3). The first term, Ψ, represents the probability of every element other than the latent qualities (e.g., the candidate arrival, link formation, generation of the budget size, and investment decisions). Because none of these elements depends on ξ_Q, I can write Ψ as a function of $\mathcal{F}_t$ only. However, Ψ does depend on ξ_R.

By the definition of conditional density, I have

$$\begin{aligned}\Pr(\xi_Q|\mathcal{F}_t) &= \Pr(\mathcal{F}_t, \xi_Q) \cdot \left[\int \Pr(\mathcal{F}_t, \xi_Q) d\xi_Q\right]^{-1} \\ &= \left[\prod_{k\in Q\cup R} \Pr(\xi_k|x_k, y_k, \mathcal{S}_{a_k})\right] \\ &\quad \cdot \left[\int \prod_{k\in Q\cup R} \Pr(\xi_k|x_k, y_k, \mathcal{S}_{a_k}) d\xi_Q\right]^{-1}.\end{aligned}$$

Notice that, as specified in (3), each $\Pr(\xi_k|x_k, y_k, \mathcal{S}_{a_k})$ is a normal distribution that depends on some other earlier entries in $\xi_{Q\cup R}$. As a result, when seen as a function of $\xi_{Q\cup R}$, the product term $\prod_{k\in Q\cup R} \Pr(\xi_k|x_k, y_k, S_{a_k})$ coincides with the probability density of a multivariate normal distribution (Ben-Gal 2007). Let me denote this density by $p(\xi_{Q\cup R})$, so I have $\Pr(\xi_Q|\mathcal{F}_t) = p(\xi_Q|\xi_R)$.

One representation of the density $p(\xi_{Q\cup R})$ can be derived as follows. Index movies in $Q \cup R$ by arrival time so that $a_1 \le \ldots \le a_{k-1} \le a_k \le \ldots$ Let V be a diagonal matrix and W be a lower triangular matrix, in which, for all $\ell, k \in Q \cup R$ and $\ell < k$,

$$\begin{aligned}V_{kk} &= \left(1 + \lambda \sum_{\ell\in Q\cup R} Y_{\ell,k}\phi^{|r_\ell - r_k|}\right)^{-1}, \\ W_{k\ell} &= \lambda Y_{\ell,k}\phi^{|r_\ell - r_k|} V_{kk}.\end{aligned}$$

Then, for all $k \in Q \cup R$,

$$\xi_k = \sum_{\ell\in Q\cup R} W_{k\ell}\xi_\ell + v_k, \quad v_k \sim \mathcal{N}(0, \sigma^2 V_{kk}).$$

In matrix form, we can write $\xi = W\xi + v$, which implies that

$$p(\xi_{Q\cup R}) \sim \mathcal{N}\left[0, \sigma^2 (I - W)^{-1} V (I - W')^{-1}\right].$$

Both $p(\xi_Q|\xi_R)$ and $p(\xi_j|\xi_R)$ can be readily computed from $p(\xi_{Q\cup R})$ using the standard formula for conditional normal distribution.

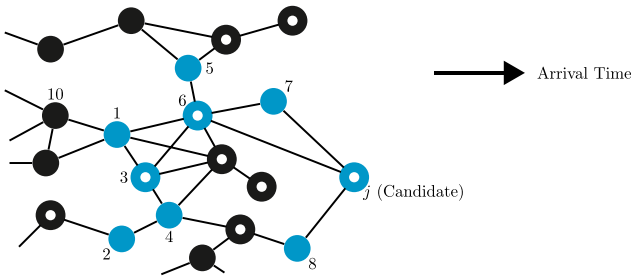
For the computation of $p(\xi_j|\xi_R)$, the inversion of $(I - W)$ can be a very time-consuming step, especially when the size of $Q \cup R$ is large. There is a way to significantly simplify the computation. The idea is to restrict attention to a subset of $Q \cup R$ such that the nodes outside this subset provide no further information toward ξ_j.

