Quant Mark Econ (2012) 10:111–150
DOI 10.1007/s11129-011-9105-4

Impact of social network structure on content propagation: A study using YouTube data

Hema Yoganarasimhan

Received: 9 November 2010 / Accepted: 12 May 2011 / Published online: 29 September 2011

Abstract We study how the size and structure of the local network around a node affects the aggregate diffusion of products seeded by it. We examine this in the context of YouTube, the popular video-sharing site. We address the endogeneity problems common to this setting by using a rich dataset and a careful estimation methodology. We empirically demonstrate that the size and structure of an author's local network is a significant driver of the popularity of videos seeded by her, even after controlling for observed and unobserved video characteristics, unobserved author characteristics, and endogenous network formation. Our findings are distinct from those in the peer effects literature, which examines neighborhood effects on individual behavior, since we document the causal relationship between a node's local network position and the global diffusion of products seeded by it. Our results provide guidelines for identifying seeds that provide the best return on investment, thereby aiding managers conducting buzz marketing campaigns on social media forums. Further, our study sheds light on the other substantive factors that affect video consumption on YouTube.

Keywords Social network · YouTube · Diffusion · Social media · User-generated content · Network structure · Online video · Social influence · Contagion

JEL C36 · C33 · M3 · O33 · L14

1 Introduction

In mid 2009, as Ford was preparing to launch its new subcompact car Ford Fiesta in the United States, it eschewed the traditional marketing approach and instead adopted a buzz campaign. It selected 100 social media savvy video bloggers (vloggers), gave them a Fiesta each, and asked them to document their experiences through videos, tweets, and blog entries (Barry 2009). This marketing campaign, called the 'Ford Fiesta Movement',[1]

[1]See http://www.fiestamovement2.com.

H. Yoganarasimhan (✉)
Graduate School of Management, University of California Davis, Davis, CA, USA
e-mail: hyoganarasimhan@ucdavis.edu

was very successful; by March 2010, Ford had generated 6.2 million YouTube views, over 750,000 Flickr views, and about 4 million Twitter impressions. More importantly, Fiesta received 100,000 hand-raisers and 6,000 reservations, half of which came from consumers who had never bought a Ford before (Greenberg 2010). An interesting aspect of this campaign was the choice of vloggers. Cadell, the main strategist behind the Fiesta Movement, states that their objective was to "find twenty-something YouTube storytellers who've learned how to earn a fan community of their own. [People] who can craft a true narrative inside video" (McCracken 2010). In short, Ford picked web-based influentials, gave them information, and encouraged them to spread it to their larger social network.

Seeding information in social media outlets through handpicked agents is now becoming a common strategy in buzz marketing campaigns. The identification of effective seeds is therefore not only key to the success of these campaigns, but also an important factor in estimating the return on investment (ROI) from a manager's perspective. Essentially, a good seed is someone who is capable of influencing others and spreading information efficiently. While many factors such as the expertise, experience, and the personality of a seed can influence her effectiveness, her position in the social network is arguably the most important factor. Notice that the first metric that Cadell mentions in his quote is the size of a vlogger's fan community. In other words, Cadell seems to consider well-connected seeds to be better disseminators of information than poorly connected ones. Size apart, the 'structure' of a seed's local network may also play a significant role in determining her influence. For example, two seeds may have the same number of connections, but one may be more influential or dominant in the network compared to the other. One may belong to a close-knit community, while the other could come from a sparsely connected one. Further, one may be situated close to the rest of the network, while the other could be structurally removed from the larger network. More generally, consider two nodes that occupy positions A and B in an arbitrary network (see Fig. 1)—if the same information were seeded at node A versus node B, how would its overall diffusion be different? In this paper, we seek to answer this question, i.e., we examine how the size and structure of a seed's local network affect its ability to disseminate information.

Though simple to state, this is a tricky question to answer empirically because of the endogeneity problems common to this setting. First, a node's social network position is likely to be correlated with other unobserved person-specific characteristics that also affect her ability to disseminate information. For example, we might find that a node with a large number of friends is a more effective seed than one with fewer friends. However, this does not establish a causal relationship between the number of friends a seed has and its effectiveness, because nodes with many friends are also likely to have more engaging personalities, greater expertise and experience in the product category, and an overall better reputation for dispensing good information—all of which also contribute to their effectiveness. Unless these correlated (and unobserved) personal and reputational attributes are explicitly controlled for, any results on the role of network position are likely to be biased. A second source of endogeneity stems from the correlation between a node's network position and unobserved product-specific attributes, especially if a seed's social network evolved as a result of her past activities. For example, consider a node that seeds high quality products. Such a node is likely to have garnered many

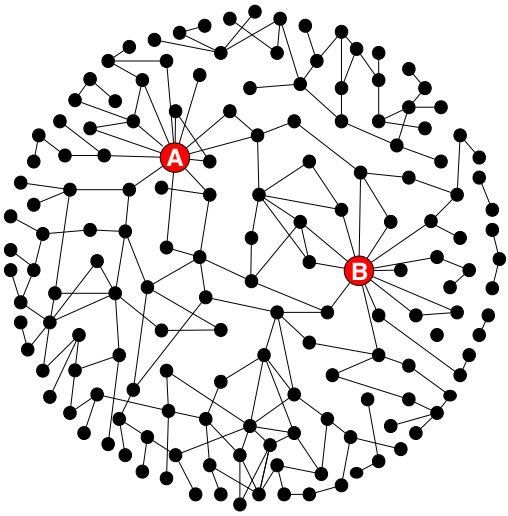


Fig. 1 Impact of the seed's network position on product diffusion

friends or become more central to the network over time. Moreover, a new product seeded by such a node is also likely to be of high quality and therefore has a higher chance of being more widely adopted. Hence, not controlling for unobserved product quality could also bias our results on the impact of network position.[2] These endogeneity problems make it econometrically difficult to infer a causal relationship between the network position of a node and the overall performance of products seeded by it.

These challenges can be addressed by using a rich dataset and a careful estimation methodology. We employ an extension of the dynamic panel data estimator developed by Blundell and Bond (1998) to resolve our endogeneity issues. A key advantage of this method is that it does not require external instruments. Instead, it allows us to use the lags and lagged differences of endogenous explanatory variables as instruments. This methodology has been successfully applied by researchers in a wide variety of fields within marketing and economics (see Acemoglu and Robinson 2001; Durlauf et al. 2005; Clark et al. 2009). Further, in our context, we extend this method to show that lagged differences of explanatory variables can be used as instruments for time-invariant endogenous variables also. Hence, we are able to control for both endogenous network structure and video properties.

This methodology can however be used only in a panel data framework. Hence, to establish causality, we need data on the diffusion of a large number of products seeded at different points in the network. Moreover, for each product

[2] In fact, unobserved product quality is problematic in other respects too. A high quality product is more likely to have a higher price, higher consumer ratings, and higher advertising expenditure, i.e., unobserved and observed product attributes are likely to be correlated. Hence, if the former is not controlled for, then our results on the role of observed product attributes are also likely to be biased.

we need multiple observations on its adoption. Further, we also require information on its network position of the seeds of all the products. While these are heavy requirements, data from YouTube, the popular video-sharing site, satisfies these conditions.

While YouTube is often perceived as a simple video-sharing site, in reality it also consists of an active social networking community of YouTube members. According to Hitwise Experian (2010), YouTube is the third most popular online social network (after Facebook and Myspace). In terms of functionalities, it resembles other online networks, with the added advantage of the video-sharing functionality. For example, YouTube members can friend other members and interact with them through tools such as comment-boxes, messages, and activity feed subscriptions.

In this setting, videos can be interpreted as products and users (or authors) who post them as seeds. Moreover, data on video performance and quality (in the form of views, ratings, and comments) and data on authors' social network (in the form of friendship ties) is also publicly available from YouTube. These factors make it an ideal setting for our study.

Our analysis reveals that the size and structure of an author's local network is a significant driver of the popularity of videos seeded by her, even after controlling for observed and unobserved video characteristics, unobserved seed characteristics, and endogenous network formation. We present four key results in this context. First, we find that both the first and second-degree connectivity of a seed has a positive impact on her product's diffusion. Further, our analysis suggests that the marginal benefit of a second-degree friend is higher than that of a first-degree friend. These results are in contrast to a recent simulation study by Watts and Dodds (2007), which found that the degree of influence of a node has no impact on the size of cascades generated by it. This discrepancy likely stems from the authors' use of simulated network structure and propagation rules, which might not reflect the structure and behavior of real life networks. Our findings, in contrast, are derived from careful empirical analysis and are based on the outcomes in a real network (YouTube).

Second, we find that high clustering in the author's local network is associated with low video popularity. High clustering around a node implies that she belongs to a close-knit community. While such a position guarantees the commitment and interest of the local peer-group, it can damage the global adoption of the product as members of a tight-knit community are less likely to interact with outsiders and inform them of the author's video. Third, we find that the local Betweenness of a node has a negative impact on the aggregate adoption of videos seeded by it. Betweenness embodies two opposing concepts: *network dominance* and *path diversity*; nodes with high Betweenness are dominant in their local network, which increases their ability to generate views. However, they also have fewer paths to reach the outer network, which decreases their influence over the larger network. Interestingly, we find that the latter effect dominates the former. Fourth, we find that the impact of network properties changes over time. First-degree friends of a seed are essential for initial take-off, but second-degree friends are responsible for later spread. Moreover, both Clustering and Betweenness dampen later growth, but do not harm initial growth. Further, specific to our context, we find that lagged video attributes such as ratings and comments have no impact on video viewership in the long run, though they aid initial diffusion.

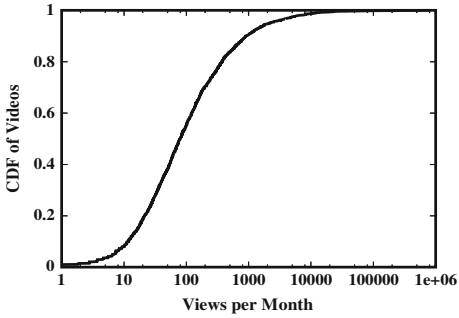


Fig. 2 Popularity distribution of YouTube videos

In sum, our key contributions are as follows. First, we empirically show that the network structure of a node affects the overall diffusion of the products seeded by it. Specifically, we demonstrate these results in the context of YouTube videos. Note that these findings are distinct from those on peer-effects. While there exist many studies on individual-level peer-effects, to our knowledge this is the first empirical study that documents the causal role of a seed's local network on macro-level diffusion (see Section 2 for details). Moreover, our focus on global diffusion allows us to explore the temporal differences in the impact of network properties, i.e., we show that network properties that drive early diffusion are fundamentally different from those that affect later diffusion. Second, we discuss and clarify the data requirements and methodological strategies required to overcome endogeneity problems in such settings. Specifically, we consider an extension of the system GMM estimation proposed by Blundell and Bond (1998) and demonstrate its effectiveness in resolving the endogeneity issues in the context of network data. Third, we use our estimates to help managers identify seeds that provide the best ROI. This is important because random selection of seeds is unlikely to fetch a good ROI; note that less than ten percent of videos get 1000 views or more in the first one month (see Fig. 2). Finally, our study sheds light on the substantive factors that affect video consumption in YouTube. While the online video market has grown tremendously in the last few years (e.g.,9.4 billion videos were streamed in April 2010 alone, Nielsen 2010), there are few formal studies on the subject, and our paper represents an important first step in this area.

The rest of the paper is organized as follows. Section 2 discusses the related literature. Section 3 describes the setting, data, and the social network properties of the authors. Sections 4 and 5 describe the model and estimation, while Section 6 discusses the main results. Section 7 examines factors that affect early and later growth. Section 8 discusses the managerial implications of the study and presents some counterfactual results. Finally Section 9 concludes with a discussion of the main findings, limitations, and suggestions for future research.

2 Related literature

Our paper relates to a large body of literature on social interactions from a wide variety of disciplines including economics, marketing, computer science, and

sociology. Two specific streams within this broader context are of particular relevance, and we discuss each in turn below.[3]

First consider the literature on peer-effects, which seeks to understand how friendship ties affect consumers' choices. In a seminal study, Coleman et al. (1966) showed that doctors were influenced by social ties in their decisions to adopt tetracycline, thereby providing the first empirical evidence on the existence of peer-effects. More recently, researchers have examined the impact of peer-effects in a variety of contexts—welfare participation (Bertrand et al. 2000), obesity (Trogdon et al. 2008), and workplace performance (Bandiera et al. 2009). In general, identification of peer-effects is problematic because of endogeneity problems such as endogenous group formation and peers' exposure to similar unobserved environmental factors (Manski 1993; Hartmann et al. 2008). This has led to some disagreement over the existence and magnitude of peer-effects. However, new methods to address these endogeneity problems have been developed (Brock and Durlauf 2007; Bramoullé et al. 2009) and most recent studies find some evidence in support of peer-effects (Sacerdote 2001; Bandeira and Rasul 2006; Nair et al. 2009).

There are several points of divergence between our paper and the peer-effects literature. First, the latter focuses on inter-personal social influence. It seeks to establish causality between the actions of two connected nodes: if C and D are connected, is C influenced by D's behavior and vice-versa? In contrast, we seek to establish causality between a node's network position and the overall diffusion of the products seeded by it. Second, the endogeneity issues that we address are very different from those faced by the peer-effects researchers. Third, we use aggregate panel data on multiple products to tackle our identification issues, whereas the peer-effects literature uses individual-level choice data to address its endogeneity problems. In sum, both the research question that we pose and the solution that we offer are fundamentally different from those in the peer-effects literature.

The second stream of literature that relates to our paper is that on opinion leaders. Past research defines opinion leaders as a small minority that exerts a strong influence on the opinions and decisions of the majority (Katz and Lazarsfeld 1955). The theory gained prominence in mid-twentieth century and continues to be influential today (Rogers 2003). Researchers apart, managers have been particularly fascinated with the idea of opinion leaders. For example, it is common practice in the pharmaceutical industry to recruit Key Opinion Leaders (KOLs) to promote new drugs (Moynihan 2008). In spite of this interest, there is no consensus on who are opinion leaders or how one identifies them (Valente and Pumpuang 2007). In general, the task of identifying opinion leaders and measuring their impact is tricky because there are many different qualities that can make someone an opinion leader—expertise, heavy usage, personality, demographics, and network position. In this paper, we focus on one of these factors—network position, i.e., we explain how the network structure around a node affects its opinion leadership.

[3] While exists a large stream of literature on Bass models (Bass 1969; Mahajan et al. 2000), these models cannot be used to establish causality between network structure and product diffusion because they assume random mixing or interactions over a fully connected network.

In general, recent studies on opinion leadership and network position have focused on identifying key players instrumental in the diffusion process. For example, Tucker (2008) documents the role of boundary-spanning players in the adoption of an intra-firm messaging technology, and Goldenberg et al. (2009) study the role of hubs in the diffusion of virtual goods in a Korean social network. However, these papers do not focus on the causal relationship between the point of origin of information and its cumulative diffusion. On a related note, using simulations, Watts and Dodds (2007) show that a node's degree (number of friends/influence) is not a key driver of diffusion. However, to the best of our knowledge, we know of no empirical studies that demonstrate causality between a node's local network position and the aggregate adoption of products seeded by it.

3 Setting and data

3.1 The setting—YouTube

We now provide some background on YouTube. YouTube was launched in 2005 and soon emerged as the most popular video-sharing site. In 2006, it was acquired by Google Inc. and has since become the 6th most visited website in the United States. In April 2010 alone, YouTube received 97 million unique visitors and streamed 4.9 billion videos (Nielson Online 2010).

