Socially Embedded Knowledge Networks and the Making of Opinion Leaders: Evidence from Twitter[1]

Hao Bo
University of Maryland

September 2020

Abstract

This paper focuses on knowledge markets exploring how network relationships between knowledge consumers impact the equilibrium number of opinion leaders. Both a theoretical model and empirical analysis show that there'll be more opinion leaders in a knowledge market if the most active knowledge consumers occupy more central positions in a social network connecting consumers. The model formalizes the following story. Knowledge consumers are embedded in network relationships through which they influence each other on which opinion providers they pay attention to. If the most active (thus most capable to influence) knowledge consumers occupy more central network positions, consumer attention gravitates toward some opinion providers, and this turns more opinion providers into opinion leaders. The model inspires and is supported by empirical analysis using a Twitter network and associated tweets. First, unsupervised machine learning is used to define knowledge markets: topic modeling finds 45 topics in tweets, network community detection yields 4 nearly isolated Twitter sub-networks, and a knowledge market is then defined by a combination of one topic and one sub-network. Second, with each knowledge market being a unit of observation, we define variables and test our theoretical predictions. This is the first paper to formally define opinion leaders, knowledge markets, and consumer attention. While the existing literature emphasizes the role of opinion providers' network positions on the making of opinion leaders, this work shows the network positions of active consumers matter because active consumers serve as a propagation machine.

Key words: Social networks, Opinion leaders, Topic modeling, Knowledge Market

[1] I thank Petter Murrell, Sebastian Galiani, and Ben Li for very valuable comments.

1. Introduction: motivation, related literature, summary, and contributions

The importance of knowledge markets is acknowledged in economics (Coase and Wang 2012), but research on knowledge markets is scarce compared with work on commodity, service, or factor markets. The most important economic insight on knowledge markets is still from Hayek's 1945 paper "The Use of Knowledge in Society". Hayek (1945) argues that the decentralized use of knowledge in a knowledge market is essential to the workings of a society and no centralized mechanism can substitute. In that paper, the underlying network structure of the knowledge market is assumed symmetric and regular: each agent is both a knowledge supplier and a knowledge consumer, and their interactions are homogenous, that is, well-stirred. This assumption helps Hayek and generations of his readers to focus on his specific insights, but in the social networks underlying real knowledge markets, it is often the case that there are "a handful of very influential celebrities on one side and many millions of people with just a couple of followers on the other" (Bruner 2013). This leads to the topic of opinion leaders; this paper falls under this rubric.

To whom people pay attention is an extensively studied topic (Jackson 2019; Katz and Lazarsfeld 1955). A common theme in knowledge markets is self-organized concentration of to whom knowledge consumers pay attention, which arises from a decentralization process where social influence between consumers plays an important role (Salganik, Dodd, and Watts 2006). However, there has been no or few formal treatments of the question of how the number of opinion leaders on a topic (or in a knowledge market) is determined. This paper shows that the structure of knowledge-consumer interactions has a significant impact on the number of opinion leaders. To put it in another way, we explore how knowledge market structure is affected by network effects through pre-existing relationships between consumers.

This is an important step in understanding the development and performance of knowledge markets. First, more opinion leaders lead to more competition in a knowledge market; more competition usually implies higher information quality and more responsiveness to consumer demand. Second, more opinion leaders can also indicate more opinion diversity; such diversity keeps non-mainstream opinions alive, enhancing a society's adaptive efficiency, a concept emphasized by North (1990) and North, Wallis, and Weingast (2009).

Since we our paper evolves around the concept of network effects, let's explain what we mean by network effects in our paper. Network effects, especially for information products and technologies, are well documented in the field of industrial organization (Belleflamme and Peitz 2015). In this paper we specifically define network effects as existing when a consumer of a product is connected with other consumers in a network and the consumption activities of his connected consumers increase his consumption activity. This is a kind of social influence. Our definition is slightly different from the usual definition one, as given by Belleflamme and Peitz (2015): "a product is said to exhibit network effects if each user's utility is increasing in the number of other users of that product or of products compatible with it." Our definition is slightly different with two considerations: (1) it's often the case that not all other users' consumption would have an impact on a user, but rather his connected consumers would; (2) the number of other users is only one facet of "other users' consumption activities" and the amount of others' consumption should also matter.

Consistent with Belleflamme and Peitz's (2015) definition, the current literature on network effects mainly focuses on well-stirred interactions: it is usually implicitly assumed that the probability of consumers to interact with and influence each other depends only on the proportions of consumer types. To put it in another way, the existing literature mainly focuses on complete networks: everybody is interacting or adjacent to everybody. Such an assumption of symmetry usually makes things tractable and thus facilitates many models yielding important and interesting economic insights. (See Belleflamme and Peitz (2015) for various examples.) In this paper we nevertheless obtain insights from networked social influence, i.e., consumers influence each other based on where they are embedded in a network connecting them, rather than homogenously; this is why we have "socially embedded knowledge network" in the title. As will be seen, pre-existing relationships between consumers have an impact on the number of suppliers. In our paper we thus emphasize the role played by specific social embeddedness of economic agents, the importance of which is also emphasized by Grannovetter (2017) and Thurner, Hanel, and Klimek (2018).

In emphasizing social embeddedness, we put network positions foremost in our story. There is a large amount of literature on various implications of agents' network positions as surveyed by Easley and Kleinberg (2010), Jackson (2019), and Newman (2018). Major studies related to ours are on the relationship between the social

network position of a knowledge supplier and the supplier's influence (Grannovetter 1973) and on the concentration of attention to a few opinion leaders (Barabási and Albert 1999). Variations under these two themes are surveyed by Easley and Kleinberg (2010), Page (2018), and Newman (2018). The importance of social networks or social embeddedness (Grannovetter 1985 and 2017) is a common theme. However, formal studies that aggregate individual network positions to understand macro-level implications are few. Engle, Macy, and Claxton (2010) provide one example; so is our paper, given our third motivation to connect micro-level structure to macro-level performance (Schelling 2006).

Having talked about motivation and existing literature, we now intuitively summarize our theoretical insight, empirical methodology and findings, and our contributions.

Consumers are connected with each other in a social network. If the most active knowledge consumers occupy more central network positions, they will, thanks to their intensive consumption activity and network effects, generate more influence on other consumers with respect to which opinion producers to pay attention to[2]. Then there'll be more concentration of consumer attention to some opinion producers. Opinion producers are uniformly distributed everywhere and can become opinion leaders with enough number of followers. The concentration process leads to more opinion leaders, in contrast to uniformly distributed attention that is not sufficient anywhere to give rise to opinion leaders. Another theoretical implication of the concentration process is that opinion leaders' similarity in terms of who follow them increases when the most active consumers occupy more central social network positions. Such concentration phenomenon is also found in (Salganik, Dodd, and Watts 2006), where consumers under social influence tend to concentrate their song downloads to a few music providers.

To make it clearer, let's illustrate the intuition using an extreme scenario. If consumers have no impact on adjacent consumers, consumers follow unrelated sets of opinion producers. Let's assume the extreme that each consumer follows a unique opinion producer. Then no opinion producer would be an opinion leader, each having only a very small number of followers. In contrast, when there is social influence and

[2] The central position is central because it is contagious and thus generates influence on all other positions in all directions. In contrast, periphery positions are only contagious to a few other positions in a few directions.

therefore people concentrate most of their attention on several opinion producers, some opinion producers become opinion leaders.

The above theoretical insights inspire and are supported by our empirical analysis using a Twitter network and related tweets from (Hodas and Lerman 2014). We view a knowledge market as consisting of consumers and opinion producers focusing on a topic and within a sub-network relatively isolated from other parts of the Twitter network. To identify topics, we estimate a topic model and find 45 topics in tweets. To identify sub-networks, we apply a network community detection method and find four nearly isolated Twitter sub-networks. Each combination of one topic and one sub-network constitutes a knowledge market. Then with each knowledge market being a unit of observation, we formally define variables including the number of opinion leaders (the main dependent variable) and the network position centrality of the most active knowledge consumers (the main explanatory variable). Controlling for topic fixed effects, sub-network fixed effects, and the total amount of consumer attention to a topic in a subnetwork, we find that when the most active knowledge consumers occupy more central network positions, the number of opinion leaders increases and opinion leaders' similarity increases in terms of who follow them. A variety of robustness tests support these findings.

We use a state-of-the-art topic model for text analysis over tweets, the correlated topic model (Blei and Lafferty 2006; Roberts, Stewart, and Tingley 2014). Gentzkow, Kelly, and Taddy (2017) provide a good survey of the economic analysis using text as data. To define knowledge markets, the topic modeling outcomes are combined with the network analysis explained in what follows. A sub-field of network science focuses on dividing a network into sub-networks within which nodes are densely connected and between which nodes are sparsely connected (Thurner, Hanel, and Klimek 2018; Newman 2018). We use a state-of-the-art sub-network detection algorithm for very large networks (Clauset, Newman, and Moore 2004; Csardi 2019), and divide the Twitter network into nearly isolated sub-networks. Using the detected sub-networks and the modeled topics, a knowledge market is defined as a combination of a topic and a sub-network and we define topic-subnetwork specific variables for regression analysis. The unit of observation in the regression analysis corresponds to a topic in a community, i.e., to a knowledge market.

