Journal of Financial Economics 131 (2019) 323–344

Contents lists available at ScienceDirect

Journal of Financial Economics

journal homepage: www.elsevier.com/locate/jfec

Inferring latent social networks from stock holdings☆

Harrison Hong[a,*], Jiangmin Xu[b]

[a] Department of Economics, Columbia University, 1022 International Affairs Building, Mail Code 3308, 420 West 118th Street, New York, NY 10027, United States

[b] Guanghua School of Management, Peking University, 5 Yiheyuan Road, Haidian District, Beijing 100871, China

ARTICLE INFO

Article history:
Received 3 August 2015
Revised 15 May 2017
Accepted 15 August 2017
Available online 31 August 2018

JEL classification:
G11
G23
G32

Keywords:
Social networks
Poisson regressions
Investor behavior

ABSTRACT

We infer the latent social networks of investors using data on their stock holdings. We map linkages to portfolio weights using a portfolio-choice model. The precision of an investor's private signal about firm value is assumed to increase with his connections in the city where the firm is headquartered. Using money-manager data, we find that managerial linkages to a city are overly dispersed relative to the Erdös–Rényi model of i.i.d. connections. Managers at the tail of this distribution with non-i.i.d. linkages have more university alumni in that city. Their stock holdings there outperform their holdings in other cities.

1. Introduction

In this paper, we develop a strategy to infer the structure of latent social networks from investors' stock holdings. To gain an intuition for our strategy, we plot in Fig. 1 an illustrative graph of the social networks of mutual fund managers in different cities. The table on the right of this figure is an example of managers' portfolio weights in different cities that manifest from these latent social networks. As we discuss below, many recent studies find that such networks matter greatly for these investment decisions (see, e.g., Hong et al., 2005; Cohen et al., 2008).

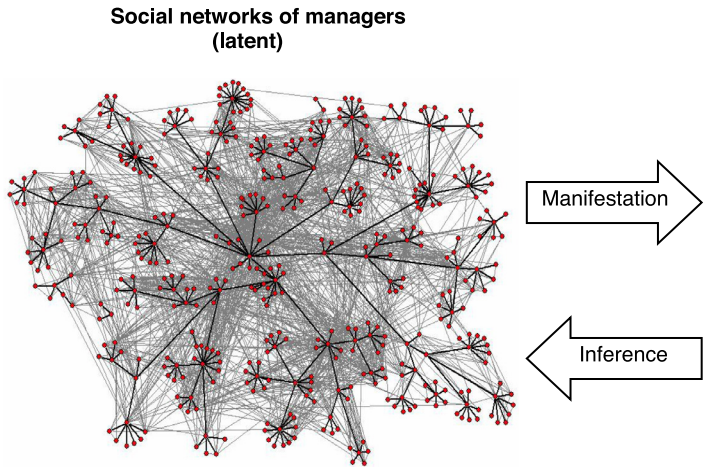
We use a standard portfolio model, where risk-neutral fund managers face short-sales constraints and quadratic trading costs, to infer from these stock holdings a count of the latent managerial connections. The key assumption is that the precision of the manager's private signal about firm value increases with the size of a manager's network in the city where the firm is headquartered. This greater precision, as we show below, leads to a higher expected portfolio weight of stocks headquartered in that city because investors are short-sales constrained. We can then infer linkages from the geographic dispersion of portfolio weights among the managerial population.

Despite the advent of Facebook and LinkedIn, data on the social networks of investors are still extremely limited. Such data are still partial and noisy at best and friends on social media also need not be the most valuable sources of information when it comes to investing decisions. Thus,

☆ We are deeply appreciative to the referee (Christopher Malloy) for many helpful comments. We want to thank Johannes Ruf, Peter Bossaerts, Soohun Kim, Motohiro Yogo, Xavier Gabaix, Ralph Koijen, Bruno Biais, and Thomas Chaney for helpful comments. We also thank seminar participants and discussants at the University of Connecticut, University of Utah, AFA 2015 Meetings, University of Minnesota Carlson School, Singapore Management University, China International Finance Conference 2014, NYU Stern School of Business, Hong Kong University of Science and Technology, and Bank of International Settlements for helpful comments. This paper was previously circulated under the title "Count Models of Social Networks in Finance".

* Corresponding author.
E-mail addresses: hh2679@columbia.edu (H. Hong), jiangminxu@gsm.pku.edu.cn (J. Xu).

https://doi.org/10.1016/j.jfineco.2017.08.005
0304-405X/



Portfolio weights of managers (observed)

Manager	City			
	NY	LA	Boston	...
John	25%	10%	8%	...
Steve	20%	0%	20%	...
Laura	12%	15%	15%	...
Jennifer	10%	30%	5%	...
...	...	...	...	...

Fig. 1. Latent social networks of fund managers and their portfolio weights on stocks headquartered in different cities. The figure on the left is an illustrative graph of the social networks of mutual fund managers in different cities. The table on the right is an illustrative example of managers' possible portfolio weights on stocks headquartered in different cities.

being able to infer latent social networks using investor holdings, which are often available, can yield a potentially valuable new source of data with which to understand the structure of investors' social networks and its impact on investment performance.

To get started, we propose a tractable statistical model of random networks (see Zheng et al., 2006), which nests the null or benchmark Erdös and Rényi (1959) model of independently and identically distributed (i.i.d.) connections as a special case. Agents have a binomial probability of being connected to other agents but some subgroups of the population can have a non-i.i.d. probability of being connected to each other. This model has been used to fit data from surveys (Killworth et al., 1998) about the count of friends a person has in different groups within the general population, such as "how many people named Nicole do you know?" or "how many prisoners do you know?"

While this model is not the most general one possible, it yields a tractable framework when embedded in our portfolio choice program and at the same time captures a defining aspect of social networks, which is overdispersion or the non-i.i.d. probability of having certain subgroups of the population being connected to each other.[1] For instance, the count distribution for the number of friends named Nicole is Poisson (the mean and variance of the distribution are equal to each other), because the chance of knowing a Nicole is i.i.d. across agents. But the count distribution for the number of friends who are prisoners is an Overdispersed Poisson – that is, most people know zero but there are a small group of people who know many. This group in the tail of the count distribution are presumably people who have been to prison and hence have non-i.i.d. chances of knowing a prisoner, i.e., they are part of the prisoner network.

We then embed this model into an otherwise standard portfolio choice program where fund managers are risk neutral but face short-sales constraints and quadratic trading costs. Both of these assumptions are realistic for fund managers. They are prohibited from shorting by institutional charter (see, e.g., Chen et al., 2002) and large risk positions have to incur price impact or hit institutional risk management constraints (Gârleanu and Pedersen, 2013). Managers form their portfolios based on both priors and private signals. The precision of the private signal of the value of a stock headquartered in a given city is assumed to increase with his non-i.i.d. connections in that city, i.e., the size of his network in that city.

These non-i.i.d. connections might be relatives or friends from a manager's hometown who work in the local companies there (Pool et al., 2015). They might be alumni from a university who have knowledge of local companies because they work there or might know other people who work there (Cohen et al., 2008; Engelberg et al., 2012). A manager who is an alumnus of Harvard University is likely to have disproportionate contacts in Boston compared to a typical manager in the population. Since people self-select into undergraduate and graduate programs with classmates and professors having similar interests, these school relationships tend to be more persistent and hence are likely to play an important role in job-related decisions (see, e.g., Lazarsfeld and Merton, 1954; McPherson et al., 2001).

We model these signals in the standard linear-normal signal extraction framework. The signal might end up being positive or negative, but we prove that the expected portfolio weight of a stock increases in the number of connections in that city. The reason is censoring due to short-sales constraints: as a manager's private-signal precision increases, he places more weight on these signals relative to his prior. Since there are shorting constraints, we only see the positive latent demand from the increased precision and hence the higher expected portfolio weight. Thus, our model provides a monotonic transformation of the social connections in a city into the expected portfolio weight in a city.

[1] Networks often follow certain power laws that are absent from this model (see, e.g., Jackson and Rogers, 2007; Gabaix, 2009).

To estimate the model, we use the well-known Expectation-Maximization (EM) algorithm (Dempster et al., 1977; Wu, 1983), which is the leading statistical method to infer parameters of latent variables based on observables derived from these variables. The key to successfully employing this method is being able to obtain closed-form solutions for likelihoods, which we are able to do for our model. We estimate our model using panel data on the holdings of mutual fund managers in different cities over the period of 1991 to 2015. In other words, we can infer the parameters of the underlying latent social networks that produce the portfolio weights across different cities by using observations on these weights (see Fig. 1). Our approach is similar to a strand of literature in computer science that tries to infer latent social networks from Internet data (see, e.g., Adamic and Adar, 2005).

Using the 20 largest Metropolitan Statistical Areas (MSAs) in terms of where stocks are headquartered as groups and the holdings of actively managed mutual funds, we find significant overdispersion of the count of connections to a given city.[2] That is, the distribution of the number of stocks in a given city held by managers and the associated managerial linkages distribution are too overdispersed or fat-tailed to be generated by the Erdös and Rényi (1959) model of i.i.d. connections. We can compare the degree of overdispersion of our money-manager connections to different cities to the degree of overdispersion in connections to other subgroups of the population such as people named Nicole, postal workers, or prisoners that are studied in Zheng et al. (2006). The degree of overdispersion in managerial connections to a given city is in the middle of the range of estimates obtained in this earlier work: it is much larger than connections to people named Nicole, comparable to knowing postal workers, but less than having a connection to prisoners.

These results are not driven by manager covariates as we can extend the model to account for a variety of manager-fund characteristics. This accounts for the worry, for instance, that growth funds might like to concentrate holdings of stocks headquartered in San Jose. These results are also not due to local bias of mutual fund managers (Coval and Moskowitz, 1999). We can alternatively re-estimate our model using just managers based in New York City and Boston and model their portfolio weights of stocks in other cities and find similar results. We can also account for whether our results are driven by a thy-neighbor's-effect of Hong et al. (2005) or family effect whereby managers from the same family invest in the same stocks geographically. We do this by accounting for each manager the average holdings of the other managers located in the same city or of other managers in the same family.

不平均分布

Given that our parameter estimates of the latent network point to significant overdispersion in connections to cities, this means that some managers in the population have non-i.i.d. ties to a given city. We would ideally like to identify these managers and see if they indeed have non-i.i.d. ties and better fund performance. To this end, our model gives a prediction for the expected number of contacts any manager should have in a given city under the Erdös and Rényi (1959) null model. We can then calculate for each manager his relative propensity to have contacts (RPC) in a city, $\eta_{i,k}$, which is a manager's actual number of contacts implied from stock holdings divided by the predicted number under the Erdös and Rényi (1959) null model. A manager who has a higher number of contacts than under Erdös and Rényi (1959) null (i.e., $\eta_{i,k} > 1$) is more likely to have a network in that city.

We find that our relative propensity to connect measure, which is a measure of the size of non-i.i.d. connections in a city or the size of the network in that city, predicts out-of-sample actual managerial networks in different cities. As a proxy for networks, we follow Cohen et al. (2008) in constructing for each manager his friendship network to the board of directors of the publicly traded companies based on the undergraduate/graduate institution that the manager shares with different members of the board of directors (i.e., an alumni measure of social connection). We can construct for each city k a manager i's alumni connections in that city, $Alumni_{i,k}$. Our interpretation is that these social connections are non-i.i.d. and so we expect and show that our RPC measure in any given city is highly correlated with the alumni connections measure in that city.

We also show that our measure is useful for understanding fund performance even if there is partial data on managerial networks. Importantly, we use our model to prove that the expected returns of stocks held by the manager are higher for those stocks located in cities where the manager has more non-i.i.d. connections due to this improved precision effect of his private signal. We find that the portfolio of stocks located in cities held by managers with higher RPC scores there outperform the other stocks in the managerial portfolios. We show that this outperformance is robust to performance benchmarks and is there in all fund style classes. In the Internet Appendix, we consider a number of robustness exercises including estimating our model for a hedge fund sample, excluding micro-cap funds, and including small MSAs.

Our findings are also related to the literature on modeling the holdings of smart money managers and how it reflects their preferences and information (Falkenstein, 1996; Zheng, 1999; Kacperczyk et al., 2005). Our focus is on the information contained about their social networks. We use mutual fund manager data to exposit and test our methodology. But our methodology could easily be applied to retail investor holdings data (Barber and Odean, 2001). Our approach only requires investor holdings that are observed even in a limited number of cross-sections as opposed to needing higher frequency trading data. For instance, with very high frequency trading data, one could try to categorize connections by looking at the similarity in timings of trades. Yet such high frequency data are also limited. Moreover, our approach does not require that the mutual fund managers know each other, only that a manager has connections, even outside the industry, in different cities.

The paper is organized as follows. We describe the model in Section 2 and the data and estimation procedures

[2] In our empirical analysis, we also drop sector funds along with index funds.

in Section 3. We collect the results in Sections 4 and 5. We conclude in Section 6.

2. The model

2.1. Structure of social network

We model the latent random social networks of investors (e.g., mutual fund managers) using the structure proposed in Zheng et al. (2006), which nests Erdös and Rényi (1959) as a special case. Consider a population of N persons, which encompasses our investors of interest as a subset. Let p_{ij} be the probability that a person i is connected to person j. Then

$$a_i \equiv \sum_{j,\ j\neq i} p_{ij} \tag{1}$$

denotes the gregariousness (the expected total number of connections) of person i. We let

$$b_k \equiv \frac{\sum_{i\in S_k} a_i}{\sum_{i=1}^{N} a_i} \tag{2}$$

then denote the proportion of total social connections that involves group k where S_k stands for "group k." A group in our context is a city k to which each person primarily resides. Person i's expected number of connections in city k is given by

$$\mu_{ik} \equiv \sum_{j\in S_k} p_{ij}. \tag{3}$$

Person i's expected relative propensity to be connected to a person in city k is defined as

$$g_{ik} \equiv \mu_{ik}/(a_i b_k). \tag{4}$$

This relative propensity will play a key role in our analysis below as it captures the extent to which a person's, i.e., our manager's, connections are concentrated in a city.

2.1.1. The Erdos–Renyi null (Poisson) model

Let y_{ik} denote the number of connections a person i has in city k. If social connections are independently and identically formed as in Erdös and Rényi (1959), y_{ik} follows a Poisson distribution with a mean $\mu_{ik} = a_i b_k$ and a probability function

$$\rho(y_{ik}|a_i, b_k) = \frac{(a_i b_k)^{y_{ik}} \exp(-a_i b_k)}{y_{ik}!}. \tag{5}$$

The variance of the Poisson distribution is equal to its mean.

As a result, a key feature of the model which drives our empirical analysis is the following. Some people may be more gregarious and have more social ties in expectation (i.e., different a_i's); but to the extent links are formed randomly, the expected number of connections in a given city is simply the gregariousness of that individual (a_i) times the fraction of aggregate connections that city accounts for (i.e., b_k). In other words, the relative propensities g_{ik} are all equal to one in Erdös and Rényi (1959).

In the context of our managers, consider a manager with an expected number of connections equal to 100. Suppose that Los Angeles accounted for 10% of aggregate connections while San Diego accounted for 5%. Under Erdös and Rényi (1959), this manager should have ten connections in Los Angeles and five in San Diego. His expected connections in any given city should scale with the aggregate links in that city since his ties are formed randomly. This is the defining characteristic of what we will refer to as our null [Poisson or Erdös and Rényi (1959)] model.

2.1.2. The overdispersed model

An important departure from the null model is likely to occur if there are structured social networks formed in a non-i.i.d. fashion. To be more precise, we distinguish between being part of a network from being merely gregarious. Being part of a network would mean that some people have a non-i.i.d. relative propensity to make connections to people from certain cities, i.e., $\{g_{ik}\} \neq 1$. As we discussed in the Introduction, these non-i.i.d. connections may be due to where the person grew up, where the person went to school, or where the person's friends happen to work. In the same way prisoners are much more likely to know prisoners, a manager who went to school at Harvard or grew up in San Jose will have concentrated or abnormal numbers of connections in those cities relative to what is predicted by Erdös and Rényi (1959).