The idea can be illustrated with the example in Figure A.1. The nodes are placed from left to right by arrival date. A hollow node indicates that the movie has not been released yet; a solid node indicates otherwise. Only the blue-colored nodes are needed to compute $p(\xi_j|\xi_R)$. For example, the value of ξ_{10} is not needed because the only way that the realization of ξ_{10} can affect ξ_j is through ξ_1, whose value is already known by t. On the other hand, ξ_2 is needed because it is informative about the unknown value of ξ_3, which is informative about the unknown ξ_6, which is then informative about ξ_j. Instead of working with the original definitions of R and Q, I can redefine R as the set of blue solid nodes and Q as the set of blue hollow nodes. It can be shown that $p(\xi_j|\xi_R)$ remains the same after the redefinitions.

A.3. Monte Carlo and Bootstrap

I use a Monte Carlo experiment to check whether my estimation algorithm (Section 5) is able to recover the model parameters. First, I set the "true" parameter values at their point estimates in Tables 7 and 8. Next, I simulate the model from 1995 to 2012 under these parameter values, conditional on the real data before 1995. Finally, I apply the estimation algorithm in Section 5 to this simulated data set. The experiment is repeated 25 times to evaluate the distribution of the estimator. The biases (i.e., the difference between the estimator's mean and the true parameter value) are found to be small; the ratios between the biases and

Figure A.1. (Color online) Which Nodes Are (Not) Needed to Compute $p(\xi_j|\xi_R)$?



Notes. Solid nodes represent released movies; hollow nodes represent movies that are currently in production. The horizontal position of a node indicates its arrival time.

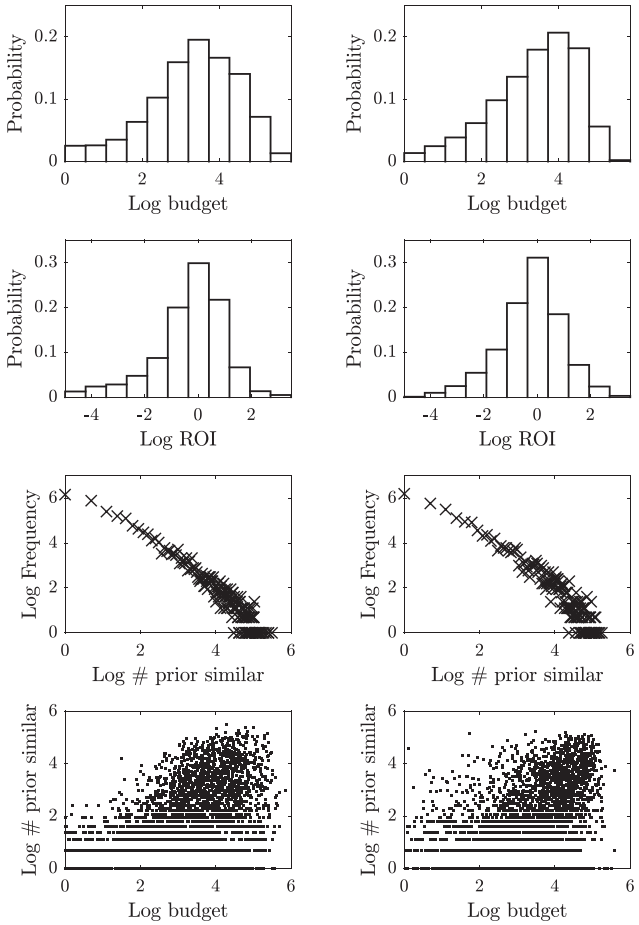
the parameter values all fall between −5% and 5%. The standard deviations of the estimator are reported as the bootstrapped standard errors in Tables 7 and 8.

A.4. Model Fit

I examine how well the estimated model reproduces four key data distributions: (i) the budget distribution, (ii) the distribution of ROI, (iii) the distribution of the number of prior similar movies, and (iv) the joint distribution between the budget and the number of prior similar movies. Specifically, for each movie j in/after 1995 in the data, I simulate a model-predicted counterpart of j conditional on the data up to time a_{j-1}. The simulation starts at a_{j-1} and takes the first accepted candidate as the model-predicted counterpart of j. In Figure A.2, the plots on the left side show the distributions in the real data, and the plots on the right side show the distributions of the simulated counterparts.

