Count Models of Social Networks in Finance *

Harrison Hong† Jiangmin Xu‡

First Draft: November 29, 2013
This Draft: December 24, 2014

ABSTRACT

Social networks are thought to be important for the investment and performance of mutual fund managers. We propose a measure of whether a manager is part of a network using only data on the distribution of the number of stocks headquartered in a given city that are held by managers. For some cities, the count distribution is roughly Poisson. However, for a significant fraction of cities, the count distribution is highly overdispersed Poisson — where most managers have a couple of picks but a few managers have many picks. We show that the degree of overdispersion is a theoretically well-motivated measure of network influence and that managers with concentrated stock picks in a city are likely to be part of a network in that city. These managers indeed significantly outperform other managers by around 1.6% per annum.

*We want to thank Johannes Ruf at the Oxford-Man Institute for introducing us to the statistics and sociology models of social networks. We also thank seminar participants and discussants at Singapore Management University, China International Finance Conference, NYU Stern School of Business, Hong Kong Science and Technology, and Bank of International Settlements for helpful comments.

†Princeton University, NBER, CAFR (e-mail: hhong@princeton.edu),

‡Guanghua School of Management, Peking University (e-mail: jiangminxu@gsm.pku.edu.cn)

1 Introduction

A growing literature points to the importance of latent social networks for the investment and performance of mutual fund managers. For instance, Hong, Kubik, and Stein (2005) finds that managers located in the same city, say Boston, are more likely to own and trade similar stocks headquartered in Kansas City, than managers located in other cities. They argue that this excessive covariance is indicative of word-of-mouth communication among managers in the same network. Cohen, Frazzini, and Malloy (2008) finds that managers' most profitable stock picks are of companies where the CEOs are likely to have been an alumni of the same university as the manager, thereby suggesting that social networks not only influence investments but improves performance as well.

Needless to say, being able to systematically identify whether a manager is part of a social network would be very valuable for understanding managerial investments and performance. However, even with the advent of Facebook and Linkedin, data on managerial networks is still limited. As such, we develop a measure of whether a manager is part of a network that relies only on data on the stock picks of managers, which are widely available due to regulatory disclosure requirements. Specifically, we propose a measure of whether a manager has a network in different cities by counting the number of her picks of stocks headquartered in these cities.

To get an intuition for our measure, we plot in Figure 1 a histogram of the count of manager-held stocks headquartered in Seattle (left-panel) and San Diego (right-panel) using the holdings of 1066 mutual fund managers reported in the fourth quarter of 2004. The x-axis is the number of stocks held by a manager. The y-axis is the frequency of managers. For Seattle, the mean of the count of stocks held by the mutual fund manager population is 2 with a variance of 4.2. For San Diego, the mean is 1.3 and the variance is 12.7.

The count distribution for Seattle is close to a Poisson distribution as the variance of the distribution is only somewhat larger than the mean of the distribution. Yet for San Diego, the variance is many times greater than the mean, i.e. the distribution is an Overdispersed Poisson

Distribution (or Negative Binomial). The degree of overdispersion captures the extent to which a few managers own many stocks in San Diego while most own few. We then make the inference that there is a San Diego network of investors while there is none or a small one for Seattle. That is, the few managers who own many stocks in San Diego are more likely to be part of a network that guides them toward San Diego stocks, while no such network presence exists for Seattle.

We show that this overdispersion measure is both theoretically well-motivated and empirically powerful in predicting mutual fund performance. Our point of departure is the work by statisticians Zheng, Salganik, and Gelman (2006) on modeling answers to survey questions from sociology (Killworth, Johnsen, McCarty, Shelley, and Bernard (1998); Killworth, McCarty, Bernard, Shelley, and Johnsen (1998), McCarty, Killworth, Bernard, Johnsen, and Shelley (2001)) about the count of friends a person has in different groups within the general population. A surprising answer is that the mean number of prisoners people know is one. Needless to say, most of us do not know even one prisoner. Indeed, the median is zero but there is a fat right tail to this count distribution where some people know a lot of prisoners. In contrast, answers to "how many people you know named Nicole" is roughly Poisson. Zheng, Salganik, and Gelman (2006) show the fat right tail of this count distribution is informative of the presence of social networks — presumably people who have been in prison probably know a lot of prisoners.

To derive our measure, we extend a standard portfolio choice model with risk-neutral managers who face a fixed cost of owning a stock. The fixed cost for a given stock is uniformly distributed across managers. But a manager's fixed cost of owning a stock headquartered in a given city falls with the number of contacts the manager has in that city. Stocks are assumed to have positive expected returns. Hence when a manager's fixed cost adjusted by the number of contacts is low enough, she owns the maximum allowable shares subject to exogenous regulatory limits which typically cap ownership at a small percentage of shares outstanding. As a result,

shares need not be informative of network presence in our setting.

Rather, the number of stocks in a city held by a manager, which is proportional to the manager's number of contacts in that city, becomes the key metric to measure managerial social networks in a city. Zheng, Salganik, and Gelman (2006) make use of an important result from Erdös and Rényi (1959) — namely, if connections are formed randomly, then the count of the number of friends a person has in any group (in our context a city) follows a Poisson distribution. In contrast, being part of a network means that people from certain groups have non-i.i.d. (i.i.d.: independently and identically distributed) propensities to form ties with each other. Excessive variance or overdispersion in contacts in a city then captures social connections or being part of a network in a city, in the same way that the prisoner population are formed in a non-i.i.d. manner as some people have a non-i.i.d. propensity to know prisoners.

We do not observe the count of contacts in a city but our model provides a monotonic transformation of the count of contacts y into the count of stock picks z in a city, which we approximate using a linear function of the form $z = cy$. We then extend the inference strategy of Zheng, Salganik, and Gelman (2006) to test whether or not social connections are formed randomly or are independently and identically distributed (i.i.d.) as in the random networks model literature following Erdös and Rényi (1959). Employing panel data on the holdings of institutional investors in different cities, we use our linear transformation and Zheng, Salganik, and Gelman (2006)'s model to estimate the parameters of the latent managerial networks. In other words, from Figure 1, we can infer the parameters of underlying latent social networks that produce the count of stock picks across different cities by using observations on these counts.

If we have N managers, K cities, we end up estimating $N + 2K + 1$ parameters with $N \times K$ number of observations reflecting the number of stock holdings that managers have in different cities. The first N parameters allow for different managers to have different degrees of gregariousness or propensity to make a contact. Gregariousness as we explain below is not the same as being a part of a network. The other $2K$ parameters allow for different city sizes

(as captured by the number of stocks located in the city) and the degree of overdispersion to vary across cities. That is, we can estimate a different overdispersion parameter for each city, and the degree of overdispersion does capture the influence of networks on stock picks in each city. $N + 2K$ are the parameters governing the latent social networks. In addition, the last degree of freedom is needed to estimate the transformation parameter c of the number of social connections into the number of stocks.

The idea of our inference strategy is similar to the one behind the profile maximum likelihood technique (see, e.g. Murphy, Rossini, and van der Vaart (1997), Murphy and van der Vaart (2000)) used in transformation models where the underlying variable of interest y is a increasing transform of the observable variable z. It can be understood as follows. For each possible value of c, we first compute the maximum likelihood estimate of θ (denoting the $N + 2K$ parameters governing the social network distributions) and the corresponding maximal value of the log-likelihood, then we find the value of c such that the log-likelihood attains *the* maximum with the associated θ estimate.