The number of opinion leaders in a knowledge market is the main dependent variable and the network position centrality of the most active knowledge consumers in a knowledge market is the main explanatory variable. The regression outcomes support the theoretical insights. To address endogeneity concerns on the explanatory variable, we first control for potential confounding variables including topic fixed effects and sub-network fixed effects (please note how we define a knowledge market). Second, we construct the network position centrality of the most active consumers based only on a friendship network and then use the constructed variable as an instrument.[3] The rationale is this: if I follow you and you don't follow me, I treat you as an opinion provider; but if I follow you and you also follow me, it's much probable that we're just friends. Under the assumption that the friendship network is subject to endogeneity concerns to a lesser degree than the original network, the estimates based on the friendship network would be significantly different from those based on the whole network, were endogeneity really a serious concern. The contrapositive of this logic gives this: under the assumption that the friendship network is subject to endogeneity concerns to a lesser degree than the original network[4], if the estimates based on the friendship network are not significantly different from those based on the whole network, then endogeneity is not a concern. Later you'll see Wald tests fail to reject the equivalence of estimates based on both networks and endogeneity is not a problem. If this contrapositive argument sounds too strange to you, please regard the 2SLS regressions as a robust check.

This paper makes the following contributions. First, this is the first paper to formally define opinion leaders, knowledge markets, and consumer attention. Second, while the existing literature emphasizes the role of opinion providers' network positions on the making of opinion leaders, this work shows the network positions of active consumers matter because active consumers serve as a propagation machine. In general, this paper shows to explore the topic of opinion diversity, a scrutiny on social network structure can be fruitful. Third, compared with most literature on network effects, the network effects explored in this paper are non-homogenously networked rather than well-stirred. This is a feature of both our theoretical and empirical work, which may inspire ideas for future research. Last but not least, as summarized in section 5, the

[3] We get the friendship network by dropping all the directed links in the whole network that don't have a reciprocal (this turns out to be a large change of the network, see section 3 for detail).

[4] The intuition behind this assumption: If I follow you and you don't follow me, it's probable that I treat you as an knowledge supplier; if I follow you and you also follow me it's possible but less probable that I treat you as a knowledge supplier because the mutual following may simply due to our friendship.

paper makes some methodological contributions by constructing IV from a multi-layer network perspective, using sub-network detection method to construct panel data, emphasizing the perspective of social embeddedness (Grannovetter 1985 and 2017).

The presentation proceeds as follows. In section 2 we derive insights via a mathematical model. In section 3 we introduce the data, explain and present the text analysis and network analysis, and define key variables used in the regression analysis. In section 4 we do regression analyses and explain the major empirical findings. In section 5 we talk about methodology contributions and conclude.

2 Theory

2.1 Environment and opinion producers

A knowledge market is modeled as consisting of opinion producers and knowledge consumers, where consumers embedded in a social network connect and influence each other. Before detailing behavioral assumptions for relevant economic agents, let's introduce the environmental setup, especially the social embeddedness assumption. (Grannovetter (2017) emphasizes the importance of being clear about the social embeddedness of economic agents in understanding many social phenomena.)

There is a line of length 3 (see figure 1 below on page 8) and agents (knowledge consumers and opinion producers) are distributed along the line in a way that will be specified soon. We regard the line 0-3 as a social space. At this moment it suffices to know that knowledge consumers near each other influence each other and that knowledge consumers follow (pay attention to) opinion producers near to them in the social space. For example, as a student in an economics department in the US, my social network position makes me subject to the influence from many US students and familiar with many US economists, while it's harder for me to be subject to the influence from Japanese students and familiar with Japanese economists.

We assume there is a continuum of opinion producers of measure 3, who are uniformly distributed along the line. They are made into opinion leaders by the process of consumers choosing to follow them: the probability for an opinion producer to be an opinion leader increases in the number of consumers following them. Thus,

opinion producers are not necessarily opinion leaders. For concreteness, we assume the probability for an opinion producer to be an opinion leader is $\frac{2}{1+e^{\delta(1-q)}}$, where q is the total number of knowledge consumers paying attention to (following) the opinion producer and δ governs the slope of the increasing curve. We assume a continuum of consumers of which the measure is one, so $q \in [0,1]$. We model the behavior pattern of opinion producers in this simplified form because the pivot of our story rests on the demand side that features the interaction between social network structure and social influence among consumers.

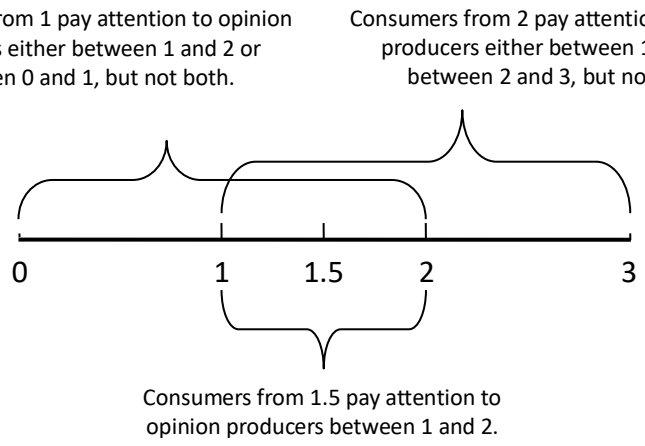
Actually, any convex curve instead of $\frac{2}{1+e^{\delta(1-q)}}$ will do. We need convexity because it captures the following intuition: removing an amount of followers from opinion producers receiving low attention will decrease the number of opinion leaders, but the decrease is fewer than the increase of the number of opinion leaders from adding the same amount of followers to opinion producers already receiving high attention. Where comes this intuition? Please recall in section 1 we talk about the following concentration effect due to social influence among consumers. If consumers have no impact on adjacent consumers, consumers follow unrelated sets of opinion producers. Then let's assume the extreme that each consumer follows a unique opinion producer. Then no opinion producer would be an opinion leader, each having only a very small number of followers. In contrast, when there is social influence and therefore people concentrate most of their attention on several opinion producers, some opinion producers become opinion leaders.

Knowledge consumers rest only at points 1, 1.5, and 2 (no consumers are from other points). There is a continuum of consumers of which the measure is 1. A half of the consumers are at 1.5 and the other half are evenly divided for 1 and 2. Besides, a half of the consumers are active and willing to pay H units of attention on the topic in this knowledge market (i.e., each of them spreads the H units of attention over a number of opinion producers they choose to follow), and the other half are inactive and willing to pay L units of attention. H>L>0 and we assume (H+L)/2 =1 to streamline notations without loss of generality.

We assume $\frac{\pi}{2}$ H-consumers and $\frac{1-\pi}{2}$ L-consumers (and thus 1/2 consumers in total)

are at point 1.5 (1.5 is the center); $\frac{1-\pi}{4}$ H-consumers and $\frac{\pi}{4}$ L-consumers are at point 1 and the same at point 2 (1 and 2 are the periphery). $\pi \in (0,1)$. When π increases, there'll be more active consumers at the center of the social space with the total attention of all consumers kept constant. So π measures the degree to which active knowledge consumers occupy the central network position. Consumers occupying the central position generate influence in two directions (from 1.5 to 1 and to 2), while those occupying peripheral positions only generate influence in one direction (from 1 to 1.5; from 1 to 2). Furthermore, a consumer's influence is proportional to how active he is. In section 2.2 we provide rigorous and mathematical definitions of consumer influence. Here the central idea is that the high ability of active consumers to influence is fully realized when they occupy central positions in a social network of consumers. Then there is greater amount of social influence leading to more opinion leaders through a concentration effect that will be clear when we complete the model.

Figure 1



__ Consumers are only from points 1, 2, and 1.5.

__ At each point between 0 and 3 rest a opinion producer.

__ $\frac{\pi}{2}$ H-consumers and $\frac{1-\pi}{2}$ L-consumers rest at point 1.5; $\frac{1-\pi}{4}$ H-consumers and $\frac{\pi}{4}$ L-consumers rest at point 1 and the same at point 2. Half are at center and half are at periphery.

We make the following assumptions, to capture the interaction of social positions and consumer activeness and to model social influence between people close to each other

in a social space. Consumers from 1.5 are only able to pay attention to opinion producers distributed between 1 and 2. The decision to pay how much attention to each opinion producer is explained in section 2.2. Consumers from 1 are able to pay attention to opinion producers distributed either between 1 and 2 or between 0 and 1, but not both. Specifically, each of the consumers from 1 chooses whether to pay attention to the opinion producers on his left (between 0 and 1) or to those on his right (between 1 and 2); in the meanwhile, he also chooses how much attention to pay to each opinion producer on the chosen side. Consumers from 2 are able to pay attention to opinion producers distributed either between 1 and 2 or between 2 and 3, but not both. Specifically, each of the consumers from 2 chooses whether to pay attention to the opinion producers on his left (between 1 and 2) or to those on his right (between 1 and 2); in the meanwhile, he also chooses how much attention to pay to each opinion producer on the chosen side. How the choices by consumers from 1 and 2 are made is explained in section 2.2. Finally, consumers from 1.5 are more central in the sense that either half of the area covered by their attention overlaps with what're covered by consumers from 1 or 2: as will be seen soon, this overlap in two directions means that consumers at 1.5 generate influence in two directions.