The fit is exceptional considering the simplicity of the model. Admittedly, there are patterns in the data that the model fails to perfectly reproduce. For example, the model seems to under-produce the very big-budget as well as the very small-budget movies. This is probably because the model assumes a single coefficient of risk aversion although, in reality, there is an array of production companies that are heterogeneous in terms of financial and risk capacities. Keep in mind that the data set contains a diverse set of movies over a long period of time as well as a complex network structure among these movies. Thus, it is not

Figure A.2. Model Fit



Notes. The four plots on the left side are based on the data, and the four plots on the right side are based on the simulation. From top to bottom, each plot shows, respectively, (i) the distribution of the log budget, (ii) the distribution of log ROI, (iii) the distribution of the log number of prior similar movies, and (iv) the scatter plot of the log number of prior similar movies against the log budget.

Table A.3. Demand-side Parameter Estimates with Data/Model Variations

	Benchmark	Ex ante network	Sampled network	Sparser network	Denser network	Alternative imitation
Log budget	1.05 (0.03)	1.04	1.06	1.06	1.05	1.05
Trend	−0.0085 (0.008)	−0.0090	−0.0111	−0.0108	−0.0062	−0.0085
Seasonality	0.166 (0.04)	0.149	0.152	0.179	0.163	0.166
Balanced atypicality	0.472 (0.13)	0.509	0.473	0.448	0.442	0.472
Parameters for ξ:						
Similarity weight (λ)	0.274 (0.06)	0.295	0.261	0.268	0.271	0.274
Discounting (ϕ)	0.911 (0.04)	0.920	0.929	0.919	0.904	0.911
Standard deviation (σ)	1.74 (0.04)	1.75	1.73	1.72	1.75	1.74
R^2 (log box office)	0.564	0.57	0.556	0.562	0.565	0.564
R^2 (log ROI)	0.147	0.158	0.138	0.143	0.149	0.147

Notes. The benchmark estimates are copied from column (3) of Table 7. Numbers in parentheses are standard errors.

Table A.4. Supply-side Parameter Estimates with Data/Model Variations

Parameters	Benchmark	Ex ante network	Sampled network	Sparser network	Denser network	Alternative imitation
Yearly arrivals (η)	407.4 (19.3)	423.8	366.1	403.1	406.6	407.9
Similarity:						
Innovation prob. (ι)	0.189 (0.014)	0.168	0.192	0.206	0.168	0.191
Older movies (θ)	−0.153 (0.004)	−0.150	−0.158	−0.158	−0.144	−0.153
Degree (ν)	0.887 (0.014)	0.861	0.892	0.896	0.882	0.889
Cluster (ω)	1.568 (0.067)	1.617	1.520	1.610	1.528	1.493
Atypicality (γ)	0.762 (0.008)	0.742	0.756	0.755	0.762	0.765
Budget distribution:						
Mean for novel (μ)	75.8 (9.8)	78.6	67.3	67.3	75.4	75.4
Dispersion (χ)	6.32 (0.19)	5.86	6.54	6.30	6.37	6.34
Go/no-go decision:						
Risk aversion (α)	0.0163 (0.003)	0.0174	0.0167	0.0156	0.0165	0.0161
Scale of shocks (ρ)	0.299 (0.034)	0.295	0.292	0.284	0.293	0.301

Note. The benchmark estimates are copied from Table 8.

surprising, given the level of model abstraction, that there are some details in the data that the model does not capture. Enriching the model is left for future research.

A.5. Robustness Checks

I estimate several variations of the data/model and recompute the counterfactual exercises accordingly. The goal is to see how sensitive the paper's conclusions are with respect to these variations. The results are displayed in Tables A.3–A.5. In each table, the first column reproduces the benchmark results, and each subsequent column displays one of the data/model variations. In summary, in all the variations considered, there are no qualitative changes in the model estimates or counterfactual results compared with the benchmark. I give the details of each variation.