In our empirical analysis of the mutual fund holdings data from 1980 until now, we are careful to drop index and sector funds. Using the top 20 biggest Metropolitan Statistical Areas (MSAs) in terms of where stocks are headquartered as groups, we find some overdispersion of the count of contacts in most cities. Nevertheless, there is only pronounced overdispersion in San Jose, Los Angeles, New York, and San Diego, where the overdispersion parameter is around 2. Under the null Poisson random social connections setting, the overdispersion parameter for any given city should be 1. These results are robust to excluding managers located in a given city (and hence our results are not simply a manifestation of local bias in Coval and Moskowitz (1999)) or controlling for fund styles of the managers.

Importantly, we use our model to calculate for each manager his relative propensity to have contacts in a city (RPC) and relate these managerial RPC scores to the manager's fund performance. Our model gives a prediction for the expected number of stocks any manager

should hold in a given city. A manager who holds a higher number of stocks than predicted is more likely to be part of that city (i.e. be part of that network). We sum up the manager's scores across all the cities to get our geographic concentration scores for each manager.

We regress fund performance on these managerial RPC scores (geographic concentration scores), while controlling for a host of the usual explanatory variables for fund performance. We find that managers with higher RPC scores outperform those with lower RPC scores by around 1.6% a year. Our findings here are reminiscent of the Industry Concentration Index of Kacperczyk, Sialm, and Zheng (2005). They find that managers who hold concentrated positions out-perform those that do not. Their interpretation is one of closet indexing as those with concentrated positions are less likely to be closet indexers. However, our measures and ICI are not very correlated and including ICI in the performance regression does not change the coefficient in front of our RPC score.

A simpler, though less theoretically well-motivated, rendition of our RPC scores is Herfindahl index of concentration of stock picks across cities. The two measures are highly correlated at around 0.5. However, our RPC score nonetheless contains substantially more forecasting power for managerial performance than this simpler Herfindahl measure. This is comforting and speaks to the value of the model.

We then relate our model's outputs to demographic information about the fund managers beyond the performance of their investments. If our RPC score is a result of manager social network effects, then the RPC score should be informative about the demographics of managers. The idea is that networks are correlated with demographics such as being female or male or being a Democrat or Republican. We have data on such demographic attributes of managers. We can then combine our RPC score with the demographic attributes of cities (e.g. being a republican-affiliated city) to then forecast the likelihood of managers having the same set of attributes. Thus a manager with concentrated picks in a Republican city or a younger city ought to predict that the manager is more likely to be a young Republican. We verify that this

is indeed the case. 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)).

Our contributions are two-fold. First, we introduce a new approach to the modeling of investors' social networks using their stock picks. The existing approach in economics and finance in modeling social interactions focused on excessive correlation of investors' actions due to them being part of the same group and sharing information (see, e.g., Glaeser and Scheinkman (2002)). Second, in applying these models to mutual fund manager networks, we tap into a vast and rich database of information about managers including their investment performance. As a result, we can systematically assess the value of such networks.

The paper is organized as follows. We describe the model in Section 2 and the data and estimation procedures in Section 3. We collect the result for mutual fund investors in Section 4, 5 and 6. We conclude in Section 7.

2 The Model

2.1 Stock Picks

Consider a one-period model in which $i = 1, \ldots, N$ investors choose to allocate their money to $j = 1, \ldots, J$ stocks (headquartered) across $k = 1, \ldots, K$ cities. For simplicity, the risk-free rate is taken to be zero. The net excess return of a stock j, given by r_j, has a binomial distribution: with probability π_j, $r_j = r_d$, and with probability $1 - \pi_j$, $r_j = r_u$. Hence the expected excess return of stock j, $\mathbb{E}(r_j) = \pi_j r_d + (1 - \pi_j) r_u$. We further assume that $r_d < 0 < r_u$ (so that there are no arbitrage opportunities) and $\mathbb{E}(r_j) > 0$ for all j (so that the expected excess return of any stock is positive).

Investors have an initial wealth of w and are risk-neutral. We normalize the initial wealth w to 1 for simplicity. To invest in a stock j in city k, an investor i has to pay a fixed cost of participation $C_{i,j,k}$. The cost is measured relative to the initial wealth since the latter is

normalized to 1. For an investor i, let $h_{i,j,k}$ denote the fraction of his wealth allocated to stock j in city k. Then the objective of the risk-neutral investor i is then to maximize with respect to $\{h_{i,j,k}\}$ (over j across k) the following

$$1 + \sum_{k} \sum_{j,\, j \in k} \big(h_{i,j,k}\, \mathbb{E}(r_j) - C_{i,j,k} \big).$$

Because of risk-neutrality, if there is no fixed participation cost, an investor i would simply choose $h_{i,j,k}$ to be the maximum value that is allowed for any j as $\mathbb{E}(r_j) > 0$. We denote this max value by $\overline{h}_j$ for stock j. However, if there are participation costs, then an investor i would only invest in stock j if $\overline{h}_j\, \mathbb{E}(r_j) - C_{i,j,k} > 0$, or $\overline{h}_j\, \mathbb{E}(r_j) > C_{i,j,k}$, i.e. the gain from participating in stock j outweighs the associated participation cost. We write $c_{i,j,k} = C_{i,j,k} \,/\, \overline{h}_j$, so that an investor i would choose to participate in stock j (headquartered in city k) if $\mathbb{E}(r_j) > c_{i,j,k}$.

Each investor knows a number of friends in different cities. We denote $y_{i,k}$ as the number of friends that an investor i has in city k. For any investor i, the existence of his friends in city k is assumed to reduce his participation costs for stocks that are headquartered in city k, and the more friends an investor has in city k, the lower his participation costs for stocks in that city. To be more specific, we assume that the friends $y_{i,k}$ an investor i has in city k reduces his participation cost for stock j from $c_{i,j,k}$ to

$$\frac{c_{i,j,k}}{f(y_{i,k})}, \text{ for stock } j \in \text{city } k,$$

where $f(\cdot)$ is an increasing function. This assumption is used in the social interaction literature on stock market participation whereby costs falls because of peer effects through observational learning for instance (see, e.g. Hong, Kubik, and Stein (2004)). Consequently, in the presence of friends, an investor i would participate (invest) in stock j at city k if

$$\mathbb{E}(r_j) > \frac{c_{i,j,k}}{f(y_{i,k})}, \text{ for stock } j \in \text{city } k$$

We let $c_{i,j,k}/f(y_{i,k}) = c_{i,j,k}$ if $y_{i,k} = 0$ (i.e. participation costs stay the same if an investor i does not have any friends in city k), and $f(y_{i,k}) > 1$ if $y_{i,k} > 0$ (which is to ensure that having friends in a city lowers the participation costs for stocks in that city).

Next, we suppose that the information acquisition (participation) cost $c_{i,j,k}$ follows a uniform distribution over $[0, \bar{c}]$, where $\mathbb{E}(r_j) \in (0, \bar{c}]$ for all j. For ease of exposition, we further let the maximum participation cost $\bar{c}$ be such that $f(y_{i,k})\mathbb{E}(r_j) \in (0, \bar{c}]$ for all $\{i, j, k\}$ triplets. Hence the probability that an investor i invests (participates) in stock j headquartered in city k, is

$$\mathbb{P}\bigg(\mathbb{E}(r_j) > \frac{c_{i,j,k}}{f(y_{i,k})}\bigg) = \mathbb{P}\Big(c_{i,j,k} < f(y_{i,k})\mathbb{E}(r_j)\Big) = \frac{f(y_{i,k})\mathbb{E}(r_j)}{\bar{c}}$$

We now turn to the number of stock picks an investor has in a city. Let $z_{i,k}$ be investor i's number of picks of stocks that are headquartered in city k. Then $z_{i,k}$ is simply the total number of city k's stocks that investor i has invested in, i.e.

$$z_{i,k} = \sum_{j,\, j\in k} \mathbf{1}\bigg\{\mathbb{E}(r_j) > \frac{c_{i,j,k}}{f(y_{i,k})}\bigg\} = \sum_{j,\, j\in k} \mathbf{1}\big\{c_{i,j,k} < f(y_{i,k})\mathbb{E}(r_j)\big\}, \tag{2.1}$$

where $\mathbf{1}\{\cdot\}$ is an indicator function.