To model these non-i.i.d. propensities, we allow people to differ not only in their gregariousness $\{a_i\}$, but also in their relative propensity $\{g_{ik}\}$ to be connected to different cities: y_{ik} is distributed as Poisson with a mean $\mu_{ik} = a_i b_k g_{ik}$. We then let g_{ik} follow a gamma distribution with a mean equal to one and a variance equal to $(\omega_k - 1)$ where $\omega_k > 1$.[3] Such a Poisson-gamma mixture leads to a (marginal) distribution/density for y_{ik} that is negative binomial (after integrating out g_{ik} and using an appropriate reparameterization)[4]

$$\rho(y_{ik}|a_i, b_k, \omega_k) = \frac{\Gamma(y_{ik} + \zeta_{ik})}{\Gamma(\zeta_{ik})\,\Gamma(y_{ik}+1)} \left(\frac{1}{\omega_k}\right)^{\zeta_{ik}} \left(\frac{\omega_k - 1}{\omega_k}\right)^{y_{ik}}, \tag{6}$$

where $\Gamma(\cdot)$ is the gamma function and $\zeta_{ik} = a_i b_k/(\omega_k - 1)$. y_{ik} then has a mean equal to $a_i b_k$ and a variance $\omega_k a_i b_k$ that is greater than its mean ($\omega_k > 1$).

We call this our overdispersed model. This is because variations in the relative propensities $\{g_{ik}\}$ have resulted in overdispersions, i.e., y_{ik}'s variance exceeds its mean, in contrast to our Poisson null model with equal mean and variance $a_i b_k$. Moreover, the ω_k's are called overdispersion parameters. They measure people's non-identicalness in forming ties to certain groups/cities and being part of structured social networks. The mean of the gamma distribution is set to one, which means that most people have

[3] In the most general form, $\{g_{ik}\}$ varies for each (i, k) pair. But this leads to an unparsimonious model that is difficult to estimate. The reason is because we only have $N \times K$ number of observations of investors' stock holdings. It is then not feasible to estimate $N \times K$ number of g_{ik}'s with only $N \times K$ number of data points unless we also use the time dimension T and assume that these propensities are constant across time, which may not be a valid assumption. Thus we opt for the most parsimonious model possible in our analysis.

[4] For a reference on this type of Poisson-gamma mixture, see Cameron and Trivedi (2005), Chapter 20.

i.i.d. ties. But those in the tail of the distribution, with $g_{ik} > 1$ have a higher relative propensity to connect to people from city k than an average person in the population.

In the context of our managers setting, if we observed the number of friends a manager has in different cities, i.e., $y_{i,k}$, it is straightforward to estimate the parameters of the social network, a_i, b_k, and ω_k, as executed in Zheng et al. (2006). The key parameter of interest is ω_k—the larger is ω_k, the more likely it is that certain managers have non-i.i.d. contacts in that city. But the motivation for our paper is that $y_{i,k}$ is not readily observable for managers. What is observable is the managers' portfolio weights of stocks headquartered in different cities. As such, we need to articulate a model that maps these linkages to portfolio choice so as to infer these latent networks based on readily observable portfolio weights.

2.2. Portfolio choices of investors

To do this, we model in every period t, for each city k, an investor i that is considering investing in the set of stocks $\mathcal{J}_k$ that is headquartered there. The set of stocks over all cities is denoted as $\mathcal{J} = \bigcup_k \mathcal{J}_k$. The friends that investor i has in city k inform the investor about the stocks there through the local information they have (private signals). Investor i also has priors on the stocks. As a result, the expected excess return (net of the risk-free rate R_f) of stock j headquartered in city k for investor i is modeled as:

$$\mathbb{E}_i\left(\tilde{R}_{j,k} - R_f\right) = (\alpha + \boldsymbol{\beta}\boldsymbol{X}_j) + \lambda_{i,k}\xi_{i,j,k}, \tag{7}$$

where $\boldsymbol{X}_j$ is a vector of stock j's usual style characteristics that include firm size, book-to-market ratio, and momentum. Eq. (7) can be micro-founded through a belief-updating model with linear-normal priors and private signals, the details of which are in the Appendix. Intuitively, the first term, $(\alpha + \boldsymbol{\beta}\boldsymbol{X}_j)$, is coming from the prior on expected returns that is common to all investors regarding stock j.[5] The second term, $\lambda_{i,k}\xi_{i,j,k}$, is coming from the weight that the investor puts on the private signal (relative to the prior). $\xi_{i,j,k}$, which is i.i.d. across $\{i, j, k\}$, follows a normal distribution with mean zero and variance normalized to one. $\lambda_{i,k}$ is an increasing function of the friends $y_{i,k}$. The higher the number of friends $y_{i,k}$ that investor i has in city k, the more precise the private signal he receives from his friends regarding the expected return of the stock j headquartered in city k, thus the more weight $\lambda_{i,k}$ investor i will put on the term $\xi_{i,j,k}$ from his private signal.

For tractability, we model the weight $\lambda_{i,k}$ as a linear function of the number of friends $y_{i,k}$:

$$\lambda_{i,k} = \lambda_0 + \lambda_1 y_{i,k}, \tag{8}$$

where λ_0 and λ_1 are both positive constants.

Otherwise, the portfolio choice problem is standard. We assume that all investors are risk neutral. Consequently, under the presence of a risk-free asset, investor i's objective is to maximize the expected excess return of his investment portfolio[6]:

$$\mathbb{E}_i\left(\tilde{R}_i - R_f\right) = \sum_k \sum_{j\in\mathcal{J}_k} \mathbb{E}_i\left(\tilde{R}_{j,k} - R_f\right) w_{i,j,k}, \tag{9}$$

where $\mathbb{E}_i(\cdot)$ is investor i's conditional expectations operator (conditional on the information (signals) he receives on the expected return of the stocks), $\tilde{R}_i - R_f$ is the excess return of investor i's portfolio, and $w_{i,j,k}$ is investor i's portfolio weight on stock $j \in \mathcal{J}_{i,k}$. We assume that investors are short-sale constrained such that $w_{i,j,k} \geq 0$ for all stocks. We assume that investors face quadratic costs from transactions for portfolio adjustments. For our case, we specify the quadradicity of transaction costs in terms of acquired market value. That is, when investor i acquires $n_{i,j,k}$ shares of stock j headquartered in city k, he pays assuming a stock price P_j:

$$TC_{i,j,k} = \frac{1}{2}\Lambda_{i,j}\left(P_j n_{i,j,k}\right)^2. \tag{10}$$

We parameterize $\Lambda_{i,j}$ as:

$$\Lambda_{i,j} = \frac{c}{W_i^2}, \tag{11}$$

where c is a transaction cost parameter common for all investors and stocks and W_i is investor i's total asset (portfolio) value. Under the parameterization in (11), the transaction cost in (10) becomes:

$$TC_{i,j,k} = \frac{c}{2} w_{i,j,k}^2. \tag{12}$$

Combining Eqs. (7)–(9), and (12), an investor i's objective can be expressed as follows:

$$\max_{\{w_{i,j,k}\geq 0\}_{j\in\mathcal{J}}} \left\{ \sum_k \sum_{j\in\mathcal{J}_k} \left[(\alpha + \boldsymbol{\beta}\boldsymbol{X}_j + (\lambda_0 + \lambda_1 y_{i,k})\xi_{i,j,k}) w_{i,j,k} - \frac{c}{2} w_{i,j,k}^2 \right] \right\}. \tag{13}$$

The Karush–Kuhn–Tucker (KKT) conditions imply that:

$$w_{i,j,k} = \left(\frac{\alpha + \boldsymbol{\beta}\boldsymbol{X}_j + \left(\lambda_0 + \lambda_1 y_{i,k}\right)\xi_{i,j,k}}{c} \right)^+, \tag{14}$$

where $(\cdot)^+ \equiv \max\{\cdot, 0\}$ simultaneously captures investor i's decision of whether and how much to invest in every stock headquartered in different cities under short-sale constraints. According to Eq. (14), investor i places a positive portfolio weight on stock j headquartered in city k provided that the expected excess return of the stock is positive based on the common prior and the private signal from his friends. The transaction costs help to pin down the level of the portfolio weight. The transaction cost parameter c in Eq. (14) is a scale parameter and cannot be identified separately from the parameters $\{\alpha, \boldsymbol{\beta}, \lambda_0, \lambda_1\}$.[7]

[5] This parametric specification is in the spirit of Brandt et al. (2009).

[6] Investor i's budget constraint requires the sum of his stock-portfolio weights to equal one minus the weight on the risk-free asset. The latter is assumed to be perfectly adjustable to the needs of the portfolio optimization problem.

[7] If we multiply α, $\boldsymbol{\beta}$, λ_0, and λ_1 by any positive constant in the numerator of (14) and c by the same constant in the denominator, it will yield a condition that is observationally equivalent to (14) and is not separately identifiable from (14).

Therefore, we normalize the transaction cost parameter c to one.

With the optimal portfolio weight in Eq. (14), the conditional probability density of portfolio weight $w_{i,j,k}$ for investor i, given the number of friends $y_{i,k}$ and $\xi_{i,j,k} \overset{i.i.d}{\sim} \mathcal{N}(0,1)$, is therefore

$$\begin{aligned} & f(w_{i,j,k}|y_{i,k}, \boldsymbol{X}_j, \theta) \\ & = (f(w_{i,j,k}|y_{i,k}, \boldsymbol{X}_j, \theta))^{d_{i,j,k}} (\mathbb{P}(w_{i,j,k}=0))^{1-d_{i,j,k}} \\ & = \left(\frac{1}{(\lambda_0+\lambda_1 y_{i,k})} \varphi\left(\frac{w_{i,j,k} - (\alpha+\beta \boldsymbol{X}_j)}{(\lambda_0+\lambda_1 y_{i,k})} \right) \right)^{d_{i,j,k}} \\ & \times \left(\Phi\left(\frac{-(\alpha+\beta \boldsymbol{X}_j)}{(\lambda_0+\lambda_1 y_{i,k})} \right) \right)^{1-d_{i,j,k}}, \end{aligned} \tag{15}$$

where $d_{i,j,k}$ is an indicator variable such that

$$d_{i,j,k} = \begin{cases} 1 & \text{iff } w_{i,j,k} > 0 \\ 0 & \text{iff } w_{i,j,k} = 0 \end{cases}$$

and φ, Φ are respectively the probability density function (pdf) and the cumulative distribution function (cdf) of the standard normal distribution. We use θ to denote all the relevant model parameters. The conditional density is intuitive given the linear-normal filtering set-up and short-sales constraints which lead to censoring in holdings when there are negative signals.

For investor i, the expected value of his portfolio weight $w_{i,j,k}$, given $y_{i,k}$, is then

$$\begin{aligned} \mathbb{E}[w_{i,j,k}|y_{i,k}, \boldsymbol{X}_j, \theta] &= \int w_{i,j,k} f(w_{i,j,k}|y_{i,k}, \boldsymbol{X}_j, \theta) dw_{i,j,k} \\ &= (\alpha+\beta \boldsymbol{X}_j)\, \Phi\left(\frac{\alpha+\beta \boldsymbol{X}_j}{(\lambda_0+\lambda_1 y_{i,k})} \right) \\ &\quad + (\lambda_0+\lambda_1 y_{i,k})\, \varphi\left(\frac{\alpha+\beta \boldsymbol{X}_j}{(\lambda_0+\lambda_1 y_{i,k})} \right) \end{aligned} \tag{16}$$

for which we leave the derivation to the Appendix.

Now we show two propositions that provide the basis for our later analysis.

Proposition 1. The expected portfolio weight in (16), $\mathbb{E}[w_{i,j,k}|y_{i,k}, \boldsymbol{X}_j, \theta]$, is monotonically increasing in the number of friends $y_{i,k}$.

Proposition 1 justifies our inference strategy whereby we can extract information about the count of friends in a city from the geographical dispersion of portfolio weights.

Proposition 2. The expected total excess return of investing in stock j headquartered in city k given $y_{i,k}$, where the expected total excess return is defined as

$$\mathbb{E}\Big[\big(\alpha+\beta \boldsymbol{X}_j + (\lambda_0+\lambda_1 y_{i,k})\xi_{i,j,k}\big) w_{i,j,k} - \frac{c}{2} w_{i,j,k}^2 \,|\, y_{i,k}, \boldsymbol{X}_j, \theta \Big] \tag{17}$$

and $w_{i,j,k}$ is the optimal portfolio weight given in (14), is monotonically increasing in the number of friends $y_{i,k}$.

Proposition 2 is the motivation for our portfolio return exercise that will be illustrated later. The proofs of both propositions are in the Appendix.

2.3. Maximum likelihood estimation with EM algorithm

2.3.1. Likelihood function

Let J_k denote the total number of stocks headquartered in city k. The number of stocks over all cities is denoted by J and $J = \sum_{k=1}^{K} J_k$. Then for each investor i in a city k, the (joint) conditional probability density of $\mathbf{w}_{i,k} = \{w_{i,j,k}\}_{j \in \mathcal{J}_k}$ (the vector of portfolio weights of investor i over all stocks headquartered in city k), given y_{ik}, is

$$f(\mathbf{w}_{i,k}|y_{i,k}, \{\boldsymbol{X}_j\}_{j\in\mathcal{J}_k}, \theta) = \prod_{j=1}^{J_k} f(w_{i,j,k}|y_{i,k}, \boldsymbol{X}_j, \theta) \tag{18}$$

since given $y_{i,k}$, the density of $w_{i,j,k}$, $f(w_{i,j,k}|y_{i,k}, \boldsymbol{X}_j, \theta)$ in (15) is i.i.d. across stocks j for each investor i in each city k. This conditional density $f(\mathbf{w}_{i,k}|y_{ik}, \{\boldsymbol{X}_j\}_{j\in\mathcal{J}_k}, \theta)$ will be used later in our estimation strategy. Intuitively, each investor i in each city k generates J_k observations of portfolio weights $w_{i,j,k}$. Thus, for an investor i in a city k, the likelihood will depend on all of the portfolio weights $\mathbf{w}_{i,k}$ in the city.

Given our latent random network model, the marginal probability density of friends $y_{i,k}$ is negative binomial given by $\rho(y_{i,k}|\theta)$ in Eq. (6). Therefore, the joint density of $(\mathbf{w}_{i,k}, y_{i,k})$ for investor i in city k is

$$\begin{aligned} & f(\mathbf{w}_{i,k}, y_{i,k}|\{\boldsymbol{X}_j\}_{j\in\mathcal{J}_k}, \theta) \\ & = f(\mathbf{w}_{i,k}|y_{i,k}, \{\boldsymbol{X}_j\}_{j\in\mathcal{J}_k}, \theta) \rho(y_{i,k}|\theta), \end{aligned} \tag{19}$$

where $f(\mathbf{w}_{i,k}|y_{i,k}, \{\boldsymbol{X}_j\}_{j\in\mathcal{J}_k}, \theta)$ is specified in Eq. (18).

The marginal density of $\mathbf{w}_{i,k} = \{w_{i,j,k}\}_{j\in\mathcal{J}_k}$ is

$$f(\mathbf{w}_{i,k}|\{\boldsymbol{X}_j\}_{j\in\mathcal{J}_k}, \theta) = \sum_{y_{i,k}=0}^{+\infty} f(\mathbf{w}_{i,k}, y_{i,k}|\{\boldsymbol{X}_j\}_{j\in\mathcal{J}_k}, \theta) \tag{20}$$

since $y_{i,k}$ has a negative binomial distribution that is a discrete distribution with its support on $\{0, 1, 2, \ldots, +\infty\}$. By Bayes' theorem, the conditional density of $y_{i,k}$, given $\mathbf{w}_{i,k}$, is therefore

$$\begin{aligned} f(y_{i,k}|\mathbf{w}_{i,k}, \{\boldsymbol{X}_j\}_{j\in\mathcal{J}_k}, \theta) &= \frac{f(\mathbf{w}_{i,k}, y_{i,k}|\{\boldsymbol{X}_j\}_{j\in\mathcal{J}_k}, \theta)}{f(\mathbf{w}_{i,k}|\{\boldsymbol{X}_j\}_{j\in\mathcal{J}_k}, \theta)} \\ &= \frac{f(\mathbf{w}_{i,k}, y_{i,k}|\{\boldsymbol{X}_j\}_{j\in\mathcal{J}_k}, \theta)}{\sum_{y_{i,k}=0}^{+\infty} f(\mathbf{w}_{i,k}, y_{i,k}|\{\boldsymbol{X}_j\}_{j\in\mathcal{J}_k}, \theta)}. \end{aligned} \tag{21}$$

Now we can define the complete data likelihood for $\{\{y_{i,k}\}, \{\mathbf{w}_{i,k}\}\}$ for all investors across all cities. Because $y_{i,k}$ and $\mathbf{w}_{i,k}$ are i.i.d. across all cities k and all investors i, the complete-data likelihood is:

$$\begin{aligned} & \mathcal{L}\big(\theta|\{y_{i,k}\}, \{\mathbf{w}_{i,k}\}, \{\boldsymbol{X}_j\}\big) \\ & = \prod_{i=1}^{N}\prod_{k=1}^{K} f(\mathbf{w}_{i,k}, y_{i,k}|\{\boldsymbol{X}_j\}_{j\in\mathcal{J}_k}, \theta) \\ & = \prod_{i=1}^{N}\prod_{k=1}^{K} \left(\prod_{j=1}^{J_k} f(w_{i,j,k}|y_{i,k}, \boldsymbol{X}_j, \theta) f(y_{i,k}|\theta) \right). \end{aligned} \tag{22}$$

In this baseline setup, we have in total $(N+K+K+4+2)$ model parameters in θ to estimate in a cross-section:

- parameters governing the negative binomial distribution of $y_{i,k}$ as shown in (6): N gregariousness parameters $\{a_i\}$, K city group size parameters $\{b_k\}$, and K overdispersion parameters $\{\omega_k\}$,
- parameters governing the censored normal distribution of $w_{i,j,k}$ as shown in (15): four parameters for the style loadings (α and $\boldsymbol{\beta}$), and two parameters λ_0 and λ_1 for the weight on the private information from the friends;

and we use $N \times J$ observations of stock-level portfolio weights of investors $\{w_{i,j,k}\}$ to do the estimation.