The first variation ("ex ante network") is motivated by a concern over the revealed-preference approach to construct the network. Recall that the main reason for me to adopt this approach is that it captures similarities comprehensively, particularly with regards to the unobserved (or difficult-to-quantify) movie characteristics. However, to the extent that the consumer ratings are generated after movie releases, the revealed similarities may be shaped by postrelease shocks and, thus, different from the ex ante similarities that studios could perceive at the time of green-lighting decisions. In this sense, the revealed similarities may be "too comprehensive." Given this, a natural robustness check is to modify the network so that it relies less on the revealed similarities but more on the observed characteristic similarities. This makes the network less prone to postrelease shocks if there are any. (At the same time, it adds a bias toward observed characteristic similarities over unobserved characteristic similarities, so there is a trade-off.)

More specifically, the modified network is constructed as follows. First, I regress the similarity estimate $\mathbf{E}(s|\hat{s},g)$ (as defined in Section A.1) on the observed characteristics of movie pairs, such as whether the pair share leading actors and whether the pair's genres overlap (as in Table 1). I use the regression to compute an adjusted value for $\mathbf{E}(s|\hat{s},g)$ by scaling up the coefficients for the regressors by 50%. Rather than assigning the top percentiles of the original values of $\mathbf{E}(s|\hat{s},g)$ as network links, I assign the top percentiles of the adjusted values as the links. The difference between this network and the benchmark network is nontrivial (Jaccard index 0.85).

The motivation for the second variation ("sampled network") is to see how data incompleteness would affect the conclusions of the paper. It reestimates the model with a large 90% random sample of the 4,445 movies used in the benchmark estimation. Intuitively, the network becomes more fragmented (some paths

Table A.5. Counterfactual Risk-adjusted ROI with Data/Model Variations

Probability of atypical combination (γ)	Benchmark (%)	Ex ante network	Sampled network	Sparser network	Denser network	Alternative imitation
0	8.78 (0.10)	8.13	8.32	9.64	9.38	8.74
0.05	7.82 (0.09)	7.54	7.37	8.89	8.4	7.81
0.2	6.09 (0.10)	5.81	5.66	7.23	6.64	6.28
0.5	3.12 (0.07)	2.95	2.64	2.02	1.4	3.32
0.9	−2.92 (0.08)	−2.87	−3.08	−1.12	−1.78	−2.38

Notes. The table displays the average risk-adjusted ROI (expressed in percentage return) of the accepted movies in the counterfactual. The numbers in the benchmark column are copied from the first column in Table 9.

connecting nodes in the full-sample network are broken), and the average atypicality will increase. This is indeed reflected in the model estimates. For an average movie candidate, the ratio between the number of atypical imitations (calibrated by γ) and the total number of imitations (given by η and ν) is increased compared with the benchmark. However, the main conclusions of the paper are robust: (i) the demand effect of balanced atypicality is still significant (Table A.3), and (ii) in the long run, the risk-adjusted ROI is still decreasing in γ (Table A.5).

The purpose of the third and fourth variations is to see whether the results are sensitive to the cutoff choice in the construction of the network. This cutoff applies to the similarity estimate $\mathbf{E}(s|\hat{s}, g)$ and defines which level of similarity warrants a link between a movie pair. The "sparser network" uses a higher cutoff so that its density is 10% lower than the benchmark network. The "denser network" uses a lower cutoff so that its density is 10% higher than the benchmark.

The last variation tests the robustness with respect to how atypical combination is modeled. Recall that, in the benchmark case, the atypical imitations are modeled as rewiring of links after the typical imitations. As an alternative, here I specify that atypical and typical imitations happen in parallel and independently of each other. Specifically, atypical links are formed by randomly selecting a set of existing nodes with the probability of selecting a particular node k proportional to $\exp(\theta|t - a_k|)$. The cardinality of the set is drawn from the geometric distribution (the discrete analog of exponential distribution) with parameter $1 - \gamma$. Typical links are formed in the same way as before. The number of typical links equals the difference between m^* and the number of atypical links. This robustness check shows that the conclusions of the paper are not restricted to the specific way of modeling atypical combinations in the benchmark model.