To simplify things, let us assume that $f(\cdot)$ is linear, so that $f(y_{i,k}) = \theta y_{i,k}$ for some constant θ. Then the number of stock picks an investor i has in city k becomes

$$z_{i,k} = \sum_{j,\, j\in k} \mathbf{1}\bigg\{\mathbb{E}(r_j) > \frac{c_{i,j,k}}{f(y_{i,k})}\bigg\} = \sum_{j,\, j\in k} \mathbf{1}\big\{\frac{c_{i,j,k}}{\theta\mathbb{E}(r_j)} < y_{i,k}\big\}, \tag{2.2}$$

It is easy to see from the above that $z_{i,k} = h(y_{i,k})$ is monotonically increasing in $y_{i,k}$. In our baseline model, we will approximate the relationship using a linear function $z_{i,k} = cy_{i,k}$.[1]

[1]We can also use a spline function to approximate $h(\cdot)$ in our inference strategy. Both strategies yield similar results. These results are available upon request from the authors.

2.2 Structure of Social Network

Following Zheng, Salganik, and Gelman (2006), we use the following notations for the social networks between investors and their acquaintances in different cities, which determines $y_{i,k}$. Here, "group" is used interchangeably with "city". Furthermore, we have

$$
\begin{aligned}
p_{ij} &: \text{probability that investor } i \text{ knows person } j \ , \\
a_i \equiv \textstyle\sum_{j,\, j\neq i} p_{ij} &: \text{gregariousness (the } \textit{expected} \text{ total number of connections) of investor } i \ , \\
b_k \equiv \frac{\sum_{i=1}^{N} a_i}{\sum_{i\in S_k} a_i} &: \text{proportion of total social connections that involves group } k, \\
&\quad \text{where } S_k \text{ stands for "group } k\text{"} \ , \\
\lambda_{ik} \equiv \textstyle\sum_{j\in S_k} p_{ij} &: \text{investor } i\text{'s } \textit{expected} \text{ number of connections in group } k \ , \\
g_{ik} \equiv \lambda_{ik}/(a_i b_k) &: \text{investor } i\text{'s } \textit{expected} \text{ relative propensity to befriend with people in group } k \ .
\end{aligned}
$$

2.2.1 The Null (Poisson) Model

If investors' social connections are independently and identically formed as in the classical model of Erdös and Rényi (1959), the probability p_{ij} of a link between an investor i and a person j from any particular group is the same for all pairs (i, j). It then implies that y_{ik} follows a Poisson distribution with a probability function:

$$
f(y_{ik}|a, b_k) = \frac{(ab_k)^{y_{ik}} \exp(-ab_k)}{y_{ik}!} \, ,
$$

where its mean $\lambda_{ik} = ab_k$ is equal to its variance. Furthermore, this model results in equal expected gregariousness a_i for all investors and relative propensities g_{ik} all equal to one.

However, some investors may be more gregarious and have more social ties in expectation. To account for the variability in gregariousness, we let parameters $\{a_i\}$ vary across individual investors. Hence y_{ik} follows a Poisson distribution with a mean $\lambda_{ik} = a_i b_k$ and a probability

function

$$f(y_{ik}|a_i, b_k) = \frac{(a_i b_k)^{y_{ik}} \exp(-a_i b_k)}{y_{ik}!},$$

but relative propensities g_{ik} are still all equal to one. We call this our null model.

2.2.2 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 being part of a network from being merely gregarious. Being part of a network would mean that some investors have a non-i.i.d. relative propensity $\{g_{ik}\}$ to make connections to certain groups since the people in those groups constitute a structured network. As a result, we allow investors to differ not only in their gregariousness $\{a_i\}$, but also in their relative propensity $\{g_{ik}\}$ to accommodate for the effect of social influence. Consequently, $g_{ik} > 1$ if investor i has a higher relative propensity to connect to people from group k than an average investor in the population.

In the most general form where $\{g_{ik}\}$ varies for each (i, k) pair, y_{ik} is distributed as Poisson with a mean $\lambda_{ik} = a_i b_k g_{ik}$. Since it is not possible to identify each g_{ik} later in the estimation if they are all different, for each group k, we let g_{ik} follow a gamma distribution with a mean equal to 1 and a variance equal to $(\omega_k - 1)$ where $\omega_k > 1$.[2] As a standard result, 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)[3]

$$f(y_{ik}|a_i, b_k, \omega_k) = \frac{\Gamma(y_{ik} + \zeta_{ik})}{\Gamma(\zeta_{ik})\,\Gamma(y_{ik} + 1)} \left(\frac{1}{\omega_k}\right)^{\zeta_{ik}} \left(\frac{\omega_k - 1}{\omega_k}\right)^{y_{ik}}, \tag{2.3}$$

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

[2]The reason that it is not possible to identify all of the g_{ik}'s if each one of them is a different constant is because we only have $N \times K$ number of observations of investors' stock picks. It is then not feasible to estimate $N \times K$ number of g_{ik}'s with only $N \times K$ number of data points.

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

$a_i b_k$ and a variance $\omega_k a_i b_k$ that is greater than its mean ($\omega > 1$). Therefore, 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 investors' non-identicalness in forming ties to certain groups and being part of structured social networks.

2.3 Transformation Parameter and Likelihood Function in Terms of Stock Picks

From our linear approximation of $y_{ik} = z_{ik}/c$, we can then rewrite the above negative binomial density of friends y_{ik} in terms of stock picks z_{ik} and the transformation parameter c:

$$f(y_{ik}|a_i, b_k, \omega_k) = f(z_{ik}/c \,|\, a_i, b_k, \omega_k, c) = \frac{\Gamma(z_{ik}/c + \zeta_{ik})}{\Gamma(\zeta_{ik})\,\Gamma(z_{ik}/c + 1)} \left(\frac{1}{\omega_k}\right)^{\zeta_{ik}} \left(\frac{\omega_k - 1}{\omega_k}\right)^{z_{ik}/c}, \quad (2.4)$$

Our primary goal is to estimate the overdispersion parameters $\{\omega_k\}$ from our overdispersed model and thus learn about diversities that exist in the formation of investors' social networks. As a byproduct, we also estimate the gregariousness parameters $\{a_i\}$ that represent the expected number of acquaintances known by investor i, the group size parameters $\{b_k\}$ that gauge the proportion of social connections involving group k, and the transformation parameter c that approximates the increasing relationship between the number of stock picks and the number of acquaintances.