2.2 Behavioral and interaction assumptions for knowledge consumers

Every knowledge consumer distributes the attention he is willing to pay among opinion producers he is able to and chooses to reach. Let a_{ik} denote the amount of attention paid by consumer i to opinion producer k (who rests at point k). For active consumers, attention budget constraints are $a_{ik} \geq 0$ and $\int_0^3 a_{ik}dk \leq H$. For inactive consumers, attention budget constraints are $a_{ik} \geq 0$ and $\int_0^3 a_{ik}dk \leq L$. Note $a_{ik} = 0$ if opinion producer k is at a position not reachable by i or if i chooses $a_{ik} = 0$ when k is reachable. Let $\int a_{ik}di \equiv A_k$ be the total amount of attention received by opinion producer k.

Consumers at the central position in the social space (at point 1.5) cannot reach opinion producers from $[0,1) \cup (2,3]$ and the utility function for a generic consumer (i) at 1.5 is $\int_1^2 (1 + A_k)^{1-\rho}(a_{ik})^{\rho}dk$. A_k is taken as given and a_{ik} for each k is chosen to maximize the utility under attention budget constraints. Consumers at 1.5 have the same utility function form, whether they are willing to pay H or L units of attention.

We have a few reasons to use this utility function. First, we use this utility function to capture social influence: the higher A_k, the more attention is received by opinion producer k[5], and then the higher is the marginal benefit for a consumer to pay attention to opinion producer k. Second, we use $(1+A_k)^{1-\rho}$ rather than $(A_k)^{1-\rho}$ because we assume there is still some utility for a consumer to follow an opinion producer who is followed by nobody else.[6] Third, per this utility function, with A_k fixed, consumer i has a decreasing marginal utility from paying more attention to k and thus tends to spread attention over many opinion producers. This is a diversification incentive. Finally, this function form makes the model tractable. Please note in this paper we don't model specific mechanisms underlying why increasing A_k increases marginal utility form following producer k, which can be an interesting topic for further research.

Consumers at point 1 get utility from opinion producers from [0,1] or [1,2] but not both. To maximize utility, consumers at 1 choose whether to pay attention to opinion producers from [0,1] or to those from [1,2], and in the meanwhile choose the amount of attention to each opinion producer (chooses a_{ik} for each k). We make this assumption because we try to model social influence: consumers at 1 cannot pay attention in both directions and which direction to choose is subject to social influence through $\{A_k\}$. The utility function for a generic consumer (i) at 1 is

$$\max\{\int_0^1 (1+A_k+s_i)^{1-\rho}(a_{ik})^{\rho}dk\,,\int_1^2 (1+A_k)^{1-\rho}(a_{ik})^{\rho}dk\}\ .$$

A_k is taken as given and a_{ik} for each k is chosen to maximize the utility under attention budget constraints. Let's explain why there is s_i in the utility function. For each consumer at the periphery an s is drawn from a uniform distribution over [-t, t] with t>0. This number magnifies the marginal utility of attention to opinion producers in [0, 1] if s>0, and lessens the marginal utility if s<0. Due to symmetry, it does not matter if s (positive or negative) is added for the opinion producers from [1, 2] or from [0, 1]. Without loss of generality, we set t=1. As will be seen later, s smoothes the equilibrium[7]. Consumers at 1 have the same utility function form,

[5] Actually in the utility function for consumer i, A_k should be the total amount of attention excluding i's attention received by the opinion producer, but since there is a continuum of knowledge consumers, we can safely drop the "excluding i's attention" in defining A_k and assume each consumer takes $\{A_k \mid k \in [0,3]\}$ as given, though $\{A_k \mid k \in [0,3]\}$ is endogenously determined in equilibrium.

[6] Here 1 can be changed into any positive number. If consumer utility from an opinion producer who is followed by nobody else is zero, there'll be an infinite number of uninteresting equilibria.

[7] Without s, we have the same theoretical conclusion: the number of opinion leaders and their similarity increase in π which, as mentioned above, measures the degree to which enthusiastic knowledge consumers occupy the central

whether they are willing to pay H or L units of attention.

Using analogous assumptions, the utility function for a generic consumer (i) at 2 is

$$\max\{\int_2^3 (1 + A_k + s_i)^{1-\rho}(a_{ik})^{\rho} dk\,, \int_1^2 (1 + A_k)^{1-\rho}(a_{ik})^{\rho} dk\} \ .$$

Consumers at 2 have the same utility function form, whether they are willing to pay H or L units of attention.

2.3 Equilibrium and comparative statics

With the behavioral assumption and social embeddedness assumption, we now solve for the equilibrium. In the equilibrium, (1) each consumer chooses how much attention to pay to each opinion producer, taking $\{A_k\}$ as given; (2) $\{A_k\}$ is endogenously determined such that no consumers want to change their choices; (3) each opinion producer becomes an opinion leader with the probability $\frac{2}{1+e^{\delta(1-q)}}$, where q is the total number of knowledge consumers paying attention to (following) the opinion producer. Details and comparative statics are given in what follows.

(1) In the equilibrium A_k's are the same for all $k \in [1, 2]$ and let's denote this common value by $\overline{A}_c$ (c for center); A_k's are the same for all $k \in [0, 1) \cup (2, 3]$ and let's denote this common value by $\overline{A}_p$ (p for periphery). A_k is the total attention received by opinion producer k.

(2) In the equilibrium, consumer choices are given in what follows. H-consumers (or L-consumers) from 1.5 pay H (or L) attention to each opinion producer between 1 and 2[8]. H-consumers (or L-consumers) from 1 with $s < s^* \equiv \overline{A}_c - \overline{A}_p$ pay H (L) attention to each opinion producer between 1 and 2; H-consumers (or L-consumers) from point 1 with $s > s^* \equiv \overline{A}_c - \overline{A}_p$ pay H (or L) attention to each opinion producer between 0 and 1. H-consumers (or L-consumers) from point 2 with $s < s^* \equiv \overline{A}_c - \overline{A}_p$ pay H (or L) attention to each opinion producer between 1 and 2; H-consumers (or L-consumers) from point 2 with $s > s^* \equiv \overline{A}_c - \overline{A}_p$ pay H (or L) attention to each opinion producer between 2 and 3.

---

network position (point 1.5) (and thus it corresponds to r1 and r2 in our regression analysis). But the increasing function is a step function, with either no increase (for π below a threshold) or increase to the full extent (for π above or equal to a threshold); it's also possible for the threshold to be zero depending on the relative magnitude of H and L. This ad hoc jumpy behavior is avoided by s providing heterogeneity for consumers, which we think is more realistic and consistent with the empirical findings and, more importantly, elucidates our insight better. Note s plays a similar role as does the random part of the utility function in standard discrete choice models.

[8] Note the measure of opinion producers share H (or L) units of attention is 1, so each opinion producer gets H (or L) units of attention.

(3) We have assumed that $\frac{\pi}{2}$ H-consumers and $\frac{1-\pi}{2}$ L-consumers are from 1.5 and that $\frac{1-\pi}{4}$ H-consumers and $\frac{\pi}{4}$ L-consumers are from 1 and the same from 2. With the assumption and consumer decisions in (2) combined, the following equations determine $\bar{A}_p$ and $\bar{A}_c$:

$$\bar{A}_p = \left(\frac{1-\pi}{4} \times H + \frac{\pi}{4} \times L\right)\left(1 - \left[\frac{1}{2} \times \left(\bar{A}_c - \bar{A}_p + 1\right)\right]\right)$$

$$\bar{A}_c = \left(\frac{1-\pi}{4} \times H + \frac{\pi}{4} \times L\right)\left[\frac{1}{2} \times \left(\bar{A}_c - \bar{A}_p + 1\right)\right] + \left(\frac{\pi}{2} \times H + \frac{1-\pi}{2} \times L\right)$$

H>L>0 and (H+L)/2 =1, as assumed above, ensures $s^* \equiv \left(\bar{A}_c - \bar{A}_p\right) \in (-1, 1)$ consistent with the distribution of s.

(4) Combining $\left(\bar{A}_c - \bar{A}_p\right) \equiv s^*$ and the two equations in (3), we solve for s^* and get $s^* = 2 - \frac{1}{1-\frac{1-\pi}{4}\times H - \frac{\pi}{4}\times L}$ which is increasing in π and belongs to (-1, 1). As will be seen soon, s^* determines how many consumers at the periphery pay attention to peripheral (and thus central) opinion producers.

(5) To simplify notation, let's assume the CDF of the uniform distribution over [-1, 1] is G(s). Per (2) and the model setup, the number of consumers paying attention to the opinion producers in the central region (between 1 and 2) is $\frac{1}{2} + \frac{1}{2}G(s^*)$, while the number of consumers paying attention to the opinion producers in peripheral regions (between 0 and 1 and between 2 and 3) are both $\frac{1}{4} - \frac{1}{4}G(s^*)$. With the assumption on opinion producers, the number of opinion leaders is given by $\frac{2}{1+e^{\delta(\frac{1}{2}-\frac{1}{2}G(s^*))}} + \frac{4}{1+e^{\delta(\frac{3}{4}+\frac{1}{4}G(s^*))}}$, which is increasing in $G(s^*)$. Together with $G(s^*)$ increasing in s^* and s^* increasing with π, we have the following comparative statics.