Since $y_{i,k}$ is unobservable, a direct application of maximum likelihood estimation on (22) is not feasible. To overcome this hurdle, we will use the Expectation-Maximization (EM) algorithm.

2.3.2. EM algorithm

The Expectation-Maximization (EM) algorithm is a two-step recursive estimation procedure to maximize the log-likelihood in the presence of latent variables $y_{i,k}$. Each iteration is guaranteed to increase the log-likelihood and the algorithm is guaranteed to converge to a local maximum of the log-likelihood function. There are many convergence papers in the literature, see, for example, Dempster et al. (1977) (who first introduced the EM algorithm) and Wu (1983). Details are available in the manual that accompanies our code online.

The first step is called the E-step (expectation step). We first find (evaluate) the *expected value* of the complete-data log-likelihood $\log\Big(\mathcal{L}\big(\theta|\{y_{i,k}\},\{\mathbf{w}_{i,k}\},\{\boldsymbol{X}_j\}\big)\Big)$ with respect to the latent, random $\{y_{ik}\}$, given the observable data $\big\{\mathbf{w}_{i,k}=\{w_{i,j,k}\}_{j\in J_k},\{\boldsymbol{X}_j\}\big\}$ and the current parameter estimates θ':[8]

$$Q(\theta,\theta')=\mathbb{E}_{\{y_{i,k}\}}\Big[\log\Big(\mathcal{L}\big(\theta|\{y_{i,k}\},\{\mathbf{w}_{i,k}\},\{\boldsymbol{X}_j\}\big)\Big)\,|\,\{\mathbf{w}_{i,k}\},\{\boldsymbol{X}_j\},\theta'\Big].$$

By integrating out the latent variable $y_{i,k}$ (in our case it is a summation instead of integration since the negative binomial is a discrete distribution for $y_{i,k}$) in (22), the expected value of the complete-data log-likelihood, $Q(\theta,\theta')$, is now a deterministic function of θ that can be maximized. Notice the meaning of the two arguments in the function $Q(\theta,\theta')$. The first argument θ corresponds to the parameters that ultimately will be optimized in an attempt to maximize the log-likelihood. The second argument θ' corresponds to the current parameter values that we use to evaluate the expectation.

We can write out this expected value $Q(\theta,\theta')$ in detail:

$$Q(\theta,\theta')=\mathbb{E}_{\{y_{i,k}\}}\Bigg[\log\Bigg(\prod_{i=1}^{N}\prod_{k=1}^{K}f(\mathbf{w}_{i,k},y_{i,k}|\{\boldsymbol{X}_j\}_{j\in\mathcal{J}_k},\theta)\Bigg)\,|\,\{\mathbf{w}_{i,k}\},\{\boldsymbol{X}_j\},\theta'\Bigg]$$

$$=\sum_{i=1}^{N}\sum_{k=1}^{K}\Bigg[\sum_{y_{i,k}=0}^{+\infty}\log\big(f(\mathbf{w}_{i,k},y_{i,k}|\{\boldsymbol{X}_j\}_{j\in\mathcal{J}_k},\theta)\big)f(y_{i,k}|\mathbf{w}_{i,k},\{\boldsymbol{X}_j\}_{j\in\mathcal{J}_k},\theta')\Bigg] \tag{23}$$

where the joint density $f(\mathbf{w}_{i,k},y_{ik}|\{\boldsymbol{X}_j\}_{j\in\mathcal{J}_k},\theta)$ is given in (19) and the conditional density of $y_{i,k}$ given $\mathbf{w}_{i,k}$, $f(y_{i,k}|\mathbf{w}_{i,k},\{\boldsymbol{X}_j\}_{j\in\mathcal{J}_k},\theta')$, is given in (21). The above follows from the definition of the expectation over $y_{i,k}$ (as the negative binomial distribution of $y_{i,k}$ is discrete).

The second step of the EM algorithm is called the M-step (maximization step). This amounts to maximizing the expectation $Q(\theta,\theta')$ we computed in the first E-step. That is, we find

$$\hat{\theta}=\arg\max_{\theta}Q(\theta,\theta').$$

The EM algorithm then repeats the E-step and the M-step recursively. That is, the EM algorithm

1. Starts with some initial parameter values θ'.
2. E-step: Computes the expected value of the complete-data log-likelihood, $Q(\theta,\theta')$.
3. M-step: Maximizes the expectation $Q(\theta,\theta')$ computed from the E-step over θ and obtain the estimate $\hat{\theta}$.
4. Iteration: Uses $\hat{\theta}$ as the new input value for θ' in the E-step, computes the new $Q(\theta,\theta')$, and maximizes it in the M-step to obtain a new estimate for θ.
5. Iterates until convergence.

2.4. Incorporating covariates

Because the estimation of the latent network involves portfolio decisions, we need to account for other non-network rationales that drive portfolio choice and which might affect our network inferences. To do this, we extend our baseline overdispersed Poisson model to allow for investor covariates.[9]

To be more precise, we will denote the investor-specific covariates by $mgrvar_i$, which could include investment style and other manager-fund attributes that vary over investors but do not vary over cities. For instance, these covariates would include fund styles and deal with confounds such as a growth-stock mutual fund will have larger portfolio weights in Silicon Valley than other areas irrespective of any network ties per se. We want to purge out of our estimates of latent networks these sorts of confounding factors.

Moreover, for our mutual fund application, we have two investor-city covariates that vary over both investors and cities. These will help address other network-related explanations in the literature. The first of these is the "know thy neighbor effect" $Neighbor_{i,k}$ a la Hong et al. (2005), which is the average of the portfolio weights allocated to city k across all funds that are headquartered in the same city as fund i but not in the same family. This earlier work found that managers located in the same city have correlated stock picks. So we want to separate our non-i.i.d.

[8] The meaning of the current parameter estimates is either an initial guess for the parameter values to start the EM algorithm, or the estimates from the previous recursive loop when the algorithm is running (as the algorithm is a recursive procedure). Details for the selection of initial estimates is in the manual accompanying our codes online.

[9] Using covariates in a Poisson-type count model also has a history in the literature of Industrial Organization to explain such as the patent-R&D relationship among individual firms (see, e.g., Hausman et al., 1984).

connections in different cities effect from this thy-neighbor effect. The second is the "same family effect," which is the average of the portfolio weights allocated to city k of all other funds in the same family as fund i. Funds in the same family have correlated positions as well due to sharing of buy-side analysts, just to name one rationale.

We integrate both of these types of covariates into our model in the following manner. We now let the number of friends $y_{i,k}$ follow a Poisson distribution with a mean equal to $a_i \exp(mgrvar_i' \psi) \exp(\kappa_1 Neighbor_{i,k} + \kappa_2 Family_{i,k}) b_k$, where ψ is the vector of coefficients on the investor-specific covariates *mgrvar*. $g_{i,k}$ will follow a gamma distribution with a mean equal to one and a variance equal to $(\omega_k - 1)$ where $\omega_k > 1$ as before. Consequently, such a Poisson-gamma mixture again leads to a (marginal) density for $y_{i,k}$ that is negative binomial (after an appropriate reparameterization):

$$\rho(y_{ik}|\theta, mgrvar_i, Neighbor_{i,k}, Family_{i,k}) = \frac{\Gamma(y_{ik} + \tilde{\zeta}_{ik})}{\Gamma(\tilde{\zeta}_{ik})\,\Gamma(y_{ik}+1)} \left(\frac{1}{\omega_k}\right)^{\tilde{\zeta}_{ik}} \left(\frac{\omega_k - 1}{\omega_k}\right)^{y_{ik}}, \tag{24}$$

where $\tilde{\zeta}_{ik} = \delta_i \exp(\kappa_1 Neighbor_{i,k} + \kappa_2 Family_{i,k}) b_k/(\omega_k - 1)$, and $\delta_i = a_i \exp(mgrvar_i' \psi)$. Thus y_{ik} has a mean equal to $\mu_{ik} = \delta_i \exp(\kappa_1 Neighbor_{i,k} + \kappa_2 Family_{i,k}) b_k$, and a variance equal to $\omega_k \delta_i \exp(\kappa_1 Neighbor_{i,k} + \kappa_2 Family_{i,k}) b_k = \omega_k \mu_{ik}$ that is greater than its mean $(\omega_k > 1)$. $\{\omega_k\}$ are the overdispersion parameters as before. In addition, because of the new term $\exp(mgrvar_i' \psi)$ involving covariates, the gregariousness parameters in this model have become $\{\delta_i = a_i \exp(mgrvar_i' \psi)\}$ instead of the $\{a_i\}$ we had before.[10]

Combining this new density of $y_{i,k}$ with covariates and the conditional density of portfolio weight $f(w_{i,j,k}|y_{i,k}, \boldsymbol{X}_j, \theta)$ in (15) as before, the complete-data likelihood now becomes:

$$\begin{aligned}
&\mathcal{L}\big(\theta|\{y_{i,k}\}, \{\mathbf{w}_{i,k}\}, \{\boldsymbol{X}_j\}, \{mgrvar_i\}, \{Neighbor_{i,k}\}, \{Family_{i,k}\}\big) \\
&= \prod_{i=1}^{N}\prod_{k=1}^{K} f\big(\mathbf{w}_{i,k}, y_{i,k}|\{\boldsymbol{X}_j\}_{j\in\mathcal{J}_k}, mgrvar_i, Neighbor_{i,k}, Family_{i,k}, \theta\big) \\
&= \prod_{i=1}^{N}\prod_{k=1}^{K} \left(\prod_{j=1}^{J_k} f(w_{i,j,k}|y_{i,k}, \boldsymbol{X}_j, \theta) \times \rho(y_{i,k}|mgrvar_i, Neighbor_{i,k}, \right. \\
&\qquad \left. Family_{i,k}, \theta) \right).
\end{aligned} \tag{25}$$

We call the above overdispersed Poisson model our extended model with covariates. This extended model will also be estimated via the EM algorithm and the associated results will be presented together with the results from our baseline model (i.e., the overdispersed Poisson model without covariates) later on.

2.4.1. Filtering for latent friends $y_{i,k}$

Once we have estimated the model parameters θ, we can apply the following filter to obtain an estimate of the number of friends y_{ik} for each investor i in each city k:

$$\begin{aligned}
\hat{y}_{i,k} &= \mathbb{E}\big(y_{i,k}\,|\,\mathbf{w}_{i,k}, \{\boldsymbol{X}_j\}_{j\in\mathcal{J}_k}, \hat{\theta}_{MLE}\big) \\
&= \sum_{y_{i,k}=0}^{+\infty} y_{i,k} f\big(y_{i,k}|\mathbf{w}_{i,k}, \{\boldsymbol{X}_j\}_{j\in\mathcal{J}_k}, \hat{\theta}_{MLE}\big)
\end{aligned} \tag{26}$$

where $\hat{\theta}_{MLE}$ denotes the maximum likelihood estimates of the model parameters θ obtained via our EM algorithm, and the density $f(y_{i,k}|\mathbf{w}_{i,k}, \{\boldsymbol{X}_j\}_{j\in\mathcal{J}_k}, \theta)$ is given in (21).

We will use these inferred latent connections later on in the paper to carry out our exercises of predicting actual alumni connections and predicting portfolio returns. That is, our methodology estimates the underlying parameters of the latent social network and portfolio choice problem given the data on portfolio weights of mutual fund managers. Using these parameters and the data on portfolio weights, we can predict (recover) the latent connections for the managers in different cities and then use these data on the predicted connections rather than the actual observations of these connections which are unavailable.

3. Data and estimation

3.1. Data

Our data on stock holdings of mutual funds are obtained from the CDA/Spectrum Mutual Fund Common Stock Holdings database provided by Thomson Reuters for the period of 1991–2015. The database sources from semi-annual mandatory filings and quarterly voluntary disclosures to the Securities and Exchange Commission (SEC) by mutual funds. We then merge the CDA/Spectrum database with the survivorship-bias free Center for Research in Security Prices (CRSP) mutual fund database. The CRSP mutual fund database provides information on a variety of mutual fund characteristics such as fund locations, investment objectives, monthly fund returns, and assets under management.

Since we are interested in understanding the demographics and networks of an individual manager for her decisions, we focus, following Hong and Kostovetsky (2012), on active domestic funds managed by a single manager as opposed to team-managed funds, where it is often difficult to establish the chain of command. We then augment our data with Hong and Kostovetsky (2012) data, which contain managerial demographic information on age, gender, name and location of undergraduate/graduate college, median SAT score of the undergraduate college attended, having a graduate degree or not, and political affiliation.[11]

In order to keep only actively managed, non-sector domestic equity funds in our sample, we apply the following detailed screening procedures. First, to exclude international, bond, and index funds, we require (1) funds' investment objective code reported by CDA/Spectrum to be aggressive growth, growth, or growth and income, (2) their

[10] The exponential mean parameterization of $\exp(mgrvar_i' \psi)$ and $\exp(\kappa_1 Neighbor_{i,k} + \kappa_2 Family_{i,k})$ for the covariates is standard in the literature of count regression models, see Chapter 20 of Cameron and Trivedi (2005).

[11] We also search on the Internet to fill in any information that is not found in the CRSP database or the data set of Hong and Kostovetsky (2012) whenever possible.

investment objectives in CRSP to be equity (E) and domestic (D) at the first two levels, (3) their CRSP objectives not to be EDCL, which indicates Standard & Poor's (S&P) 500 index fund, and (4) their names not to contain the word "index." Second, to exclude sector funds, we require funds' CRSP investment objectives at the third level to be either (C) or (Y). Third, to exclude the possible presence of hedge funds, we require funds' CRSP investment objectives not to be (H) or (S) at the last level. This screening leaves us with a sample of about 1,700 unique actively managed, non-sector domestic equity funds, or about 117,000 fund-quarter observations on stock holdings.

Next, we categorize the stocks held by mutual funds into city groups. We use the information on companies' headquartered cities that are available from the CRSP stock database. To obtain city groups for stocks, we match the city information of companies with the location information from Compustat, which maps cities into metropolitan statistical areas (MSAs).

We only consider the 20 largest cities (MSAs). The reason is because the 20 largest groups already cover approximately 80% of all the stocks held by mutual funds in our sample. There is no significant value added by allowing for more groups in our study.[12] Hence in what follows, the number of groups K is fixed at 20.

For stock characteristics, we obtain them from the CRSP–Compustat merged database. We construct the following variables: *SIZE* (size), defined as the log of the market capitalization of a stock; *BTM* (book-to-market), defined as the log of one plus book equity divided by market equity; *MOM* (momentum), defined as the past 12-month return.[13] We also obtain the quarterly return *Retq* for each stock that will be used later in our return analysis.

In Table 1, we report the summary statistics for the funds and fund managers in our sample. The mean portfolio weight on a stock held by a manager in a given city, $w_{i,j,k}$, is 1.33% with a rather large standard deviation of 1.51%. The median is 0.92%. The other fund characteristics are similar to the literature. Notice that the *FinCenter* dummy variable has a mean of 56.14%, so that roughly 60% of the funds are based in the six financial center cities.[14] The *ICI* variable has a mean of 14.85% and a standard deviation of 12.43%. The *logSAT* score has a mean of 7.11. Around 73% of the managers have an advanced degree and 11% are female. The mean age of the managers is about 49 and 25% are identified as Republicans. A similar fraction is identified as Democrats but the majority have no affiliation. Turning to stock characteristics, the mean *SIZE* is 6.2 with a standard deviation of 1.8 and the mean *BTM* is 0.45 with a standard deviation of 0.26. The average quarterly return of a stock is 2.42% with a large standard deviation of about 25%. The median is 1.24%.