To make notations clear, we will write the likelihood function of our model directly in terms of z_{ik} and the transformation parameter c. Following from the density expression in (2.4) (or (2.3)), the likelihood function of $z = \{z_{ik}\}$ in our overdispersed model is

$$p(z_{ik}/c \,|\, a, b, \omega, c) = \prod_{i=1}^{N} \prod_{k=1}^{K} \frac{\Gamma(z_{ik}/c + \zeta_{ik})}{\Gamma(\zeta_{ik})\,\Gamma(z_{ik}/c + 1)} \left(\frac{1}{\omega_k}\right)^{\zeta_{ik}} \left(\frac{\omega_k - 1}{\omega_k}\right)^{z_{ik}/c},$$

and the log-likelihood

$$\mathcal{L} = \sum_{i=1}^{N}\sum_{k=1}^{K}\Big(LG\big(z_{ik}/c + \zeta_{ik}\big) - LG(\zeta_{ik}) - LG\big(z_{ik}/c + 1\big) \\ - \zeta_{ik}\log(\omega_k) + z_{ik}/c\big[\log(\omega_k - 1) - \log(\omega_k)\big]\Big). \tag{2.5}$$

where $LG(\cdot)$ here denotes the log-gamma function $\log(\Gamma(\cdot))$ and $\zeta_{ik} = a_i b_k/(\omega_k - 1)$ as stated before.

The parameters of interest in our model are $\theta = \big(\{\omega_k\}_{k=1}^K, \{a_i\}_{i=1}^N, \{b_k\}_{k=1}^K, c\big)'$, a $(N + 2K + 1) \times 1$ vector. We will estimate these parameters using $N \times K$ observations of number of stock picks z_{ik}. We estimate our model parameters using the method of maximum likelihood (MLE) based on (2.5), and we normalize $\sum_{k=1}^K b_k$ to one to separately identify $\{a_i\}$ and $\{b_k\}$.[4] The estimation procedure is further discussed in the next section.

Before moving on to the model estimation stage, we provide some intuition for our MLE method. The idea is similar to the one behind the profile maximum likelihood technique (see, e.g. Murphy, Rossini, and van der Vaart (1997), Murphy and van der Vaart (2000)) used in transformation models where the underlying variable of interest y is a increasing transform of the observable variable z. It can be understood as follows. For each possible value of c, we first compute the maximum likelihood estimate of θ and the corresponding maximal value of the log-likelihood, then we find the value of c such that the log-likelihood (2.5) attains *the* maximum with the associated θ estimate.

[4]This normalization is needed because $\{a_i\}$ and $\{b_k\}$ enter the log-likelihood function together only as a joint entity $a_i b_k$.

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 Thompson Reuters for the period 1980–2011. The database sources from semi-annually mandatory filings to the SEC and quarterly voluntary disclosure by mutual funds. We then merge the CDA/Spectrum database with survivorship-bias free 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. Additionally, we augment our mutual fund data with the database used in Hong and Kostovetsky (2012), which contains managerial demographic information on age, gender, name and location of undergraduate college, median SAT score of the undergraduate college attended, having a graduate degree or not, and political affiliation.

In order to keep only actively managed, non-sector domestic equity funds in our sample, we apply the following detailed screening procedures. Firstly, 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 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 S&P500 index fund, and (4) their names not to contain anything in the vicinity of the word "index". Secondly, to exclude sector funds, we require funds' CRSP investment objectives at the third level to be either (C) or (Y). Thirdly, 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 1744 unique actively managed, non-sector domestic equity funds, or 117467 fund-quarter observations on stock holdings.[5]

[5]On average, approximately 920 funds reported their portfolio holdings information in a single quarter, The frequency of reporting peaked at 2005Q2 when around 1550 funds filed their holdings information.

Next, we categorize the stocks held by mutual funds investors 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).[6]

We shall only consider the largest 20 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. Hence in what follows, the number of groups K is fixed at 20.

3.2 Rolling Estimation

We shall conduct a rolling maximum likelihood estimation on the model's parameters $\theta = (\{\omega_k\}, \{a_i\}, \{b_k\})'$ and the transformation parameter c using the mutual fund holdings data. To be more precise, at each point in time (quarter for mutual funds), we will use the past 12 quarters of holdings data as a rolling subsample to estimate θ and c based on the log-likelihood (2.5). The observations z_{ik} are then the number of unique stock picks from a city k made by an investor i during the past 12 quarters. Therefore, our rolling estimates start at 1983Q1 and end at 2011Q4 for mutual funds.

After obtaining the rolling estimates, we will follow Fama and MacBeth (1973) in taking the time series means of the rolling estimates to form our overall estimates of θ and c. We denote these Fama-MacBeth estimates as our estimated parameter values.

[6]We would like to thank Hyun-Soo Choi from the Singapore Management University for providing us with the MSA information.

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 city groups.

4.1 Transformation Parameter

Table 1 presents the estimates (Fama-MacBeth means of the quarterly rolling estimates) and related summary statistics of the transformation parameter. It shows that the mean of the transformation parameter c is 1.39 with a standard deviation of .09 over time. There is not much variation over time in this parameter. This parameter estimates suggest that the number of contacts in a group is the number of holdings in that group divided by 1.39.

4.2 Gregariousness Parameters

Next, Table 2 shows the summary statistics of the estimated values of the gregarious parameters a_i and Figure 3 illustrates the histogram of their Fama-MacBeth averages. We observe that the mean of a_i is 102. This estimate can be interpreted literally as the typical manager having around 102 friends in the mutual fund industry overall and just in our sample. But there is a fairly sizeable standard deviation of around 113 or so friends. The estimate does not seem out of bounds relative to results in the sociology literature on the number of friends people have more generally.

Nevertheless, we view the estimates of gregariousness parameters as more akin to investor fixed effects for some investors having more stocks than others. 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 since it could also be affected by other factors such as investment style.

4.3 City Group Size Parameters

We then report the parameter estimates for b_k that gauge the relative sizes of cities. Table 3 and Figure 4 demonstrate the values of b_k for the 20 cities. Two aspects of the estimates are noticeable. First, there are a few groups that have a much larger number of potential social connections attached to them comparing to the rest, for example, New York and LA. However, a group having a larger b_k does not imply that the degree of overdispersion in the group would necessarily be higher. To put it another way, just because a city has a substantial (relative) size does not mean that investors are more likely to form structured social networks with individuals from that group. Second, most of the standard deviations of the Fama-MacBeth b_k estimates are small, implying that the sizes of various groups are stable across time.

4.4 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 groups. 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. For groups where ω_k is closer to 1, there is not much variation in these relative propensities. In contrast, larger values of ω_k imply dissimilarities in individuals' relative propensities to make connections.

Table 4 and Figure 5 display the estimated overdispersion parameter ω_k for the cities. There are three evident features. First, New York, Los Angeles, San Jose and San Diego stand out as the most overdispersed cities compared to the rest. This suggests that investors are more likely to form and be part of structured networks with acquaintances from these cities. There is a greater standard deviation of these estimates but the rankings of cities remain fairly stable over time.

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 0.305, and the rank correlation between them is merely about 0.236.

Third, and most importantly, although the majority of the cities do not exhibit a substantial degree of overdispersion, the t-statistics of testing the null Poisson distribution of $\omega = 1$ are all significant at the 5% level. This clearly indicates that our null hypothesis (model) of randomly formed managerial social networks is firmly rejected. Hence it implies that some managers do belong to certain integrated social networks even in the smaller cities (in terms of b_k) such as Miami or Minnesota. The overdispersion estimates therefore signify that a number of managers live among some intricate social networks. They do not have to be the most gregarious managers, nor are they necessarily tied to the largest cities or industries.

4.5 Robustness Checks: Controlling for Local Bias and Fund Styles

Lastly, we report the results of two robustness checks on the overdispersion estimates using mutual fund data with city groups. The first check is a local-bias check, where we exclude managers' local stock holdings from the estimation to ensure that our overdispersion results are not due to local biases. The other one is a verification where we dropped all growth funds from the estimation to ensure our results are not driven by fund styles.