An interesting aspect of YouTube is that it is not only a platform for sharing videos, but also a social network. While any Internet user can watch videos, YouTube members can also post videos and become friends with other members. These friendship ties are undirected, i.e., both parties have to be willing for the tie to be made. YouTube has many features that enhance interactions between friends and contribute to the community feel of the forum. When a member posts a video, her friends are notified immediately (as an update on their own YouTube pages). Similarly, when a member rates a video, comments on it, or 'Favorites' it, her friends are informed of her action and provided a link to the relevant video. Moreover, YouTube also allows friends to recommend videos to each other and share information by sending private messages or commenting publicly on each other's YouTube pages. These features have enabled a vibrant and active social network to thrive on YouTube. It is thus a unique forum, which provides data on both product diffusion and network structure, and therefore an ideal setting for our study.

3.2 Data

We collect two types of data—longitudinal data on a random panel of videos and data on the social network of the authors of the videos in the panel.

3.2.1 Data on videos

All YouTube videos are available to the general public, unless classified as private by the author. Typically, thousands of videos are posted every hour, and a public

listing of videos posted in the last few minutes is available on YouTube's website. YouTube also enables certain types of user feedback on videos. After viewing a video, members can comment on it and rate it on a scale of 1 to 5, with a higher rating implying higher appreciation. They can also 'Favorite' videos. Videos favorited by a user are displayed prominently on her YouTube page and remain there until she un-favorites them. Popular videos are 'Honored' by YouTube, i.e., they get tags like *6th Most Discussed (this week)*—indicating that it is the 6th most discussed video of the week or *29th Top Rated (today)*—indicating that it is the 29th highest rated video of the day. Honored videos are highlighted on YouTube's website and appear prominently in searches. Figure 3 is a screenshot of a video titled 'Application Review #44', which reviews applications for iPod Touch. The video statistics discussed above are seen in it.

To generate our dataset, we randomly picked 1939 videos from the list of recently uploaded videos in November 2007, and monitored them daily for 38 days (see Appendix A.1 for details). We chose 24 h as the interval of observation to account for the diurnal patterns in viewership. During the data collection process, some videos were removed or declared private. However, the overall process was

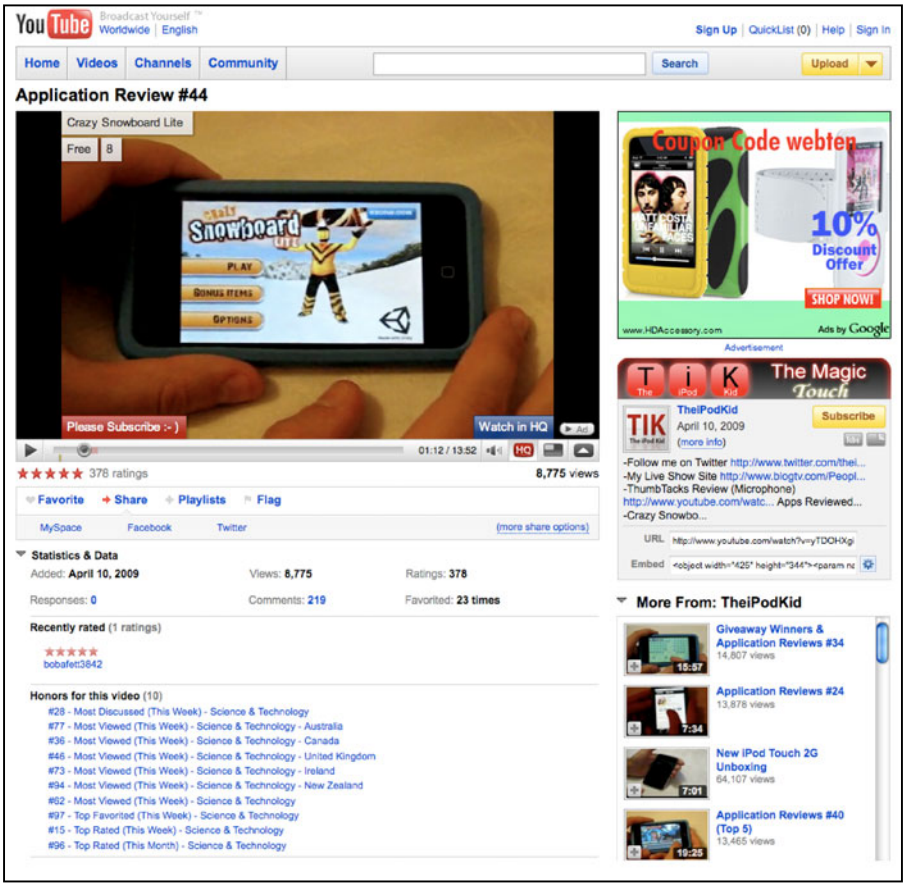


Fig. 3 Screenshot of a YouTube video's page

relatively smooth, and 85% of the videos have data for 31 days or more. During each observation of video i, we collected data on the following variables:

1) Views ($V_{i,t}$)—the total number of views that i has received at time t since its launch.
2) Num. Ratings ($Nr_{i,t}$)—the total number of ratings that i has received at time t since its launch.
3) Avg. Rating ($Ar_{i,t}$)—the average rating of video i at time period t.
4) Comments ($C_{i,t}$)—the total number of comments received by video i at time t since its launch.
5) Favorited ($f_{i,t}$)—the number of people who indicate that video i is a 'Favorite' at time period t.
6) Honors ($h_{i,t}$)—the number of Honors that a video has at time period t.

Note that Favorited and Honors are contemporaneous variables unlike Views, Num. Ratings, and Comments, which are cumulative variables. Videos often lose Honors received in the past; for example, 'Application Review #44', which was one of the most discussed videos a few days after its launch, lost this Honor after a couple of weeks. $h_{i,t}$ thus reflects the number of Honors video i has at period t and is not a count of all the Honors received until t. Favorited is also contemporaneous because a video's Favorited count takes into account only those users that are currently favoriting it.

From the primary variables discussed above, we constructed the following new variables:

7) Daily Views ($v_{i,t}$)—number of new views that i receives during time period t.

$$v_{i,t} = V_{i,t} - V_{i,t-1} \tag{1}$$

8) Indicator no Rating ($Inr_{i,t}$)—is an indicator of whether i has been rated at least once by period t.

$$Inr_{i,t} = \begin{cases} 1, \text{if} & Nr_{i,t} = 0 \\ 0, \text{if} & Nr_{i,t} > 0 \end{cases} \tag{2}$$

We use $(1 - Inr_{i,t}).Ar_{i,t}$ and $\text{Inr}_{i,t}$ together to capture the effect of ratings. This ensures that Avg. Rating ($Ar_{i,t}$) is not treated as a missing variable when a video hasn't been rated.

9) Daily Num. Ratings—the number of ratings that i receives during time period t.

$$\text{nr}_{i,t} = Nr_{i,t} - Nr_{i,t-1} \tag{3}$$

10) Daily Comments ($c_{i,t}$)—the number of comments that i receives during time period t.

$$c_{i,t} = C_{i,t} - C_{i,t-1} \tag{4}$$

Contemporaneous variables are denoted with small letters and cumulative variables with capital letters. Table 1 presents the summary statistics of the data at t = 10, 20, and 30. Table 2 presents the correlations between the video attributes.

3.2.2 Data on social network

The web pages of all videos contain links to the authors' YouTube pages. Further, the YouTube pages of authors contain information on their social network, i.e., an author's page has a listing of all the friends of the author. In our dataset of 1939 videos, we found that 1806 authors had publicly listed their

Table 1 Summary statistics of video data

Variables		t=10	t=20	t=30
Number of Videos		1618	1587	1547
Views	Min, 25th, 50th, 75th, Max	1, 15, 36, 85, 29707	1, 20, 50, 127, 54368	1, 23, 58, 145, 77222
	Mean, Std. Dev.	181.06, 1226.15	257.26, 1834.25	300.09, 2470.52
Daily Views	Min, 25th, 50th, 75th, Max	0, 0, 1, 4, 6128	0, 0, 1, 3, 2256	0, 0, 1, 2, 2152
	Mean, Std. Dev.	14.42, 189.29	7.51, 67.87	6.2, 58.11
Num. Ratings	Min, 25th, 50th, 75th, Max	0, 0, 0, 1, 834	0, 0, 0, 1,931	0, 0, 0, 1, 988
	Mean, Std. Dev.	1.34, 21.04	1.56, 23.67	1.57, 25.29
Indicator no Rating	Min, Max	0, 1	0, 1	0, 1
	Freq. of 0, Freq. of 1	473, 1145	531, 1046	499, 1048
Daily Num. Ratings	Min, 25th, 50th, 75th, Max	0, 0, 0, 0, 4	0, 0, 0, 0, 2	0, 0, 0, 0, 1
	Mean, Std. Dev.	0.02, 0.21	0.01, 0.09	0.01, 0.08
Avg. Rating (For videos that have been rated)	Min, 25th, 50th, 75th, Max	1, 4, 5, 5, 5	1, 4, 5, 5, 5	1, 4, 5, 5, 5
	Mean, Std. Dev.	4.14, 1.30	4.1, 1.33	4.08, 1.35
Comments	Min, 25th, 50th, 75th, Max	0, 0, 0, 1 255	0, 0, 0, 1, 338	0, 0, 0, 1, 375
	Mean, Std. Dev.	0.91, 7.27	1.18, 9.4	1.19, 9.98
Daily Comments	Min, 25th, 50th, 75th, Max	-1, 0, 0, 0, 3	0, 0, 0, 0, 4	0, 0, 0, 0, 7
	Mean, Std. Dev.	0.02, 0.16	0.01, 0.16	0.02, 0.22
Favorited	Min, 25th, 50th, 75th, Max	0, 0, 0, 0, 281	0, 0, 0, 0, 207	0, 0, 0, 0, 219
	Mean, Std. Dev.	0.64, 8.7	0.63, 6.0	0.62, 6.18
Honors	Min, 25th, 50th, 75th, Max	0, 0, 0, 0, 5	0, 0, 0, 0, 4	0, 0, 0, 0, 2
	Mean, Std. Dev.	0.02, 0.22	0.02, 0.24	0.005, 0.09

Table 2 Correlations between video properties

	Views	Daily Views	Num. Ratings	Ind. no Ratings	Daily Num. Ratings	Comments	Daily Comments	Favorited	Honors
Views	1.000								
Daily Views	0.493	1.000							
Num. Ratings	0.553	0.268	1.000						
Ind. no Rating	–0.134	–0.092	–0.096	1.000					
Daily Num. Ratings	0.082	0.6	0.213	–0.035	1.000				
Comments	0.598	0.271	0.938	–0.152	0.166	1.000			
Daily Comments	0.134	0.667	0.265	–0.059	0.847	0.258	1.000		
Favorited	0.555	0.370	0.840	–0.135	0.188	0.79	0.225	1.000	
Honors	0.083	0.159	0.138	–0.097	0.16	0.217	0.261	0.179	1.000

No. of observations=50994

friends, whereas 133 had chosen not to (i.e., friends list is not public).[4] For these 1806 authors, we first obtained a list of the author's friends. We then visited the pages of these friends and obtained a list of their friends. So for each video, we reconstructed the social network of the author up to two hops. This data was collected during Nov 23rd–26th 2007.

The 1806 authors had a total of 15,861 first-degree friends and 1,627,091 second-degree friends. Some first and second-degree friends were common to multiple authors, so the total unique first and second-degree friends are lower at 12,361 and 745,176 respectively. Figure 4 shows the CDFs of the first and second-degree friends of the 1806 authors.

An important caveat is that we don't have data on the complete social network. Note that collecting complete data on large social networks is a time-consuming process that takes months. Crawling the two-hop network of 1806 authors (about 750,000 unique users) from a single computer with a five second latency takes about 45 days. We deployed our crawler on multiple computers and collected the data within 4 days. However, continued deployment on multiple computers was not viable due to infrastructural issues. So collecting complete network data or collecting two-hop data at multiple time points was infeasible. In fact, the study of global network properties of large networks is usually a research agenda in and of itself, and therefore outside the scope of this paper (see Barabasi et al. 2000 and Mislove et al. 2007).Furthermore, it should be noted that even if we undertook such a large-scale data collection task, the

[4] The fact that we are only able to analyze videos whose authors have publicized their friendships may cause some selection bias. There is no direct test to confirm or refute this. However we can test for selection bias indirectly by comparing the viewership distributions of the two samples of authors (i.e., the ones who publicized their friendship links and the ones who didn't). Since an author's social network affects her video's viewership, if we find the viewership distributions of both samples to be similar, then we can infer that both samples are drawn from the same social network distribution. Therefore we compared the viewership distributions of the two samples using Kolmogorov-Smirnov tests at t = 10, 20, and 30. In all three cases, the two distributions were statistically indistinguishable. Hence, we can safely say that any selection bias, if it exists, is not substantial.

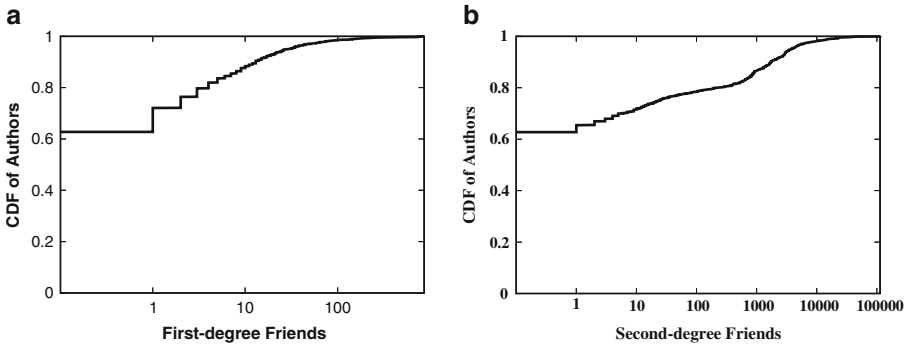


Fig. 4 **a**—CDF of First-degree Friends. **b**—CDF of second-degree friends

resulting data would not reflect the network at the time of observation since networks evolve over time. It would also suffer from reverse causality concerns because authors' networks are likely to be influenced by the performance of their videos. Moreover, managers conducting marketing campaigns on YouTube are unlikely to have complete network information. So a study of local networks is not only practical, but also managerially relevant. Nevertheless, lack of data on the complete social network presents challenges for inference, and we discuss this issue in detail in Section 4.

3.3 Social network properties

We now quantify the local network position of the authors in our dataset using metrics that capture three fundamental concepts—connectivity, clustering, and centrality. Our graph metrics are local or local approximations of global properties since our social network data is two-hop. First, we introduce some notation to aid exposition. Let $G = \{N, E\}$ be a network, where N is the set of nodes and E is the set of undirected edges such that $E = \{(i, j) \mid i \text{ and } j \text{ are connected}\}\ \forall i, j \in N$.

3.3.1 Connectivity metrics

Connectivity refers to the ease with which a node i can access others in the network or the ease with which information from i can flow to the rest of the network. We derive three connectivity metrics.

Degree or First-degree friends (d_i): is the simplest measure of connectivity and refers to the total number of first-degree friends of i.

$$d_i = |F(\{i\})| \tag{5}$$

where $F(\{i\}) = \{j|(i,j) \in E\}$. However, Degree is an imperfect measure of connectivity because it ignores subsequent connections. For example, a node A may have very few friends, but one of these friends (say node B) could be a hub. In this case, A's Degree is a deceptive measure of its connectivity because A's ability to access the larger network is greatly enhanced by its connection to B. Our next two metrics address this issue.