3.2. Estimation

We estimate our model parameters θ through our EM algorithm for the baseline model and the extended model with covariates quarter by quarter using the mutual fund holdings data from 1991Q1 to 2015Q3. After obtaining the quarterly estimates, we will follow Fama and MacBeth (1973) in taking the time-series means of the quarterly estimates to form our overall estimates of θ. We denote these Fama-MacBeth estimates as our estimated parameter values.

4. Are managerial social connections randomly formed?

In this section, we report our main estimation results based on the mutual fund data with the 20 cities for the baseline model and the extended model with covariates. The manager covariates include fund characteristics, manager demographics, and dummy variables for fund styles, including being a mid-cap fund (*MEDIUM*), a growth fund (*GROWTH*), an income fund (*INCOME*), or a balanced fund (*BALANCED*). The excluded style is being a micro-cap fund.[15]

4.1. Gregariousness parameters

Table 2 shows the summary statistics of the estimated values of the gregariousness parameters $\{a_i\}$ and $\{\delta_i\}$ for our baseline and extended models, respectively. We observe that the mean of a_i is 64.0 in the baseline model, while that of δ_i is 72.5 in the extended model. These gregariousness parameter estimates can be interpreted literally as the typical manager having around 70 friends across the 20 cities in our sample. However, there is a fairly sizeable standard deviation of around 27 (baseline model) or 25 (extended model) friends.

In Table 3 we report the summary statistics of the estimates of the coefficients ψ on manager covariates and the controls $\{a_i\}$ in our extended model, where the gregariousness parameters $\delta_i = a_i \exp(mgrvar_i' \psi)$. We also report the coefficient estimates of κ_1 and κ_2 on the thy-neighbor (*Neighbor*) and same family (*Family*) effects, respectively. The inclusion of covariates helps to explain the variation in gregariousness across managers, comparing to the baseline model where gregariousness is measured by the constant a_i only.

Typical studies find that people have an expected number of connections in the low hundreds. But there is a large amount of uncertainty around these estimates (see, e.g., Zheng et al., 2006). We are only considering connections in the top 20 cities. Hence our estimates do not seem out

[12] In the Robustness section later, we repeat our analysis by adding the next ten cities into our sample so that we have the 30 largest groups. The results are similar.

[13] Following Brandt et al. (2009), we omit stocks with negative book-to-market ratio and take logs of market capitalization and book-to-market ratio to reduce the effect of outliers. We also follow standard practice in the return literature by winsorizing the two return variables *MOM* and *Retq* at the 1st and the 99th percentiles to reduce the impact of outliers.

[14] There are six financial centers in total following Christoffersen and Sarkissian (2009), which include Boston, Chicago, Los Angeles, New York, Philadelphia, and San Francisco.

[15] The style of a fund, i.e., small-cap, mid-cap, growth, income, or balanced, is from the CRSP fund objective code information. Here the benchmark is small-cap and micro-cap (the style that we do not include as a dummy variable), and using the CRSP fund objective code, we exclude index funds.

Table 1
Summary statistics of fund characteristics and fund manager demographics. This table reports the summary statistics for the funds and the fund managers in our sample. Panel A shows the summary statistics of the portfolio weights on stocks and the predicted number of connections of the funds. $w_{i,j,k}$ is a fund's portfolio weight on a stock j headquartered in a city k. $\hat{y}_{ik}$ is the predicted number of contacts a fund has in a city from our baseline model. Panel B presents the summary statistics of the characteristics of the funds. *logTNA, logFamSize, ExpRatio, Turnover*, and *FundAge* denote, respectively, the log of the total net asset, the log of one plus the total net asset of other funds in the family, the expense ratio, the turnover ratio, and the number of years since the establishment of a fund. *FinCenter* is a dummy variable that equals one if a fund is located in either Boston, Chicago, Los Angeles, New York, Philadelphia, or San Francisco. *ICI* denotes the Industry Concentration Index of a fund based on the raw industry weights of a manager's portfolio. Panel C shows the summary statistics of the demographics of the fund managers. *logSAT* is the log of the median SAT score (in 2005) of the undergraduate school that a fund manager attended. *Adv* and *Female* are dummy variables that equal one if a fund manager holds a graduate degree and if a manager is a female, respectively. *Age* is the age of a fund manager, and *Rep* is a dummy variable that equals one if a manager is Republican-affiliated. Panel D reports the characteristics of the stocks held by the funds. *SIZE* is the log of the market capitalization of a stock, *BTM* is the log of one plus the book-to-market ratio, *MOM* is the past 12-month return, and *Retq* is the quarterly return. The sample period is from 1991 to 2015.

Panel A: Portfolio weights and connections

	Mean	S.D.	Median	p10	p90
$w_{i,j,k}$ (%)	1.33	1.51	0.92	0.11	3.02
$\hat{y}_{ik}$	3.29	4.34	2.03	0.00	7.06

Panel B: Fund characteristics

	Mean	S.D.	Median	p10	p90
logTNA ($ million)	5.59	1.90	5.63	3.13	7.97
logFamSize ($ million)	7.12	3.41	7.88	0.00	10.81
ExpRatio (%)	1.31	1.58	1.19	0.80	1.90
Turnover (%)	87.76	95.10	67.00	18.00	175.00
FundAge (years)	17.50	15.41	12.99	3.75	38.75
FinCenter (%)	56.14				
ICI (%)	14.85	12.43	13.73	8.38	19.64

Panel C: Manager demographics

	Mean	S.D.	Median	p10	p90
logSAT	7.11	0.14	7.12	6.94	7.27
Adv (%)	73.49				
Female (%)	10.82				
Age (years)	49.47	10.37	48.00	37.00	64.00
Rep (%)	24.91				

Panel D: Stock characteristics

	Mean	S.D.	Median	p10	p90
SIZE	6.22	1.79	6.11	4.02	8.57
BTM	0.45	0.26	0.41	0.16	0.76
MOM	0.13	0.61	0.04	-0.45	0.73
Retq (%)	2.42	24.86	1.24	-26.09	30.31

of bounds relative to results in the network literature. Nevertheless, we view the estimates of the gregariousness parameters as more akin to investor fixed effects. They are separate from and do not affect our inference on whether investors belong to a network. In other words, having a lot of friends is not the same as being part of a network.

4.2. City group size parameters

We then report the parameter estimates for b_k that gauge the relative sizes of cities. Table 4 demonstrates the values of b_k for the 20 cities in our baseline model. The summary statistics of the b_k estimates in the extended model are shown in Table 4 as well. Two aspects of the estimates are noticeable. First, there are a few cities such as New York that have a much larger number of potential social connections attached to them compared to the rest. The b_k estimates are correlated with city population. Second, they are largely unaffected by fund covariates since $\{b_k\}$ are similar to city sizes and incorporating covariates for individual managers has minimal impact on those estimates on city sizes.

Table 2
Summary statistics of estimates of gregariousness parameters for mutual funds. The table shows the summary statistics of the estimated values of the gregariousness parameters $\{a_i\}$ in the baseline model (Baseline) and $\{\delta_i = a_i \exp(mgrvar_i' \psi)\}$ in the extended model with manager covariates (With covariates), using our mutual fund data with 20 cities. We first compute the time-series average of the quarterly gregariousness estimates for each fund, then we report the summary statistics of these time-series averages.

	Mean	S.D.	Median	p10	p90
Baseline	63.96	27.23	69.58	22.76	82.76
With covariates	72.53	25.45	67.19	24.15	102.62

Table 3
Summary statistics of estimates of coefficients on fund manager covariates. This table reports the summary statistics of the quarterly estimates of ψ, the coefficients on manager covariates *mgrvar* in the gregariousness parameter specification $\delta_i = a_i \exp(mgrvar_i' \psi)$ in the extended model. The manager covariates include dummy variables for whether a fund is a mid-cap fund (*Medium*), a growth fund (*Growth*), an income fund (*Income*), or a balanced fund (*Balanced*). The benchmark is being a small- or micro-cap fund, i.e., the dummy variable we do not include. The covariates also include all of the fund characteristic and manager demographic variables. Specifically, *logTNA, logFamSize, ExpRatio, Turnover*, and *FundAge* denote, respectively, the log of the total net asset, the log of one plus the total net asset of other funds in the family, the expense ratio, the turnover ratio, and the number of years since the establishment of a fund. *FinCenter* is a dummy variable that equals one if a fund is located in a financial center. *ICI* is the Industry Concentration Index of a fund. *logSAT* is the log of the median SAT score (in 2005) of the undergraduate school that a fund manager attended. *Adv* and *Female* are dummy variables that equal one if a fund manager holds a graduate degree and if a manager is a female, respectively. *Age* is the age of a fund manager. *Rep* is a dummy variable that equals one if a manager is Republican-affiliated. *Neighbor* for a fund is the average of the portfolio weights allocated to city k of all other funds that are headquartered in the same city as the fund but not in the same family (the "know-thy-neighbor effect"). *Family* for a fund is the average of the portfolio weights allocated to city k of all other funds in the same family of the fund (the "same family effect").

	Mean	S.D.	Median
a_i	23.72	8.84	24.57
MEDIUM	-0.029	0.279	-0.062
BALANCED	0.078	0.174	0.033
GROWTH	0.079	0.501	0.025
INCOME	0.049	0.182	0.032
logTNA	-0.008	0.040	0.003
logFamSize	-0.008	0.029	0.000
ExpRatio	0.054	0.079	0.030
Turnover	-0.002	0.001	-0.001
FundAge	0.004	0.009	0.005
FinCenter	-0.015	0.070	0.006
ICI	0.037	0.023	0.030
SAT	0.053	0.075	0.023
Adv	0.005	0.045	0.015
Female	0.031	0.212	0.014
Age	0.005	0.011	0.010
Rep	0.002	0.118	0.016
Neighbor	0.005	0.009	0.006
Family	0.003	0.006	0.002

4.3. Overdispersion parameters and rejecting the null model

Now we turn to the estimates of our main parameter of interest—the degree of overdispersion ω_k among different cities. Recall that we introduced the overdispersions in our model in an attempt to estimate the variability in investors' relative propensities to form ties to members of different groups.

Table 4
Summary statistics of estimates of prevalence (group size) parameters b_k for 20 cities. This table shows the summary statistics of the quarterly estimates of the relative group size parameters $\{b_k\}$ for the 20 cities. The full names for the city abbreviations are as follows. NY: New York, LA: Los Angeles, Bos: Boston, SF: San Francisco, Chi: Chicago, SJ: San Jose, Dal: Dallas, Hou: Houston, Phi: Philadelphia, Was: Washington, Mia: Miami, Atl: Atlanta, Min: Minnesota, Den: Denver, SD: San Diego, Stfd: Stamford, Sea: Seattle, Phx: Phoenix, SL: St. Louis, Det: Detroit.

	Baseline			With covariates		
	Mean	S.D.	Median	Mean	S.D.	Median
NY	0.107	0.051	0.101	0.120	0.044	0.132
LA	0.057	0.034	0.057	0.065	0.025	0.074
Bos	0.054	0.023	0.053	0.067	0.025	0.070
SF	0.049	0.021	0.050	0.050	0.021	0.053
Chi	0.058	0.033	0.052	0.064	0.026	0.067
SJ	0.053	0.025	0.051	0.057	0.030	0.060
Dal	0.051	0.023	0.055	0.053	0.018	0.056
Hou	0.052	0.024	0.051	0.058	0.045	0.055
Phi	0.050	0.052	0.036	0.049	0.021	0.052
Was	0.048	0.034	0.035	0.042	0.017	0.042
Mia	0.045	0.023	0.047	0.039	0.043	0.029
Atl	0.039	0.018	0.038	0.040	0.022	0.036
Min	0.042	0.020	0.042	0.044	0.023	0.041
Den	0.043	0.034	0.039	0.035	0.025	0.030
SD	0.043	0.021	0.042	0.030	0.021	0.029
Stfd	0.042	0.027	0.044	0.049	0.048	0.029
Sea	0.046	0.036	0.041	0.036	0.031	0.026
Phx	0.047	0.038	0.037	0.037	0.038	0.023
SL	0.041	0.036	0.035	0.033	0.027	0.025
Det	0.036	0.025	0.039	0.033	0.040	0.025

Table 5 and the left panel of Fig. 2 display the estimated overdispersion parameters $\{\omega_k\}$ for the cities in our baseline model. There are two features. First, every city displays a statistically significant overdispersion parameter ω that is greater than one. The t-statistics of testing the null Poisson distribution of $\omega = 1$ are all significant at the 1% level. This indicates that our null hypothesis (model) of randomly formed managerial social networks is rejected.

Second, cities being larger (in terms of b_k) does not necessarily imply cities being more overdispersed. The correlation between the Fama-MacBeth estimates of ω_k and those of b_k is about 0.48, and the Spearman rank correlation between them is merely about 0.31. New York, Los Angeles, San Jose, Miami, and San Diego stand out as the most overdispersed cities compared to the rest. Our paper is concerned with the inference strategy for a given city but it would be interesting to try to relate these cross-city variations in overdispersion to sociological or demographic factors in future research.

Additionally, we depict the overdispersion estimates for our extended model in Table 5 and the right panel of Fig. 2. These sets of estimates are similar to those from the baseline model. Involving covariates allows the extended model to better explain the variations in the propensities of managers to form networks among different cities. Thus, the means and the standard deviations of the overdispersion estimates from the extended model have decreased across all cities in general compared to those from the baseline model. The reason is that the covariates soak up

Table 5
Summary statistics of estimates of overdispersion parameters ω_k for 20 cities. The table shows the summary statistics of the quarterly estimates of the overdispersion parameters $\{\omega_k\}$ for the 20 cities. Baseline stands for our baseline model, and With covariates stands for the extension on the baseline model where we allow for manager covariates. The t-statistics are Newey–West adjusted. They test the null hypothesis of $\omega_k = 1$ (Poisson) against the alternative of $\omega_k > 1$ (overdispersion). For an explanation of the abbreviated city names, please refer to Table 4.

	Baseline				With covariates			
	Mean	S.D.	Median	t-stat	Mean	S.D.	Median	t-stat
NY	2.301	0.945	2.223	15.01	1.973	0.680	1.963	12.19
LA	2.072	0.956	1.999	12.03	1.695	0.461	1.838	10.19
Bos	1.836	0.888	1.774	8.80	1.556	0.523	1.740	8.04
SF	1.932	0.876	1.917	10.20	1.643	0.465	1.822	9.09
Chi	1.811	0.927	1.766	10.91	1.646	0.541	1.771	9.84
SJ	2.089	0.894	2.160	10.07	1.868	0.557	2.008	8.29
Dal	1.672	0.922	1.510	12.52	1.465	0.506	1.391	8.72
Hou	1.834	0.899	1.725	12.34	1.560	0.559	1.606	9.17
Phi	1.630	0.869	1.551	13.93	1.470	0.563	1.407	12.18
Was	1.576	0.716	1.538	8.90	1.405	0.457	1.423	10.25
Mia	2.248	0.821	2.046	9.81	1.816	0.521	1.892	14.40
Atl	1.810	0.941	1.634	7.41	1.652	0.402	1.737	8.98
Min	1.774	0.795	1.677	12.78	1.687	0.492	1.623	8.49
Den	1.892	0.694	1.803	10.81	1.612	0.460	1.687	15.46
SD	2.159	0.786	2.055	12.98	1.807	0.356	1.959	19.92
Stfd	1.633	0.718	1.647	11.19	1.578	0.463	1.571	10.16
Sea	1.615	0.727	1.640	10.61	1.635	0.495	1.794	21.56
Phx	1.665	0.886	1.477	8.38	1.564	0.554	1.718	11.65
SL	1.783	0.812	1.592	20.96	1.632	0.370	1.770	10.18
Det	1.784	0.814	1.666	19.26	1.639	0.560	1.778	15.09

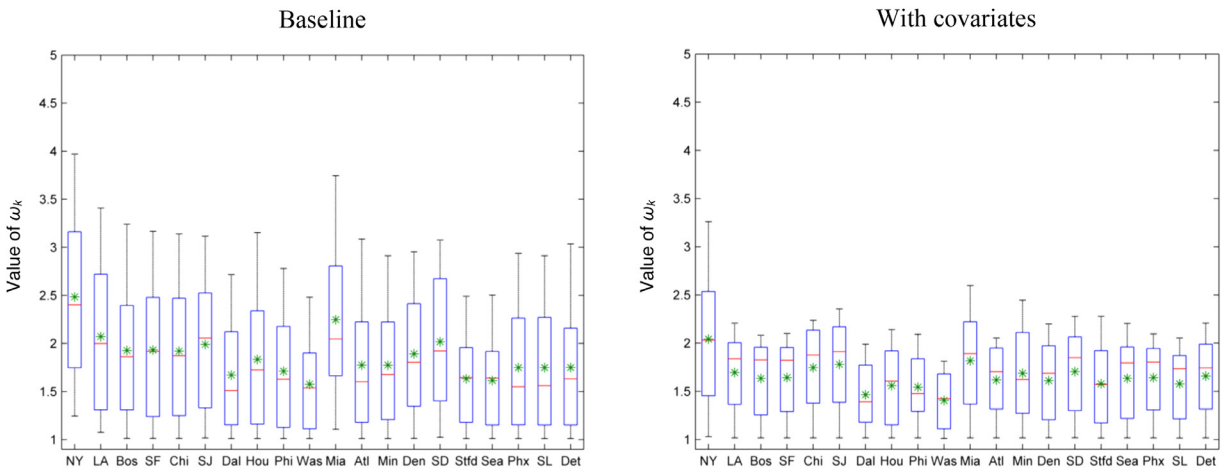


Fig. 2. Boxplot of ω_k estimates of 20 cities. The figure gives the boxplot of the quarterly estimates of the overdispersion parameters $\{\omega_k\}$ for the 20 cities from our baseline model on the left panel and from our extended model with covariates on the right panel. The * marker is the mean, the dash marker inside the box is the median, the box is the interquartile range, and the tails extend to the min and the max. For an explanation of the abbreviated city names, please refer to Table 4.

some of the variations in the propensities to form social ties from the overdispersion parameters. But the reduction is small and the overdispersion parameters are all still economically and statistically significant.