Comparative Statics 1: Increasing π results in more opinion leaders and a higher concentration of opinion leaders between 1 and 2.

(6) Per (5), as π increases, $\frac{1}{2} + \frac{1}{2}G(s^*)$ (the number of consumers paying attention to the opinion producers in the central region) increases, $\frac{1}{4} - \frac{1}{4}G(s^*)$ (the number of consumers paying attention to the opinion producers in peripheral regions)

decreases, and the amount of increase is greater than the amount of decrease. Consequently, there are more opinion leaders in the central region, who are followed by similar sets of consumers, and less opinion leaders in the peripheral regions, who are followed less similar sets of consumers. So, opinion leaders on average have more similarity with each other in terms of who follow them.

Comparative Statics 2: The higher concentration of opinion leaders implies more similarity in terms of who follow them.

Thanks to their high ability to increase A_k, active consumers (H-consumers) generate more social influence on other consumers' choices than inactive consumers (L-consumers) do. Consumers at the center make choices influencing consumers at the periphery by attracting them to pay attention to opinion producers at the central region. Thus, when there are more active consumers at the center (when π is higher), there'll be higher concentration of consumers' attention into the central region. This concentration leads to more opinion leaders and more similarity among them in terms of who follow them, as shown in (5) and (6) above. π corresponds to the key explanatory variables in the empirical exercise to which we now turn.

3 Empirical analysis I: data, topic modeling, network analysis, and defining knowledge markets and key variables

3.1 Data

We use data from Hodas and Lerman (2014).[9] Hodas and Lerman (2014) collected tweets over three weeks in the fall of 2010 and then retained tweets containing a URL in the message body. A URL, uniform resource locator, is usually in the form of a string of characters or symbols (e.g., http://gd.is/4nfm) that references a web resource and specifies its location on a computer network. When a twitter user shares a web resource in his tweet, the web address is coded into a short string of characters and symbols and shown in the tweet. Thus, a URL can be regarded as a meaningful key word, more informative about the content or topic of a tweet than usual words such as good, the, or, increase, etc. Hodas and Lerman (2014) then retrieved all tweets

[9] One can download data and find data details at https://www.isi.edu/~lerman/downloads/twitter/twitter2010.html

containing the URLs they collected, ensuring the complete tweeting history of all the URLs[10], and resulting in 3 million tweets in total. They also collected friend and follower information for all tweeting users at the time, resulting in a network with almost 700K nodes and over 36M directed edges (an edge is directed from user A to user B if user A follows user B in Twitter). From Hodas and Lerman (2014), we thus have as our raw data a social network of who follows whom, with each Twitter user being a node in the network, and a set of URLs for each user that ever appear in his or her tweets. Each user is represented by a user id. Twitter has changed their way of coding user accounts, so we don't know who a user is from his or her id.

Using the data, in sections 3.2 and 3.3 we do topic modeling and network analysis. In section 3.4 we use the topic modeling and network analysis outcomes to define knowledge markets and to define key variables characterizing knowledge markets. The variables will be used in regression analysis in section 4, with each knowledge market corresponding to a unit of observation.

3.2 Topic modeling

The inputs of the topic modeling are documents and words. We regard each user as a document and URLs as words informative about the topics of documents. We then model topics using the correlated topic model (CTM) (Blei and Lafferty 2006). The CTM is a state-of-the-art topic modeling method.[11] It models a document generating process that exploits more subtleties in the data to capture topic correlation[12] (Blei and Lafferty 2006; Roberts, Stewart, and Tingley 2014). Below we briefly introduce the CTM before showing the topic modeling outcome.

The CTM models every document as a distribution over topics and every topic as a distribution over words. The parameters of these distributions are estimated, as the CTM searches for a document generating process that best fits a set of documents. We need to specify the number of topics before fitting a model, and we'll choose the number that corresponds to the best fit. The estimated distribution parameters can be used to calculate topic proportions for each document (how much a document is about each topic). Thus, the outputs of the topic modeling are the number of topics and

[10] They have URLs in tweets but no full text contents in the tweets. Twitter has forbidden using large amounts of text contents by external researchers.

[11] Besides, the CTM can be seen as an improved version of the LDA, a very commonly used topic modeling method; Blei, an author of the CTM, is also one of the researchers independently discovering the LDA.

[12] For example, a document about sports is more likely to also be about health than international finance.

topic proportions for each document (i.e., for each user).

For a document with m words (m distinct URLs in our case), the CTM models the following generative process. For technical expositions, see Blei and Lafferty (2006) and Roberts, Stewart, and Tingley (2014); the CTM can be regarded as a kind of unsupervised machine learning.

(1) A K-by-1 vector is drawn. K is the number of topics specified by researchers. The vector is drawn from a logistic normal distribution[13] and represents topic frequencies (proportions) for a document to be generated.

(2) The K-by-1 vector is used as the parameters of a multinomial distribution over K topics. An m-by-1 vector is generated by randomly drawing topics m times from the multinomial distribution, where m is the document length. Word order in a document in not concerned. Each element in the m-by-1 vector corresponds to one topic from the K topics.

(3) A distribution of words is used for each topic. Specifically, an N-by-1 vector is used for each of the K topics and represents the distribution of words in each topic. N is the total number of words in all documents. The elements in the vector for a topic correspond to the frequency of each word in the topic.

(4) For each element that indicates a topic from the m-by-1 vector in (2), a word is randomly drawn based on the topic's word distribution in (3). Then the m words thus drawn form a document.

Topic modeling outcomes

We regard a Twitter user as a document and URLs as words informative about the topics of documents. Since we're interested in the number of opinion leaders of different topics, in topic modeling we only focus on the documents that are widely read (i.e., Twitter users that are intensively followed). Let's call Twitter users whose numbers of followers are among the top 0.5% as the top 0.5% followees. Then to have inputs for the CTM, we regard the top 0.5% followees as documents and their URLs as words.[14] The number of the top 0.5% followees is 3498, and the followees among them with the fewest followers have 1118 followers (these two numbers for the top 1% followees are 6985 and 648). Note a Twitter user is both a follower and a followee: when we focus on how many users follow a Twitter user, the user is a followee; when

[13] The use of the logistic normal distribution allows the CTM to capture topic correlation (Blei and Lafferty 2006). For example, a document about sports is more likely to also be about health than international finance.

[14] The conclusions are the same if we instead us the top 1% as shown in the appendix.

we focus on who and how many are followed by a Twitter user, the user is a follower.

With inputs defined, 45 topics result from the CTM. The number of topics is chosen using the method introduced by Roberts, Stewart, and Tingley (2014). Specifically, we experiment with different numbers of topics for the above generative process (from 30 to 100 with a step of 5) and randomly leave out documents for out-of-sample evaluation. Among the numbers (of topics) that best explain the out-of-sample documents, we choose the one that best fits the in-sample documents. Like using the top 0.5% followees, using the top 1% followees also results in the topic number of 45.

With the number of topics fixed at 45, estimates of the CTM (parameter estimates of the logistic normal distribution used in the above step (1)) give topic proportions for each document (each of the top 0.5% Twitter followees). Many documents have very small topic proportions over different topics. We treat a topic proportion of a document as 0 if the CTM estimates the topic proportion of the document as less than 10%.[15] We have two reasons for this censoring: a very small topic proportion is usually not taken seriously in text mining (Robinson and Silge 2017) since it's largely a model artifact for a better fit. Without the small proportions being neglected, each document corresponds to almost all topics, and this means the numbers of opinion leaders (or widely read documents; we'll formally define opinion leaders in section 3.4) are almost the same for all topics, which is neither interesting nor realistic.

Next we present our network analysis, which in section 3.4 will be combined with the topic modeling outcome to define knowledge markets and variables for regression analysis in section 4.

3.3 Network analysis: sub-network detection

How things work in Hobbits' Shire is quite different from that among Gandalf's wizard friends (Tolkien, 1955). A very large social network may consist of several nearly isolated sub-networks each of which is better to be regarded as an isolated kingdom. A field of the network science, sub-network detection (also often called community detection), concerns dividing a network into nearly isolated sub-networks (communities) within which nodes are densely connected and between which nodes are sparsely connected (Thurner, Hanel, and Klimek, 2018; Newman, 2018). We use a

[15] The conclusions are the same when we use 20% instead of 10%, as shown in the appendix.

state-of-the-art sub-network detection method (Clauset, Newman, and Moore, 2004; Csardi, 2019) for very large networks as is the present case. We find the Twitter network mainly consists of 4 nearly isolated sub-networks. About 94% of the links that are directed toward or away from the nodes in these four sub-networks are within sub-networks. Of course, 94% is not 100%, so in section 4 we use some econometric techniques to deal with this concern. Besides, nearly all the top 0.5% followees (more than 95%) are in these four sub-networks. 1494, 439, 798, and 601 of the top 0.5% followees are respectively in the four sub-networks.

Before we explain the sub-network detection method, let's explain the purpose of this network analysis. First, each sub-network detected is to be combined with each topic found in section 3.2 to define a knowledge market. In section 4 each knowledge market corresponds to a unit of observation in regression analysis. Thus, variables (defined in section 3.4) used in the regression analysis are topic-sub-network specific; this gives us a panel data structure. Second, it's reasonable to assume agents from different sub-networks behave differently, as shown by Reddy, Kitsuregawa, Sreekanth, and Rao (2002) who use sub-network detection to identify consumers with similar interests and purchasing habits.