Acknowledgments

This paper was developed from a chapter of the author's PhD dissertation at the University of Pennsylvania, and an earlier version was presented under the title "Imitation vs. Innovation: Product Similarity Network in the Motion Picture Industry." The author thanks the author's committee for guidance: Holger Sieg, Eric Bradlow, Joseph Harrington, Katja Seim, and Christophe Van den Bulte. For helpful suggestions and also not blaming them for any mistake in the paper, the author thanks Ron Berman, Anand V. Bodapati, Anthony Dukes, Jehoshua Eliashberg, Hanming Fang, Devin Reilly, Hongxun Ruan, Jagmohan Raju, Francisco Silva, Qiaowei Shen, Rakesh Vohra, Pinar Yildirim, and Brian Uzzi; as well as seminar participants at the 37th Marketing Science Conference, UPenn, Rochester, University of Pennsylvania, Rochester, University of California at Los Angeles, University of Southern California, University of California at San Diego, New York University, Hong Kong University of Science and Technology, University of California at Berkeley.

Endnotes

[1] See, for example, Mednick (1962), Price (1976), Becker (1982), Finke et al. (1992), Weitzman (1998), Uzzi et al. (2013), Toubia and Netzer (2016), and Strang (2016).

[2] In marketing, studies on product networks (in contrast to social networks) include Dellarocas et al. (2010), Goldenberg and Reuchman (2012), Oestreicher-Singer and Sundararajan (2012), and Wei (2018). Outside marketing, a notable study on product networks is Hidalgo et al. (2007). Generally speaking, much less attention has been given to the networks of products compared with social networks.

[3] There have been many discussions on movie imitation. For example, see Squire (2005). In media, see "Hollywood Learns Originality Does Not Pay" in *Financial Times* (May 29, 2015) and "Are Blockbusters Destroying the Movies?" in *the New York Times* (January 6, 2015). On movie uncertainty, see De Vany and Walls (1996, 1999).

[4] There is a large literature in marketing on the movie industry, but the vast majority of studies focus on the postproduction phase; few focus on modeling the earlier investment decisions (Delre et al. 2017).

[5] See Eliashberg et al. (2007) for an application of text analysis to quantify a movie's synopsis.

[6] Goettler and Leslie (2005) include microbudget movies, which results in a somewhat larger sample size per year.

[7] There might be the question of how time-varying consumer preferences factor into the construction of the similarity network. In terms of the characteristic space, the time-varying feature of consumer preferences corresponds to the drifting of the distribution of the consumers' ideal points. The locations of movies, however, do not change over time. So it holds true that if two movies are close (i.e., similar), a consumer who likes one of them will tend to like the other even though the aggregate consumer distribution may change.

[8] Outside marketing, see Hidalgo et al. (2007), which also applies the idea of using outcome data to uncover relations between products.

[9] It is common in the literature to rely on domestic box-office revenues and production budgets to measure ROI although movies also collect revenues from subsequent markets (e.g., international and home video). One reason is that the revenues in subsequent markets are highly correlated with domestic box-office numbers (Goettler and Leslie 2005, Einav 2007). The other reason is that the data coverage on subsequent markets is relatively poor.

[10] If the log box-office revenue is used as the dependent variable in Table 5, all the coefficients stay the same except for the one in front of log budget, which increases by exactly one. The values of R^2 will rise above 0.5. However, I choose to focus on the ROI because it is a closer measure for the success of a movie; a movie making a large revenue is still considered unsuccessful if the costs are even larger (often known as "box-office bombs").

[11] For a survey on studies of firm learning in marketing, see Ching et al. (2017).

[12] Here I treat a movie's release as a point in time. In reality, a movie typically stays in theaters for six to eight weeks with the first two weeks being most important, collecting about 60% of the movie's lifetime domestic box-office revenue. Given that my data spans a period of decades, treating a few weeks as a time point seems reasonable.

[13] I treat m_t as an exogenous time series. It is a known fact (as well as a puzzle) that ticket price hardly varies across seasons and movies; see

Orbach and Einav (2007) for more discussions. In addition, results are not sensitive to the choice of the moviegoer population size.

[14] A technical issue here is that unless $\nu = 1$, m^* is generally not an integer. In this case, I pick one of the two integers closest to m^*; the probability of picking the larger integer is proportional to the distance between m^* and the smaller integer.

[15] Taken from Eliashberg et al. (2006).