We can compare the degree of overdispersion of our money-manager networks to those studied in Zheng et al. (2006). This study considers many subgroups from people named Nicole (which have an ω close to one) to postal workers (which have an ω close to two) to prisoners (which have an ω close to four). The inferred overdispersion in network links for mutual fund managers to different cities, which is around two, is comparable in economic magnitude to those of postal workers, somewhere in the moderate range of non-i.i.d.-ness compared to other types of connections.

4.4. Controlling for local bias

One type of confounding explanation which our covariate approach cannot account for is the local bias of the stockholding of mutual fund managers (Coval and Moskowitz, 1999), i.e., fund manager portfolios are slightly

Table 6
Robustness checks on estimates of ω_k of 20 cities. This table shows the results of two robustness checks on the overdispersion estimates for the 20 cities. Baseline stands for our baseline model, and With covariates stands for the extended model with manager covariates. "NY&Bos funds" stands for the case where we only consider funds located in New York or Boston, and drop their picks of stocks that are headquartered in New York or Boston from the estimation. "No local response" denotes the case where managers' local holdings have been dropped from the estimation. The t-statistics are Newey–West adjusted. They test the null hypothesis of $\omega_k = 1$ (Poisson) against the alternative of $\omega_k > 1$ (overdispersion). For an explanation of the abbreviated city names, please refer to Table 4.

	Baseline				With covariates			
	NY&Bos funds		No local response		NY&Bos funds		No local response	
	Mean	t-stat	Mean	t-stat	Mean	t-stat	Mean	t-stat
NY	–	–	3.71	22.58	–	–	2.68	17.12
LA	2.42	11.15	2.41	23.71	1.75	8.67	1.56	19.29
Bos	–	–	2.55	8.93	–	–	1.75	9.72
SF	2.51	11.52	2.37	14.87	1.82	8.18	1.74	8.45
Chi	2.38	12.62	2.37	11.32	1.79	13.06	1.86	8.11
SJ	2.61	11.33	2.61	9.08	1.80	9.85	1.63	9.92
Dal	2.29	23.60	2.41	24.31	1.97	16.08	1.77	14.08
Hou	2.30	15.30	2.37	26.60	1.66	14.08	1.64	18.38
Phi	2.03	14.95	2.22	23.78	1.45	12.86	1.72	27.38
Was	2.30	11.37	2.21	10.13	1.51	12.67	1.61	9.73
Mia	2.52	10.72	2.82	12.18	1.40	16.14	1.66	21.85
Atl	2.41	7.87	2.54	7.36	2.16	10.03	2.08	10.89
Min	2.53	14.82	2.45	12.66	2.05	11.36	1.81	6.86
Den	2.78	9.52	3.00	13.39	2.16	18.64	2.12	15.30
SD	3.00	16.71	2.87	16.76	2.12	19.42	1.57	26.35
Stfd	2.15	12.33	1.97	13.14	1.61	11.77	1.60	12.71
Sea	2.17	9.12	2.18	10.09	2.03	27.31	1.83	22.60
Phx	2.40	9.15	2.25	11.91	1.96	13.89	1.64	18.78
SL	2.19	23.57	2.22	20.20	1.45	10.61	1.56	11.28
Det	2.32	22.69	2.24	24.94	1.84	16.02	1.65	22.95

tilted to stocks headquartered near where they reside. This might be driven by proximity effects independent of networks per se. To address this concern, we first re-estimate our model using just managers based in New York City and Boston and model their portfolio weights in other cities in the same way. Managers in these cities control a substantial amount of assets and typically invest all over the country. Second, we exclude managers' local stock holdings from the estimation as another way to ensure that our overdispersion results are not due to local biases.

As can be seen clearly from Table 6, the results from the two robustness checks echo our earlier findings in Table 5 for both the baseline and the extended models. They suggest that the overdispersions we document are not significantly subject to the influence of either local biases, and once more the social connections of managers are not formed in an i.i.d. manner.

4.5. Portfolio choice parameters

Along with the network parameters, we also obtain portfolio choice parameters in the process. Tables 7 and 8 report the estimates for our portfolio choice parameters. Recall that $(\alpha + \boldsymbol{\beta X}_j)$ represents the part of the expected stock return coming from the common prior (where $\boldsymbol{X}_j$ are the stock characteristics), and $\lambda_{i,k}\xi_{i,j,k} = (\lambda_0 + \lambda_1 y_{i,k})\xi_{i,j,k}$ the part from an investor's private signal (friends). They constitute the optimal portfolio weight condition shown in Eq. (14).

Table 7 shows the parameters of factor loadings (β_{size}, β_{btm}, β_{mom}) on stock characteristics that include size (log of the market cap), book-to-market (log of one plus the book-to-market ratio), and momentum (past 12-month return), respectively. We can see that in both the baseline and the extended models, the Fama–MacBeth estimate of the loading parameter on size (β_{size}) is positive, while the loading parameters on book-to-market and momentum (β_{btm} and β_{mom}) are negative. This says on average, the funds in our sample prefer large-cap stocks, prefer growth stocks, and are not momentum traders. These results are similar to what is found in the literature on mutual fund portfolios that is discussed in the Introduction.

Table 8 shows the parameters of λ_0 and λ_1 representing the weight investors put on the private signal. The Fama–MacBeth estimate of λ_0 is about 0.1, while that of λ_1 is about 0.02. To interpret these estimates, we can take the mean value of estimated $y_{i,k}$ (shown in Table 1 of summary statistics), which is about 3.0. Then on average, $\lambda_{i,k} = \lambda_0 + \lambda_1 y_{i,k} = 0.16$. Thus the loading of a fund manager on his private information (the $\xi_{i,j,k}$ term coming from the manager's private signal) is 0.16.

We can compare the economic effect of this private signal on portfolio weights relative to previously studied stock characteristics in X_j. From the optimal portfolio weight in Eq. (14), when $w_{i,j,k}$ is positive, a one standard deviation increase in one of the characteristics, say *SIZE*, would increase $w_{i,j,k}$ by $0.011 \times 1.79 = 0.02$ (the estimated $\hat{\beta}_{SIZE} = 0.011$ and the standard deviation of *SIZE* is 1.79). In contrast, a one standard deviation increase in the private signal term $\xi_{i,j,k}$ (which is equal to one as it is a standard normal random variable), for an investor i having an estimated $\hat{y}_{i,k}$ equal to the average (which is

Table 7
Summary statistics of estimates of factor loadings for portfolio weights. This table reports the summary statistics of the factor loadings parameters $\beta = (\beta_{\text{size}}, \beta_{\text{btm}}, \beta_{\text{mom}})$ on stock characteristics and the constant α in the optimal portfolio choice condition of Eq. (14) where $\alpha + \beta X_j$ is the part coming from the prior and X_j are the stock characteristics. The characteristics include size (log of the market cap), book-to-market (log of one plus the book-to-market ratio), and momentum (past 12-month return), respectively.

	Baseline			With covariates		
	Mean	S.D.	Median	Mean	S.D.	Median
β_{size} (size)	0.011	0.022	0.013	0.004	0.023	−0.003
β_{btm} (book-to-market)	−0.084	0.123	−0.036	−0.078	0.185	−0.012
β_{mom} (momentum)	−0.005	0.022	−0.002	−0.009	0.030	−0.004
α (constant)	−0.233	0.218	−0.187	−0.162	0.258	−0.053

Table 8
Summary statistics of estimates of λ_0 and λ_1 for the weight $\lambda_{i,k} = \lambda_0 + \lambda_1 y_{i,k}$ on the private signal. This table reports the summary statistics of the weight parameters λ_0 and λ_1 in the optimal portfolio choice condition of Eq. (14), where $\lambda_{i,k} = \lambda_0 + \lambda_1 y_{i,k}$ is the weight on the private signal that a fund manager receives in city k and $y_{i,k}$ is the number of (latent) friends the manager has in that city.

	Baseline			With covariates		
	Mean	S.D.	Median	Mean	S.D.	Median
λ_0	0.097	0.058	0.081	0.103	0.067	0.071
λ_1	0.020	0.013	0.013	0.017	0.010	0.014

3.29), would increase $w_{i,j,k}$ by 0.16. That is, the effect from the private signal mediated by the connections has a larger economic effect on portfolio weights than previously studied fund characteristics. It is exactly to this novel link between connections and portfolio weights that we now study in more depth.

5. Using relative propensity to connect to predict university alumni networks and portfolio returns

Given that our parameter estimates of the latent network point to significant overdispersion in connections to cities, this means that some managers in the population have non-i.i.d. ties to a given city. We would ideally like to identify these managers and see if our inferred network connections can actually predict actual fund manager social networks and presumably also fund performance.

5.1. Measures of relative propensity to connect

To this end, recall that in our model, investors' *expected* relative propensities to know a member in city k is $g_{ik} = \mu_{ik}/(a_i b_k)$ (or $g_{ik} = \mu_{ik}/(\delta_i \exp(\kappa_1 Neighbor_{i,k} + \kappa_2 Family_{i,k}) b_k)$ in the extended model). We can use our model's estimates, $\hat{a}_i$, $\hat{b}_k$, $\hat{\delta}_i$, $\hat{\kappa}_1$, $\hat{\kappa}_2$, and $\hat{y}_{ik}$, to generate for each manager his relative propensity to be connected (RPC or $\eta_{i,k}$) in a non-i.i.d. way to acquaintances in different cities, where

$$\eta_{i,k} = \hat{y}_{ik}/(\hat{a}_i \hat{b}_k) \tag{27}$$

for the baseline model and

$$\eta_{i,k} = \hat{y}_{ik}/(\hat{\delta}_i \exp(\hat{\kappa}_1 Neighbor_{i,k} + \hat{\kappa}_2 Family_{i,k}) \hat{b}_k) \tag{28}$$

Table 9
Summary statistics of RPC measure $\eta_{i,k}$ for mutual funds. This table shows the summary statistics of our managerial RPC measure $\eta_{i,k}$ of mutual funds over all quarters. Baseline denotes our baseline model, and With covariates stands for the extended model with manager covariates. $\log(\eta_{i,k})$ is the log of one plus the value of $\eta_{i,k}$.

Baseline	Mean	S.D.	Median
$\eta_{i,k}$	1.03	0.88	0.81
$\log(\eta_{i,k})$	0.60	0.41	0.59
With covariates	**Mean**	**S.D.**	**Median**
$\eta_{i,k}$	1.05	0.85	0.81
$\log(\eta_{i,k})$	0.64	0.38	0.59

for the extended model. In other words, the $\eta_{i,k}$ manager-city RPC measures can be thought of as investors' *realized* relative propensities to know a member from a specific city.

Our baseline model (resp. extended model) predicts that an investor should have an expected number of $a_i b_k$ (resp. $\delta_i \exp(\kappa_1 Neighbor_{i,k} + \kappa_2 Family_{i,k}) b_k$) connections in a given city, and that y_{ik} should be very close to $a_i b_k$ (resp. $\delta_i \exp(\kappa_1 Neighbor_{i,k} + \kappa_2 Family_{i,k}) b_k$) if connections are formed in an Erdös and Rényi (1959) i.i.d. manner. On the other hand, an investor who places a (much) larger portfolio weight on stocks and hence on average knows a (much) larger number of acquaintances than expected in a city is more likely to be part of and has $\eta_{i,k} > 1$ in that city, i.e., being part of that network.[16]

The summary statistics of our RPC measures by manager-city pair $\eta_{i,k}$ for both models are shown in Table 9. We notice that the mean of these measures is around 1.0 for the two models. The standard deviation is sizeable for both models, 0.88 in the baseline and 0.85 in the extended model with covariates. These imputed values are consistent with our underlying latent network model where most managers are presumed to have i.i.d. connections to a given city but a subset of managers might have non-i.i.d. ones. In addition, we also report the estimates of $\log(\eta_{i,k})$ (the log of one plus the value of $\eta_{i,k}$) that will be used later. In both models, $\log(\eta_{i,k})$ has a mean about 0.6 and a standard deviation about 0.4.

[16] This is because of our Proposition 1, i.e., the expected portfolio weight is monotonically increasing in the number of friends.

5.2. University alumni networks

We are interested in whether our model's relative propensity to connect measure $\eta_{i,k}$ is informative of managers' non-i.i.d. connections in any given city. One important source of non-i.i.d. connections is alumni networks. To construct these alumni connections, we follow Cohen et al. (2008) and construct mutual fund manager connections to members of the board of directors of different companies. We first use the RiskMetrics database to acquire information on boards of directors of US companies each year. The RiskMetrics database contains the following director information: name, age, employment title, and primary employer's name.[17] Companies in the database are identified by CUSIP number, and we focus on the period of 2000–2015. The coverage universe of RiskMetrics is the S&P 1500 companies.

We define social networks between fund managers and senior officials of firms [Chairman, Chief Executive Officer (CEO), and Chief Financial Officer (CFO)] using similar educational institutions. RiskMetrics does not have educational information on company directors, thus we use the S&P Capital IQ People Intelligence database to extract the educational backgrounds of directors. The Capital IQ database has college and graduate education (name of undergraduate/graduate school) as well as degree information (i.e., Bachelors, Masters, JD, MBA, PhD, and MD).[18] We match this information to our RiskMetrics data set using the names of the directors. Furthermore, we use director ages from RiskMetrics to extrapolate the years of graduation with various degrees that we cannot obtain from Capital IQ. To do this, we assume a person obtains an undergraduate degree at 22, a masters at 24, a JD at 25, a PhD at 27, an MBA at 29, an MD at 30. If someone has multiple graduate degrees, we assume that the MBA comes after any other degree and we add a five-year working gap between the first graduate degree and the MBA. Furthermore, we assume that a PhD would come five years after a JD or an MD, and an MD would come eight years after a JD.

We are now able to match educational institution and degree information from our fund manager database [with information on manager educational background provided by Hong and Kostovetsky (2012)] to that from our Capital IQ-RiskMetrics data set constructed above on company senior officials (Chairman, CEO, and CFO).

We specify two measures of network connections between fund managers and company officials, one based on whether a portfolio manager and a senior official of a firm (Chairman, CEO, or CFO) attended the same school and received the same degree (School-Degree), and the other on whether they have graduated within a year from the same school and received the same degree (School-Degree-Year).[19] The second connection measure is thus a stronger type of link compared to the first measure. Therefore, in our data set of managerial networks, we have two metrics on whether a fund manager is connected to a senior official of a company headquartered in different cities. The two metrics will give rise to two different sets of linkages for fund managers, and we call these our alumni connections data.

The summary statistics on the alumni connections are shown in Table 10. Panel A shows the summary statistics of the top five universities between 2000 and 2015 in our sample, which have the highest number of alumni presence in terms of firms (i.e., top company officials) and fund managers. Harvard, UPenn, and Columbia are in the top five of both lists, with Harvard having significantly more alumni at firms and funds than any other university.

Panel B shows the distribution of degrees and graduation years among company senior officials (Chairman, CEO, CFO) and fund managers in our sample between 2000 and 2015. The percentage numbers are averaged over the years. Almost all company senior officials and fund managers have an undergraduate degree. Thus the undergraduate degree will be the pivotal degree which makes connections between firm officials and fund managers.