Having reported the sub-network detection outcome, let's briefly talk about how we do the detection. For technical expositions, see Thurner, Hanel, and Klimek (2018), Newman (2018) and Clauset, Newman, and Moore (2004). We first transform the Twitter network into an undirected network such that an undirected link between Twitter user i and Twitter user j exists when there is a directed link from i to j or from j to i. Let's say $\widetilde{A}_{ij} = 1$ if there is an undirected link between i and j; otherwise, $\widetilde{A}_{ij} = 0$. We use k_i to denote the number of network neighbors of node i in the undirected network. (Two nodes are network neighbors of each other if they are connected in a network.) If a network is generated randomly lacking sub-network structure, the probability that $\widetilde{A}_{ij} = 1$ is given by $\frac{k_i k_j}{2L}$. Let's say $\delta_{ij} = 1$ if i and j are assigned into the same sub-network; otherwise, $\delta_{ij} = 0$. The sub-network detection method in this paper chooses which Twitter users in what sub-network and the number of sub-networks by maximizing a modularity score, Q, as defined in the following (Clauset, Newman, and Moore 2004): $Q = \frac{1}{2L}\sum_{ij}\left(\widetilde{A}_{ij} - \frac{k_i k_j}{2L}\right)\delta_{ij}$, where L is the number of links in the undirected network. Intuitively this score compares the actual

network to random networks that lack sub-network structure. For different sub-network membership assignments and numbers of sub-networks, this score compares $\widetilde{A}_{ij}$ (the actual presence or absence of a link) and $\frac{k_i k_j}{2L}$ (the presence probability of a link if the network is generated randomly lacking sub-network structure), when $\delta_{ij} = 1$, i.e., for pairs of nodes assigned in the same sub-network.

3.4 Defining knowledge markets and key variables for regression analysis

With 45 topics from section 3.2 and 4 sub-networks from section 3.3, we define each combination of a topic and a sub-network as a knowledge market. Each knowledge market will be a unit of observation in regression analysis in section 4, where you'll see this definition allows for a panel data structure. Please note units of observation used for regression analysis are different from those used in the topic modeling. In the topic modeling, each of the top 0.5% followees is a unit of observation: the modeling inputs are these users as documents and the URLs in their tweets as words.

Let's at first be clear about what we mean by opinion leaders: a Twitter user is an opinion leader in the knowledge market defined by topic j and sub-network k, if he or she (1) belongs to the top 0.5% followees,[16] (2) has a topic proportion greater than 10% on topic j, and (3) belongs to sub-network k. With opinion leaders defined, in the rest of this section we define variables that characterize knowledge markets and are thus topic-sub-network specific.

1. $\mathbf{N_{jk}}$: the number of opinion leaders in the knowledge market defined by topic j and sub-network k (j = 1, 2, …, 45 and k = 1, 2, 3, 4).

The number of opinion leaders in the knowledge market (jk) is the number of top 0.5% followees who are in sub-network k and whose topic proportions are greater than 10% for topic j. (In section 4.3 and in the appendix, we report outcomes with other choices of these two thresholds.)

2. $\mathbf{S^{in}_{jk}}$: the average similarity of opinion leaders in the knowledge market defined by topic j and sub-network k (j = 1, 2, …, 45 and k = 1, 2, 3, 4), in terms of who follow them.

Based on Adamic and Adar (2003), we define the average similarity of opinion leaders

[16] Note the top 0.5% followees are used in CTM to mine topics and topic proportions.

in a knowledge market in terms of who follow them, to be the average in-degree Jaccard similarity over all pairs of opinion leaders in the knowledge market. The in-degree Jaccard similarity for a pair of opinion leaders in the knowledge market (jk) is the number of users in sub-network k following both opinion leaders divided by the number of users in sub-network k following at least one of the opinion leaders.

3. $\mathbf{f_{jk}}$: the total amount of consumer attention (divided by 1000) to all opinion leaders in the knowledge market defined by topic j and sub-network k (j = 1, 2, …, 45 and k = 1, 2, 3, 4).

First, a consumer in the knowledge market (jk) is defined as a Twitter user who is not an opinion leader, is located in sub-network k, and follows at least one opinion leader from the knowledge market (jk). Second, for each consumer in the knowledge market (jk), say, for consumer m, we sum the topic j proportions of the opinion leaders followed by him or her in the market (jk). The sum is then called consumer m's attention in the knowledge market (jk) and denoted by f_{mjk}. Finally, we sum the attention of all the consumers in the market (jk) and divide it by 1000, to get $f_{jk} = \frac{\sum_m f_{mjk}}{1000}$. The denominator 1000 will make estimates in section 4 look simple.

4. $\mathbf{r1_{jk}}$: the weighted average in-degree centrality (divided by 1000) of consumers who pay more than median attention in the knowledge market defined by topic j and sub-network k (j = 1, 2, …, 45 and k = 1, 2, 3, 4).

First, the in-degree centrality of a Twitter user in the knowledge market (jk) is the number of users following him or her in sub-network k. Second, from 3 we have individual consumer's attention in each knowledge market (f_{mjk}), and let's use Ω_{jk} to denote the set of consumers whose attention is greater than the median consumer attention in the knowledge market (jk). Thirdly, for a consumer in Ω_{jk}, say, consumer $m \in \Omega_{jk}$, we denote his in-degree centrality by ID_{mk}. Finally, we define $r1_{jk}$ by averaging in-degree centralities over consumers in Ω_{jk}, weighted by consumer attention per the following equation, and then divide the average by 1000 to make estimates in section 4 look simple.

$$r1_{jk} = \frac{\sum_{m\in\Omega_{jk}}(f_{mjk} \times ID_{mk})}{\sum_{m\in\Omega_{jk}} f_{mjk}} \times \frac{1}{1000}$$

$r1_{jk}$ is one of the major explanatory variables in regression analysis in section 4.

Intuitively, $r1_{jk}$ measures how central are the network positions occupied by active knowledge consumers. It thus corresponds to π in the model in section 2, the degree to which active consumers take central social network positions.

5. **$r2_{jk}$**: the weighted average eigenvector centrality of consumers who pay more than median attention in the knowledge market defined by topic j and sub-network k (j = 1, 2, …, 45 and k = 1, 2, 3, 4).

Compared with $r1_{jk}$, we use eigenvector centrality rather than in-degree centrality to define $r2_{jk}$. In-degree centrality of a user only counts how many users follow him or her in a network, but we can exploit more subtlety in network structure: some of a user's followers may be more central in the network (e.g., may be followed by many who are also central) while others of the user's followers are less central (e.g., may be followed by few who are also less central). Eigenvector centrality extends in-degree centrality by taking into account this heterogeneity in the centralities of followers, when evaluating user centrality in a network: users are more central not only if they have more followers, but also if their followers are more central.[17] In what follows, we show how we define $r2_{jk}$. With f_{mjk} defined in 4 and with consumer m's eigenvector centrality (technically defined in footnote 14) in sub-network k denoted by EC_{mk}, we define $r2_{jk}$ per the following equation.

$$r2_{jk} = \frac{\sum_{m\in\Omega_{jk}}(f_{mjk} \times EC_{mk})}{\sum_{m\in\Omega_{jk}} f_{mjk}}$$

This is the same as $r1_{jk}$ except with EC_{mk} substituted for ID_{mk} and without $\frac{1}{1000}$.

[17] Formally, when an user's centrality is regarded as correlated with the sum of his followers' centralities, the eigenvector centrality, x_i, of user i can be defined from $x_i = \kappa^{-1}\sum_j A_{ji}\, x_j$, where $A_{ji} = 1$ if j follows i (to calculate eigenvector centrality for undirected networks, the same formula is used with $A_{ji} = 1$ if there is a link between j and i). κ is a scalar whose role will be seen soon. Putting the eigenvector centralities of all users into a vector, x, one gets $A^T x = \kappa x$. x is thus a eigenvector of A^T. Per Perron-Frobenius theorem, for an adjacency matrix from a connected network only the leading eigenvector is non-negative. Note each community is a connected network. Eigenvector centralities, usually assumed to be non-negative, are thus given by the leading eigenvector of A^T. However, there is an undesirable property for using eigenvector centrality for directed networks. People without followers have zero eigenvector centrality. Per the definition of eigenvector centrality, anyone who is followed only by such followers has zero eigenvector centrality. Iteratively, users who only receive incoming links from zero-centrality users have zero eigenvector centrality. However, in our case the impact from this property can be ignored since such users account for only around 1% of users in each community. For technical details, please see Newman (2018), Thurner, Hanel, and Klimek (2018), and Bonacich (1987).

We'll use $r2_{jk}$ and $r1_{jk}$ separately as the major explanatory variable in regression analysis in section 4: this can be viewed as a robustness check.

6. **$r1iv_{jk}$**: a variation of $r1_{jk}$, with the only difference in that we calculate in-degree centrality based on friendship networks instead of on the original sub-networks.