As can be seen clearly from Table 5, the results from the two robustness checks echo our earlier findings in Table 4. Thus it implies that the overdispersions we find are not subject to the influence of either local biases or fund styles, and once more the social connections of managers are not formed in an i.i.d. manner.

5 Predicting Managerial Performance Using Model's Estimates of Managerial Relative Propensity to Connect (RPC)

We use our model's output to generate for each manager his relative propensity to be connected (RPC) in a non-i.i.d. way to acquaintances in different cities. Recall that in our model, investors' *expected* relative propensities to know a member in city k, $g_{ik} = \lambda_{ik}/(a_i b_k)$, cannot be identified or estimated individually. The RPC measures that we construct, $RPC_{ik} = y_{ik}/(a_i b_k)$, can then be considered as a proxy for g_{ik}. In other words, the RPC measures can be thought of as investors' *realized* relative propensities to know a member from a specific city. The RPC *measure* for any investor in a particular city k is computed as $g_{ik} = y_{ik}/(a_i b_k)$.[7] Our model predicts that an investor should have an expected number of $a_i b_k$ connections in a given city, and that y_{ik} should be very close to $a_i b_k$ if connections are formed in an Erdös and Rényi (1959) i.i.d. manner. On the other hand, an investor who holds a (much) higher number of stocks and hence knows a (much) larger number of acquaintances than expected in a city is more likely to be part of and has $g_{ik} > 1$ in that city, i.e. being part of that network.

Then we sum up investors' RPC measures across all the cities, i.e. $gsum_i = \sum_{k=1}^{K} [y_{ik}/(a_i b_k)]$. We shall label this the RPC *score* for each investor and will use $gsum_i$ interchangeably with RPC score. Furthermore, if social connections are formed in an i.i.d. fashion so that $y_{ik}/(a_i b_k)$ are around 1 for each (i, k) pair, we would expect all the RPC scores $\{gsum_i\}$ to be close to 20 as we have $K = 20$ cities. However, if there are structured social networks among various cities, we would anticipate $gsum_i > 20$ for an investor i who is part of networks. This is because his underlying *true* $\sum_k g_{ik} = \lambda_{ik}/(a_i b_k)$ is likely to be greater than 20 as a result of social influences.

[7] Strictly speaking, this should be denoted as $\hat{g}_{ik} = y_{ik}/(\hat{a}_i \hat{b}_k)$ (where $\hat{a}_i$ and $\hat{b}_k$ are our estimates) since it is not the real g_{ik} that equals $\lambda_{ik}/(a_i b_k)$. However, as stated before, we do not estimate individual g_{ik} value in our model. Hence this notation is unlikely to cause any major confusion in what follows and we will denote g_{ik} to mean $y_{ik}/(\hat{a}_i \hat{b}_k)$. In addition, we will use g_{ik} and RPC measure interchangeably.

Table 6 illustrates the correlations between our RPC measures g_{ik} and our gregariousness parameter estimates a_i, using their respective Fama-MacBeth averages. It is clear from the table that the correlations between g_{ik} and a_i are rather mild for city groups. Such weak correlations further confirm that being gregarious and being part of a network are not one and the same.

The summary statistics for our RPC scores $gsum_i$ are demonstrated in Table 7 as well as in Figure 6. We notice that the mean of RPC scores are close to 19, yet the standard deviation (around 8.48) is sizeable. Once more, this is another piece of evidence showing that certain managers have non-i.i.d. propensities to form ties with members from different cities. Furthermore, we find in Table 8 that for investors who have RPC scores greater than 20 (i.e. they are part of certain networks), the number of cities in which they have RPC measures larger than 1 is approximately eight. It indicates that for investors who are part of networks, they have higher propensities of making connections to certain cities only but not to all of the cities.

5.1 Managerial RPC and Fund Performance

Now we turn our attention to a more important question, which is how social networks, i.e. our RPC scores $gsum$, are related to mutual fund performances. There is a range of existing literature suggesting that social networks could exert positive values on investment performances, e.g. Hong, Kubik, and Stein (2005), Cohen, Frazzini, and Malloy (2008) and Feng and Seasholes (2008). Networks, such as knowing someone who is the CEO of a company, are not easy to obtain and may contain valuable investment information not accessible by the common public. Based on these ideas, the presence of structured networks in our model would imply that investors with RPC scores (much) larger than 20 should earn higher returns on their investment portfolios. Consequently, active equity funds with larger RPC scores should enjoy higher performances than their counterpart with smaller scores.

To test such implications, we utilize the following regression specification from Chen, Hong,

Huang, and Kubik (2004) to examine the effect of social networks on mutual fund performances:

$$pfm_{i,t} = \alpha + \beta RPCdummy_{i,t-1} + x'_{i,t-1}\gamma + \varepsilon_{i,t} \ , \tag{5.1}$$

where the dependent variable $pfm_{i,t}$ is fund i's net return in quarter t. $RPCdummy_{i,t-1}$ is a dummy variable that equals one if fund i's RPC score $gsum_i$ is greater than 20 in quarter t. Furthermore, $x'_{i,t-1}$ is a vector of standard fund characteristic controls at (quarter) $t-1$. They include: (1) fund i's lagged net return, (2) log of the total net asset of fund i, (3) log of one plus the total net asset of other funds in fund i's family, (4) the expense ratio of fund i, (5) the turnover ratio of fund i, and (6) fund i's age. Additionally, we also control for the gregariousness of a manager via his $\log(a_i)$ and for whether a fund is located in a financial center (which is found by Christoffersen and Sarkissian (2009) to be associated with superior performance).[8] They are contained in the regressor x and are both measured at $t-1$ as well. Finally, α is a constant term and $\varepsilon_{i,t}$ is a generic error term uncorrelated with all other explanatory variables in equation (5.1). We will carry out the regression (5.1) quarter by quarter and then take the Fama-MacBeth time-series means and Newey-West standard errors of the quarterly estimates.

Table 9 depicts our fund performance regression results. Most of the regression coefficients come in with the expected signs given the results in Chen, Hong, Huang, and Kubik (2004). For instance, fund size (log TNA) is associated with poor returns. There is persistence in performance and expense ratio is associated with poor returns. Moreover, we find consistent with Christoffersen and Sarkissian (2009) that a fund located in a financial center has superior performance.

Most relevant for us, it is evident that fund managers with higher RPC scores (i.e. with $gsum > 20$) outperform substantially, by close to 1.6% a year. However, we notice that being gregariousness does not necessarily lead to outperformance, as the coefficient on $\log(a_i)$ is close

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

to zero and is insignificant. Thus this difference in generating superior performance supports our prediction that being gregarious is not the same as being part of networks.

The findings on the influence of RPC scores on mutual fund performances here are reminiscent of the Industry Concentration Index (ICI) of Kacperczyk, Sialm, and Zheng (2005). They find that managers who hold concentrated positions outperform those that do not. Their interpretation on ICI is along the lines of closet indexing as those with concentrated portfolio holdings are less likely to be index-fund mimickers. However, our RPC scores and ICI are not very correlated and including ICI in the performance regression does not change the coefficient in front of our RPC scores. This is shown in column (2) of Table 9 where ICI is included as an extra explanatory variable in the regression specification of (5.1). In addition, our result that social networks are valuable to the tune of 1.6% a year for mutual fund returns is evocative of earlier studies documenting the value of investor and CEO networks such as Cohen, Frazzini, and Malloy (2008) and Engelberg, Gao, and Parsons (2012).