With our alumni connections data, we can proceed to examine the out-of-sample predictive power of our model on these (non-random) alumni linkages. To be more specific, in any period, let the alumni connection number $Alumni_{ik}$ be the number of unique company officials from a city k a manager i is connected to. The summary statistics for this variable are given in Table 11. Using School-Degree as the measure of non-i.i.d. connections for any given city, the mean is 1.38, the 25th percentile is zero, the 75th percentile is 1.25, and the max is 32.6 friends. In other words, there is a significant right tail in the non-i.i.d. connections of managers to cities, consistent with the premise of our analysis that certain managers have significant connections in a city. The numbers for School-Degree-Year have a mean of 0.59 and a max of 16.6.

We would like to see whether our model's RPC measure $\eta_{i,k}$ can forecast the number of non-random social connections $Alumni_{ik}$ for managers in different cities. To this end, we employ the following regression specification:

$$\log(Alumni_{ik}) = \alpha + \gamma \log(\eta_{i,k}) + x_i'\beta + v_k'\rho + \epsilon_i, \qquad (29)$$

where x_i is a vector of controls for manager characteristics (including log of the median SAT score of the manager's undergraduate school, graduate degree holding status, gender (being a female), age, and political affiliation (being Republican-affiliated)) and v_k is a vector of controls for city characteristics (including income, population, and demographic attributes such as race, sex, and age).

The RPC measure $\eta_{i,k}$ of our model has forecasting power on the number of observable linkages $Alumni_{ik}$ of

[17] There is a slight issue with the Chairman information in RiskMetrics. When RiskMetrics recorded a company director's employment title as Chairman, the director might not be the Chairman of the actual company in question, since the director could be employed as the independent Chairman for another different company. However, this issue can be easily resolved by looking at the primary employer's name of the director. We will know that the director is not the Chairman of the company in question if the director's primary employer is another different company. We also double-check this by cross-referencing with the board of directors information from the other company whenever such an issue arises.

[18] Company senior officials or fund managers without a university or an advanced degree are dropped from the sample.

[19] For the School-Degree-Year, two people are connected if they are within one year, plus or minus one, of each other.

Table 10
Descriptive statistics on alumni data. This table reports the descriptive statistics of our alumni data constructed using RiskMetrics and Capital IQ databases. Panel A shows the statistics of the top five universities between 2000 and 2015 in our sample, ranked by the average number of alumni presence in firms and funds. A firm has an alumnus from a university if a senior official (Chairman, CEO, or CFO) of the firm received any degree from the university. A fund has an alumnus from a university if the fund manager received any degree from the university. Panel B shows the distribution of degrees and graduation years among company senior officials (Chariman, CEO, CFO) and fund managers in our sample between 2000 and 2015. The percentage numbers are averaged over the years.

Panel A: Universities with the highest number of alumni presence in firms and funds

University	Ave no. of firms	Ave % of all firms	University	Ave no. of fund managers	Ave % of all fund managers
Harvard	113	10.6%	Harvard	75	12.2%
Stanford	48	4.4%	UPenn	59	9.5%
UPenn	37	3.5%	Columbia	36	5.9%
MIT	30	2.8%	NYU	34	5.7%
Columbia	25	2.4%	UChicago	32	5.2%

Panel B: Degree and graduation year distributions

Degree	Company senior official	Fund manager	Graduation year	Company senior official	Fund manager
Undergrad	84.9%	96.8%	<1960	10.2%	2.1%
Masters	7.5%	8.9%	1960–1969	34.1%	19.2%
MBA	16.5%	61.3%	1970–1979	40.8%	20.4%
PhD	3.7%	1.7%	1980–1989	13.2%	41.9%
Law	5.1%	1.4%	≥ 1990	1.7%	16.3%

Table 11
Summary statistics of $Alumni_{ik}$ for mutual funds. This table shows the summary statistics of the number of alumni connections $Alumni_{ik}$ in each city of mutual funds over all quarters, based on our alumni data constructed using RiskMetrics and Capital IQ databases. We first compute the time-series average of the quarterly $Alumni_{ik}$ values (pooled across cities) for each individual fund, then we report the summary statistics of these time-series averages. School-Degree stands for the case where the alumni connection is based on a portfolio manager and a senior official of a firm (Chairman, CEO, or CFO) having attended the same school and received the same degree, and School-Degree-Year stands for the other case where the alumni connection is based on a portfolio manager and a senior official of a firm having graduated within a year from the same school and received the same degree. Since the values of $Alumni_{ik}$ can be zero, $\log(Alumni_{ik})$ is the log of one plus the value of $Alumni_{ik}$.

Summary statistics of $Alumni_{ik}$

School-Degree	Mean	S.D.	Median	p25	p75	Min	Max
$Alumni_{ik}$	1.38	3.42	0.00	0.00	1.25	0.00	32.60
$\log(Alumni_{ik})$	0.52	0.82	0.00	0.00	0.79	0.00	3.51
School-Degree-Year	**Mean**	**S.D.**	**Median**	**p25**	**p75**	**Min**	**Max**
$Alumni_{ik}$	0.59	1.58	0.00	0.00	0.50	0.00	16.60
$\log(Alumni_{ik})$	0.28	0.53	0.00	0.00	0.35	0.00	2.86

managers if the coefficient γ is positive and significant. We average the variables $Alumni_{ik}$ and $\eta_{i,k}$ across all time periods, so that regression (29) is run as a single cross-section regression.[20] The estimation results are summarized in Table 12.

From Table 12, the estimates of γ are statistically significant under both measures of alumni connections between fund managers and company officials (School-Degree and School-Degree-Year). Moreover, they are economically significant as well, even after we control for the median SAT score of a manager's undergraduate school and whether a manager holds a graduate degree.[21] A one standard deviation rise in $\log(\eta_{i,k})$ would lead to approximately a 6.5% rise in $Alumni_{ik}$ based on the School-Degree connection measure. This is 8% of the standard deviation of the left-hand side variable $\log(Alumni_{ik})$. The corresponding number is close to a 4.5% rise based on the stronger School-Degree-Year connection measure. Hence, our RPC measure is informative of alumni connections in a city.

5.3. Portfolio returns

Now we turn our attention to Proposition 2, which shows that a manager's stock positions in cities where he is connected should outperform his stock positions in cities where he is not connected. To test this implication, we follow the empirical strategy in Cohen et al. (2008) and construct stock portfolios using our RPC measure $\eta_{i,k}$. Recall that in the baseline model with i.i.d. connections, $g_{i,k}$

[20] Almost all of the manager-level controls x_i and city controls v_k in Eq. (29) do not vary across time, except for manager's age, which we average across time too.

[21] A higher median SAT score of the undergraduate school would imply the manager is more likely to have graduated from Ivy-league-like colleges, and holding a graduate degree would imply the manager is more likely to have an MBA degree. Both of these would lead to a higher number of observable network connections for managers (see the summary statistics in Table 10).

Table 12
Using model's RPC measure to predict alumni connections. This table reports the estimation result in the regression model $\log(Alumni_{ik}) = \alpha + \gamma \log(\eta_{i,k}) + x_i'\beta + v_k'\rho + \epsilon_i$ where the dependent variable is the log of one plus the number of alumni connections a fund manager i has in city k (based on our alumni connections data constructed using the Capital IQ-RiskMetrics data set). $Alumni_{ik}$ and $\eta_{i,k}$ are averaged across time for each fund manager i in city k. Baseline stands for our baseline model, and With covariates stands for the extended model with manager covariates. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively. School-Degree stands for the case where the alumni connection is based on a portfolio manager and a senior official of a firm (Chairman, CEO, or CFO) having attended the same school and received the same degree, and School-Degree-Year stands for the other case where the alumni connection is based on a portfolio manager and a senior official of a firm having graduated within a year from the same school and received the same degree. The main explanatory variable is $\log(\eta_{i,k})$, the log of one plus the RPC measure $\eta_{i,k}$ of manager i in city k. x_i denotes the vector of controls for fund manager demographics. For a given manager, *logSAT* is the log of the median SAT score (in 2005) of the undergraduate school that the manager attended, *Adv* and *Female* denote whether the manager holds a graduate degree and whether the manager is a female, *Age* is the age of the manager, and *Rep* denotes whether the manager is Republican-affiliated. v_k is the vector of controls for city attributes. For a given city, $\hat{b}_k$, $logIncome_k$, $logPop_k$, $NonWhite_k$, $SexRatio_k$, and $OldRatio_k$ denote, respectively, the estimated value of the relative city size parameter, the log of the per capita income, the log of the population per square mile of land area, the fraction of the population that is not white, the male-to-female sex ratio, and the fraction of the population that is above 55 years old. These city attributes are obtained from the 2000 U.S. Census data. *const* is the constant term.

	Baseline		With covariates	
	School-Degree	School-Degree-Year	School-Degree	School-Degree-Year
$\log(\eta_{i,k})$	0.172***	0.133***	0.162***	0.110**
	(4.30)	(3.13)	(3.74)	(2.27)
logSAT	0.204***	0.160***	0.207***	0.154***
	(8.13)	(5.91)	(8.02)	(5.87)
Adv	0.164***	0.129***	0.166***	0.129***
	(7.60)	(6.27)	(7.54)	(6.21)
Female	0.012	0.008	0.011	0.008
	(0.77)	(1.05)	(0.71)	(1.01)
Age	0.003***	0.001*	0.003***	0.001
	(8.40)	(1.75)	(7.51)	(1.54)
Rep	0.064***	0.026**	0.067***	0.027**
	(4.83)	(2.28)	(4.83)	(2.39)
$\log(\hat{b}_k)$	0.838***	0.441***	0.470***	0.242***
	(9.13)	(7.94)	(10.74)	(7.74)
$logIncome_k$	0.706***	0.444***	0.338**	0.257**
	(4.70)	(5.17)	(2.29)	(2.53)
$logPop_k$	0.102**	0.077***	0.083*	0.067**
	(2.66)	(3.48)	(1.98)	(2.50)
$NonWhite_k$	0.137	−0.050	0.098	−0.065
	(0.52)	(−0.35)	(0.58)	(−0.70)
$SexRatio_k$	−1.353***	−0.871***	−0.489	−0.444
	(−5.12)	(−5.12)	(−0.88)	(−1.35)
$OldRatio_k$	−1.167***	−0.552***	−0.100	0.361
	(−4.96)	(−3.14)	(−0.14)	(0.93)
const	−3.248***	−1.913***	−2.171***	−1.296***
	(−4.49)	(−4.62)	(−4.66)	(−4.13)

should be equal to one. Our model hence suggests a natural grouping. For manager i, cities where he is connected are those cities k for which $\eta_{i,k} > 1$. Those cities for which $\eta_{i,k} \leq 1$ are those where he is not connected. So, in each quarter, for each fund manager i, we sort their portfolio holdings into two groups: stock holdings from cities where they have $\eta_{i,k} > 1$ and stock holdings from cities where they have $\eta_{i,k} \leq 1$.

Then we form a value-weighted (based on the dollar value of the holdings) or equal-weighted portfolio for each of the two groups, and call them $\eta_{i,k} > 1$ portfolio and $\eta_{i,k} \leq 1$ portfolio for manager i. Finally, we construct the value-weighted or equal-weighted $\eta_{i,k} > 1$ portfolio across all managers by value-weighting (based on the total assets under management of a manager in the previous quarter) or equal-weighting all managers' $\eta_{i,k} > 1$ portfolios, and similarly the $\eta_{i,k} \leq 1$ portfolio across all managers. The $\eta_{i,k} > 1$ portfolio and the $\eta_{i,k} \leq 1$ portfolio are held for one quarter, and then the two portfolios are rebalanced based on the $\eta_{i,k}$'s of managers and their holdings. The mean quarterly net returns of the two portfolios (net of risk-free rate), their return differences Diff (return of $\eta_{i,k} > 1$ portfolio minus $\eta_{i,k} \leq 1$ portfolio), and their associated Carhart four-factor alphas are shown in Table 13. There are a range of results but they all point toward the same conclusion.

In Panel A, we report the quarterly net returns and Carhart four-factor adjusted alphas for the $\eta_{i,k} > 1$ and $\eta_{i,k} \leq 1$ portfolios from our baseline model, as well as their

Table 13
Returns for portfolios constructed using RPC measure. This table reports the returns of the portfolios constructed based on our RPC measure $\eta_{i,k}$ at the quarterly level. In each quarter, the $\eta_{i,k} > 1$ portfolio of a manager i is the portfolio based on his stock holdings from those cities where he has $\eta_{i,k} > 1$, and the $\eta_{i,k} \leq 1$ portfolio is the portfolio based on his stock holdings from those cities where he has $\eta_{i,k} \leq 1$. We then construct the value-weighted or equal-weighted $\eta_{i,k} > 1$ portfolio across all managers (the $\eta_{i,k} > 1$ portfolio) by value-weighting or equal-weighting all fund managers' $\eta_{i,k} > 1$ portfolios, and similarly the $\eta_{i,k} \leq 1$ portfolio. The $\eta_{i,k} > 1$ portfolio and the $\eta_{i,k} \leq 1$ portfolio are held for one quarter, and then the two portfolios are rebalanced based on the $\eta_{i,k}$'s. The quarterly net returns of the two portfolios (net of risk-free rate), their return differences Diff (return of $\eta_{i,k} > 1$ portfolio minus $\eta_{i,k} \leq 1$ portfolio), and their associated Carhart four-factor alphas are shown, together with their t-statistics (Newey–West adjusted).

Panel A: Baseline

Value-weighted	Net return		Carhart alpha	
Portfolio	Mean (%)	t-stat	Mean (%)	t-stat
$\eta_{i,k} > 1$	1.46	2.63	1.03	2.20
$\eta_{i,k} \leq 1$	0.98	1.67	0.58	1.01
Diff	0.48	2.07	0.45	2.10
Equal-weighted	Net return		Carhart alpha	
Portfolio	Mean (%)	t-stat	Mean (%)	t-stat
$\eta_{i,k} > 1$	1.37	3.49	0.77	2.19
$\eta_{i,k} \leq 1$	0.93	1.99	0.39	0.98
Diff	0.44	2.21	0.38	2.29

Panel B: With covariates

Value-weighted	Net return		Carhart alpha	
Portfolio	Mean (%)	t-stat	Mean (%)	t-stat
$\eta_{i,k} > 1$	1.42	2.53	1.02	2.07
$\eta_{i,k} \leq 1$	0.99	1.71	0.60	1.23
Diff	0.43	2.14	0.42	2.24
Equal-weighted	Net return		Carhart alpha	
Portfolio	Mean (%)	t-stat	Mean (%)	t-stat
$\eta_{i,k} > 1$	1.34	3.27	0.76	2.10
$\eta_{i,k} \leq 1$	0.95	1.95	0.40	1.00
Diff	0.39	2.15	0.36	2.37

return difference (Diff). For value-weighted returns, the $\eta_{i,k} > 1$ portfolio has a mean net return of 1.46% ($t = 2.63$) and a Carhart four-factor alpha of 1.03% ($t = 2.20$) per quarter, while the $\eta_{i,k} \leq 1$ portfolio has a net return of 0.98% ($t = 1.67$) and an alpha of 0.58% ($t = 1.01$) per quarter. The return difference between the two portfolios is 0.48% per quarter ($t = 2.07$) with an associated Carhart alpha of 0.45% ($t = 2.10$), both of which are statistically significant. Hence the $\eta_{i,k} > 1$ portfolio outperforms the $\eta_{i,k} \leq 1$ counterpart by about 1.8% per year in terms of Carhart alpha. For equal-weighted returns, the results are quite similar. In particular, the $\eta_{i,k} > 1$ portfolio generates an outperformance of 0.44% per quarter ($t = 2.21$) over the $\eta_{i,k} \leq 1$ portfolio, which translates into an alpha of 0.38% per quarter ($t = 2.29$), or about 1.5% per year. This is consistent with our Proposition 2 that connections which increase the precision of private signals are valuable for the performance of stock holdings in that city.

In Panel B, we report the corresponding results for the $\eta_{i,k} > 1$ and $\eta_{i,k} \leq 1$ portfolios from our extended model with covariates, which are comparable to the ones from the baseline model. For value-weighted returns, the outperformance of the $\eta_{i,k} > 1$ portfolio over the $\eta_{i,k} \leq 1$ portfolio is 0.43% ($t = 2.14$) per quarter with an associated Carhart alpha of 0.42% ($t = 2.24$), which is about 1.7% per year. For equal-weighted returns, the corresponding outperformance and alpha are 0.39% ($t = 2.15$) and 0.36% ($t = 2.37$) per quarter, respectively. Overall, the $\eta_{i,k} > 1$ portfolio outperforms the $\eta_{i,k} \leq 1$ portfolio by around 1.5% to 1.8% per year.