Second-degree friends (sd_i): refers to the number of unique second-degree friends of i.

$$sd_i = |F(F(\{i\})) - F(\{i\}) - \{i\}| \tag{6}$$

Nodes that are friends of i's friends, but not friends with i are referred to as second-degree friends. Also, sd_i is a count of unique second-degree friends, i.e., if two first-degree friends of i are friends with the same second-degree friend, then that second-degree friend is not double counted.

Average friends of first-degree friends (aff_i): average number of friends of i's first-degree friends.

$$aff_i = sd_i/d_i \tag{7}$$

Note that aff_i is defined and positive only for nodes that have at least one first and second-degree friend. So we use it in conjunction with indicator variables I_i^F and I_i^S, where $I_i^F = 1$ if and only if $d_i = 0$, and $I_i^S = 1$ if and only if $sd_i = 0$.

3.3.2 Clustering

Clustering characterizes the density of connections in a network. Highly clustered networks are usually close-knit and well-defined communities (Girvan and Newman 2002). Figure 5(a) depicts a highly clustered network, while Fig. 5(b) depicts a network with low clustering. We follow Watts and Strogatz (1998) to quantify the clustering in i's local network as follows:

$$C_i = \frac{|Q|}{d_i(d_i - 1)/2} \tag{8}$$

where $Q = \{(m, n) \mid m, n \in F(\{i\}) \ \& \ (m, n) \in E\}$. The total number of edges that can exist between all of i's friends is $d_i(d_i - 1)/2$. However, the number of edges that actually exist is $|Q|$. The Clustering coefficient C_i is thus the fraction of possible edges that actually exist.

3.3.3 Centrality

Centrality captures how central a node is to the network and is a measure of the node's power or social capital. Power can come from different structural positions in different networks. So centrality is synonymous with many constructs such as influence, brokerage,

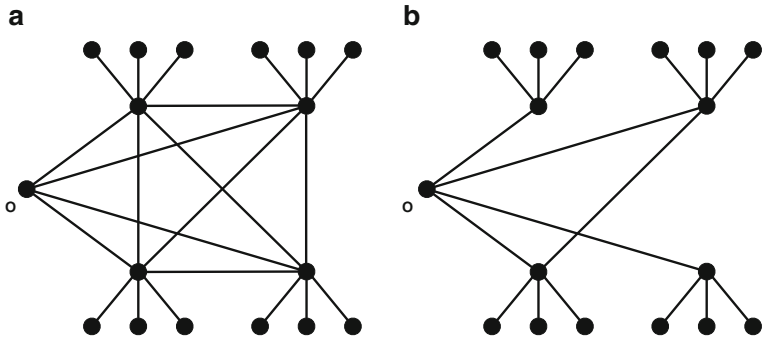
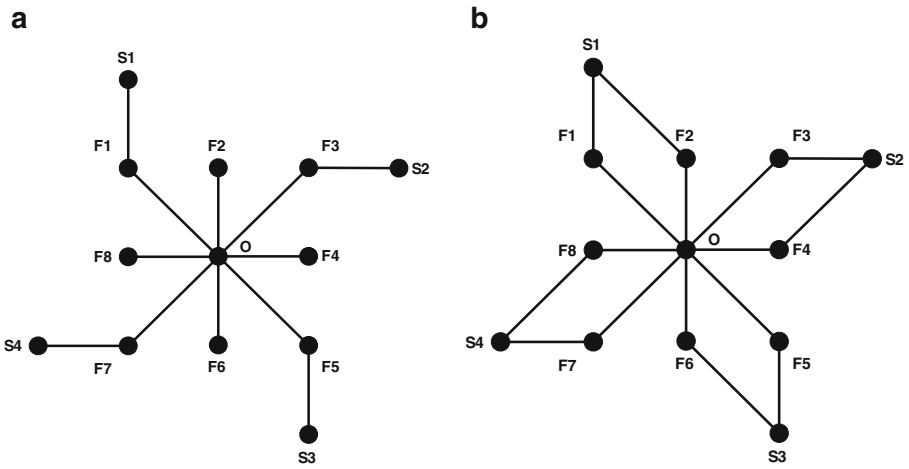


Fig. 5 Networks with high and low clustering



Network dominance of O – High
Path Diversity in O's network–Low

Network dominance of O – Low
Path Diversity in O's network–High

Fig. 6 Betweenness of O in **a** > Betweenness of O in **b**

exposure, control, and attention. Since it is impossible to capture all these concepts in a single metric, a variety of centrality metrics have been proposed (see Borgatti and Everett 2006 for a review), of which the three most prominent ones are Degree, Closeness, and Betweenness. Degree has been already discussed earlier. With two-hop data, Closeness (cs_i) provides no new information on network structure over and above aff_i because $cs_i = 2 - (1/(1 + aff_i))$.[5] So we use Betweenness as our centrality measure.

Betweenness captures the idea that a strategically placed node that lies on paths between other nodes has the potential to control communication in the network and command attention (Freeman 1979). The removal of O from Fig. 6(a) makes the network much more disconnected than the removal of O from Fig. 6(b). Hence, the Betweenness of O in Fig. 6(a) is higher than that of O in Fig. 6(b). Betweenness can either be calculated for the whole network or up to k hops, in which case it is called k-Betweenness (Borgatti and Everett 2006). Since we have two-hop data, we use '2-Betweenness' B_i, which is derived as follows: Let $F(\{i\}) = \{n_1, n_2, ..., n_{d_i}\}$ be the set of i's friends. For all pairs (n_j, n_k), let g_{jk} be the number of geodesics (shortest paths) between n_j and n_k.[6] For a pair (n_j, n_k), let $g_{jk}(i)$ be the number of geodesics that pass through i. Then, $p_{jk}(i) = g_{jk}(i)/g_{jk}$ is the proportion of geodesics between n_j and n_k that contain i and $B_i = \sum_{j=1}^{d_i} \sum_{k=j+1}^{d_i} p_{jk}(i)$.

Since the magnitude of Betweenness depends on network size, a central node in a small network might appear less central than a non-central node in a large network. Hence, a normalized measure is appropriate when comparing across networks. For a network of x

[5] Closeness is defined as the average geodesic distance of a node to the rest of the network. In a two-hop network, Closeness is $cs_i = (d_i + 2sd_i)/(d_i + sd_i)$. This can be rewritten as $cs_i = 2 - (1/(1 + aff_i))$.

[6] Note that there can be more than one geodesic between two nodes. For example, in Fig. 6(b), there are two shortest paths between nodes F3 and F4: F3-S2-F4 and F3-O-F4.

Table 3 Summary statistics of social network properties

Variable		No. Obs.	Min.	25th Percentile	50th Percentile	75th Percentile	Max.	Mean	Std. Dev.
Degree or Num. of first-degree friends (d_i)		1806	0	0	0	2	866	8.78	52.11
Second-degree friends (sd_i)		1806	0	0	0	19	112362	892.15	4888.94
Avg. friends of first-degree friends (aff_i) (For authors who have at least one first-degree friend)		674	0	4	60.58	162.67	871	113.09	154.02
Clustering coefficient (C_i)	All authors	1806	0	0	0	0	1	0.023	0.104
	Authors with more than one first-degree friend	505	0	0	0.017	0.068	1	0.083	0.184
Normalized 2-Betweenness (B_i^N)	All authors	1806	0	0	0	0.517	1	0.221	0.375
	Authors with more than one first-degree friend	505	0	0.709	0.844	0.958	1	0.789	0.232

nodes, Freeman (1979) suggested normalization by $B(x) = (x^2 - 3x + 2)/2$. Though all nodes in our data belong to the same network, the use of 2-Betweenness necessitates normalization by the size of the local neighborhood. So we use 'Normalized-2-Betweenness' (B_i^N), which is defined in (10). In future, we refer to Normalized-2-Betweenness as simply Betweenness for convenience.

$$B_i^N = B_i / B(d_i + 1) \tag{9}$$

A pleasant byproduct of this normalization is that it eliminates the inherent correlation between Degree and Betweenness. This allows us to use Betweenness purely as a measure of structural centrality without conflation with connectivity.[7]

3.3.4 Summary statistics

Table 3 presents the summary statistics of the network properties[8] and Table 4 describes the correlations between them. Figure 7(a) and (b) show the CDFs of Clustering and Betweenness forall the authors and for authors with Degree > 1, respectively.

[7] While 2-Beteweenness can be interpreted as a measure of local centrality, many have argued that it is in fact superior to global Betweenness. In 2-Betweenness, only geodesics of length two or less are considered, while global Betweenness considers geodesics of all lengths. However, lengthy paths are seldom used for communication. So taking them into account can result in a distorted picture of centrality. Therefore, some researchers advocate the use of 2-Betweenness even when complete network data is available. See Gould and Fernandez (1989), Friedkin (1991), and Borgatti and Everett (2006) for a comprehensive discussion of these issues. Moreover, both Everett and Borgatti (2005) and Borgatti et al. (2006) have shown that local Betweenness is highly correlated with global Betweenness.

[8] Notice that the degree distribution of first-degree friends looks very different from that of the author's degree distribution (see Table 3), i.e., Mean (Friend of friends) >Mean (Friends). As noted by Feld (1991), this property is common in social networks because well-connected people (with large Degree) tend to show up disproportionately more often in everyone's friends lists. Hence, in any social network, random sampling of authors (or nodes) will give rise to a sample of first-degree friends, which will contain some high-degree nodes. In Section 6.4, we perform robustness checks to confirm that our results are not driven by such high-degree nodes.

Table 4 Correlations between Network Properties (The first number denotes the correlations for all the videos and the second number denotes the correlation value for only those videos whose authors have more than one first-degree friend and zero second-degree friends)

	First deg. friends	Second deg. friends	Avg. friends of first degree friends	Clustering coefficient	Normalized 2-Betweenness
First deg. friends	1.000, 1.000				
Second deg. friends	0.83, 0.815	1.000, 1.000			
Avg. friends of first deg. friends	NA, 0.078	NA, 0.196	NA, 1.000		
Clustering coefficient	0.024, –0.079	0.014, –0.101	NA, –0.249	1.000, 1.000	
Normalized 2-Betweenness	0.245, –0.018	0.255, –0.053	NA, –0.108	0.096, –0.788	1.000, 1.000

3.4 Classification of variables

Broadly speaking, we have two types of independent variables: time varying and time-invariant. The time varying variables consist exclusively of video characteristics and the time-invariant ones consist of the social network characteristics (see Table 5). In most of our model specifications, we will use some subset of these variables and refer to them using the following notation: $X_{i,t}$ - vector of time varying variables and Z_i - vector of time-invariant variables.

4 Model

In this section, we develop a descriptive dynamic model of video growth. Let $y_{i,t} = \ln(v_{i,t} + 1)$, where $v_{i,t}$ is Daily Views. For $t > K$, $y_{i,t}$ is modeled as follows:

$$y_{i,t} = c + \sum_{k=1}^{K} \alpha_k y_{i,t-k} + \gamma X_{i,t-1} + \beta Z_i + \eta_i + \varepsilon_{i,t} \tag{10}$$

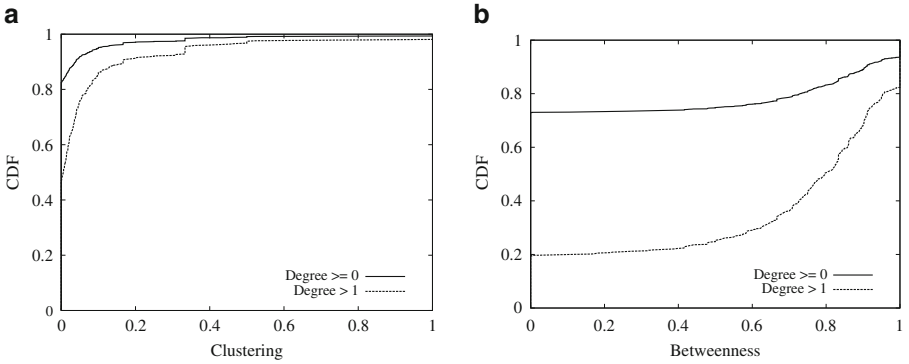


Fig. 7 Cumulative distribution functions of Clustering and Betweenness. Shown for all authors (Degree ≥ 0) and for authors with at least one first-degree friend (Degree > 1)

Table 5 Classification of variables used

Time varying variables	Time invariant variables
Daily Views ($v_{i,t}$)	Degree or Num. of first-degree friends (d_i)
Daily Num. Ratings ($nr_{i,t}$)	Second-degree friends (sd_i)
Indicator no Rating ($Inr_{i,t}$)	Total unique first and second-degree friends (fsd_i)
Avg. Rating ($Ar_{i,t}$)	Avg. friends of first-degree friends (aff_i)
Daily Comments ($c_{i,t}$)	Clustering coefficient (C_i)
Favorited ($f_{i,t}$)	Normalized 2-Betweenness (B_i^N)
Honors ($h_{i,t}$)	Closeness (cs_i)

According to (10), the number of views received by video i in period t depends on the views it received in the past K periods $(y_{i,t-1}, y_{i,t-2}, ..., y_{i,t-K})$, the lagged time varying video characteristics ($X_{i,t-1}$), the social network properties of its author (Z_i), a mean-zero error term ($\varepsilon_{i,t}$), and a time-invariant unobserved effect (η_i) that embodies three types of unobservables – a_i, v_i, and g_i. a_i refers to the unobservables about the author that affect her video's performance, such as her popularity, ability to post good videos, past success and ensuing reputation, and offline brand equity. v_i refers to the unobservable attributes of the video that affect its performance, such as its intrinsic quality and its relevance to popular culture. g_i captures the impact of the unobserved global social network, which includes the third, fourth, and farther hops of the YouTube network, and people outside YouTube's social network. Since we don't have data on multiple videos for the same author and since we don't observe the complete social network, we can't separately identify these three effects. Therefore, we collapse them into a cumulative unobserved effect η_i and refer to it as the fixed effect.

Further, we make the following assumptions regarding the model.

10.1 $E(\varepsilon_{i,t}) = E(\eta_i) = E(\varepsilon_{i,t}.\eta_i) = 0 \ \forall \ i,t$

10.2 $$E(\varepsilon_{i,t}.\varepsilon_{i,s}) = \begin{cases} 0 \text{ if } s \neq t \\ \sigma_i^2 \text{ if } s = t \end{cases} \forall \ i,s,t$$

10.3 $$E(\eta_i.\eta_j) = \begin{cases} 0 \text{ if } i \neq j \\ \sigma_\eta^2 \text{ if } i = j \end{cases} \forall \ i,j$$

10.4 $$\begin{array}{l} E(X_{i,t}.\varepsilon_{i,s}) = 0 \text{ if } s > t \\ E(X_{i,t}.\varepsilon_{i,s}) \neq 0 \text{ if } s \leq t \end{array} \forall \ i,s,t$$

10.5 $E(X_{i,t}.\eta_i) \neq 0 \ \forall \ i,t$

10.6 $E(Z_i.\eta_i) \neq 0 \ \forall \ i$

10.7 Initial Conditions Assumption:

$$y_{i,1} = c + \eta_i\left(\frac{1 + \kappa_X\gamma + \kappa_Z\beta A_K}{1 - A_K}\right) + \beta Z_i + \varepsilon_{i,1} \quad \forall \ i \tag{10.7a}$$

$$y_{i,t} = c + \eta_i\left(\frac{1+(\kappa_X\gamma+\kappa_Z\beta)A_K}{1-A_K}\right) + \gamma X_{i,t-1} + \beta Z_i + \varepsilon_{i,t}, \forall\, 2 \le t \le K, i \quad (10.7b)$$

where $A_K = \sum_{k=1}^{K} \alpha_k$ and κ_X, κ_Z are system-wide parameters specified in Appendix A.2. Equation 10.7a gives $y_{i,t}$ for $t = 1$ and for the remaining initial periods ($2 \le t \le K$). We now discuss the assumptions in detail.