5.2 A Related Herfindahl Measure in Predicting Performance

In order to further understand the fund performance result, we develop a simpler measure of the geographic concentration of managers' stock picks as a complement to our RPC score. In essence, the related measure that we have is a Herfindahl index on the number of stock picks over the 20 biggest cities for fund managers. The idea is that, if a manager has a high RPC score, then he is likely to skew his stock picks towards some cities because of the influence of social networks. As a result, the manager's uneven numbers of stock selections would lead to a high Herfindahl index that is computed based on his stock picks across cities.

Our Herfindahl index H_i of the stock picks across the 20 cities for a manager i in any quarter is constructed as follows:

$$H_i = \sum_{k=1}^{20} s_{i,k}^2 \ ,$$

where $s_{i,k}$ is equal to the manager i's number of stock picks in city k divided by the manager i's total number of stock picks across the 20 cities. $s_{i,k}$ is basically the "share" of the manager's number of picks in city k within his total number of picks.

Table 10 then illustrates the summary statistics of the Herfindahl index for all managers, while Table 11 shows the summary statistics of the correlation between the index and our RPC score $gsum$ across all quarters. The noticeable correlation of around 0.5 demonstrates the complementarity of the Herfindahl index of stock picks and our RPC score. In other words, our model provides a rationale for using Herfindahl-like indices of concentration in stock picks.

Next, we run fund-performance regressions using the same specification as in (5.1), but with the RPC score (dummy) replaced by the Herfindahl index (dummy). The corresponding results are depicted in Table 12. It is clear that fund managers with Herfindahl index numbers above the upper quartile value would outperform others by close to 60 basis points per year on average, which is consonant with the superior performance number that we found earlier using the social-network measure $gsum$. However, this outperformance number is noticeably smaller than the counterpart produced by the RPC score, which is around 1.6%. Furthermore, when we perform the regression with both the RPC score and the Herfindahl index included as shown in column (3) of Table 12, we also find that the RPC score does a better job at forecasting superior investment performance than the simpler Herfindahl index. Therefore, this validates the usefulness of our RPC score in capturing the effect of social networks on the geographic concentration of stock picks of fund managers. More importantly, it showcases the predictive power of RPC score on investment performance.

6 Using RPC Score to Predict Managers' Demographic Characteristics

If our RPC score is a result of manager social network effects, then the RPC score should be informative about the demographics of managers. Our approach is similar to a literature in computer science that tries to infer latent social networks from internet data (see, e.g. Adamic and Adar (2005)). The idea is that networks are correlated with demographics such as being female or male or being a Democrat or Republican. We have data on such demographic attributes of managers. We can then combine our RPC score with the demographic attributes of cities (e.g. being a republican-affiliated city) to then forecast the likelihood of managers having the same set of attributes. Therefore, a manager with concentrated picks in a Republican city or a younger city ought to predict that the manager is more likely to be a young Republican.

We focus on four types of sociodemographic characteristics for managers: being a female (gender), being younger than 45 years old (age), being a Republican (political affiliation), and whether they went to college in cities where most of the managers received their undergraduate education (school).[9] Next, for each of these characteristics, we reconstruct the RPC score $gsum$ for managers in the following way. Instead of adding up the values of g_{ik} over all 20 cities to reach our original $gsum$ score for each manager, we sum the g_{ik} values only from those cities that have the desired demographic attribute. To be more precise, for the gender attribute, we use g_{ik}'s from cities that have a relatively high female-to-male sex ratio. For the age attribute, we look at cities that have a relatively high proportion of adults between the age of 25 and 45. And for the political affiliation attribute, we consider cities that are republican-affiliated. Finally, for the school attribute, we tabulate the summary statistics of proportions of managers who attended undergraduate schools in different cities in Table 13, and consider cities that have a relatively high college-attendance proportion ("university cities").

[9] We call these "university" cities that will be made more precise in a moment.

Consequently, for each type of manager demographic characteristics, we have a corresponding reconstructed $gsum$ score based on cities that are of the desired demographic nature. We will denote (the logs of) these four sets of "demographic" scores respectively as $gsumFEM$ for gender ("female"), $gsumYNG$ for age ("young"), $gsumREP$ for republican, and $gsumSCH$ for school (attended college in a "university city").[10]

We are now in a position to examine whether managers having a high demographic $gsum$ score of a particular type would increase the likelihood of them having that particular demographic attribute. Since the demographic attributes of managers are all binary variables, we will use logit regressions with the demographic $gsum$ score as an regressor and the manager demographic attribute as the dependent variable. Our logit regression models have the following form of probability distribution

$$\mathbb{P}(Y_i|\alpha, x_i, gsumY_i) = f(\alpha + x_i'\beta + gsumY_i\,\gamma)^{Y_i}\big(1 - f(\alpha + x_i'\beta + gsumY_i\,\gamma)\big)^{1-Y_i},$$

where the binary outcome variable Y_i denotes whether manager i has the specific demographic attribute Y, $gsumY_i$ denotes the *log* of manager i's demographic $gsum$ score with the nature Y, x_i is a set of other manager demographic controls except the attribute Y and $f(\cdot)$ denotes the standard logistic function $f(z) = 1/(1 + e^{-z})$.[11] As discussed above, if our underlying model is

[10]We use the 2000 US census data to determine which cities have a relatively high female-to-male sex ratio or have a relatively large proportion of adults between the age of 25 and 45. To be fair, for the 20 cities, we take the top 10 cities with the highest female-to-male sex ratio to construct $gsumFEM$ or with the largest proportion of adults between the age of 25 and 45 to construct $gsumYNG$. The 10 cities with the highest female-to-male sex ratio are New York, Philadelphia, St. Louis, Stamford, Boston, Miami, Detroit, Washington, Chicago and Atlanta. The 10 cities with the largest proportion of adults between the age of 25 and 45 are San Diego, Los Angeles, Houston, San Jose, New York, Washington, Seattle, Phoenix, Dallas and Atlanta. Finally, we determine republican-affiliated cities based on the election data in1996, 2000 and 2004 from David Leip's atlas of U.S. presidential elections at $http://uselectionatlas.org/$. There are 4 republican cities only throughout the years: Houston, Dallas, Phoenix and San Diego. Lastly, we use the statistics from Table 13 and take the top 10 cities with the highest median value of proportion of college-attending managers to construct $gsumSCH$. These "university" cities are New York, Los Angeles, Boston, San Francisco, Chicago, San Jose, Philadelphia, Washington, Miami and Minneapolis.

[11]Other manager demographic variables include the log of the median SAT score of the undergraduate school that a manager attended and whether a manager has a graduate degree. Depending on the logit model, they also include two of the following: being a female, being younger than 45 years old, being a republican and having attended college in a university city.

capturing manager social network effects, we would expect the sign of the coefficient γ in front of the demographic $gsum$ score to be positive. We shall estimate the logit model for each of the four demographic attributes in every quarter first, then take the Fama-MacBeth time-series means and Newey-West standard errors of the quarterly estimates. The estimation results are summarized in Table 14.

What we can see from Table 14 is that the four demographic $gsum$ variables all have positive coefficient estimates that are statistically significant. This means managers having a high demographic $gsum$ score of a particular type increases the probability of them having that particular demographic attribute. Moreover, we compute the marginal effect of these demographic $gsum$ variables in the logit model by evaluating the median SAT score and the demographic $gsum$ variable at their mean values, and set other demographic controls to 1. The marginal effects are 0.061, 0.173, 0.093 and 0.067 respectively for $gsumFEM$, $gsumYNG$, $gsumREP$ and $gsumSCH$. This suggests, for example, if a manager attended an undergraduate school with a median SAT score at the average level, and the manager has a graduate degree, is a female, is a Republican and attended college in a "university city", then she is almost 0.20 percentage point more likely to be below 45 years old than above when her $gsumYNG$ score increases by 1%. If instead the manager is below 45 years old but we do not know her political affiliation, the probability of her being a Republican would increase by close to 0.10 percentage point when her $gsumREP$ score rises by 1%.