We construct $r1iv_{jk}$ in the same principle as $r1_{jk}$ with the in-degree centrality calculated based on four friendship networks. Each friendship network is constructed from one of the four original sub-networks, by dropping any directed link without a reciprocal of it, i.e., a link from i to j is dropped if there is no link from j to i. The numbers of user pairs connected in the friendship networks are respectively 30%, 44%, 20%, and 55% of the numbers of user pairs connected (with at least one directed link) in the original sub-networks. When we run regressions in section 4, we'll explain why and how ***r1iv*** can serve as an instrument for ***r1***.

7. **$r2iv_{jk}$**: a variation of $r2_{jk}$, with the only difference in that we calculate eigenvector centrality based on friendship networks instead of on the original sub-networks.

We construct $r1iv_{jk}$ in the same principle as $r2_{jk}$ with the eigenvector centrality calculated based on the four friendship networks defined in 6. In section 4 we'll explain why and how ***r2iv*** can serve as an instrument for ***r2***.

We'll use the above-defined variables in regression analysis in section 4. Their summary statistics are given in table 1.

Table 1: summary statistics of the variables characterizing knowledge markets

Variable Name	Mean	Standard Deviation
N	50.3	127.9
S^{in}	0.111	0.115
r1	0.212	0.131
r2	0.066	0.061
f	69.4	192.9
r1iv	0.156	0.109
r2iv	0.062	0.059

4 Empirical analysis II: panel data regression analysis without and with IV

Each knowledge market corresponds to a unit of observation in regression analysis. In section 3 we have defined each knowledge market as a combination of a topic and a sub-network, which allows for a panel data structure. Compared with standard panel data, topic corresponds to individual and community to time. This panel data is unbalanced because of the following facts. The top 0.5% followees are distributed across four sub-networks and the top 0.5% followees within each sub-network do not cover all topics: not all topics appear in all of the four sub-networks. Actually 9 topics appear in all of the four sub-networks, 9 topics in only three of the four sub-networks, 13 topics in only two of the four sub-networks, and 10 topics in only one of the four sub-networks.[18]

4.1 Panel data analysis without IV

In this section we have regression equations as what follows:

$$\mathbf{y_{jk} = \beta \times x_{jk} + \gamma \times f_{jk} + \mu_j + \rho_k + \varepsilon_{jk}.}$$

Each knowledge market is a unit of observation in the regression analysis and is uniquely pinned down by combining a topic with a sub-network: for example, the knowledge market jk is about topic j and in sub-network k. β and γ are coefficients. ε_{jk} is the error term. We run two-way fixed effect regressions controlling for topic fixed effects (μ_j) and sub-network fixed effects (ρ_k). We next explain how the variables defined in section 3.4 are to be used.

y_{jk} will be either N_{jk} (the number of opinion leaders in the knowledge market jk) or S^{in}_{jk} (the similarity of opinion leaders in market jk in terms of who follow them).

x_{jk} will be either $r1_{jk}$ (weighted average in-degree centrality of active knowledge consumers in the knowledge market jk, divided by 1000) or $r2_{jk}$ (weighted average eigenvector centrality of active consumers in market jk). **r1** and **r2** correspond to π in the theory in section 2, measuring in different ways how central are the network positions occupied by active consumers. It can be viewed as a robustness check that we separately use the two alternative variables as the major explanatory variable. The

[18] Note the sum of these four numbers is 41 rather than 45. This is because for some topics in some communities, there is only one opinion leader. The values for S^{in}, S^{in}, r1, r2, r1iv, and r2iv are NAs when there is only one opinion leader. So we drop such incomplete cases.

theoretical prediction in section 2 implies significant positive coefficients of x_{jk} for both dependent variables N_{jk} and S_{jk}^{in}. Recall the theoretical prediction in sections 2: in a knowledge market where active consumers on average occupy more central positions in the consumer social network, more opinion producers become opinion leaders and there is more similarity among opinion leaders in terms of who pay attention to them.

In addition to the fixed effects we also control for f_{jk} (the total amount of consumer attention in the knowledge market jk, divided by 1000), because in the model we fix this amount and focus on comparative statics of changing π. Moreover, more total consumer attention may result in a larger number of opinion leaders due to high demand, while the total attention may also correlate with network structure in unknown ways such that it may in turn correlate with our major explanatory variable. Thus, we treat it as a potential confounding variable and control for it.

It's important to note that time in a standard panel data corresponds to sub-network membership in our panel data, and individual to topic. The sub-network detection has done its best to detect four nearly isolated sub-networks, but there are still a small number of links between sub-networks[19] and error terms can be correlated across sub-networks. To estimate standard errors, we use the Arellano covariance estimator that takes into account serial correlation of arbitrary form (Croissant and Millo 2008; Arellano 1987). An arbitrary form is allowed for, because, though time in standard panel data has an order where some AR process can be imposed, sub-network membership cannot be ordered. Also, the Arellano estimator is advisable for short panel (note only 4 sub-networks in our data) (Croissant and Millo 2008).

Regression results with the two dependent variables and with the alternative major explanatory variables are shown in table 2 and table 3.

[19] As noted above, about 94% of links among all the links directed toward to and away from the nodes in these four communities are within communities.

Table 2: The effect of active consumers' centrality on the number of opinion leaders and their similarity in terms of who follow them: OLS and r1 (in-degree centrality)

	Dependent Variables	
	N	S^{in}
r1 (how central active consumers are in the consumer social network)	84.6** (p=0.03)	0.238*** (p=0.002)
f (total consumer attention in each knowledge market, divided by 1000)	5.46*** (p<0.001)	0.0003 (p=0.247)

1. Here we use the top 0.5% followees for topic modeling, treat any less than 10% topic proportion as zero, use median attention to define active consumers.
2. We control for the total amount of consumer attention (f_{jk}) in market (jk), topic fixed effects, and sub-network fixed effects. The data structure is an unbalanced panel with 99 observations. Compared with standard panel data, topic corresponds to individual and sub-network to time.
3. We use the Arellano covariance estimator to estimate standard errors (Croissant and Millo 2008; Arellano 1987)
4. Numbers in parentheses are p-values for the null hypothesis of zero effect; significance code: * for 0.1, ** for 0.05, *** for 0.01.

Table 3: The effect of active consumers' centrality on the number of opinion leaders and their similarity in terms of who follow them: OLS with r2 (eigenvector centrality)

	Dependent Variables	
	N	S^{in}
r2 (how central active consumers are in the consumer social network)	175* (p=0.074)	0.889*** (p=0.0001)
f (total consumer attention in each knowledge market, divided by 1000)	5.43*** (p<0.001)	0.000209 (p=0.43)

1. Here we use the top 0.5% followees for topic modeling, treat any less than 10% topic proportion as zero, use median attention to define active consumers.
2. We control for the total amount of consumer attention (f_{jk}) in market (jk), topic fixed effects, and sub-network fixed effects. The data structure is an unbalanced panel with 99 observations. Compared with standard panel data, topic corresponds to individual and sub-network to time.
3. We use the Arellano covariance estimator to estimate standard errors (Croissant and Millo 2008; Arellano 1987)
4. Numbers in parentheses are p-values for the null hypothesis of zero effect; significance code: * for 0.1, ** for 0.05, *** for 0.01.

Controlling for topic fixed effects, sub-network fixed effects, and total consumer attention in each knowledge market, we find there tend to be more opinion leaders (larger N_{jk}) and opinion leaders tend to be more similar in terms of who follow them

(larger S^{in}_{jk}), when active consumers (who pay more than median attention) occupy more central positions in a social network (larger r1 and r2).

A remark on effect magnitudes

A back-of-the-envelope calculation based on the information from Table 1-3 gives the following effect magnitudes. Increasing r1 by one standard deviation increases N by approximately 0.1 standard deviations and S^{in} by approximately 0.3 standard deviations. Increasing r2 by one standard deviation increases N by approximately 0.1 standard deviations and S^{in} by approximately 0.5 standard deviations. The scale of these effects suggests that the estimates capture economically important phenomena. In future investigation, researchers can study how important a 0.1-standard-diviation increase in the number of opinion leader is for opinion diversity, and how opinion diversity is associated with other variables such as social welfare.

4.2 Panel data analysis with IV: 2SLS

In this section we use r1iv and r2iv as instruments for r1 and r2 respectively and do a 2SLS version of the above panel data analysis. The second stage regression equations are the same as the regression equations in section 4.1 except that x_{jk} (r1 or r2) are substituted with their first stage fitted values. The first stage regression equations are as what follows:

$$\mathbf{x_{jk} = \tilde{\beta} \times xiv_{jk} + \tilde{\gamma} \times f_{jk} + \tilde{\mu}_j + \tilde{\rho}_k + \tilde{\varepsilon}_{jk}}$$

As in section 4.1, x_{jk} will be either $r1_{jk}$ (weighted average in-degree centrality of active knowledge consumers in the knowledge market jk, divided by 1000) or $r2_{jk}$ (weighted average eigenvector centrality of active consumers in market jk), with xiv_{jk} being r1iv and r2iv respectively. $\tilde{B}$ and $\tilde{\gamma}$ are coefficients and $\tilde{\varepsilon}_{jk}$ is the error term in the first stage. $\tilde{\mu}_j$ and $\tilde{\rho}_k$ are topic and sub-network fixed effects in the first stage. We next explain why the outcomes from using r1iv and r2iv as instruments are informative.