Assumption 10.1: We follow the familiar error component structure, i.e., η_i and $\varepsilon_{i,t}$ are mean-zero and uncorrelated for all i and t.

Assumption 10.2: $\varepsilon_{i,t}$ s are allowed to be heteroskedastic across videos, but assumed to be serially uncorrelated. In Section 6, we test the validity of this assumption using the Arellano-Bond (2) test.

Assumption 10.3: We assume that there is no cross-panel correlation, that is, the unobserved fixed effect η_i is assumed to be an independent draw for each video.

Assumption 10.4: This allows for correlation between the error-term $\varepsilon_{i,t}$ and both future and current $X_{i,t}$s ($X_{i,t}, \ldots, X_{i,T}$). For instance, a positive shock to the number of visitors in period t also increases the probability of video i getting rated during t, implying that $E(\varepsilon_{i,t}.Inr_{i,t}) \neq 0$. In fact, since future viewership depends on current viewership, the probability of getting rated in future is also affected by $\varepsilon_{i,t}$, that is, $E(\varepsilon_{i,t}.Inr_{i,k}) \neq 0$ for $k \geq$ t. Similarly, a shock to viewership at t is also likely to affect other time-varying covariates such as Avg. Ratings in both current and future periods. Hence, we cannot assume $X_{i,t}$s to be strictly exogenous. So we impose the weaker restriction that $X_{i,t}$ is exogenous only to future shocks.

Assumption 10.5: $X_{i,t}$s may also be correlated with η_i because the unobserved attributes that affect video popularity ($y_{i,t}$) may also affect the time varying covariates $X_{i,t}$. For example, a high quality video that receives a large number of views is also likely to receive more Favorites and Honors, implying that $E(X_{i,t}.\eta_i) \neq 0$. This correlation is assumed to be linear.

Assumption 10.6: We also allow for correlation between the time-invariant network properties Z_i and η_i because unobserved attributes that affect video popularity ($y_{i,t}$) may also affect the authors' social network (Z_i). For example, authors who make high quality videos that receive many views are also likely to be popular and have a large social network, implying that $E(Z_i.\eta_i) \neq 0$.This correlation is also assumed to be linear.

Assumption 10.7: This is similar to the initial conditions assumption in Blundell and Bond (1998), albeit modified to accommodate multiple lags and endogenous variables. The basic idea is that the realizations of $y_{i,t}$ in the initial periods are centered around its long-term mean and the deviations from the mean are uncorrelated to the mean itself. Intuitively, after controlling for covariates, videos with larger fixed effects are not systematically further or closer to their steady states than those with smaller fixed effects in the initial periods. This is a reasonable assumption in most settings, including ours, because it is essentially a form of stationarity assumption on the initial conditions. A key advantage of this assumption is that, it allows us to express $y_{i,t}$ as $y_{i,t} = \eta_0\eta_i + f_t(.)$ for all time-periods, where η_0 is a constant and $f_t(.)$ is a

time varying function of covariates independent of η_i (see Lemma 1 in Appendix A.2 for details). This in turn ensures that $\Delta y_{i,t}$ is independent of the fixed effect η_i, a regularity condition that can be exploited to specify moment conditions necessary to estimate the model (as we will see in Section 5).In our context, this is equivalent to assuming that the unobserved author and video attributes (η_i) are uncorrelated to changes in log-viewership ($\Delta y_{i,t}$).

Finally, note that even though we haven't specified any correlation between $y_{i,t}$ and $\varepsilon_{i,t}$, the current error term affects both current and future views by definition, i.e., $E(\varepsilon_{i,t}.y_{i,k}) \neq 0$ for $k \geq t$. Naturally, past views are not affected by future shocks, so $E(\varepsilon_{i,t}.y_{i,k}) = 0$ for $k < t$. Also, by construction, all $y_{i,t}$s are correlated with η_i. So $E(y_{i,t}.\eta_i) \neq 0 \ \forall \ i,t$. In sum, we impose very mild exogeneity conditions, and our model specification is fairly general and accommodating.

A key identification issue is that we only know the total views received by a video per day. We don't know how many of these views came from the measured local network and how many from the unmeasured global network that consists of the third, fourth, and farther hops of the YouTube network and people outside YouTube's social network. Note that viewers outside the author's local network can visit the video through two mechanisms. First, they may visit it through mechanisms uncorrelated to the author's local network. For example, they may find the video through search engines (e.g., Google, YouTube), news websites, or blogs. Lack of data on these views should not bias the estimates of Z_i because they are uncorrelated to the author's local network on YouTube. If we believe that the majority of views from the unmeasured global network are of this kind, then the inference of Z_i remains unbiased by lack of data on g_i. Second, viewers from outside the local network may come through mechanisms correlated to the author's local network. For example, a predominant fraction of the global views may come from the author's third-degree network, whose size is likely to be correlated to the size of her second-degree network. In such cases, $E(Z_i.g_i) \neq 0$, and we allow for this correlation through Assumption 10.6. (Recall that $E(Z_i.\eta_i) \neq 0$, where η_i is a function of g_i.) However, for valid inference, we require the impact of the unobserved global network on $y_{i,t}$ (log views) to remain constant over time, though we allow the impact of observed network properties to vary with time. This is done through the initial conditions assumption (Assumption 10.7) similar to that used in Blundell and Bond (1998).This in turn ensures that $\Delta y_{i,t}$ is independent of η_i (and hence g_i) but not Z_i, allowing it to function as an instrument for Z_i and other endogenous variables in Equation (10). See Appendix A.4 for details.

5 Estimation

In Section 4, we saw that the model exhibits three types of endogeneity:

1) Time varying video characteristics are correlated with the unobserved fixedeffect.
2) Time varying video characteristics are correlated with shocks to past (and current) views.

3) Time invariant network properties are correlated with the unobserved fixed effect.

We need an estimation strategy capable of handling all these endogeneity problems. Unfortunately, we cannot use standard VAR-based estimation strategies such as SUR because these estimators require error terms to be uncorrelated with all the explanatory variables. Further, the two commonly used methods of estimating panel data models, Random-effects estimation and Fixed-effects estimation cannot be used in a dynamic setting. The former requires explanatory variables to be strictly exogenous to $\varepsilon_{i,t}$ and η_i. This is far from true in our case. The latter allows for correlation between η_i and explanatory variables, but since it uses a within-transformation, it requires all time-varying variables to be strictly exogenous to $\varepsilon_{i,t}$. This is impossible in a dynamic setting with finite T (Nickell 1981). This rules out both Fixed-effects estimation and other methods that use the within-transformation.

An obvious solution is to find external instruments for the endogenous variables. However, it is difficult to find variables that affect network properties, lagged viewership and video properties, but do not affect current viewership. Therefore, we turn to the GMM style estimators of dynamic panel data models that exploit the lags and lagged differences of explanatory variables as instruments. This method was pioneered by Anderson and Hsaio (1981), who showed that in the absence of serial correlation in error-terms, lags of explanatory variables can be used to instrument for the endogenous explanatory variablesin first-differenced equations of interest. This method was further developed by Arellano and Bond (1991), who also provided a specification test, the Arellano-Bond test for serial correlation, to check the validity of the underlying assumption of serially uncorrelated errors in the data. These earlier papers rely only on the first-differenced equations. More recently, Blundell and Bond (1998) have proposed a system GMM approach that uses both first-differenced and level equations. Specifically, they showed that if we are willing to assume that the initial deviations of the dependent variable are independent of its long-term average, then lagged differences of the dependent variable can be used to instrument for the endogenous explanatory variables in level equations. We follow their approach and extend it to include multiple lags and endogenous time varying and invariant variables. Below, we outline our approach.

Moment Conditions for First-Differenced Equations Consider Equation (11), where Δ is the first-difference operator, i.e., $\Delta y_{i,t} = y_{i,t} - y_{i,t-1}$.

$$\Delta y_{i,t} = \sum_{k=1}^{K} \alpha_k \Delta y_{i,t-k} + \gamma \Delta X_{i,t-1} + \Delta \varepsilon_{i,t} \tag{11}$$

Notice that first differencing has eliminated the time-invariant social network metrics (Z_i) and the video-author specific effect η_i. So the correlation between the explanatory variables and the η_i is not an issue any more. However, by first differencing we have introduced another kind of bias. Now the error term $\Delta\varepsilon_{i,t}$ is correlated with the explanatory variables $y_{i,t-1}$ and $X_{i,t-1}$. Therefore, straightforward estimation is still not feasible. However, we can show that $y_{i,p}$ and $X_{\mathrm{i},p}$ are

not correlated to $\Delta\varepsilon_{i,t}$ for $p \leq 2$, but correlated with$\Delta y_{i,t-k}$ and $\Delta X_{i,t-1}$ making them good instruments for (11). See Proposition 1 in Appendix A.3 for details. We therefore specify the following two sets of moment conditions for (11).

$$E(y_{i,p}.\Delta\varepsilon_{i,t}) = 0, \text{where } p \leq t-2 \tag{12a}$$

$$E(X_{i,p}.\Delta\varepsilon_{i,t}) = 0, \text{where } p \leq t-2 \tag{12b}$$

The advantage of moment conditions (12a, 12b) is that they don't require Assumption 10.7. They only require $\varepsilon_{i,t}$s to be serially uncorrelated. However, methods that rely only on (12a) and 12(b) suffer from two drawbacks: 1) large finite sample biases if the dynamic process is persistent or if the variance of η_i is high (Blundell and Bond 1998), and 2) inability to recover the coefficients of Z_i. These drawbacks are particularly debilitating in our case because our model exhibits persistence, and our key parameters of interest are the network properties Z_i. Thereforewe consider level equations too.

Moment Conditions for Level Equations Consider the level equations (10):

$$y_{i,t} = c + \sum_{k=1}^{K} \alpha_k y_{i,t-k} + \gamma X_{i,t-1} + \beta Z_i + \eta_i + \varepsilon_{i,t} \tag{10}$$

Here the video-author specific effect η_i is correlated with all the explanatory variables ($y_{i,t-k}$s, $X_{i,t-1}$, and Z_i). So we need to instrument for all of three of them. Recall that all $y_{i,p}$s have a constant η_i term (Assumption 10.7), implying that $\Delta y_{i,p}$ is independent of $(\eta_i+\varepsilon_{i,t})$ for all $p \leq t-1$. Further, we can show that $\Delta y_{i,p}$s are correlated with all three sets of explanatory variables, $y_{i,t-k}$s, $X_{i,t-1}$, and Z_i. Together, these two properties make $\Delta y_{i,p}$s good instruments for all the explanatory variables in (10). See Proposition 2 in Appendix A.4 for details. Therefore, the first set of moment conditions that we specify for (10) is:

$$E\big(\Delta y_{i,p}.(\eta_i + \varepsilon_{i,t})\big) = 0, \text{where } p \leq t-1 \tag{13a}$$

Similarly, we can show that $\Delta X_{i,p}$ is uncorrelated to $(\eta_i+\varepsilon_{i,t})$, but correlated with both $X_{i,t-1}$ and $y_{i,t-k}$s for all $p \leq t-1$, which makes them good instruments for both these endogenous variables (see Proposition 2 in Appendix A.4). Hence, the second set of moment conditions that we specify is:

$$E\big(\Delta X_{i,p}.(\eta_i + \varepsilon_{i,t})\big) = 0, \text{where } p \leq t-1 \tag{13b}$$

System GMM Estimator Stacking (12a) and 12(b) over (13a) and 13(b) gives us a system GMM estimator that provides consistent estimates of both time-varying and time-invariant variables, even when the dynamic process is persistent. Generally, if the disturbances are heteroskedastic (as in our case), the two-step GMM is more efficient. However, the standard errors of the two-step GMM estimator are known to be biased. Windmeijer (2005) proposed a correction for this bias, and we follow his method to obtain robust standard errors. Also, following Arellano and Bond (1991), we test the validity of the instruments using the Arellano-Bond (2) test for serial correlation, as described below.

Serial Correlation and Lagged Dependent Variables A key assumption in the method outlined above is that the error terms are *not* serially correlated (Assumption 10.2). If they were, then the restrictions that we apply would not hold. Consider a scenario where the errors follow a MA(1) process such that $\varepsilon_{i,t} = \rho\varepsilon_{i,t-1} + u_{i,t}$, where $E(u_{i,t}) = 0$ and $E(u_{i,t}.u_{i,s}) = 0$ for all $t \neq s$. In that case, for $p = t-2$, the moment condition (12a) can be expanded as:

$$E\left(y_{i,t-2}.(\varepsilon_{i,t} - \rho\varepsilon_{i,t-2} + u_{i,t-1})\right) = 0 \tag{14}$$

Similarly, for $p = t-1$, the moment condition (13a) can be expanded as:

$$E\left(\Delta y_{i,t-1}.(\eta_i + \rho\varepsilon_{i,t-1} + u_{i,t})\right) = 0 \tag{15}$$

However, notice that both (14) and (15) are invalid. In (14), $y_{i,t-2}$ is correlated with $\varepsilon_{i,t-2}$, which implies that $E\left(y_{i,t-2}.(\varepsilon_{i,t} - \rho\varepsilon_{i,t-2} + u_{i,t-1})\right) \neq 0$. Similarly in (15), $\Delta y_{i,t-1}$ is correlated with $\varepsilon_{i,t-1}$because $\Delta y_{i,t-1}$ contains a $y_{i,t-1}$ term which is correlated with $\varepsilon_{i,t-1}$. So $E\left(\Delta y_{i,t-1}.(\eta_i + \rho\varepsilon_{i,t-1} + u_{i,t})\right) \neq 0$. Hence, in the presence of serial correlation, our restrictions break down.

There are two ways to solve this problem. The first is to add sufficient lags of $y_{i,t}$ on the right hand side; this alleviates serial correlation because lags capture past shocks. The second is to allow serial correlation, but use farther removed lags as instruments. For instance, in the case of MA(1) correlation, we can drop $y_{i,t-2}$, $X_{i,t-2}$ as instruments for Equation (11) and just use $y_{i,p}$s and $X_{i,p}$s, where $p \leq t-3$. If T is sufficiently long, the first approach is better because instruments farther from t are usually weak. However, if T is short, upon adding too many lags of $y_{i,t}$ on the right hand side, we may not have enough data to work with. In our case, we find that using five lags of $y_{i,t}$ is sufficient to rule out serial correlation. Since 85% of our videos have 31 observations or more, losing five lags is not problematic. We therefore use the first approach.

We confirm the absence of serial correlation using the Arellano-Bond (2) test which tests for second-order serial correlation in the first-difference of error-terms. By construction $\Delta\varepsilon_{i,t}$ and $\Delta\varepsilon_{i,t-1}$ are correlated (through the common $\varepsilon_{i,t-1}$ term). However, in the absence of serial correlation, $\Delta\varepsilon_{i,t}$ and $\Delta\varepsilon_{i,t-2}$ should be uncorrelated and the Arellano-Bond (2) test examines if this is indeed the case.