[16] Strictly speaking, the studio is uncertain of m_{r_j}. However, this is a very small source of uncertainty compared with ξ_j. So I assume that the studio can perfectly foresee m_{r_j} at the time of a_j. Alternatively, I may assume that the studio uses m_{a_j} as the prediction of m_{r_j}, which turns out to make little difference in the estimates or counterfactuals.

[17] See Luo (2014) for a study on the market of movie scripts (which, unfortunately, also lacks data on rejected scripts).

References

Ainslie A, Drèze X, Zufryden F (2005) Modeling movie life cycles and market share. *Marketing Sci.* 24(3):508–517.

Anderson C (2006) *The Long Tail: Why the Future of Business Is Selling Less of More* (Hyperion Books, New York).

Barabási AL, Albert R (1999) Emergence of scaling in random networks. *Science* 286(5439):509–512.

Becker HS (1982) *Art Worlds* (University of California Press, Berkeley).

Ben-Gal I (2007) Bayesian networks. Ruggeri F, Kennett R, Faltin F, eds. *Encyclopedia of Statistics in Quality and Reliability* (Wiley & Sons Hoboken, NJ), 179–183.

Bradlow ET, Bronnenberg B, Russell GJ, Arora N, Bell DR, Duvvuri SD, Hofstede FT, Sismeiro C, Thomadsen R, Yang S (2005) Spatial models in marketing. *Marketing Lett.* 16(3–4): 267–278.

Bramoulle Y, Currarini S, Jackson M, Pinh P, Rogers B (2012) Homophily and long-run integration in social networks. *J. Econom. Theory* 147(5):1754–1786.

Ching A, Erdem T, Keane MP (2017) Empirical models of learning dynamics: A survey of recent developments. Wierenga B, ed. *Handbook of Marketing Decision Models* (Springer, New York), 223–257.

Chintagunta P (1994) Heterogeneous logit model implications for brand positioning. *J. Marketing Res.* 31(2):304–311.

Chintagunta P, Gopinath S, Venkataraman S (2010) The effects of online user reviews on movie box office performance: Accounting for sequential rollout and aggregation across local markets. *Marketing Sci.* 29(5):944–957.

De Vany A, Walls D (1996) Bose-Einstein dynamics and adaptive contracting in the motion picture industry. *Econom. J.* 106(439): 1493–1514.

De Vany A, Walls D (1999) Uncertainty in the movie industry: Does star power reduce the terror of the box office? *J. Cultural Econom.* 23(4):285–318.

Dellarocas C, Katona Z, Rand W (2010) Media, aggregators, and the link economy: Strategic hyperlink formation in content networks. *Management Sci.* 59(10):2360–2379.

Delre SA, Panico C, Wierenga B (2017) Competitive strategies in the motion picture industry: An ABM to study investment decisions. *Internat. J. Res. Marketing* 34(1):69–99.

Desrosiers C, Karypis G (2011) A comprehensive survey of neighborhood-based recommendation methods. Ricci F, Rokach L, Shapira B, Kantor PB, eds. *Recommender Systems Handbook* (Springer, New York), 107–144.

Einav L (2007) Seasonality in the U.S. motion picture industry. *RAND J. Econom.* 38(1):127–145.

Einav L (2010) Not all rivals look alike: Estimating an equilibrium model of the release date timing game. *Econom. Inquiry* 48(2): 369–390.

Elberse A, Eliashberg J (2003) Demand and supply dynamics for sequentially released products in international markets: The case of motion pictures. *Marketing Sci.* 22(3):329–354.

Eliashberg J, Elberse A, Leenders M (2006) The motion picture industry: Critical issues in practice, current research, and new research directions. *Marketing Sci.* 25(6):638–661.

Eliashberg J, Hui S, Zhang J (2007) From story line to box office: A new approach for green-lighting movie scripts. *Management Sci.* 53(6):881–893.

Elrod T, Keane MP (1995) A factor-analytic probit model for representing the market structure in panel data. *J. Marketing Res.* 32(1): 1–16.

Finke RA, Smith SM, Ward TB (1992) *Creative Cognition* (MIT Press, Cambridge, MA).

Fruchterman T, Reingold E (1991) Graph drawing by force-directed placement. *Software Practice Experience* 21(11): 1129–1164.