Even we control for topic fixed effects, sub-network fixed effects, and total consumer attention in each knowledge market, r1 and r2 may still be subject to endogeneity concern, though it's hard to imagine a story that the dependent variables (N and S^{in}) have causal impacts on the explanatory variables (r1 and r2), or to imagine a story that uncontrolled factors simultaneously impact both. One possibility is that the network

structure may be endogenous in a way that confounds our estimates. So, we use r1iv and r2iv as instruments which, compared with r1 and r2, are defined based on friendship networks rather than the original sub-networks (please refer to section 3 for details). We explain in what follows the logic of using these instruments.

If in Twitter I follow you and you don't follow me, I regard you as a source of knowledge. If I follow you and you also follow me, it's still possible that I regard you as a source of knowledge, but it's more possible that we are merely friends unrelated to knowledge market activity. The point is the friendship networks (as defined in section 3) should be exogenous or subject to endogeneity concern to smaller degree.[20] To put it in another way, some unknown interaction of topic characteristics and sub-network characteristics might impact both the number of opinion leaders and active consumers' positions in the original sub-networks, but the position measures based on the friendship networks are unlikely to subject to endogeneity concern, since it is mainly driven by friendship.

Of course, we don't know if the friendship networks are actually totally exogenous to knowledge market activity, but we can relax the assumption of absolute exogeneity and still have informative 2SLS estimates. Significant positive 2SLS estimates alone are not very informative, but significant positive 2SLS estimates also insignificantly different from (statistically equivalent to) the estimates in section 4.1 are informative and support that the endogeneity concern can be ignored. Let's explain why. First note this statement: (assuming that the friendship networks are subject to endogeneity concern to a smaller degree[21] than the original sub-networks) if endogeneity were really a concern, the estimates based on the friendship networks would significantly differ from those based on the original sub-networks. Now note the contrapositive of the statement: (assuming that the friendship networks are subject to endogeneity concern to a smaller degree than the original sub-networks) if the estimates based on the friendship networks did not significantly differ from those based on the original sub-networks, the endogeneity were not really concern. We report in table 4 and table 5 (on page 28) 2SLS outcomes and in table 6 (on page 29) Wald tests testing the equivalence of the estimates with and without instruments. The outcomes support the theoretical predictions in section 2 and the findings in section 4.1, as will be discussed.

[20] By a smaller degree, we mean the proportion of exogenous variation in the variation of a variable is smaller.
[21] By a smaller degree, we mean the proportion of exogenous variation in the variation of a variable is smaller.

Table 4: The effect of active consumers' centrality on the number of opinion leaders and their similarity in terms of who follow them: 2SLS and r1iv instrumenting r1 (in-degree centrality)

	Dependent Variables	
	N	S^{in}
r1 (how central active consumers are in the consumer social network)	70.2** (p=0.047)	0.243*** (p=0.002)
f (total consumer attention in each knowledge market, divided by 1000)	5.45*** (p<0.001)	0.00031 (p=0.246)
The first stage F-statistics for excluded iv is 1600		

1. Here we use the top 0.5% followees for topic modeling, treat any less than 10% topic proportion as zero, use median attention to define active consumers.
2. We control for the total amount of consumer attention (f_{jk}) in market (jk), topic fixed effects, and sub-network fixed effects. The data structure is an unbalanced panel with 99 observations. Compared with standard panel data, topic corresponds to individual and sub-network to time.
3. We use the Arellano covariance estimator to estimate standard errors (Croissant and Millo 2008; Arellano 1987)
4. Numbers in parentheses are p-values for the null hypothesis of zero effect; significance code: * for 0.1, ** for 0.05, *** for 0.01.

Table 5: The effect of active consumers' centrality on the number of opinion leaders and their similarity in terms of who follow them: 2SLS and r2iv instrumenting r2 (eigenvector centrality)

	Dependent Variables	
	N	S^{in}
r2 (how central active consumers are in the consumer social network)	166* (p=0.094)	0.918*** (p=0.0001)
f (total consumer attention in each knowledge market, divided by 1000)	5.43*** (p<0.001)	0.000207 (p=0.44)
The first stage F-statistics for excluded iv is 1000		

1. Here we use the top 0.5% followees for topic modeling, treat any less than 10% topic proportion as zero, use median attention to define active consumers.
2. We control for the total amount of consumer attention (f_{jk}) in market (jk), topic fixed effects, and sub-network fixed effects. The data structure is an unbalanced panel with 99 observations. Compared with standard panel data, topic corresponds to individual and sub-network to time.
3. We use the Arellano covariance estimator to estimate standard errors (Croissant and Millo 2008; Arellano 1987)
4. Numbers in parentheses are p-values for the null hypothesis of zero effect; significance code: * for 0.1, ** for 0.05, *** for 0.01.

In addition to the information in table 4 and 5, two things to note. First, as shown in section 3, the friendship networks, though "correlated" with the original sub-networks, are very different. Second, the first-stage F statistics for excluded IV are much greater than 10, so bias toward the OLS estimates of the 2SLS ones can be safely ignored.

In table 6 we fail to reject the equivalence of the estimates with and without instruments, so, per the above argument, endogeneity is not a problem in our case.[22]

Table 6: Equivalence tests with null hypotheses that coefficients of the primary explanatory variables are equivalent with and without instruments.

p-values for Wald tests with H0: OLS = 2SLS	N	S^{in}
r1/r1iv	0.278	0.792
r2/r2iv	0.396	0.629

Here we use the top 0.5% followees for topic modeling, treat any less than 10% topic proportion as zero, use median attention to define active consumers.

Controlling for topic fixed effects, sub-network fixed effects, and total consumer attention in each knowledge market, and using r1iv and r2iv to instrument r1 and r2 respectively, we get estimates insignificantly different from their counterparts in section 4.2. First-stages are strong. This insignificant difference is not due to imprecise estimation and implies we can ignore endogeneity concern. So, outcomes in section 4.2 supports the theoretical predictions in section 2 and the findings in section 4.1.

4.3 Robustness checks

In the main text we report outcomes from using the top *0.5%* followees for topic modeling, treating any less than *10%* topic proportion as zero (let's call this a topic proportion threshold),[23] and using *median* attention to define active consumers. In the appendix we experiment with different choices of these parameters: we repeat the above analysis (section 3-4.2) with any combination of top 1% or 0.5% followees, 10%

[22] First, the statistical equivalence is not resulted from imprecise estimations per information from Table 2-5. Second, please think about this in terms of standard Wald tests, rather than in terms of the Hausman test which is a similar exercise interpreted in a way specific for certain purpose.

[23] When a top followee's topic proportion on a topic is smaller than 10%, we treat the topic proportion as zero; also see the previous section when we define N_{jk}.

or 20% topic proportion threshold, and median or mean to define active consumers. So together with r1 and r2 skinning the same cat in two different ways and with OLS and 2SLS (if the 2SLS regressions are viewed as robustness checks), we check robustness across 32 possibilities for 64 coefficients (there are two dependent variables). The conclusions are the same, with only one exception that when we regress the number of opinion leaders on r2 with and without instrument using the top 1% followees, 20% topic proportion threshold, and median, we get two estimates with p-values greater than 0.1. Since the robustness is checked across many possibilities, we think this single fish doesn't spoil the whole pond.

5 Discussion: methodology reflection and what can be done in the future

In this paper we find that if active knowledge consumers occupy more central positions in the social network of consumers, there will be more opinion leaders and opinion leaders are more similar in terms of who pay attention to them. This is the first paper to formally define opinion leaders, knowledge markets, and consumer attention. While the existing literature emphasizes the role of opinion providers' network positions on the making of opinion leaders, this work shows the network positions of active consumers matter because active consumers serve as a propagation machine. In this last section we highlight some methodological points in this paper that we think may offer other researchers something new.

A multi-layer network perspective: IV based on friendship networks

In the regression analysis we use measures defined based on friendship networks as instrumental variables for those based on the original network. We've talked about our logic in section 4. Now let's try to be general. People live simultaneously in different networks or in a network with many layers, e.g., a network based on friendship (a friendship layer) and a network based on financial relationship (a finance layer). Though different layers can be deeply interacted with each other, each layer should to some degree be able to provide exogenous variation for another layer. Of course, the use of a multi-layer network structure for economics and econometrics is more than searching for instrumental variables. For technical tools and applications in the field of multilayer networks, please see Bianconi (2018).

Sub-network structure and panel data

Sub-network detection is a well-developed field in the network science (Thurner, Hanel, and Klimek 2018; Newman 2018). To the best of our knowledge, no one else has applied sub-network detection to construct a panel data structure for econometric analysis, as is the case in this paper. We sub-network detection in this paper not only because what happens in nearly isolated sub-networks should be disentangled, but also because panel data structure improves identification.

An embeddedness perspective

A rudimental difference between the paradigm of economics and that of sociology is that economists use rational calculation to explain human behavior and sociologists see people as social constructions whose behavior depends on their social embeddedness (Grannovetter 1985 and 2017). We believe that why and how people do maximization can be a function of social contexts and that social influence can exist due to rational calculation. In the sense that "All models are wrong, but some are useful" (Box, 1976), we don't argue which paradigm is more fundamental. To be useful, our model uses features from both paradigms: we model rational agents socially embedded in specific ways. Also, we think it's fruitful in many research contexts to be specific about agents' social embeddedness, rather than simply assuming people are interacting with each other in a well-stirred manner (Belleflamme and Peitz 2015).