In sum, our estimation strategy has two key advantages over other commonly used methods: 1) it is able to handle the three types of endogeneity (Assumptions 10.4, 10.5, and 10.6) inherent in the model, and 2) it is able to recover the coefficients of the time-invariant network properties Z_i.

6 Results

This section is organized as follows. In Section 6.1, we outline two basic model specifications. In Section 6.2, we discuss the impact of network properties on video popularity. In Section 6.3, we discuss the impact of lagged video properties and explore a few more variations of the model. Finally, in Section 6.4, we present additional robustness checks to establish the validity of our results.

6.1 Variations of the basic model

We now present the results for two basic variations of the model. In Model 1, we use $X_{i,t} = \{c_{i,t}, f_{i,t}, h_{i,t}, Inr_{i,t}, (1 - Inr_{i,t}).Ar_{i,t}\}$ to capture the video attributes and $Z_i = \{I_i^F, I_i^S, (1 - I_i^F).\ln(d_i), (1 - I_i^F).(1 - I_i^S).\ln(aff_i), C_i, B_i^N\}$ to capture the network properties. The indicator variables I_i^F and I_i^S ensure that we measure the effect of first and second-degree friends only when they exist. For this model, we use lags 2 to 6 of $y_{i,t}$ and $X_{i,t}$ ($y_{i,t-2}$,…,$y_{i,t-6}$,$X_{i,t-2}$,…,$X_{i,t-6}$) as instruments for (11) and $\Delta y_{i,t-1}$, $\Delta X_{i,t-1}$ as instruments for (10). In Model 2, we use the same instruments and network metrics as Model 1, but with a slightly different set of video properties. See Section 6.3 for details on Model 2. Table 6 presents the estimation results.

The log transformation of the connectivity metrics is necessitated by the skew and range in their distributions (Hansen 2008). Further, we avoid using sd_i directly in conjunction with d_i because of the high correlation between them (corr. = 0.83) and instead use d_i and aff_i. Also, note that the Arellano-Bond (2) tests confirm that our models are not misspecified, i.e., the tests present no evidence of serial correlation. In all the models, the p-values indicate that we cannot reject the null of no serial correlation.

6.2 Impact of network properties

The estimation suggests that all three network properties—connectivity, clustering, and Betweenness—significantly influence video popularity. We discuss the impact of each in detail below.

6.2.1 Connectivity

To aid understanding, we substitute for aff_i and express $v_{i,t}$ as follows:

$$(v_{i,t} + 1) \propto (d_i)^{\beta_1 - \beta_2} (sd_i)^{\beta_2} \tag{16}$$

where β_1 and β_2 are the coefficients of $\left(1 - I_i^F\right).\ln(d_i)$, and $\left(1 - I_i^F\right).\left(1 - I_i^S\right).\ln(aff_i)$. Equation (16) gives us the marginal impact of first-degree friends as $\beta_1 - \beta_2$ and that of second-degree friends as β_2.

First, we find that an author's Degree has a considerable impact on the popularity of her videos. In both Models 1 and 2, the coefficient $\beta_1 - \beta_2$ is positive and significant (see Table 6). This is in contrast to the results from Watts and Dodds (2007), who find that high degree nodes don't have a significant impact on information diffusion. The discrepancy likely stems from their use of simulations, i.e., they use both simulated networks and propagation rules, which may not reflect the structure or behavior of real life networks. Moreover, it is easy to see that friends can be valuable in the context of YouTube—they are likely to visit the author's video, forward it to acquaintances, talk about it to friends, or place links to it on their blogs, and all these activities can enhance viewership. Second, we find that an author's second-degree friends also aid video diffusion. The coefficient β_2 is positive and significant in both Models 1 and2. In sum, we find that the 'size' of a node's local network (number of first and second-degree friends) has a significant impact on the diffusion of products seeded by it, and this effect persists even after controlling

Table 6 Estimation results

Dependent Variable: Log Daily Views (t)		Model 1		Model 2		Model 3		Model 4		Model 5	
		Parameter	t-stats	Parameter	t-stats	Parameter	t-stats	Parameter	t-stats	Parameter	t-stats
Lagged Dependent Variables	Log Daily views (t – 1)	0.257***	(10.17)	0.269***	(14.42)	0.271***	(11.54)	0.251***	(10.28)	0.302***	(7.20)
	Log Daily views (t – 2)	0.356***	(23.00)	0.363***	(32.15)	0.365***	(19.13)	0.353***	(19.66)	0.419***	(13.95)
	Log Daily views (t – 3)	0.042***	(4.18)	0.044***	(4.86)	0.038***	(3.09)	0.043***	(4.68)	−0.001	(−0.02)
	Log Daily views (t – 4)	0.005	(0.51)	0.006	(0.64)	0.000	(−0.01)	0.007	(0.71)	−0.041	(−1.07)
	Log Daily views (t – 5)	0.028***	(3.98)	0.047***	(6.42)	0.024	(1.48)	0.029***	(4.70)	0.045***	(2.97)
Lagged Video Characteristics	Indicator no rating (t – 1)	−0.520	(−0.71)			−0.536	(−0.45)	−0.512	(−0.71)	−0.155	(−0.85)
	Avg. rating (t – 1)	0.063	(0.36)			0.067	(0.23)	0.065	(0.39)	0.093**	(2.10)
	Daily num. ratings (t – 1)			0.003	(0.22)						
	Daily comments (t – 1)	0.019	(0.82)	0.01	(0.44)	0.016	(0.83)	0.020	(1.52)	−0.004	(0.33)
	Honors (t – 1)	−0.027	(−0.71)	−0.026	(−0.80)	−0.033	(−1.14)	−0.027	(−1.47)	−0.054*	(−1.85)
	Favorited (t – 1)	0.005***	(4.81)	0.005***	(3.29)	0.005***	(3.87)	0.005***	(4.47)	0.004***	(6.11)
Network Properties	Ind. zero first-deg friends	0.039	(0.05)	−0.557	(−0.79)	−0.029	(−0.03)	0.007	(0.01)		
	Ind. zero second-deg friends	−0.273	(−0.33)	1.042	(1.05)	−0.225	(−0.22)	−0.242	(−0.29)	−1.39*	(1.85)
	Log Degree	0.275***	(2.88)	0.526***	(3.72)	0.261***	(2.70)	0.285***	(2.90)	0.108***	(2.81)
	Log Avg. friends of first-degree friends	0.159**	(2.21)	0.436***	(3.69)	0.147*	(1.93)	0.164**	(2.22)	−0.018	(−0.39)
	Norm. 2-Betweenness	−1.435**	(−2.54)	−2.409***	(−3.30)	−1.409**	(−2.11)	−1.473***	(−2.58)	−1.045**	(−2.03)
	Clustering coefficient	−3.460***	(−2.68)	−4.548**	(−2.52)	−3.399***	(−2.69)	−3.569***	(−2.95)	−1.885***	(−2.95)
	Constant	0.742	(0.87)	−0.341	(−0.68)	0.762	(0.57)	0.733	(0.90)	−1.056*	(−1.64)
No. of observations, groups, instruments		44203, 1649, 988		44203, 1649, 915		44203, 1649, 827		44203, 1649, 1003		12009, 452, 588	
Arellano-Bond (2) test (*p-value*)		−0.514, (0.608)		−0.589, (0.556)		−0.632, (0.527)		−0.424, (0.672)		−1.522, (0.128)	
Goodness of Fit Measures	$\text{Corr}(y,\hat{y})^2$	0.679		0.492		0.689		0.671		0.789	
	MSE	0.456		0.855		0.443		0.471		0.306	
	MAD	0.485		0.620		0.481		0.491		0.422	

Note: *** ⇒ $p \leq 0.01$, ** ⇒ $p \leq 0.05$, * ⇒ $p \leq 0.1$

All Models are analogous to Model 1, with the following changes. Model 2 uses Daily Num. Ratings instead of Ind. no Ratings and Avg. Ratings. Model 3 uses lags 2–5 of $y_{i,t}$ and $X_{i,t}$ as instruments for (11) and Model 4 uses lags 2–7 of $y_{i,t}$ and $X_{i,t}$ as instruments for (11). Model 5 excludes authors with zero or one friend

for video characteristics, unobserved seed/author qualities, unobserved video quality, and endogenous network formation.

6.2.2 Relative impact of first and second-degree friends

We now examine who is more valuable from an author's perspective—first-degree friendsor second-degree friends. Specifically, how does the marginal benefit of a first-degree friend compare with that of a second-degree friend? In theory, we should expect first-degree friends to have a larger impact on viewership because they have better access to the author's video and are more likely to be interested in it. However, our analysis suggests otherwise. Notice that $\beta_1 - \beta_2 < \beta_2$ in Models 1 and 2. Though not implausible, this is certainly surprising. We therefore examine this result further and provide two potential explanations. First, it is possible that even though first-degree friends have better access to and greater interest in an author's video, second-degree friends have larger and wider networks, thereby rendering themselves more valuable. Second, it is possible that this result stems from our data limitations (as explained below).

Figure 8 depicts an example of the marginal utility comparisons, where 8(a) is the base case. To obtain the marginal benefit of a first-degree friend, we need to compare scenarios 8(a) and 8(b), where 8(b) is obtained by adding one first-degree (F3) and zero second-degree friends to 8(a). Similarly, to calculate the marginal benefit of a second-degree friend, we need to compare 8(a) and 8(c). Here, one second-degree (S5) and zero first-degree friends have been added in 8(c). Thus, comparing the marginal benefit of first and second-degree friends is equivalent to assessing whether the focal node O in 8(a) would prefer to be in 8(b) or in 8(c).

However, in practice our evaluation of a second-degree friend's marginal benefit is affected by lack of data on third-degree friends. Note that when calculating the marginal benefit of a second-degree friend, our analysis assumes that S5 is a typical second-degree friend, i.e., it has as many links as an average second-degree friend. So we might actually be comparing scenarios 8(a) and 8(d), whereas what we want is the comparison between scenarios 8(a) and 8(c). Given our data limitations, we cannot address this issue satisfactorily and therefore provide a more subdued interpretation of our result—we cannot unequivocally assert that first-degree friends are more important than second-degree friends—an interesting finding in its own right.

6.2.3 Clustering

A large stream of literature on in-group bias suggests that belonging to a tight-knit group should benefit a seed because in-group members are likely to help and reward each other (Tajfel and Turner 1986). That is, authors from close-knit groups are likely to enjoy the advantage of committed friends who view and promote their videos. In line with these arguments, in a study of individual level peer-effects, Katona et al. (2009) find that local clustering has a positive impact on an individual's adoption probability.[9]

[9] Stephen and Toubia (2010) also study the impact of local Clustering, but in the context of a sellers' network. In their setting, a node's (seller's) goal is to generate high incoming traffic, whereas in our context a node's goal is to maximize the outgoing information on video. So their findings are not applicable to our context.

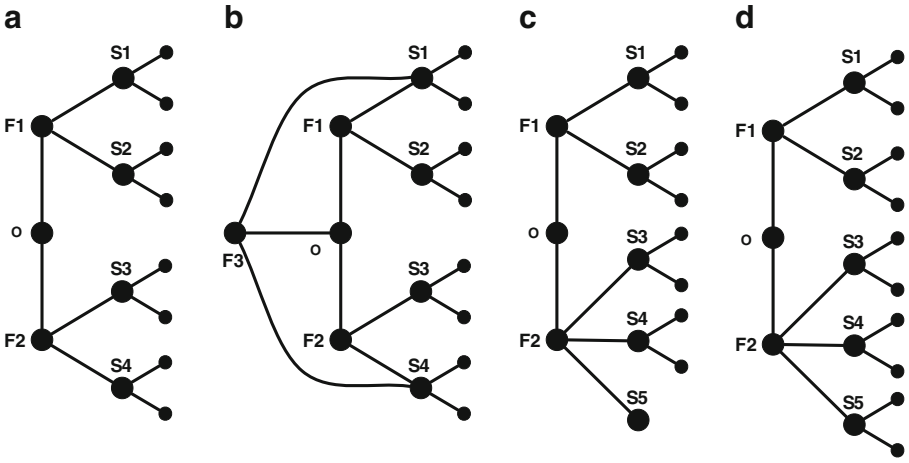


Fig. 8 Measuring the relative impact of first and second-degree friends

In contrast, we find that high clustering in an author's local network has a significant negative impact on the popularity of videos seeded by her (see Table 6). One possible reason for this effect could be that members of clustered communities have few outside connections, which in turn could decrease their ability to spread information to the wider network. However, this possibility is ruled out in our case since we have controlled for second-degree friends; we find that even when two communities have the same number of outside connections (e.g., Fig. 5(a) and (b)), the one with lower clustering (Fig. 5(b)) disseminates information better than the one with higher clustering (Fig. 5(a)). Hence, our results suggest that members of a tight-knit community may not interact much with outsiders even if they are connected to them, thereby failing to spread information about the video to the wider network.[10,11]

Note that the discrepancy between our results and that of Katona et al. (2009) stems from our focus on global diffusion as opposed to local peer effects. Clustering is thus an interesting metric that has a positive local effect and a negative global effect. This divergence highlights the perils of forecasting the global impact of a seed's network properties from individual-level studies.

6.2.4 Betweenness centrality

We find that Betweenness has a negative and significant impact on video popularity (Table 6). This is surprising since high Betweenness has always been associated with

[10] Note that authors from close-knit groups are more likely to post niche videos (videos of interest to only those close to them), which could dampen the global diffusion of their videos. We allow for this possibility within the model through Assumption 10.6, i.e., we allow for correlations between the unobserved content of the videoand its author's network properties. In the estimation, when specifying moment conditions, we ensure that this correlation is not violated. Hence, we can safely state that the negative effect of Clustering doesn't stem from the correlation between a video's content (niche or broad) and the clustering in its author's network.

[11] On a related note, Granovetter (1973) suggests that new information often comes from weak ties or acquaintances and not from close friends. Since members of close-knit groups tend to be close friends and those of loosely knit groups tend to be acquaintances, this result can also be interpreted as acquaintances being more valuable than close friends from a seed's perspective.

higher social capital (see Burt 1995; 2004, and Borgatti et al. 1998). However, we argue that Betweenness is not always a positive property. In fact, Betweenness embodies two competing concepts: *network dominance* and *path diversity.*

Consider Fig. 6(a) and 6(b): O has the same first and second-degree friends in both, but O's Betweenness in Fig. 6(a) is higher than that in Fig. 6(b). O's position in Fig. 6(a) has two advantages over that in Fig. 6(b): First, since O's friends have to always go through O to reach each other in Fig. 6(a) (unlike Fig. 6(b), where there are alternate paths between them), they are more likely to visit O's page and view its videos. So O's dominant position as the primary local connector enhances its visibility and importance. Second, since O's friends have fewer connections in Fig. 6 (a), it faces less competition. For example, in Fig. 6(b), F4 may distribute her time between videos posted by S2 and O, whereas in Fig. 6(a), F4 has more time for O's videos.[12] Thus, in line with previous research, it is clear that a seed would benefit from high Betweenness because it augments her *network dominance*. However, nodes with high Betweenness have fewer paths to reach second-degree friends. In Fig. 6(b), O has two paths to S2, O-F3-S2 and O-F4-S2, whereas in Fig. 6(a) it has only one path to S2, O-F3-S2. Information about O's videos is more likely to reach S2 in Fig. 6(b) than in Fig. 6(a). Higher Betweenness therefore implies lower *path diversity* in a node's local network, which in turn has a negative impact on its ability to disseminate information.