These results are reminiscent of the Industry Concentration Index of Kacperczyk et al. (2005) who find that among active managers, those who hold concentrated positions, measured in portfolio weights in industries as opposed to our geographic concentration measure, outperform those that do not. They view their results as a test of discriminating among closet indexers and non-closet indexers. Our model puts forth a different mechanism for why concentration is valuable that is more naturally tied to a model of social networks and portfolio choice.

In Section B of our Online Appendix (Table B1), we show the portfolio return results for subsamples of self-reported fund styles. The three categories are Small-cap, Growth, and Income. We report the value-weighted returns and their Carhart four-factor adjusted alphas. We find that the results are not concentrated in a particular fund style but are rather consistent across all self-reported styles.

5.4. Robustness

For robustness, we repeat our main analysis for three sets of additional studies. In the first one, we drop micro-cap funds from our mutual fund sample. In the second one, we include the next ten smaller cities (MSAs) into our study. In the third one, we carry out our tests on a hedge fund sample. All results can be found in Section A of our Online Appendix but we will briefly describe the main points here. For the first one, the results are virtually the same as our main case since micro-cap funds only constitute a small part of our mutual fund sample. For the second one, adding the next ten cities slightly improves the predictive power of our RPC measure $\eta_{i,k}$ for returns by about 0.1% per year in terms of the outperformance of the $\eta_{i,k} > 1$ portfolio, but other results are similar to our main case. For the third one regarding the hedge fund sample, we find that hedge funds have smaller gregariousness but have more overdispersed networks (larger overdispersion) than mutual funds in those cities. Moreover, the predictive power of our RPC measure for returns is higher, and the $\eta_{i,k} > 1$ portfolio outperforms by 2.7% per year in terms of alpha.

6. Conclusion

We develop a methodology to infer latent connections using only data on investors' holdings of stocks headquartered in different cities. By embedding a tractable statistical model of random networks into an otherwise standard portfolio model, where the precision of private

signals increases with an investor's connections in a city, we map the observed distribution of stock holdings to network connections in different cities and thereby infer the structure of latent social networks. Using mutual fund manager data, we find that the distribution of the number of connections a manager has in a given city is too overdispersed or fat-tailed to be generated by the Erdös-Rényi model of i.i.d. connections. We then show that managers in the tail of this distribution, with geographically concentrated stock holdings and a high relative propensity to be connected to a city, are more likely to have university alumni connections in that city and their stock holdings in these cities outperform their holdings in other cities. Our inference strategy can be applied to other types of investor networks, thereby allowing a comparison of network structures across investor types (retail versus institutional). We leave these other applications for future research.

Appendix A

A.1. A belief-updating model for conditional expected stock excess return of Eq. (7)

For each investor i, they have a prior $\tilde{f}_{j,k} \equiv \tilde{R}_{j,k} - R_f$ about the excess return of stock j headquartered in city k. The prior is:

$$\tilde{f}_{j,k} \overset{i.i.d.}{\sim} \mathcal{N}\left(\alpha + \boldsymbol{\beta}\boldsymbol{X}_j, \frac{1}{\tau^0}\right) \tag{30}$$

where τ^0 is the precision of the prior, i.e., the inverse of the variance. The distribution (prior) of $\tilde{f}_{j,k}$ is common for all investors. Moreover, for stock j headquartered in city k, an investor i also receives a private signal about the excess return, which comes from the friends that he knows in city k. The private signal is

$$S_{i,j,k} = \tilde{f}_{j,k} + \epsilon_{i,j,k} \tag{31}$$

where

$$\epsilon_{i,j,k} \overset{i.i.d.}{\sim} \mathcal{N}\left(0, \frac{1}{\tau^{\epsilon}(y_{i,k})}\right) \tag{32}$$

and $\tau^{\epsilon}(\cdot)$ is an increasing function of the number of friends $y_{i,k}$. This says the higher the number of friends an investor i has in city k, the more precise his private signal about the excess return of stock j headquartered in city k would be. $\epsilon_{i,j,\,k}$ is independent of $\tilde{f}_{j,k}$ (prior), thus the distribution of the private signal $S_{i,j,\,k}$ is

$$S_{i,j,k} \overset{i.i.d.}{\sim} \mathcal{N}\left(\alpha + \boldsymbol{\beta}\boldsymbol{X}_j, \frac{\tau^0 + \tau^{\epsilon}(y_{i,k})}{\tau^0 \tau^{\epsilon}(y_{i,k})}\right). \tag{33}$$

Investors form posterior beliefs about excess returns of stocks using Bayes' law based on the prior and the private signal they receive. Because the prior belief (30) and the private signal (33) both follow normal distributions, investor i's posterior distribution for the excess return of stock j headquartered in city k, conditional on the signal he receives, also follows a normal distribution. In particular, by the standard Bayesian-updating results, the posterior mean is

$$\begin{aligned}\hat{f}_{i,j,k} &\equiv \mathbb{E}[(R_{j,k} - R_f)|S_{i,j,k}] \equiv \mathbb{E}_i(R_{j,k} - R_f) \\ &= \frac{\tau^0}{\tau^0 + \tau^{\epsilon}(y_{i,k})}(\alpha + \boldsymbol{\beta}\boldsymbol{X}_j) + \frac{\tau^{\epsilon}(y_{i,k})}{\tau^0 + \tau^{\epsilon}(y_{i,k})} S_{i,j,k} \\ &= (\alpha + \boldsymbol{\beta}\boldsymbol{X}_j) + \frac{\tau^{\epsilon}(y_{i,k})}{\tau^0 + \tau^{\epsilon}(y_{i,k})}(S_{i,j,k} - (\alpha + \boldsymbol{\beta}\boldsymbol{X}_j))\end{aligned} \tag{34}$$

and the posterior variance is

$$\hat{V}_{i,j,k} \equiv Var((R_{j,k} - R_f)|S_{i,j,k}) = \frac{1}{\tau^0 + \tau^{\epsilon}(y_{i,k})}. \tag{35}$$

Since the private signal $S_{i,j,\,k}$ is unobservable to outsiders, we can define

$$\Xi_{i,j,k} \equiv S_{i,j,k} - (\alpha + \boldsymbol{\beta}\boldsymbol{X}_j) \overset{i.i.d.}{\sim} \mathcal{N}\left(0, \frac{\tau^0 + \tau^{\epsilon}(y_{i,k})}{\tau^0 \tau^{\epsilon}(y_{i,k})}\right) \tag{36}$$

based on the distribution of $S_{i,j,\,k}$ in Eq. (33). Then

$$\begin{aligned}\frac{\tau^{\epsilon}(y_{i,k})}{\tau^0 + \tau^{\epsilon}(y_{i,k})}\Xi_{i,j,k} &\equiv \frac{\tau^{\epsilon}(y_{i,k})}{\tau^0 + \tau^{\epsilon}(y_{i,k})}(S_{i,j,k} - (\alpha + \boldsymbol{\beta}\boldsymbol{X}_j)) \overset{i.i.d.}{\sim} \\ &\mathcal{N}\left(0, \frac{\tau^{\epsilon}(y_{i,k})}{\tau^0(\tau^0 + \tau^{\epsilon}(y_{i,k}))}\right).\end{aligned} \tag{37}$$

But if we define a variable $\lambda_{i,k}\xi_{i,j,\,k}$, where

$$\lambda_{i,k} = \sqrt{\frac{\tau^{\epsilon}(y_{i,k})}{\tau^0(\tau^0 + \tau^{\epsilon}(y_{i,k}))}} \quad \text{and} \quad \xi_{i,j,k} \overset{i.i.d.}{\sim} \mathcal{N}(0,1), \tag{38}$$

then $\lambda_{i,k}\xi_{i,j,\,k}$ is also $\overset{i.i.d.}{\sim} \mathcal{N}\left(0, \frac{\tau^{\epsilon}(y_{i,k})}{\tau^0(\tau^0+\tau^{\epsilon}(y_{i,k}))}\right)$. Therefore, $\lambda_{i,k}\xi_{i,j,\,k}$ is equivalent to the term $\frac{\tau^{\epsilon}(y_{i,k})}{\tau^0+\tau^{\epsilon}(y_{i,k})}(S_{i,j,k} - (\alpha + \boldsymbol{\beta}\boldsymbol{X}_j))$. Hence the posterior mean in Eq. (34) can also be written as

$$\mathbb{E}_i(R_{j,k} - R_f) = (\alpha + \boldsymbol{\beta}\boldsymbol{X}_j) + \lambda_{i,k}\xi_{i,j,k} \tag{39}$$

where $\lambda_{i,k}\xi_{i,j,\,k}$ is specified in Eq. (38). This is exactly the same specification for investor i's conditional expected excess return of stock j headquartered in city k as stated in Eq. (7) of the main article.

Clearly, $\lambda_{i,k}$ is a monotonically increasing function of $\frac{\tau^{\epsilon}(y_{i,k})}{\tau^0 + \tau^{\epsilon}(y_{i,k})}$, the posterior weight that investor i puts on the private signal $S_{i,j,\,k}$ (relative to the prior). Furthermore,

$$\begin{aligned}\frac{\partial \lambda_{i,k}}{\partial y_{i,k}} &= \frac{1}{2}\left(\frac{\tau^{\epsilon}(y_{i,k})}{\tau^0(\tau^0 + \tau^{\epsilon}(y_{i,k}))}\right)^{-\frac{1}{2}} \\ &\quad \times \frac{\tau^0(\tau^0 + \tau^{\epsilon}(y_{i,k}))\tau^{\epsilon\prime}(y_{i,k}) - \tau^{\epsilon}(y_{i,k})\tau^0\tau^{\epsilon\prime}(y_{i,k})}{(\tau^0(\tau^0 + \tau^{\epsilon}(y_{i,k})))^2} \\ &= \frac{1}{2}\left(\frac{\tau^{\epsilon}(y_{i,k})}{\tau^0(\tau^0 + \tau^{\epsilon}(y_{i,k}))}\right)^{-\frac{1}{2}} \frac{(\tau^0)^2\tau^{\epsilon\prime}(y_{i,k})}{(\tau^0(\tau^0 + \tau^{\epsilon}(y_{i,k})))^2} \\ &> 0\end{aligned} \tag{40}$$

as $\tau^{\epsilon\prime}(y_{i,k}) \equiv \frac{\partial \tau^{\epsilon}(y_{i,k})}{\partial y_{i,k}} > 0$ since $\tau^{\epsilon}(\cdot)$ is an increasing function of the number of friends $y_{i,k}$. Hence $\lambda_{i,k}$ is an in-

creasing function of the number of friends $y_{i,k}$, which is the statement that we made under Eq. (7) in the main article.

For tractability, we parameterize $\lambda_{i,k}$ as

$$\lambda_{i,k} = \lambda_0 + \lambda_1 y_{i,k} \tag{41}$$

with $\lambda_0 > 0$ and $\lambda_1 > 0$. This is Eq. (8) stated in the main article.

A.2. Proofs

Derivation of $\mathbb{E}[w_{i,j,k}|y_{i,k}, \boldsymbol{X}_j, \theta]$ in Eq. (16):

$$\begin{aligned}
&\mathbb{E}[w_{i,j,k}|y_{i,k}, \theta] \\
&= \mathbb{E}[w_{i,j,k}|w_{i,j,k} > 0, y_{i,k}, \boldsymbol{X}_j, \theta]\mathbb{P}(w_{i,j,k} > 0|y_{i,k}, \boldsymbol{X}_j, \theta) \\
&\quad + \mathbb{E}[w_{i,j,k}|w_{i,j,k} = 0, y_{i,k}, \boldsymbol{X}_j, \theta]\mathbb{P}(w_{i,j,k} = 0|y_{i,k}, \boldsymbol{X}_j, \theta) \\
&= \mathbb{E}[w_{i,j,k}|w_{i,j,k} > 0, y_{i,k}, \boldsymbol{X}_j, \theta]\mathbb{P}(w_{i,j,k} > 0|y_{i,k}, \boldsymbol{X}_j, \theta) \\
&= \left(\alpha + \boldsymbol{\beta}\boldsymbol{X}_j + (\lambda_0 + \lambda_1 y_{i,k}) \frac{\varphi\left(-\frac{\alpha+\boldsymbol{\beta}\boldsymbol{X}_j}{(\lambda_0+\lambda_1 y_{i,k})} \right)}{1 - \Phi\left(-\frac{\alpha+\boldsymbol{\beta}\boldsymbol{X}_j}{(\lambda_0+\lambda_1 y_{i,k})} \right)} \right) \\
&\quad \times \left(1 - \Phi\left(-\frac{\alpha + \boldsymbol{\beta}\boldsymbol{X}_j}{(\lambda_0 + \lambda_1 y_{i,k})} \right) \right) \\
&= (\alpha + \boldsymbol{\beta}\boldsymbol{X}_j)\, \Phi\left(\frac{\alpha + \boldsymbol{\beta}\boldsymbol{X}_j}{(\lambda_0 + \lambda_1 y_{i,k})} \right) + (\lambda_0 + \lambda_1 y_{i,k}) \\
&\quad \varphi\left(\frac{\alpha + \boldsymbol{\beta}\boldsymbol{X}_j}{(\lambda_0 + \lambda_1 y_{i,k})} \right)
\end{aligned} \tag{42}$$

where the first line (equal sign) comes from the definition of the expectation, the second line from the fact that $\mathbb{E}[w_{i,j,k}|w_{i,j,k} = 0, y_{i,k}, \boldsymbol{X}_j, \theta]$ is identically zero, the third line from the fact that $w_{i,j,k}$ given $y_{i,k}$ follows a normal distribution with mean $\alpha + \boldsymbol{\beta}\boldsymbol{X}_j$ and variance $(\lambda_0 + \lambda_1 y_{i,k})^2$ and properties of truncated normal distributions (truncated from below at zero), and the last line from the properties of the standard normal cdf Φ and pdf φ functions.

Proof of Proposition 1. To simplify notations, let us denote $\alpha + \boldsymbol{\beta}\boldsymbol{X}_j = r_j$. Note that

$$\begin{aligned}
\frac{\partial \mathbb{E}[w_{i,j,k}|y_{i,k}, \boldsymbol{X}_j, \theta]}{\partial y_{i,k}} &= \frac{\partial \mathbb{E}[w_{i,j,k}|y_{i,k}, \boldsymbol{X}_j, \theta]}{\partial \lambda_{i,k}} \frac{\partial \lambda_{i,k}}{\partial y_{i,k}} \\
&= \frac{\partial \mathbb{E}[w_{i,j,k}|y_{i,k}, \boldsymbol{X}_j, \theta]}{\partial \lambda_{i,k}} \lambda_1
\end{aligned} \tag{43}$$

where $\lambda_{i,k} = \lambda_0 + \lambda_1 y_{i,k}$ as defined in Eq. (8). If we can show $\frac{\partial \mathbb{E}[w_{i,j,k}|y_{i,k}, \boldsymbol{X}_j, \theta]}{\partial y_{i,k}} > 0$, then we will have proved that $\mathbb{E}[w_{i,j,k}|y_{i,k}, \boldsymbol{X}_j, \theta]$ is monotonically increasing in $y_{i,k}$. Since $\lambda_1 > 0$, this is equivalent to showing

$$\frac{\partial \mathbb{E}[w_{i,j,k}|y_{i,k}, \boldsymbol{X}_j, \theta]}{\partial \lambda_{i,k}} > 0. \tag{44}$$