What can be done in the future?

First, in this paper we don't formally study how the number of opinion leaders relates to opinion diversity or how opinion diversity matters, e.g. how it enhances adaptive efficiency for a society. These can be topics for future research, and an important next step in the direction is to design criteria to qualify opinions, which we haven't seen in the academic world. Such a step may need a combination of text analysis and social network analysis, like what we do this paper. Second, it's important for future research to explore subtleties on the supply side of knowledge markets; knowledge suppliers face a different incentive structure than suppliers in commodity, service, or factor markets. Finally, in the future it can be very interesting to study what forces can drive active consumers to and away from central social network positions.

References

Adamic, L, and Adar, E., 2003. Friends and neighbors on the Web. Social Networks, 25(3):211-230.

Akerlof, G.A., 2018. Sins of Omission and the Practice of Economics. Journal of Economic Literature.

Arellano, M., 1987. PRACTITIONERS' CORNER: Computing robust standard errors for within groups estimators. Oxford bulletin of Economics and Statistics, 49(4), pp.431-434.

Barabási, A.L. and Albert, R., 1999. Emergence of scaling in random networks. science, 286(5439), pp.509-512.

Belleflamme, P., and Peitz, M., 2015. Industrial Organization: Markets and Strategies. Cambridge University Press

Blei, D.M. and Lafferty, J.D., 2006, June. Dynamic topic models. In Proceedings of the 23rd international conference on Machine learning (pp. 113-120). ACM.

Bianconi, G., 2018. Multilayer Networks: Structure and Function. Oxford University Press.

Bonacich, P., 1987. Power and Centrality: A Family of Measures. American Journal of Sociology,92, 1170-1182.

Bruner, J., 2013. Tweets Loud and Quiet. https://www.oreilly.com/ideas/tweets-loud-and-quiet

Clauset, A., Newman, M.E. and Moore, C., 2004. Finding community structure in very large networks. Physical review E, 70(6), p.066111.

Coase,R., and Wang,N., 2012. How China Became Capitalist. Palgrave Macmillan.

Croissant, Y. and Millo, G., 2008. Panel data econometrics in R: The plm package. Journal of statistical software, 27(2), pp.1-43.

Csárdi, M.G., 2019. R Package 'igraph'.

Eagle, N., Macy, M. and Claxton, R., 2010. Network diversity and economic development. Science, 328(5981), pp.1029-1031.

Easley, D., and Kleinberg, J., 2010. Networks, Crowds, and Markets. Cambridge University Press.

Gentzkow, M., Kelly, B.T. and Taddy, M., 2017. Text as data (No. w23276). National Bureau of Economic Research.

Grajzl, P. and Murrell, P., 2019. Toward understanding 17th century English culture: A structural topic model of Francis Bacon's ideas. Journal of Comparative Economics, 47(1), pp.111-135.

Grannovetter, M., 2017.Society and Economy: Framework and Principles. Belknap Press: An Imprint of Harvard University Press.

Granovetter, M., 1985. Economic action and social structure: The problem of embeddedness. American journal of sociology, 91(3), pp.481-510.

Granovetter, M., 1973. The Strength of Weak Ties. American Journal of Sociology, 78(6), pp.1360-1380.

Hayek, F.A., 1945. The use of knowledge in society. The American economic review, 35(4), pp.519-530.

Hodas, N.O. and Lerman, K., 2014. The simple rules of social contagion. Scientific reports, 4, p.4343.

Jackson, M., 2019. The Human Network. Pantheon.

Katz, E, and Lazarsfeld, P, 1955. Personal Influence. Routledge.

Newman, M., 2018. Networks. Oxford University Press

North, D., 1990. Institutions, Institutional Change and Economic Performance. Cambridge University Press

North, D., Wallis, J.J., and Weingast, Barry, 2009. Violence and Social Orders: A Conceptual Framework for Interpreting Recorded Human History. Cambridge University Press

Page, S., 2018. The Model Thinker. Basic Books.

Reddy, P.K., Kitsuregawa, M., Sreekanth, P. and Rao, S.S., 2002, December. A graph based approach to extract a neighborhood customer community for collaborative filtering. In International Workshop on Databases in Networked Information Systems (pp. 188-200). Springer, Berlin, Heidelberg.

Roberts, M.E., Stewart, B.M. and Tingley, D., 2014. stm: R package for structural topic models. Journal of Statistical Software, 10(2), pp.1-40.

Robinson, David, and Silge, Julia, 2017. Text Mining with R: A Tidy Approach. O'Reilly Media.

Salganik, M.J., Dodds, P.S. and Watts, D.J., 2006. Experimental study of inequality and unpredictability in an artificial cultural market. science, 311(5762), pp.854-856.

Shelling, T.C., 2006. Micromotives and Macrobehavior. W. W. Norton & Company.

Thurner, S., Hanel, R., and Klimek, P, 2018.Introduction to the Theory of Complex Systems. Oxford University Press.

Tolkien, J.R.R., 1955. The Lord of the Rings. Allen & Unwin.

Yang, X., 2003. Economic Development and the Division of Labor. BLACKWELL PUBLISHERS

Appendix: robustness check

Table A1: the main results (top 0.5% followee, 10% threshold, mean)

	N	S^{in}	1st-stage F for excluded IV	Statistically equal to its 2sls
r1	1.28e-01 (0.014**)	3.06e-04 (0.003***)		Yes
r2	2.40e+02 (0.043**)	1.08e+00 (0.00001***)		Yes
2SLS: r1 r1iv as iv	1.09e-01 (0.027**)	3.15e-04 (0.002***)	900	
2SLS: r2 r2iv as iv	2.34e+02 (0.059*)	1.13e+00 (0.000003***)	900	

Table A2: the main results (top 0.5% followee, 20% threshold, median)

	N	S^{in}	1st-stage F for excluded IV	Statistically equal to its 2sls
r1	6.73e-02 (0.031**)	2.08e-04 (0.007***)		Yes
r2	1.34e+02 (0.077*)	7.76e-01 (0.001***)		Yes
2SLS: r1 r1iv as iv	5.70e-02 (0.044**)	2.14e-04 (0.005***)	1700	
2SLS: r2 r2iv as iv	1.26e+02 (0.094*)	7.99e-01 (0.001***)	1000	

Table A3: the main results (top 0.5% followee, 20% threshold, mean)

	N	S^{in}	1st-stage F for excluded IV	Statistically equal to its 2sls
r1	1.17e-01 (0.014**)	3.03e-04 (0.006***)		Yes
r2	2.12e+02 (0.039**)	1.04e+00 (0.0001***)		Yes
2SLS: r1 r1iv as iv	1.04e-01 (0.021**)	3.16e-04 (0.004***)	840	
2SLS: r2 r2iv as iv	2.11e+02 (0.048**)	1.10e+00 (0.00004***)	780	

Table A4: the main results (top 1% followee, 10% threshold, median)

	N	S^{in}	1st-stage F for excluded IV	Statistically equal to its 2sls
r1	6.10e-01 (0.0004***)	3.82e-04 (0.000005***)		Yes
r2	8.48e+02 (0.011**)	6.88e-01 (0.005***)		Yes
2SLS: r1 r1iv as iv	5.20e-01 (0.001***)	3.76e-04 (0.00002***)	1300	
2SLS: r2 r2iv as iv	7.22e+02 (0.022**)	6.28e-01 (0.013**)	530	

Table A5: the main results (top 1% followee, 10% threshold, mean)

	N	S^{in}	1st-stage F for excluded IV	Statistically equal to its 2sls
r1	5.97e-01 (0.001***)	3.63e-04 (0.00001***)		Yes
r2	8.76e+02 (0.011**)	6.96e-01 (0.001***)		Yes
2SLS: r1 r1iv as iv	5.29e-02 (0.001***)	3.67e-04 (0.00001***)	1400	
2SLS: r2 r2iv as iv	7.73e+02 (0.016**)	6.55e-01 (0.004***)	580	

Table A6: the main results (top 1% followee, 20% threshold, median)

	N	S^{in}	1st-stage F for excluded IV	Statistically equal to its 2sls
r1	3.11e-01 (0.004***)	3.00e-04 (0.002***)		Yes
r2	3.94e+02 (0.123)	6.28e-01 (0.047**)		Yes
2SLS: r1 r1iv as iv	2.37e-01 (0.048**)	2.95e-04 (0.006***)	1600	
2SLS: r2 r2iv as iv	2.79e+02 (0.30)	6.33e-01 (0.066*)	520	

Table A7: the main results (top 1% followee, 20% threshold, mean)

	N	S^{in}	1st-stage F for excluded IV	Statistically equal to its 2sls
r1	4.72e-01 (0.00004***)	3.18e-04 (0.002***)		Yes
r2	6.88e+02 (0.016**)	7.18e-01 (0.017**)		Yes
2SLS: r1 r1iv as iv	3.93e-01 (0.0005***)	3.25e-04 (0.002***)	1300	
2SLS: r2 r2iv as iv	5.66e+02 (0.045**)	7.47e-01 (0.023**)	530	