Our result suggests that the negative impact of low path diversity overwhelms the positive impact of network dominance. While previous research on Betweenness and social capital mostly focuses on information flow *into* a focal node, we are interested in information flow *out* of a focal node to the rest of the network. Naturally, a node that lies on many paths between other nodes ensures that the network traffic goes through it (making it an information sink), which can be a positive feature. However, such a position also makes the node a bottleneck for information flow. Here the objective is to spread the information to as many nodes as possible, not control or gather information. Thus, while in some cases high Betweenness is a positive attribute, in our setting it is a negative attribute that hampers video growth.

6.3 Impact of lagged video characteristics

Overall, we find that viewership exhibits considerable state dependence (lags 1, 2, 3, and 5 of $y_{i,t}$ are significant and positive). The first two lags are particularly influential, i.e., the viewership for any given day is significantly influenced by the viewership from the last 2 days. Interestingly however, the lag of Avg. Ratings, the primary measure of quality has no significant positive impact on viewership. While not having any ratings does have a negative impact on viewership ($Inr_{i,t}$), it is also not significant. The former might be a result of the low variation and high mean of

[12] We expect videos of central authors to receive high attention, thus leading to larger viewership. However, one might suspect otherwise if central authors are also more likely to post many videos in quick succession, thereby diluting the attention *per* video. To test if this is true, we counted the number of other videos (apart from the one in our dataset) posted by the most central authors. Specifically, of the 1806 authors in our dataset, 116 have a Betweenness value of 1 and these 116 authors posted an average of 0.53 other videos during the interval of our observation. Given our 38-day observation interval, this amounts to a mere 0.0138 videos per day. Thus, the attention fragmentation hypothesis is unlikely to be true.

the Avg. Ratings variable (see Table 1), i.e., viewers only seem to rate videos they like. So in Model 2, we evaluate the impact of ratings using the lag of Daily Num. Ratings (number of ratings received at $t-1$), but even this has no impact on viewership. We therefore conclude that both the number of ratings and the average rating received by a video do not affect viewership during the course of our observation. (It is possible that ratings are important at some stages of video growth, but not all. We explore this idea further in Section 7.) Unlike ratings, the other indicator of quality, Favorited, has a significant positive impact on viewership in both Models 1 and 2. There are two possible reasons for this. First, Favorited is a stronger endorsement than ratings—viewers may rate many videos, but may declare only a few chosen ones as Favorites. Second, videos favorited by a user are displayed prominently on her YouTube page and visible to all those who visit her page. Both these factors are likely to lead to new viewership.

Finally, we find that lagged Daily Comments and Honors have no impact on viewership.[13] This is surprising since videos with Honors are displayed prominently on YouTube and show up in searches with higher frequency. However, Honors are usually based on views, ratings, or comments received in the last few periods. Since we have already controlled for these lagged variables, it is possible that Honors has little or no new information.

6.4 Robustness checks

We now present several specification checks to demonstrate the robustness of our findings. We start by varying the number of instruments used in the estimation. In the main model (Model 1), we used lags 2 to 6 of $y_{i,t}$ and $X_{i,t}$ as instruments for Equation (12). Reducing the instruments (e.g., excluding the sixth lags of $y_{i,t}$ and $X_{i,t}$ as instruments for (12)) doesn't change the results (Model 3, Table 6). Expanding the number of instruments to include the seventh lags of $y_{i,t}$ and $X_{i,t}$ also doesn't change the results (Model 4, Table 6). Overall, any increase in the number of instruments beyond six or seven lags doesn't lead to significant improvements in consistency.[14]

Next, we investigate the results on Clustering and Betweenness. By definition, both Clustering and Betweenness are zero for nodes with zero or one friend. However, Clustering (Betweenness) is also zero when a node has more than two friends who are completely unconnected (connected). To ascertain if the significance of the results on Clustering and Betweenness stem from the metrics' inability to distinguish between such scenarios, we also estimate a model that excludes authors with less than two friends (Model 5, Table 6). We find that the results on Clustering and Betweenness remain negative and significant, providing evidence for true causality. Further, to ensure that the results for Clustering and Betweenness are not

[13] Both lagged Daily Num. Ratings and lagged Daily Comments are likely to be correlated with lagged Daily Views, which can be problematic. However, we found that normalizing these variables by lagged Daily Views also doesn't make them significant.

[14] For a sufficiently long T, the number of instruments available for Equations 10 and 11 expands rapidly. While theoretically using all instruments increases consistency, Tauchen (1986) and Ziliak (1997) have shown that there is a consistency-efficiency trade-off in finite samples. In our case, we find that one set of lagged differences and four to six sets of lagged levels are sufficient to get consistent results without any significant loss in efficiency.

driven by a small number of high-degree nodes or supernodes, we conduct a few more experiments. First, we define supernodes as nodes with 500 or more friends. We then estimate two more models. In Model 6, we exclude authors who are either supernodes or connected to supernodes, and in Model 7, we exclude authors if they have zero or one friend, or if they are themselves supernodes or connected to supernodes. The coefficients of Clustering and Betweenness remain negative and significant (see Table 7).

Next, we also investigate the impact of the centrality metric Closeness. Recall that in a two-hop setting Closeness can be expressed as $cs_i = 2 - (1/(1 + aff_i))$. The impact of Closeness on viewership $(dy_{i,t}/dcs_i)$ can therefore be expressed as $dy_{i,t}/dcs_i = \left(dy_{i,t}/daff_i\right).\left(daff_i/dcs_i\right)$. From our earlier analysis, we know that $dy_{i,t}/daff_i > 0$. Further, we can show that $daff_i/dcs_i = 1\Big/(2 - cs_i)^2 > 0$. Therefore $\left(dy_{i,t}/dcs_i\right) > 0$, i.e., even without estimating a model, it is clear that Closeness has a positive impact on viewership. Nevertheless, we confirm this empirically in Model 8 (see Table 7).

Recall that Assumption 10.3 implies that author-video fixed effects are not correlated across authors. The literature on network homophily (McPherson et al. 2001) suggests that this assumption may be unrealistic if many authors in the dataset are friends with each other. However, of the 1806 authors, only four are connected through nine links. Upon estimating the model after excluding these four authors, we find that that the results remain unchanged (see Model 9, Table 7).

Finally, we investigate whether the inclusion of additional video specific effects has any impact on the results. For this purpose, we use data on the video's category or topic. The videos in our dataset belong to three categories: Comedy, Entertainment, and News & Politics. We use this classification to estimate a model with category fixed-effects (Model 10, Table 7). We find that, on average, videos on News & Politics perform the best, followed by Entertainment videos, while Comedy videos perform the worst. However, the qualitative impact of network properties remain unchanged.

7 Early versus later video viewership worsening

We now investigate the temporal variations in the impact of network properties and lagged video characteristics on viewership. First, we divide the dataset into two parts by slicing it at t = 10. All periods up till 10 are designated as early and those after 10 form the later stage.[15] We rerun the model on both these datasets; Model 11 is estimated on the early data and Model 12 on the later data. Both are analogous to Model 1, i.e., they use the same instruments and video attributes. However, we use $Z_i = \{\mathrm{I}.(Deg < 2),\ \ln(d_i + 1),\ \ln(sd_i + 1),\ [1-\mathrm{I}.(Deg < 2)].\mathrm{C_i},\ [1-\mathrm{I}.(\mathrm{Deg} < 2)].\mathrm{B_i^N}\}$[16] as the

[15] Note that we choose 10 periods as the point of demarcation even though 'early' in the YouTube context might mean just 4–5 days. We do this primarily because we need sufficient time periods for the analysis. In Model 11, we use 6 lags of Daily views on the right hand side; this leaves us only 4 data points or less per video. If we shortened the span of the early stage, then we would have even fewer data points per video, making analysis difficult.

[16] In the estimation, $\ln(d_i+1)$ is significant in Model 11, but not in Model 12, while $\ln(sd_i+1)$ is significant in Model 12, but not in Model 11 (see Table 8). Given this pattern, the use of a composite variable like $\ln(aff_i)$ that contains both sd_i and d_i is problematic because it makes it difficult to ascertain whether $d_i$is significant in its own right (see Woolridge 2008). So we instead use $\ln(d_i+1)$and $\ln(sd_i+1)$ directly.

Table 7 Robustness to inclusion of closeness, exclusion of interconnected authors, and inclusion of category fixed effects

Dependent Variable: Log Daily Views (t)		Model 6		Model 7		Model 8		Model 9		Model 10	
		Parameter	t-stats	Parameter	t-stats	Parameter	t-stats	Parameter	t-stats	Parameter	t-stats
Lagged Dependent Variables	Log Daily views (t – 1)	0.269***	(13.60)	0.255***	(6.41)	0.257***	(9.44)	0.257***	(8.30)	0.240***	(7.76)
	Log Daily views (t – 2)	0.364***	(27.96)	0.475***	(7.97)	0.357***	(23.38)	0.357***	(17.99)	0.351***	(16.74)
	Log Daily views (t – 3)	0.046***	(4.54)	0.134**	(1.99)	0.042***	(4.02)	0.042***	(3.94)	0.045***	(4.70)
	Log Daily views (t – 4)	0.001	(0.08)	–0.144*	(−1.90)	0.006	(0.53)	0.005	(0.39)	0.003	(0.30)
	Log Daily views (t – 5)	0.023**	(2.36)	0.016	(0.50)	0.028***	(3.67)	0.028***	(3.44)	0.029***	(4.73)
Lagged Video Char-acteristics	Indicator no rating (t – 1)	–0.590	(−0.96)	0.001	(0.00)	–0.513	(−0.69)	–0.532	(−0.55)	–0.373	(−0.53)
	Avg. rating (t – 1)	0.062	(0.43)	0.116**	(2.38)	0.066	(0.37)	0.064	(0.28)	0.056	(0.36)
	Daily comments (t – 1)	0.055**	(2.09)	–0.006	(−0.08)	0.019	(0.73)	0.021	(0.91)	0.016	(0.88)
	Honors (t – 1)	–0.031	(−1.28)	0.110	(0.90)	–0.027	(−0.71)	–0.028	(−0.73)	–0.02	(−0.76)
	Favorited (t – 1)	0.007***	(9.70)	–0.001	(−0.04)	0.005***	(4.83)	0.005***	(4.19)	0.005***	(3.58)
	I. (News & Politics)									1.603***	(5.02)
	I. (Entertainment)									0.736***	(2.70)
Network Properties	Ind. zero first-deg friends	0.042	(0.07)			0.046	(0.06)	0.027	(0.03)	–1.214	(−1.24)
	Ind. zero second-deg friends	0.047	(0.07)	–0.752**	(−2.01)	2.632	(1.34)	–0.263	(−0.32)	1.140	(1.13)
	Log Degree	1.036***	(3.20)	0.017	(0.32)	0.283***	(2.57)	0.263***	(2.69)	0.213*	(1.91)
	Log Avg. friends of first-degree friends	0.215**	(2.16)	–0.015	(−0.21)			0.155*	(1.93)	0.164*	(1.83)
	Closeness					1.872**	(1.96)				
	Norm. 2-Betweenness	–2.519***	(−4.22)	–1.083**	(−2.22)	–1.540**	(−2.29)	–1.416**	(−2.17)	–1.086*	(−1.69)
	Clustering coefficient	–4.861***	(−2.89)	–1.343**	(−2.12)	–3.611**	(−2.53)	–3.184**	(−2.37)	–3.805**	(−2.29)
	Constant	0.600	(0.82)	–1.180	(1.55)	–2.186	(−1.18)	0.743	(0.69)	–0.229	(−0.27)
No. of obs., groups, instrs.		44203, 1649, 988		44203, 1649, 915		44035, 1643, 988		44085, 1645, 988		44203, 1649, 1003	
Arellano Bond (2) test (*p-value*)		–0.514, (0.608)		–0.589, (0.556)		–0.521, (0.602)		–0.496, (0.620)		–0.612, (0.540)	
Goodness of Fit Measures	$\text{Corr}(y,\bar{y})^2$, MSE, MAD	0.601, 0.560, 0.519		0.803, 0.203, 0.355		0.670, 0.476, 0.465		0.692, 0.432, 0.478		0.503, 0.942, 0.789	

Note: *** ⇒ $p \leq 0.01$, ** ⇒ $p \leq 0.05$, * ⇒ $p \leq 0.1$

Model 6 excludes both supernodes and nodes that are friends with Supernodes. Model 7 excludes supernodes, nodes that are friends with supernodes, and nodes with less than two friends. Model 8 includes Closeness in the explanatory variables. Model 9 excludes four authors who are connected to each other. Model 10 includes category fixed effects

set of network metrics, where $\mathrm{I}.(Deg < 2) = 1$ if and only if i has less than two friends. That is, we measure the effect of Clustering and Betweenness only for nodes with at least two friends. Also, in Model 11, we use six lags of $y_{i,t}$ on the right hand side because five lags are insufficient to rule out serial correlation, whereas in Model 12 we find that four lags of $y_{i,t}$ are sufficient for the same. All the results are shown in Table 8.

Network Properties We find that the impact of network properties on early growth is significantly different from that on later growth. Specifically, first-degree friends play an important role in the initial stages, though their role is negligible in later stages. The coefficient of $\ln(d_i+1)$ in Model 11 is positive and significant, while it is insignificant in Model 12. On the other hand, second-degree friends play a minimal role in the early stages, but transform into key drivers of growth later on (see Table 8). So while first-degree friends are essential for initial take off and to spread the word early on, the spread stops if there are no second-degree friends, curtailing the video's success. In sum, both first and second-degree friends are important. Deficiencies in either would hamper growth.

Further, both Clustering and Betweenness do not affect early diffusion, but have a negative effect on later viewership (see Table 8). Recall that both high Clustering and Betweenness make it difficult for information to spread beyond the local network, but this effect is visible only in the later periods. These results again highlight the difficulties in extrapolating global diffusion from individual level peer-effects studies. In sum, these results not only provide additional support for our earlier hypotheses, but also highlight the need for modeling the different stages of growth carefully.

Video Characteristics In the early stages, lagged viewership is not a good predictor of current viewership. However, this changes during the later stages (see Table 8). Next, we find that lagged video characteristics have a significant impact on viewership in the initial periods, though they become insignificant later. Specifically, videos that have not been rated perform worse in the early stages. In Model 11, the coefficient of lagged Indicator no Rating is negative and almost significant. However, lagged Avg. Ratings remains insignificant in both models. Taken together, these results suggest that while the actual rating may not matter, the mere act of getting rated can enhance the popularity of new videos. Lags of Favorited and Daily Comments have a positive impact on viewership in the early stages, but none later on. Comments on YouTube are usually of two types: 1) comments pertinent to the video, and 2) spam or flame-war comments that are irrelevant to the video. Usually, spam comments increase over time while comments relevant to the video decrease. It is possible that our results reflect this effect. Overall these results suggest that lagged video characteristics are important drivers of growth initially though their importance wears off over time.

8 Managerial implications

We now use our estimates to explore the marginal benefit of targeting nodes in key network positions. We first provide an analytical characterization of the impact of network properties on viewership over T periods and then present results from counterfactual experiments.