Goettler R, Leslie P (2005) Cofinancing to manage risk in the motion picture industry. *J. Econom. Management Strategy* 14(2):231–261.

Goettler R, Shachar R (2001) Spatial competition in the network television industry. *RAND J. Econom.* 32(4):624–656.

Goldenberg J, Oestreicher-Singer G, Reuchman S (2012) The quest for content: How user-generated links can facilitate online exploration. *J. Marketing Res.* 49(4):452–468.

Hennig-Thurau T, Houston MB, Sridhar S (2006) Can good marketing carry a bad product? Evidence from the motion picture industry. *Marketing Lett.* 17(3):205–219.

Hidalgo CA, Klinger B, Barabasi A, Hausmann R (2007) The product space conditions the development of nations. *Science* 317(5837): 482–487.

Holme P, Kim BJ (2002) Growing scale-free networks with tunable clustering. *Phys. Rev. E* 65(2):026107.

Jackson MO (2008) *Social and Economic Networks* (Princeton University Press, Princeton, NJ).

Jackson MO, Rogers B (2007) Meeting strangers and friends of friends: How random are social networks? *Amer. Econom. Rev.* 97(3): 890–915

Kleinberg JM, Kumar R, Raghavan P, Rajagopalan S, Tomkins A (1999) The web as a graph: Measurements, models and methods. Asano T, Imai H, Lee DT, Nakano S-I, Tokuyama T, eds. *Computing and Combinatorics* (Springer, New York), 1–17. (Springer).

Kraidy MM (2005) *Hybridity, or the Cultural Logic of Globalization* (Temple University Press, Philadelphia).

Kumar R, Raghavan P, Rajagopalan S, Sivakumar D, Tomkins A, Upfal E (2000) Stochastic models for the web graph. *Proc. 41st Annual Sympos. Foundations Comput. Sci.* (IEEE, Piscataway, NJ), 57–65.

LeSage J (2008) An introduction to spatial econometrics. *Revue d'economie Industrielle* 123(3):19–44.

Linden G, Smith B, York J (2003) Amazon.com recommendations: Item-to-Item collaborative filtering. *IEEE Internet Comput.* 7(1): 76–80.

Liu Y (2006) Word of mouth for movies: Its dynamics and impact on box office revenue. *J. Marketing* 70(3):74–89.

Luo H (2014) When to sell your idea: Theory and evidence from the movie industry. *Management Sci.* 60(12):3067–3086.

Mednick S (1962) The associative basis of the creative process. *Psych. Rev.* 69(3):220–232.

Newman M (2006) Modularity and community structure in networks. *Proc. Natl. Acad. Sci. USA* 103(23):8577–8582.

Newman M (2010) *Networks: An Introduction* (Oxford University Press, Oxford, UK).

Oestreicher-Singer G, Sundararajan A (2012) The visible hand? Demand effects of recommendation networks in electronic markets. *Management Sci.* 58(11):1963–1981.

Orbach B, Einav L (2007) Uniform prices for differentiated goods: The case of the movie-theater industry. *Internat. Rev. Law Econom.* 27(2):129–153.

Price DDS (1965) Networks of scientific papers. *Science* 149(3683): 510–515.

Price DDS (1976) A general theory of bibliometric and other cumulative advantage processes. *J. Assoc. Inform. Sci. Tech.* 27(5):292–306.

Schumpeter JA (2017) *The Theory of Economic Development* (Routledge, Philadelphia).

Squire J (2005) *The Movie Business Book*, 3rd ed. (Simon and Schuster, New York).

Strang D (2016) *Learning by Example: Imitation and Innovation at a Global Bank* (Princeton University Press, Princeton, NJ).

Toubia O, Netzer O (2016) Idea generation, creativity, and prototypicality. *Marketing Sci.* 36(1):1–20.

Uzzi B, Mukherjee S, Stringer M, Jones B (2013) Atypical combinations and scientific impact. *Science* 342(6157):468–472.

Wei Y (2018) Airline networks, traffic densities, and value of links. *Quant. Marketing Econom.* 16(3):341–370.

Weitzman ML (1998) Recombinant growth. *Quart. J. Econom.* 113(2): 331–360.