7 Conclusion

There is a growing use of social networks to model phenomena from all corners of financial economics, beyond simply mutual fund managers. For instance, in the aftermath of the Financial Crisis of 2007, many have turned to the modeling of networks among banks and other financial intermediaries to explain financial contagion in the hopes of discovering a more stable financial

architecture (see, e,g., Allen and Gale (2007), Boyer, Kumagai, and Yuan (2006), Allen, Babus, and Carletti (2010)). Additionally, networks have also made their way to corporate finance as networks of CEOs, venture capitalists, entrepreneurs and banks are influential in allocating resources (see, e.g., Engelberg, Gao, and Parsons (2012), Lerner and Malmendier (2013), Shue (2013), Hochberg, Ljungqvist, and Lu (2010)).

Our count models of social networks in finance can be used to study these broader sets of financial networks where investment data are available. For instance, our set-up can be applied to banking networks where one can count trades between a bank with other banks in different countries or lending volume between banks and companies in different industries. In other words, while we do not have answers to survey questions about how many people investors know in different groups, we can proxy for answers to these questions by counting their investments across different categories. We leave these other applications of count models of social networks in financial markets for future research.

References

Adamic, L. A., and E. Adar, 2005, "How to Search a Social Network," *Social Networks*, 27, 187–203.

Allen, F., A. Babus, and E. Carletti, 2010, "Financial Connections and Systemic Risk," NBER Working Paper 16177.

Allen, F., and D. M. Gale, 2007, "An Introduction to Financial Crises," Wharton Financial Institutions Center Working Paper No. 07-20.

Boyer, B. H., T. Kumagai, and K. Yuan, 2006, "How Do Crises Spread? Evidence from Accessible and Inaccessible Stock Indices," *Journal of Finance*, 61, 957–1003.

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

Chen, J., H. Hong, M. Huang, and J. D. Kubik, 2004, "Does Fund Size Erode Mutual Fund Performance? The Role of Liquidity and Organization," *American Economic Review*, 94(5), 1276–1302.

Christoffersen, S. E., and S. Sarkissian, 2009, "City Size and Fund Performance," *Journal of Financial Economics*, 92, 252–275.

Cohen, L., A. Frazzini, and C. J. Malloy, 2008, "The Small World of Investing: Board Connections and Mutual Fund Returns," *Journal of Political Economy*, 116, 951–979.

Coval, J. D., and T. J. Moskowitz, 1999, "Home Bias at Home: Local Equity Preference in Domestic Portfolios," *Journal of Finance*, 54, 2045–2073.

Engelberg, J., P. Gao, and C. A. Parsons, 2012, "Friends with money," *Journal of Financial Economics*, 103, 169–188.

Erdös, P., and A. Rényi, 1959, "On Random Graphs," *Publicationes Mathematicae*, 6, 290–297.

Fama, E. F., and J. D. MacBeth, 1973, "Risk, Return, and Equilibrium: Empirical Tests," *Journal of Political Economy*, 81, 607–636.

Feng, L., and M. S. Seasholes, 2008, "Individual investors and gender similarities in an emerging stock market," *Pacific-Basin Finance Journal*, 16, 44–60.

Glaeser, E. L., and J. A. Scheinkman, 2002, "Non-Market Interactions," *Advances in Economics and Econometrics: Theory and Applications, Eighth World Congress,*, M. Dewatripont, L.P. Hansen, and S. Turnovsky (eds.), Cambridge University Press.

Hochberg, Y. V., A. Ljungqvist, and Y. Lu, 2010, "Networking as a Barrier to Entry and the Competitive Supply of Venture Capital," *Journal of Finance*, 66, 829–859.

Hong, H., and L. Kostovetsky, 2012, "Red and Blue Investing: Value and Finance," *Journal of Financial Economics*, 103, 1–19.

Hong, H., J. D. Kubik, and J. C. Stein, 2004, "Social Interaction and Stock-Market Participation," *Journal of Finance*, 59, 137–163.

——— , 2005, "Thy Neighbor's Portfolio: Word-of-Mouth Effects in the Holdings and Trades of Money Managers," *Journal of Finance*, 60, 2801–2824.

Kacperczyk, M., C. Sialm, and L. Zheng, 2005, "On the Industry Concentration of Actively Managed Equity Mutual Funds," *Journal of Finance*, 60, 1983–2011.

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

Killworth, P. D., C. McCarty, H. Bernard, G. A. Shelley, and E. C. Johnsen, 1998, "Estimation of Seroprevalence, Rape, and Homelessness in the United States Using a Social Network Approach," *Evaluation Review*, 22, 289–308.

Lerner, J., and U. Malmendier, 2013, "With a Little Help from My (Random) Friends: Success and Failure in Post-Business School Entrepreneurship," *Review of Financial Studies*, 26, 2411–2452.

McCarty, C., P. D. Killworth, H. Bernard, E. C. Johnsen, and G. A. Shelley, 2001, "Comparing Two Methods for Estimating Network Size," *Human Organization*, 60, 28–39.

Murphy, S., A. Rossini, and A. van der Vaart, 1997, "Maximum Likelihood Estimation in the Porportional Odds Model," *Journal of the American Statistical Association*, 92, 968–976.

Murphy, S., and A. van der Vaart, 2000, "On the Profile Likelihood," *Journal of the American Statistical Association*, 95, 449–465.

Newey, W. K., and K. D. West, 1987, "A Simple, Positive Semi-definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica*, 55, 703–708.

Shue, K., 2013, "Executive Networks and Firm Policies: Evidence from the Random Assignment of MBA Peers," *Review of Financial Studies*, 26, 1401–1442.

Zheng, T., M. J. Salganik, and A. Gelman, 2006, "How Many People Do You Know in Prison?: Using Overdispersion in Count Data to Estimate Social Structure in Networks," *Journal of the American Statistical Association*, 101, 409–423.

Table 1: Summary statistics of estimates of transformation parameter c for mutual funds

This table reports the summary statistics of the transformation parameter estimates (quarterly rolling estimates) using mutual fund data with 20 city groups as described in the main article. s.d. denotes standard deviation and med denotes median.

mean	s.d.	med	min	max
1.39	0.09	1.39	1.13	1.87

Table 2: Summary statistics of estimates of a_i for mutual funds

The table shows the summary statistics of the estimated values of the gregarious parameters $\{a_i\}$ using mutual fund data with 20 city groups as described in the main article. We first compute the time-series average of the quarterly a_i estimates for each individual fund i, then we report the summary statistics of these time-series averages. s.d. denotes standard deviation and med denotes median.

mean	s.d.	med	min	max
101.9	113.0	74.0	0.6	1497.0

Table 3: Estimates of 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, 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: Pheonix, SL: St. Louis, Det: Detroit. In addition, s.d. stands for standard deviation and med stands for median.