Table 8 Estimation results: early vs. later viewership

Dependent Variable: Log Daily Views (t)		Model 11		Model 12	
		Early: 10 ≥ t		Later: t > 10	
		Parameter	t-stats	Parameter	t-stats
Lagged Dependent Variables	Log Daily views (t – 1)	0.021	(0.15)	0.212***	(6.19)
	Log Daily views (t – 2)	0.177***	(4.08)	0.381***	(16.12)
	Log Daily views (t – 3)	0.019	(0.28)	0.115***	(5.23)
	Log Daily views (t – 4)	–0.077	(−1.53)	0.12***	(3.03)
	Log Daily views (t – 5)	0.103*	(1.90)		
	Log Daily views (t – 6)	0.012	(1.13)		
Lagged Video Char-acteristics	Indicator no rating (t – 1)	–1.419	(−1.55)	–0.357	(−0.69)
	Avg. rating (t – 1)	–0.088	(−0.44)	0.028	(0.21)
	Daily comments (t – 1)	0.063**	(2.48)	–0.001	(−0.10)
	Honors (t – 1)	–0.009	(−0.45)	–0.025	(−0.43)
	Favorited (t – 1)	0.022***	(3.04)	0.002	(1.17)
Network Properties	I.(Degree<2)	–7.877	(−1.07)	–2.442*	(−1.69)
	Log Degree	2.427**	(2.17)	–0.19	(−1.50)
	Log Second-degree friends	–0.697	(−1.37)	0.134*	(1.70)
	[1- I.(Degree<2)]. Norm. 2-Betweenness	–11.403	(−1.38)	–2.689*	(−1.78)
	[1- I.(Degree<2)]. Clustering coefficient	–1.755	(−0.17)	–6.898***	(−2.73)
	Constant	9.587**	(1.16)	2.758	(0.47)
No. of Observations, groups, instruments		6545, 1648, 61		33238, 1618, 451	
Arellano-Bond (2) test (*p-value*)		0.514, (0.607)		0.206, (0.837)	
Goodness of Fit Measures	$\text{Corr}(y,\bar{y})^2$	0.13		0.699	
	MSE	3.529		0.417	
	MAD	1.178		0.392	

Note: *** ⇒ $p \leq 0.01$, ** ⇒ $p \leq 0.05$, * ⇒ $p \leq 0.1$

Equation (10) suggests that network properties have both a Direct and an Indirect impact on $y_{i,t}$. β, the coefficient of Z_i, is simply the Direct impact of network properties on $y_{i,t}$. The Indirect effect stems from the lagged variables $y_{i,t-k}$s, which are also functions of Z_i. Hence the Total impact of network properties on $y_{i,t}$ in any time-period t can be expressed as:

$$\text{Total\_impact}_{i,t} = \text{Direct\_impact}_{i,t} + \text{Indirect\_impact}_{i,t} \tag{17}$$

For a total of T periods, we can calculate the Cumulative impact of Z_i on total views as:

$$\text{Cumulative\_impact}_{i,T} = \exp\left(\sum_{t=1}^{T} \text{Total\_impact}_{i,t}\right) \tag{18}$$

The use of the exponential function in (18) is necessitated by our use of the transformed variable $y_{i,t}$ (i.e., we use logarithm of Daily views$_{i,t}$ in Equation (10)). There exists no closed-form equation for Cumulative impact, so we illustrate its magnitude using numerical experiments.

Consider a manager who wants to initiate a social media campaign by seeding a specific piece of information with a group of 200 authors on YouTube. Her objective is to reach as many people as possible in a one-month period. She has two options in the choice of seeds: she can choose seeds randomly or she can pick seeds based on their social network properties. In this context, we examine the value of network-based seed selection using counterfactual experiments. For the purpose of illustration, we consider two network metrics—Degree and Betweenness.

We start with first-degree friends. Consider two scenarios: In scenario 1, the manager picks seeds by randomly sampling from the full distribution of 1806 authors. In scenario 2, she chooses seeds by sampling authors who are in the 90th–100th percentile of first-degree friends. Figure 9 illustrates the viewership patterns for the two sets of seeds over 31 days. When the manager picks seeds from the topmost decile, the median of views obtained by her videos in one month is about 2000, whereas this number is around 750 when she chooses seeds randomly, indicating more than a two-fold increase in viewership. In this comparison, we set all other network and video attributes to zero. Hence, these gains stem from the differences in first-degree friends alone. Next, we illustrate the gains in viewership from choosing seeds with low Betweenness. In this case, we only consider the distribution defined by authors who have two or more friends. Figure 10 compares the growth pattern of videos seeded by a random draw of authors with that of videos seeded by authors whose Betweenness values are in the bottom most decile. Our results indicate a significant increase in median viewership over a one month period (see Fig. 10).

These gains can be further enhanced by picking seeds that have a combination of positive network properties, e.g., those with a large number of first and second-degree friends and low Betweenness. Further, since seeds in prominent network positions tend to have other positive traits, such as engaging personalities and good

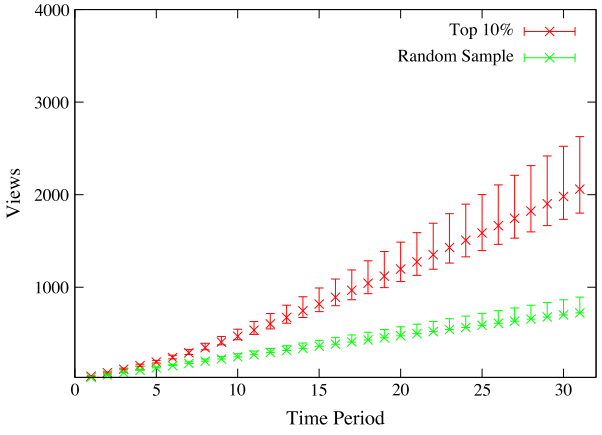


Fig. 9 Comparison of viewership for two samples of seeds—random and topmost decile of first-degree friends. The cross represents the median; the top and bottom dashes represent the 25th and 75th percentile.(We set all other network and video properties to zero. Also, we set $\eta_i = 1$ and $\kappa_X = \kappa_Z = 0.5$)

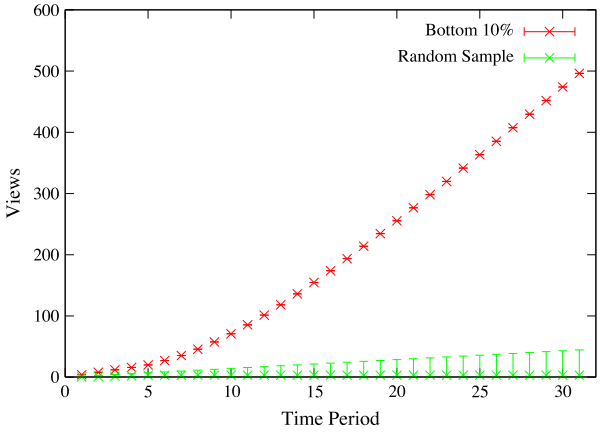


Fig. 10 Comparison of viewership for two samples of seeds—random and bottommost decile of betweenness centrality. The cross represents the median; the top and bottom dashes represent the 25th and 75th percentile. (We set all other network and video properties to zero. Also, we set $\eta_i = 1$ and $\kappa_X = \kappa_Z = 0.5$)

reputation, these benefits are likely to be larger in reality. In sum, our estimates suggest that there exist significant ROI from network-based seed selection strategies.

9 Conclusion, limitations and future research

In this paper, we examine how the size and structure of a node's local network affects the aggregate diffusion of the products seeded by it. We study this problem in the context of YouTube, the popular video-sharing website, where videos can be interpreted as products and the users (or authors) who post them as seeds. We present a descriptive dynamic model of video growth, where viewership is modeled as a function of video attributes and network properties of the author. While identification issues are common in our problem setting, we are able to control for the various sources of endogeneity using a rich data set in tandem with a sophisticated estimation methodology.

However, our paper is not without limitations. As discussed in Section 3.2.2, we are limited by data; we don't have data on the complete social network within YouTube and on authors' network beyond YouTube. This hampers our ability to explore the impact of global network properties on viewership. With the advent of YouTube widgets on forums like Facebook, it might become easier to combine data from multiple sources in the future. Studies on such composite networks may provide new and useful insights.

Nevertheless, our paper makes some important contributions to the literature. First, we empirically demonstrate that the size and structure of the local network around a node has a significant impact on the overall diffusion of products seeded by it. While there exist many studies on individual-level peer-effects, to our knowledge this is the first empirical study that documents the causal effect of a seed's local network on macro-level diffusion. Second, we discuss and clarify the data requirements and methodological strategies required to overcome the endogeneity problems in such settings. Third, we document the temporal variations in the effect of a seed's local network on product diffusion. Specifically, we find that network properties that drive early diffusion are fundamentally different from those affecting later diffusion. Fourth, our results provide guidelines to managers conducting buzz marketing by aiding them in the identification of seeds that provide the best ROI.

Finally, our study sheds light on the substantive factors that affect video consumption in YouTube. While video-sharing websites have become increasingly popular in the last few years, managers have limited information on utilizing this new medium as a marketing tool (Feed Company 2008). Our study represents an important first step towards a better understanding of the online video market.

Acknowledgement Discussions with Dina Mayzlin, Harikesh Nair, Sridhar Naryanan, and Jiwoong Shin have greatly improved this paper. Comments from the Editor, Greg Allenby, and two anonymous reviewers have also helped the paper considerably. Finally, thanks are also due to the participants of the PhD Student Research Workshop at the Yale School of Management 2009, NASMEI 2009, UT Dallas Forms Conference 2009, Marketing Science Conference 2010, Marketing Dynamics Conference 2010, Stanford Marketing Seminar 2010, Haas Marketing Seminar 2010, and University of Washington Marketing Seminar 2011, for their feedback.

Appendix

A.1. Technical details on data collection

We collected the YouTube data using a set of custom scripts written in Perl. We bootstrapped the data collection process using a Perl script to find the list of newly uploaded videos to YouTube. We then used a separate script to periodically access the statistics page corresponding to the videos, collected the relevant video characteristics, and stored them in a MySql database for later analysis. The Perl script parsed the HTML content of the statistics pages by looking for key markers in the HTML tags associated with the various video related data. We used Perl's "HTML Parser" library to perform the data extraction.

Concurrently, we used a separate set of Perl scripts deployed on a cluster of workstations to collect data on the social network of the authors that have seeded the videos. The video page provided the link to the author's page, which in turn contained data on the author and the author's social network. For instance, the author's page contains the identities of his or her directly connected friends. We used a cluster of workstations in order to collect a snapshot of the social network structure within 4 days. The entire process was managed by a centralized controller that was responsible for handing out the network crawling tasks to the individual computers, monitoring their progress, and occasionally reissuing tasks if they are not completed within a specified time interval. The social network data was also stored in a MySql database and then analyzed using custom programs written in C. The analysis yielded the various social network metrics that we use in the paper, e.g., degree, number of second-degree friends, clustering, and Betweenness centrality.

We make all of the above scripts available for researchers interested in collecting YouTube data, at http://faculty.gsm.ucdavis.edu/~hema/youtube/. We do note that YouTube changes its webpage layout and data format regularly, so it is likely that our scripts would have to be modified to account for recent changes.

A.2. Initial conditions assumption

This assumption ensures that the impact of the unobserved fixed effect η_i on growth $(y_{i,t})$ remains constant over time. Let $A_K = \sum_{k=1}^{K} \alpha_k$ and $\eta_0 = \frac{(1+\kappa_X\gamma+\kappa_Z\beta)}{1-A_K}$. Recall that

$X_{i,t}$ and Z_i are linearly correlated with η_i. Hence, they can be expressed as $X_{i,t} = \lambda_X X'_{\mathrm{i,t}} + \kappa_X \eta_i + \delta_{i,t}$ and $Z_i = \lambda_Z Z'_{\mathrm{i}} + \kappa_Z \eta_i + \xi_i$, where $X'_{\mathrm{i,t}}$ and Z'_{i} is not correlated to η_i and $\delta_{i,t}$ and ξ_i are random shocks such that $E(\delta_{i,t}) = 0$, $E(\xi_i) = 0$ and $E(\delta_{i,t}.\eta_i) = 0$, $E(\xi_i.\eta_i) = 0$. Using these expansions recursively in tandem with Assumption 10.7, we now show that the impact of η_i has a constant $y_{i,t}$ in all periods.

Lemma 1: The effect of the unobserved fixed effect η_i on $y_{i,t}$ is constant for all periods and is equal to η_0.

Proof: Period 1: Consider the growth equation for $t = 1$. By Assumption (10.7a), we have:

$$y_{i,1} = c + \eta_i \left(\frac{1 + \kappa_X \gamma + \kappa_Z \beta A_K}{1 - A_K} \right) + \beta Z_i + \varepsilon_{i,1}$$

We know that Z_i can be expressed as $Z_i = \lambda_Z Z'_{\mathrm{i}} + \kappa_Z \eta_i + \xi_i$, where Z'_i and ξ_i are not correlated with η_i. After substituting for Z_i, (10.7a) can be expressed as follows:

$$y_{i,1} = c + \eta_0 \eta_i + \lambda_Z \beta Z'_{\mathrm{i}} + \beta \xi_i + \varepsilon_{i,1}.$$

Since Z'_{i}, ξ_i, $\varepsilon_{i,1}$ and c are independent of η_i, the coefficient of η_i is given by η_0.

Periods 2 to K: Next, consider the growth equations for the remaining (K-1) initial periods, i.e., $2 \leqslant t \leqslant K$. From Assumption (10.7b), we have:

$$y_{i,t} = c + \eta_i \left(\frac{1 + (\kappa_Z \beta + \kappa_X \gamma) A_K}{1 - A_K} \right) + \gamma X_{i,t-1} + \beta Z_i + \varepsilon_{i,t}, \quad \forall\ 2 \ \leq t \leq K$$

As before, we can substitute for Z_i in (10.7b). In addition, we can also substitute for $X_{i,t-1}$ as follows: $X_{i,t-1} = \lambda_X X'_{\mathrm{i,t-1}} + \kappa_X \eta_i + \delta_{i,t-1}$). Thus, (10.7b) can be rewritten as:

$$y_{i,t} = c + \eta_0 \eta_i + \lambda_x \gamma X'_{\mathrm{i,t-1}} + \gamma \delta_{i,t-1} + \lambda_Z \beta Z'_{\mathrm{i}} + \beta \xi_i + \varepsilon_{i,t}, \ \forall\ 1 < t \leq K$$

Since $X'_{i,t-1}$, Z'_i, $\delta_{i,t-1}$, ξ_i, $\varepsilon_{i,t}$ and c are not correlated with η_i, the coefficient of η_i is given by η_0.

Period K+1: Next, consider the growth in $(K+1)^{th}$ period (from Equation 10)

$$y_{i,K+1} = c + \sum_{k=1}^{K} \alpha_k y_{i,t-k} + \gamma X_{i,K} + \beta Z_i + \eta_i + \varepsilon_{i,K+1}$$

The η_i term in $\gamma X_{i,K} + \beta Z_i + \eta_i$ is given by $(1 + \kappa_X \gamma + \kappa_Z \beta)\eta_i$. We know that each $y_{i,t-k}$ term in $\sum_{k=1}^{K} \alpha_k y_{i,t-k}$ contains η_0 and therefore, the total contribution $\sum_{k=1}^{K} \alpha_k y_{i,t-k}$ to the η_i term is $A_K \eta_0 \eta_i$. Thus, the complete coefficient of η_i in Equation (10) is given by η_0.