Now

$$\begin{aligned}
&\frac{\partial \mathbb{E}[w_{i,j,k}|y_{i,k}, \boldsymbol{X}_j, \theta]}{\partial \lambda_{i,k}} \\
&= (\alpha + \boldsymbol{\beta}\boldsymbol{X}_j) \cdot \varphi\left(\frac{\alpha + \boldsymbol{\beta}\boldsymbol{X}_j}{\lambda_{i,k}} \right) \cdot \left(-\frac{\alpha + \boldsymbol{\beta}\boldsymbol{X}_j}{\lambda_{i,k}^2} \right) \\
&\quad + \varphi\left(\frac{\alpha + \boldsymbol{\beta}\boldsymbol{X}_j}{\lambda_{i,k}} \right) + \lambda_{i,k} \left[-\left(\frac{\alpha + \boldsymbol{\beta}\boldsymbol{X}_j}{\lambda_{i,k}} \right) \cdot \varphi\left(\frac{\alpha + \boldsymbol{\beta}\boldsymbol{X}_j}{\lambda_{i,k}} \right) \right. \\
&\quad \left. \cdot \left(-\frac{\alpha + \boldsymbol{\beta}\boldsymbol{X}_j}{\lambda_{i,k}^2} \right) \right] \\
&= -\varphi\left(\frac{\alpha + \boldsymbol{\beta}\boldsymbol{X}_j}{\lambda_{i,k}} \right) \cdot \left(\frac{\alpha + \boldsymbol{\beta}\boldsymbol{X}_j}{\lambda_{i,k}} \right)^2 + \varphi\left(\frac{\alpha + \boldsymbol{\beta}\boldsymbol{X}_j}{\lambda_{i,k}} \right) + \varphi\left(\frac{\alpha + \boldsymbol{\beta}\boldsymbol{X}_j}{\lambda_{i,k}} \right) \\
&\quad \cdot \left(\frac{\alpha + \boldsymbol{\beta}\boldsymbol{X}_j}{\lambda_{i,k}} \right)^2 \\
&= \varphi\left(\frac{\alpha + \boldsymbol{\beta}\boldsymbol{X}_j}{\lambda_{i,k}} \right) > 0.
\end{aligned} \tag{45}$$

Therefore, we have

$$\frac{\partial \mathbb{E}[w_{i,j,k}|y_{i,k}, \boldsymbol{X}_j, \theta]}{\partial y_{i,k}} = \varphi\left(\frac{\alpha + \boldsymbol{\beta}\boldsymbol{X}_j}{\lambda_0 + \lambda_1 y_{i,k}} \right) \cdot \lambda_1 > 0 \tag{46}$$

which is the desired result. □

Proof of Proposition 2. Under the optimal portfolio weight from Eq. (14), the expected total excess return of investing in stock j headquartered in city k given $y_{i,k}$ is

$$\begin{aligned}
&\mathbb{E}\left[\left(\alpha + \boldsymbol{\beta}\boldsymbol{X}_j + \lambda_{i,k}\xi_{i,j,k} \right) \left(\alpha + \boldsymbol{\beta}\boldsymbol{X}_j + \lambda_{i,k}\xi_{i,j,k} \right)^+ \right. \\
&\quad \left. - \frac{1}{2} \left(\left(\alpha + \boldsymbol{\beta}\boldsymbol{X}_j + \lambda_{i,k}\xi_{i,j,k} \right)^+ \right)^2 | y_{i,k}, \boldsymbol{X}_j, \theta \right]
\end{aligned} \tag{47}$$

where $\lambda_{i,k} = \lambda_0 + \lambda_1 y_{i,k}$ as before and the expectation is taken over $\xi_{i,j,k}$. Consequently, to show that this is monotonically increasing in the number of friends $y_{i,k}$, we just need to show that its derivative with respect to $y_{i,k}$ is strictly positive. To simplify notations, let us denote $\alpha + \boldsymbol{\beta}\boldsymbol{X}_j = r_j$ and leave out the notation of conditioning on $y_{i,k}$, $\boldsymbol{X}_j$, θ as it is always there. Now this expectation term of (47) is equal to

$$\begin{aligned}
&\mathbb{E}\left(\frac{1}{2} \left(r_j + \lambda_{i,k}\xi_{i,j,k} \right)^2 | r_j + \lambda_{i,k}\xi_{i,j,k} > 0 \right) \mathbb{P}\left(r_j + \lambda_{i,k}\xi_{i,j,k} > 0 \right) \\
&\quad + \mathbb{E}\left(\frac{1}{2} \left(r_j + \lambda_{i,k}\xi_{i,j,k} \right)^2 | r_j + \lambda_{i,k}\xi_{i,j,k} = 0 \right) \mathbb{P}\left(r_j + \lambda_{i,k}\xi_{i,j,k} = 0 \right) \\
&= \mathbb{E}\left(\frac{1}{2} \left(r_j + \lambda_{i,k}\xi_{i,j,k} \right)^2 | r_j + \lambda_{i,k}\xi_{i,j,k} > 0 \right) \mathbb{P}\left(r_j + \lambda_{i,k}\xi_{i,j,k} > 0 \right)
\end{aligned} \tag{48}$$

as the term after the plus sign is identically zero. Since $\xi_{i,j,k}$ is standard normal, we have

$$\begin{aligned}
\mathbb{P}\left(r_j + \lambda_{i,k}\xi_{i,j,k} > 0 \right) &= \mathbb{P}\left(\xi_{i,j,k} > -\frac{r_j}{\lambda_{i,k}} \right) \\
&= 1 - \Phi\left(-\frac{r_j}{\lambda_{i,k}} \right) = \Phi\left(\frac{r_j}{\lambda_{i,k}} \right).
\end{aligned} \tag{49}$$

The first conditional expectation term is

$$\begin{aligned}
&\mathbb{E}\left(\frac{1}{2} \left(r_j + \lambda_{i,k}\xi_{i,j,k} \right)^2 | r_j + \lambda_{i,k}\xi_{i,j,k} > 0 \right) \\
&= \frac{1}{2} \left[Var\left(\left(r_j + \lambda_{i,k}\xi_{i,j,k} \right) | r_j + \lambda_{i,k}\xi_{i,j,k} > 0 \right) \right. \\
&\quad \left. + \mathbb{E}\left(\left(r_j + \lambda_{i,k}\xi_{i,j,k} \right) | r_j + \lambda_{i,k}\xi_{i,j,k} > 0 \right)^2 \right]
\end{aligned} \tag{50}$$

where *Var* denotes variance. By the properties of truncated normal distributions (truncated from below at zero), we have

$$\begin{aligned} &Var\left(\left(r_j+\lambda_{i,k}\xi_{i,j,k}\right)\mid r_j+\lambda_{i,k}\xi_{i,j,k}>0\right)\\ &=\lambda_{i,k}^2\left(1+\frac{\left(-\frac{r_j}{\lambda_{i,k}}\right)\varphi\left(-\frac{r_j}{\lambda_{i,k}}\right)}{1-\Phi\left(-\frac{r_j}{\lambda_{i,k}}\right)}-\left(\frac{\varphi\left(-\frac{r_j}{\lambda_{i,k}}\right)}{1-\Phi\left(-\frac{r_j}{\lambda_{i,k}}\right)}\right)^2\right)\\ &=\lambda_{i,k}^2\left(1-\frac{\left(\frac{r_j}{\lambda_{i,k}}\right)\varphi\left(\frac{r_j}{\lambda_{i,k}}\right)}{\Phi\left(\frac{r_j}{\lambda_{i,k}}\right)}-\frac{\varphi\left(\frac{r_j}{\lambda_{i,k}}\right)^2}{\Phi\left(\frac{r_j}{\lambda_{i,k}}\right)^2}\right) \end{aligned} \tag{51}$$

and

$$\begin{aligned} &\mathbb{E}\left(\left(r_j+\lambda_{i,k}\xi_{i,j,k}\right)\mid r_j+\lambda_{i,k}\xi_{i,j,k}>0\right)^2\\ &=\left(r_j+\frac{\lambda_{i,k}\varphi\left(-\frac{r_j}{\lambda_{i,k}}\right)}{1-\Phi\left(-\frac{r_j}{\lambda_{i,k}}\right)}\right)^2\\ &=r_j^2+\frac{\lambda_{i,k}^2\varphi\left(\frac{r_j}{\lambda_{i,k}}\right)^2}{\Phi\left(\frac{r_j}{\lambda_{i,k}}\right)^2}+\frac{2r_j\lambda_{i,k}\varphi\left(\frac{r_j}{\lambda_{i,k}}\right)}{\Phi\left(\frac{r_j}{\lambda_{i,k}}\right)}. \end{aligned} \tag{52}$$

Hence

$$\begin{aligned} &\mathbb{E}\left(\frac{1}{2}\left(r_j+\lambda_{i,k}\xi_{i,j,k}\right)^2\mid r_j+\lambda_{i,k}\xi_{i,j,k}>0\right)\\ &=\frac{1}{2}\left(\lambda_{i,k}^2\left(1-\frac{\left(\frac{r_j}{\lambda_{i,k}}\right)\varphi\left(\frac{r_j}{\lambda_{i,k}}\right)}{\Phi\left(\frac{r_j}{\lambda_{i,k}}\right)}-\frac{\varphi\left(\frac{r_j}{\lambda_{i,k}}\right)^2}{\Phi\left(\frac{r_j}{\lambda_{i,k}}\right)^2}\right)\right.\\ &\left.+\left(r_j^2+\frac{\lambda_{i,k}^2\varphi\left(\frac{r_j}{\lambda_{i,k}}\right)^2}{\Phi\left(\frac{r_j}{\lambda_{i,k}}\right)^2}+\frac{2r_j\lambda_{i,k}\varphi\left(\frac{r_j}{\lambda_{i,k}}\right)}{\Phi\left(\frac{r_j}{\lambda_{i,k}}\right)}\right)\right). \end{aligned} \tag{53}$$

Therefore, combining Eqs. (49) and (53), we have

$$\begin{aligned} &\mathbb{E}\left(\frac{1}{2}\left(r_j+\lambda_{i,k}\xi_{i,j,k}\right)^2\mid r_j+\lambda_{i,k}\xi_{i,j,k}>0\right)\mathbb{P}\left(r_j+\lambda_{i,k}\xi_{i,j,k}>0\right)\\ &=\frac{1}{2}\left(\lambda_{i,k}^2\left(1-\frac{\left(\frac{r_j}{\lambda_{i,k}}\right)\varphi\left(\frac{r_j}{\lambda_{i,k}}\right)}{\Phi\left(\frac{r_j}{\lambda_{i,k}}\right)}-\frac{\varphi\left(\frac{r_j}{\lambda_{i,k}}\right)^2}{\Phi\left(\frac{r_j}{\lambda_{i,k}}\right)^2}\right)\right.\\ &\left.+\left(r_j^2+\frac{\lambda_{i,k}^2\varphi\left(\frac{r_j}{\lambda_{i,k}}\right)^2}{\Phi\left(\frac{r_j}{\lambda_{i,k}}\right)^2}+\frac{2r_j\lambda_{i,k}\varphi\left(\frac{r_j}{\lambda_{i,k}}\right)}{\Phi\left(\frac{r_j}{\lambda_{i,k}}\right)}\right)\right)\times\Phi\left(\frac{r_j}{\lambda_{i,k}}\right)\\ &=\frac{1}{2}\left[\Phi\left(\frac{r_j}{\lambda_{i,k}}\right)\left(\lambda_{i,k}^2+r_j^2\right)+r_j\lambda_{i,k}\varphi\left(\frac{r_j}{\lambda_{i,k}}\right)\right]. \end{aligned} \tag{54}$$

□

Now we want to show the derivative of (54) above with respect to $y_{i,k}$ is strictly positive. Let us denote the whole term of Eq. (54) as W, then we want to show

$$\frac{\partial W}{\partial y_{i,k}}>0. \tag{55}$$

Because

$$\frac{\partial W}{\partial y_{i,k}}=\frac{\partial W}{\partial\lambda_{i,k}}\frac{\partial\lambda_{i,k}}{\partial y_{i,k}}=\frac{\partial W}{\partial\lambda_{i,k}}\lambda_1 \tag{56}$$

as $\lambda_{i,k}=\lambda_0+\lambda_1 y_{i,k}$, we just need to derive the expression for $\frac{\partial W}{\partial\lambda_{i,k}}$. This is

$$\begin{aligned} \frac{\partial W}{\partial\lambda_{i,k}}=&\frac{1}{2}\left[\left(\lambda_{i,k}^2+r_j^2\right)\varphi\left(\frac{r_j}{\lambda_{i,k}}\right)\left(-\frac{r_j}{\lambda_{i,k}^2}\right)+2\Phi\left(\frac{r_j}{\lambda_{i,k}}\right)\lambda_{i,k}\right.\\ &\left.+r_j\left(\varphi\left(\frac{r_j}{\lambda_{i,k}}\right)+\varphi\left(\frac{r_j}{\lambda_{i,k}}\right)\frac{r_j^2}{\lambda_{i,k}^2}\right)\right]\\ =&\frac{1}{2}\left[2\Phi\left(\frac{r_j}{\lambda_{i,k}}\right)\lambda_{i,k}\right]=\Phi\left(\frac{r_j}{\lambda_{i,k}}\right)\lambda_{i,k}. \end{aligned} \tag{57}$$

Therefore,

$$\frac{\partial W}{\partial y_{i,k}}=\Phi\left(\frac{r_j}{\lambda_{i,k}}\right)\lambda_{i,k}\lambda_1>0 \tag{58}$$

where $r_j=\alpha+\boldsymbol{\beta}\boldsymbol{X}_j$ and $\lambda_{i,k}=\lambda_0+\lambda_1 y_{i,k}$. Hence we have the desired result. □

References

Adamic, L.A., Adar, E., 2005. How to search a social network. Soc. Netw. 27, 187–203.

Barber, B.M., Odean, T., 2001. Boys will be boys: gender, overconfidence, and common stock investment. Q. J. Econ. 116, 261–292.

Brandt, W.M., Santa-Clara, P., Valkanov, R., 2009. Parametric portfolio policies: exploiting characteristics in the cross-section of equity returns. Rev. Financ. Stud. 22, 3411–3447.

Cameron, A., Trivedi, P.K., 2005. Microeconometrics: Methods and Applications. Cambridge University Press, UK.

Chen, J., Hong, H., Stein, J.C., 2002. Breadth of ownership and stock returns. J. Financ. Econ. 66, 171–205.

Christoffersen, S.E., Sarkissian, S., 2009. City size and fund performance. J. Financ. Econ. 92, 252–275.

Cohen, L., Frazzini, A., Malloy, C.J., 2008. The small world of investing: board connections and mutual fund returns. J. Political Econ. 116, 951–979.

Coval, J.D., Moskowitz, T.J., 1999. Home bias at home: local equity preference in domestic portfolios. J. Financ. 54, 2045–2073.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the em algorithm. J. R. Stat. Soc. Ser. B (Methodol.) 39, 1–38.

Engelberg, J., Gao, P., Parsons, C.A., 2012. Friends with money. J. Financ. Econ. 103, 169–188.

Erdös, P., Rényi, A., 1959. On random graphs. Publ. Math. 6, 290–297.

Falkenstein, E.G., 1996. Preferences for stock characteristics as revealed by mutual fund portfolio holdings. J. Financ. 51, 111–135.

Fama, E.F., MacBeth, J.D., 1973. Risk, return, and equilibrium: empirical tests. J. Political Econ. 81, 607–636.

Gabaix, X., 2009. Power laws in economics and finance. Annu. Rev. Econ. 1, 255–294.

Gârleanu, N., Pedersen, L.H., 2013. Dynamic trading with predictable returns and transaction costs. J. Financ. 68, 2309–2340.

Hausman, J., Hall, B.H., Griliches, Z., 1984. Econometric models for count data with an application to the patents – R&D relationship. Econometrica 52, 909–938.

Hong, H., Kostovetsky, L., 2012. Red and blue investing: value and finance. J. Financ. Econ. 103, 1–19.

Hong, H., Kubik, J.D., Stein, J.C., 2005. Thy neighbor's portfolio: word-of-mouth effects in the holdings and trades of money managers. J. Financ. 60, 2801–2824.

Jackson, M.O., Rogers, B.W., 2007. Meeting strangers and friends of friends: how random are social networks? Am. Econ. Rev. 97, 890–915.

Kacperczyk, M., Sialm, C., Zheng, L., 2005. On the industry concentration of actively managed equity mutual funds. J. Financ. 60, 1983–2011.

Killworth, P.D., Johnsen, E.C., McCarty, C., Shelley, G.A., Bernard, H., 1998. A social network approach to estimating seroprevalence in the United States. Soc. Netw. 20, 23–50.

Lazarsfeld, P.F., Merton, R.K., 1954. Friendship as a social process: a substantive and methodological analysis. Freedom and Control in Modern Society. Van Nostrand, New York, pp. 18–66..

McPherson, M., Smith-Lovin, L., Cook, J.M., 2001. Birds of a feather: homophily in social networks. Annu. Rev. Sociol. 27, 415–444.

Pool, V., Stoffman, N., Yonker, S., 2015. The people in your neighborhood: social interactions and mutual fund portfolios. J. Financ. 70, 2679–2732.
Wu, C.F.J., 1983. On the convergence properties of the em algorithm. Ann. Stat. 11, 95–103.
Zheng, L., 1999. Is money smart? A study of mutual fund investors' fund selection ability. J. Financ. 54, 901–933.
Zheng, T., Salganik, M.J., Gelman, A., 2006. How many people do you know in prison? Using overdispersion in count data to estimate social structure in networks. J. Am. Stat. Assoc. 101, 409–423.