Periods K+2 to T: Now consider the growth in the $(K+2)^{th}$ period:

$$y_{i,K+2} = c + \sum_{k=2}^{K+1} \alpha_k y_{i,t-k} + \gamma X_{i,K+1} + \beta Z_i + \eta_i + \varepsilon_{i,K+2}$$

As before, the η_i term from $\gamma X_{i,K+1} + \beta Z_i + \eta_i$ is $(1 + \kappa_X \gamma + \kappa_Z \beta)\eta_i$ and since all $y_{i,t-k}$s in $\sum_{k=2}^{K+1} \alpha_k y_{i,t-k}$ contain $\eta_0 \eta_i$, their contribution to the η_i term is $A_K \eta_0 \eta_i$. Thus, the total η_i term in $y_{i,K+2}$ is $\eta_0 \eta_i$, which is the same as the η_i term in $y_{i,1}, \ldots, y_{i,K+1}$. Next,

to show that the coefficient of η_i in $y_{i,K+3}$ is η_0, we use two facts: 1) the functional form of $y_{i,K+3}$ is the same as that of $y_{i,K+2}$, and 2) the coefficient of η_i in all the lagged terms $y_{i,K+2}, ..., y_{i,3}$, is η_0. So using the same technique as above, we can show that the coefficient of η_i in $y_{i,K+3}$ is also η_0. Similarly, by recursive induction, the coefficient of $y_{i,K+j}$ is also η_0 for all $j > 3$. Thus, all $y_{i,t}$s can be expressed as follows:

$$y_{i,t} = \eta_0\eta_i + f_t\left(X'_{i,t-1}, ..., X'_{i,1}, \delta_{i,t-1}, ..., \delta_{i,1}, Z'_i, \xi_i, \varepsilon_{i,t}, ..., \varepsilon_{i,1}\right)$$

A.3. Moment conditions for first-differenced equation

The first-differenced equations are given by:

$$\Delta y_{i,t} = \sum_{k=1}^{K} \alpha_k \Delta y_{i,t-k} + \gamma \Delta X_{i,t-1} + \Delta\varepsilon_{i,t} \tag{11}$$

We specify two sets of moment conditions, 12(a) and 12(b), for Equation 11. In Proposition 1, we show that these moment conditions are true.

Proposition 1: For$p = K, K+1,..., t-2$, $E(y_{i,p}.\Delta\varepsilon_{i,t}) = 0$ and $E(X_{i,p}.\Delta\varepsilon_{i,t}) = 0$.
Proof: We start with moment conditions $E(X_{i,p}.\Delta\varepsilon_{i,t}) = 0$ where $p = K$, $K+1, ..., t-2$. From Assumption (10.4), we have $E(X_{i,t}.\varepsilon_{i,s}) = 0$ if s $> t$ and $E(X_{i,t}.\varepsilon_{i,s}) \neq 0$ if $s \leq t$. This implies that $X_{i,p}$ is uncorrelated to $\Delta\varepsilon_{i,t}$ for all $p \leq t-2$. Next, consider the moment conditions $E(y_{i,p}.\Delta\varepsilon_{i,t}) = 0$, where $p = K, K+1, ..., t-2$. From Lemma 1, we know that all $y_{i,t}$ terms can be written as follows:

$$y_{i,p} = \eta_0\eta_i + f_p\left(X'_{i,p-1}, ..., X'_{i,1}, \delta_{i,p-1}, ..., \delta_{i,1}, Z'_i, \xi_i, \varepsilon_{i,p}, ..., \varepsilon_{i,1}\right)$$

From Assumption (10.1), we know that $E(\eta_i.\Delta\varepsilon_{i,t}) = 0$. Also, from Assumptions 10.1, 10.4, 10.5, and 10.6, we know that $f_p(.)$ is uncorrelated with $\Delta\varepsilon_{i,t}$ for all $p \leq t-2$. So $E(y_{i,p}.\Delta\varepsilon_{i,t}) = 0$.

A.4. Moment conditions for level equation

The level equations when $t > K$ are given by:

$$y_{i,t} = c + \sum_{k=1}^{K} \alpha_k y_{i,t-k} + \gamma X_{i,t-1} + \beta Z_i + \eta_i + \varepsilon_{i,t} \tag{10}$$

We specify two sets of moment conditions for Equation (10) (see Equations 13a and 13b). In Proposition 2, we show that these moment conditions are true.

Proposition 2: For $p = K, ..., t-1$, $E\left(\Delta y_{i,p}.(\eta_i + \varepsilon_{i,t})\right) = 0$ and $E\left(\Delta X_{i,p}.(\eta_i + \varepsilon_{i,t})\right) = 0$.
Proof: We start with $E\left(\Delta X_{i,p}.(\eta_i + \varepsilon_{i,t})\right) = 0$. From Assumption (10.4), we have $E(X_{i,t}.\varepsilon_{i,s}) = 0$ if $s > t$ and $E(X_{i,t}.\varepsilon_{i,s}) \neq 0$ if $s \leq t$. This implies that $X_{i,p}$ is uncorrelated to $\varepsilon_{i,t}$ for all $p \leq t-1$ and by extension $\Delta X_{i,p}$ is uncorrelated with $\varepsilon_{i,t}$ for $p \leq t-1$. From Assumption (10.5), we know that $X_{i,p}$s are linearly correlated with η_i, which implies that $\Delta X_{i,t-1}$ is uncorrelated with η_i. Thus, $\Delta X_{i,p}$ is uncorrelated to both η_i and $\varepsilon_{i,t}$. Next, consider the moments $E\left(\Delta y_{i,p}.(\eta_i + \varepsilon_{i,t})\right) = 0$. Recall that, for

$p \geq K+1$, $y_{i,p}$ can be written as $y_{i,p} = \eta_0 \eta_i + f_p(.)$, where $f_p(.)$ is independent of η_i. Hence, $\Delta y_{i,p}$ can be written as follows:

$$\Delta y_{i,p} = f_p\left(X'_{i,p-1}, ..., X'_{i,1}, Z'_i, \varepsilon_{i,p}, ..., \varepsilon_{i,1}\right) - f_{p-1}\left(X'_{i,p-2}, ..., X'_{i,1}, Z'_i, \varepsilon_{i,t-1}, ..., \varepsilon_{i,1}\right)$$

Thus, the moment condition, $E\left(\Delta y_{i,p}.(\eta_i + \varepsilon_{i,t})\right) = 0$, can be expressed as follows:

$$E\Big(\Big(f_p\left(X'_{i,p-1}, .., X'_{i,1}, \delta_{i,p-1}, .., \delta_{i,1}, Z'_i, \xi_i, \varepsilon_{i,p}, .., \varepsilon_{i,1}\right) - f_{p-1}\left(X'_{i,p-2}, .., X'_{i,1}, \delta_{i,p-2}, .., \delta_{i,1}, Z'_i, \xi_i, \varepsilon_{i,t-1}, .., \varepsilon_{i,1}\right)\Big).(\eta_i + \varepsilon_{i,t})\Big) = 0$$

We already know that $f_p(.) - f_{p-1}(.)$ is not correlated with η_i for all p. Following Assumptions (10.1) and (10.4), it is easy to see that $f_p(.) - f_{p-1}(.)$ is also uncorrelated to $\varepsilon_{i,t}$ for all $p \leq t-1$. Therefore, $E\left(\Delta y_{i,p}.(\eta_i + \varepsilon_{i,t})\right) = 0$. Finally, note that $f_p(.) - f_{p-1}(.)$ is correlated with $X_{i,t-1}$, Z_i and $y_{i,t-k}$s for $p \leq t-1$ ensuring that $\Delta y_{i,p}$s are good instruments for all these terms.

References

Acemoglu, D., & Robinson, J. (2001). A theory of political transition. *The American Economic Review, 91* (4), 938–963.

Anderson, T. W., & Hsaio, C. (1981). Estimation of dynamic models with error components. *Journal of the American Statistical Association, 76*(375), 598–606.

Arellano, M., & Bond, S. (1991). Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. *The Review of Economic Studies, 58*, 277–97.

Bandeira, O., & Rasul, I. (2006). Social networks and technology adoption in Northern Mozambique. *The Economic Journal, 116*, 869–902.

Bandiera, O., Barankay, I., & Rasul, I. (2009). Social connections and incentives in the workplace: evidence from personnel data. *Econometrica., 77*, 1047–94.

Barabasi, A. L., Albert, R., & Jeong, H. (2000). Scale-free characteristics of random networks: the topology of the world wide web. *Physica A, 281*, 69–77.

Barry, K. (2009). Ford bets the fiesta on social networking. *Wired*.

Bass, F. M. (1969). A new product growth model for consumer durables. *Management Science, 15*, 215–227.

Bertrand, M., Luttmer, E. F. P., & Mullainathan, S. (2000). Network effects and welfare cultures. *Quarterly Journal of Economics, 115*, 1019–1056.

Blundell, R., & Bond, S. (1998). Initial conditions and moment restrictionsin dynamic panel data models. *Journal of Econometrics., 87*, 115–43.

Borgatti, S. P. G., Jones, C., & Everett, M. G. (1998). Network measures of social capital. *Connections, 21*, 27–36.

Borgatti, S. P., Carley, K. M., & Krackhardt, D. (2006). On the robustness of centrality measures under conditions of imperfect data. *Social Networks, 28*, 124–136.

Borgatti, S. P., & Everett, M. G. (2006). A graph-theoretic perspective on centrality. *Social Networks, 28*, 466–84.

Bramoullé, Y., Djebbari, H., & Fortin, B. (2009). Identification of peer effects through social networks. *Journal of Econometrics., 150*, 41–55.

Brock, W. A., & Durlauf, S. N. (2007). Identification of binary choice models with social interactions. *Journal of Econometrics, 140*(1), 52–75.

Burt, R. (1995). *Structural holes: The social structure of competition*. Harvard University Press.

Clark, C. C., Doraszelski, U., & Draganska, M. (2009). The effect of advertising on brand awareness and perceived quality: an empirical investigation using panel data. *Quantitative Marketing and Economics, 7*, 207–236.

Coleman, J. S., Katz, E., & Menzel, H. (1966). *Medical innovation: A diffusion study.* Indianapolis: Bobb-Merrill.

Durlauf, S., Johnson, P., & Temple, J. (2005). Growth econometrics. In P. Aghion & S. Durlauf (Eds.), *Handbook of econometric growth* (Vol. 1A, pp. 555–677). Amsterdam: North-Holland.
Everett, M. G., & Borgatti, S. P. (2005). Ego-network betweenness. *Social Networks, 27*(1), 31–38.
Feed Company. (2008). Viral video marketing survey: The agency perspective.
Feld, S. L. (1991). Why your friends have more friends than you do. *The American Journal of Sociology, 96*(6), 1464–77.
Freeman, L. C. (1979). Centrality in social networks: a conceptual clarification. *Social Networks.* pp. 1–21.
Friedkin, N. E. (1991). Theoretical foundations for centrality measures. *The American Journal of Sociology, 96*, 1478–1504.
Greenberg, K. (2010). Ford fiesta movement shifts into high gear. *Marketing Daily.*
Girvan, M., & Newman, M. E. J. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Science of the United States of America, 99*(12), 7821–26.
Goldenberg, J., Sangman, H., Lehmann, D. R., & Hong, J. W. (2009). The role of hubs in the adoption process. *Journal of Marketing, 73*, 1–13.
Gould, R. V., & Fernandez, R. M. (1989). Structures of mediation: A formal approach to brokerage in transaction networks. In C. C. Clogg & A. Arbor (Eds.), *Sociological methodology* (pp. 89–126). MI: Blackwell.
Granovetter, M. (1973). The strength of weak ties. *The American Journal of Sociology, 78*(6), 1360–80.
Hansen, B. E. (2008). *Econometrics*. available at: http://www.ssc.wisc.edu/~bhansen/econometrics/Econometrics2008.pdf.
Hartmann, W. R., Manchanda, P., Nair, H., Bothner, M., Dodds, P., Godes, D., et al. (2008). Modeling social interactions: identification, empirical methods and policy implications. *Marketing Letters, 19* (3).
Hitwise Experian. (2010). Top 20 sites and engines. available at: http://www.hitwise.com/us/datacenter/main/.
Katona, Z., Zubcsek, P. P., & Sarvary, M. (2009). Network effects and personal influences: Diffusion of an online social network. Working paper.
Katz, E., & Lazarsfeld, P. F. (1955). *Personal influence: The part played by people in the flow of mass communications*. Glencoe: Free.
Mahajan, V., Muller, E., & Wind, Y. (2000). New product diffusion models: From theory to practice. In V. Majan, E. Muller, & Y. Wind (Eds.), *New product diffusion models*. Boston: Kluwer.
Manski, C. F. (1993). Identification of endogenous social effects: the reflection problem. *The Review of Economic Studies, 60*(3), 531–42.
McCracken, G. (2010). How Ford got social marketing right. *The Conversation, Harvard Business Review.*
McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a feather: Homophily in social networks.
Mislove, A., Marcon, M., Gummadi, K., Druschel, P., & Bhattacharjee, B. (2007). Measurement and Analysis of Online Social Networks. In Proceedings of the *5th ACM/USENIX Internet Measurement Conference*, San Diego, CA.
Moynihan, R. (2008). Key opinion leaders: independent experts or drug representatives in disguise. *British Medical Journal, 336*, 1402–03.
Nair, H., Manchanda, P., & Bhatia, T. (2009). Asymmetric social interactions in physician prescription behavior: The role of opinion leaders. Working paper.
Nickell, S. (1981). Biases in dynamic models with fixed effects. *Econometrica, 39*, 359–87.
Nielson Online. (2010). Nielsen net ratings April 2010.
Rogers, E. M. (2003). *Diffusion of innovations* (5th ed.). New York: Free.
Sacerdote, B. (2001). Peer effects with random assignment: results for Dartmouth roommates. *Quarterly Journal of Economics, 116*, 681–704.
Stephen A. T., & Toubia, O. (2010). Deriving value from social commerce networks. forthcoming *Journal of Marketing Research*.
Tajfel, H., & Turner, J. C. (1986). The social identity theory of inter-group behavior. In S. Worchel & W. G. Austin (Eds.), *Psychology of intergroup relations* (2nd ed., pp. 7–24). Chicago: Nelson-Hall.
Tauchen, G. (1986). Statistical properties of generalized method of moments estimators of structural parameters obtained from financial market data. *Journal of Business and Economic Statistics, 4*(4), 397–416.
Trogdon, J., Nonnemaker, J., & Pais, J. (2008). Peer effects in adolescent overweight. *Journal of Health Economics, 27*(5), 1388–1399.

Tucker, C. (2008). Identifying formal and informal influence in technology adoption with network externalities. *Management Science, 55*(12), 2024–2039.

Valente, T. W., & Pumpuang, P. (2007). Identifying opinion leaders to promote behavior changes. *Health Education & Behavior, 34*, 881–96.

Watts, D. J., & Dodds, P. S. (2007). Influentials, networks, and public opinion formation. *Journal of Consumer Research, 34*, 441–58.

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'Small-World' networks. *Nature, 393*(4), 440–42.

Windmeijer, F. (2005). A finite sample correction for the variance of linear efficient two-step GMM estimators. *Journal of Econometrics, 126*(1), 25–51.

Woolridge, J. (2008). *Introductory econometrics: A modern approach*. 4th ed., South-Western College Pub.

Ziliak, J. P. (1997). Efficient estimation with panel data when instruments are predetermined: an empirical comparison of moment-condition estimators. *Journal of Business and Economic Statistics, 15*(4), 419–31.