	mean	s.d.	med	min	max
NY	0.178	0.018	0.171	0.154	0.212
LA	0.073	0.013	0.068	0.060	0.102
Bos	0.069	0.003	0.069	0.064	0.075
SF	0.056	0.017	0.065	0.025	0.075
Chi	0.086	0.015	0.093	0.065	0.106
SJ	0.082	0.013	0.083	0.056	0.109
Dal	0.063	0.004	0.062	0.054	0.069
Hou	0.064	0.008	0.065	0.049	0.078
Phi	0.046	0.005	0.046	0.035	0.054
Was	0.045	0.005	0.045	0.036	0.054
Mia	0.019	0.002	0.020	0.014	0.023
Atl	0.036	0.003	0.037	0.025	0.042
Min	0.036	0.004	0.037	0.027	0.043
Den	0.019	0.004	0.019	0.013	0.027
SD	0.017	0.003	0.017	0.012	0.024
Stfd	0.032	0.007	0.035	0.022	0.045
Sea	0.023	0.003	0.024	0.016	0.027
Phx	0.016	0.004	0.018	0.009	0.021
SL	0.023	0.001	0.023	0.020	0.025
Det	0.016	0.003	0.016	0.012	0.021

Table 4: Estimates of 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. s.d. stands for standard deviation and med stands for median. The t-statistics are adjusted for serial correlation using Newey and West (1987) lags of order twelve since we use past twelve quarters as our rolling estimation window size. They test the null hypothesis of $\omega_k = 1$ (Poisson) against the alternative of $\omega_k > 1$ (overdispersion). For an explanation to the abbreviated city names, please refer to the note under Table 3.

	mean	s.d.	med	min	max	t-stat
NY	1.382	0.307	1.257	1.012	2.386	3.175
LA	1.322	0.297	1.217	1.012	3.167	2.826
Bos	1.147	0.174	1.075	1.002	1.660	3.220
SF	1.121	0.173	1.055	1.002	1.812	2.685
Chi	1.142	0.161	1.089	1.001	1.740	3.023
SJ	2.620	0.602	2.611	1.471	5.396	11.790
Dal	1.020	0.055	1.007	1.001	1.445	3.602
Hou	1.144	0.218	1.024	1.002	1.792	2.186
Phi	1.037	0.068	1.012	1.001	1.424	3.010
Was	1.038	0.074	1.010	1.001	1.437	3.091
Mia	1.216	0.228	1.101	1.003	2.182	3.163
Atl	1.045	0.074	1.008	1.001	1.419	2.772
Min	1.237	0.196	1.203	1.002	1.662	3.959
Den	1.190	0.155	1.149	1.001	1.696	5.103
SD	1.517	0.399	1.518	1.019	3.496	4.236
Stfd	1.015	0.051	1.005	1.001	1.407	3.474
Sea	1.069	0.073	1.041	1.002	1.476	5.563
Phx	1.053	0.078	1.023	1.002	1.518	3.241
SL	1.041	0.067	1.018	1.004	1.422	3.332
Det	1.056	0.089	1.019	1.003	1.494	3.513

Table 5: 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. "No Local Response" denotes the case where managers' local holdings have been dropped from the estimation, and "No Growth Fund" indicates that all growth funds have been dropped from the estimation. The t-statistics are adjusted for serial correlation using Newey and West (1987) lags of order twelve since we use past twelve quarters as our rolling estimation window size. They test the null hypothesis of $\omega_k = 1$ (Poisson) against the alternative of $\omega_k > 1$ (overdispersion). For an explanation to the abbreviated city names, please refer to the note under Table 3.

	No Local Response		No Growth Fund	
	mean	t-stat	mean	t-stat
NY	1.425	3.258	1.364	3.187
LA	1.384	3.093	1.335	3.058
Bos	1.127	3.844	1.177	3.014
SF	1.102	2.076	1.105	2.898
Chi	1.127	3.317	1.115	2.532
SJ	2.551	11.155	2.575	12.155
Dal	1.024	4.310	1.010	3.473
Hou	1.179	2.353	1.123	2.293
Phi	1.020	3.107	1.026	3.207
Was	1.040	2.914	1.027	3.080
Mia	1.173	3.185	1.281	3.232
Atl	1.031	2.503	1.036	3.154
Min	1.195	3.656	1.218	3.989
Den	1.204	4.833	1.206	4.088
SD	1.494	4.234	1.634	4.297
Stfd	1.012	4.265	1.012	3.260
Sea	1.055	5.080	1.058	5.832
Phx	1.047	2.834	1.058	2.704
SL	1.028	3.744	1.031	3.327
Det	1.069	2.560	1.037	3.958

Table 6: Correlations between g_{ik} and a_i for mutual funds

The table displays the correlation between our managerial RPC measures g_{ik} and our gregariousness parameter estimates a_i of mutual funds in each of the 20 cities. The correlations are calculated based on the Fama-MacBeth time-series means of g_{ik} (for each city k) and of a_i. For explanations on abbreviated city group names, please refer to the notes under Table 3.

NY	LA	Bos	SF	Chi	SJ	Dal	Hou	Phi	Was
-0.028	0.058	0.029	0.127	-0.011	0.053	0.009	0.002	0.040	-0.012
Mia	Atl	Min	Den	SD	Stfd	Sea	Phx	SL	Det
0.186	-0.019	-0.006	-0.015	0.217	-0.026	-0.025	-0.012	0.047	0.121

Table 7: Summary statistics of $gsum_i$ for mutual funds

This table shows the summary statistics of our managerial RPC scores $\{gsum_i\}$ of mutual funds over all quarters. We first compute the time-series average of the quarterly $gsum_i$ values for each individual fund i, then we report the summary statistics of these time-series averages. s.d. denotes standard deviation and med denotes median.

mean	s.d.	med	min	max
18.82	8.48	17.88	7.09	280.56

Table 8: Statistics on numbers of cities with $g_{ik} > 1$ for managers having $gsum_i > 20$

The table depicts, for mutual funds that have RPC scores $gsum_i > 20$, the summary statistics on the number of cites in which they have their RPC measures g_{ik} strictly larger than 1. s.d. denotes standard deviation and med denotes median.

mean	s.d.	med	min	max
8.33	1.44	8.04	5.04	11.44

Table 9: RPC scores and mutual fund performances

This table reports the Fama-MacBeth estimates of the regression coefficients in the specification $pfm_{i,t} = \alpha + \beta RPCdummy_{i,t-1} + x'_{i,t-1}\gamma + \varepsilon_{i,t}$, with t-statistics based on Newey-West HAC standard errors (of lag order 12) shown in parentheses. *, ** and *** denote statistical significance at the 10%, 5% and 1% levels respectively. The dependent variable is fund i's net return at quarter t. const denotes the constant term. $FundReturn$, $logTNA$, $logFamSize$, $ExpRatio$, $Turnover$ and $FundAge$ denote, respectively, fund i's lagged net return, the log of the total net asset of fund i, the log of one plus the total net asset of other funds in fund i's family, the expense ratio of fund i, the turnover ratio of fund i and fund i's age. $gsum > 20$ is a dummy variable that equals 1 if a fund's RPC score $gsum$ is larger than 20. FinCenter is a dummy variable that equals 1 if a fund is located in a financial center. The following six cities are defined to be financial centers: Boston, Chicago, Los Angeles, New York, Philadelphia, and San Francisco, in the spirit of Christoffersen and Sarkissian (2009). $\log(a_i)$ is the log of fund i's gregariousness parameter estimate. ICI denotes the Industry Concentration Index (ICI) of fund i, which is constructed in a similar manner as in Kacperczyk, Sialm, and Zheng (2005). But for simplicity, we use an equally weighted index instead. All of the non-constant regressors are measured at quarter $t-1$.

	(1)	(2)
$const$	0.016**	0.017***
	(2.57)	(3.18)
$FundReturn_{t-1}$	0.064***	0.073***
	(3.05)	(3.92)
$logTNA_{t-1}$	-0.0010**	-0.0009***
	(-2.54)	(-2.83)
$logFamSize_{t-1}$	0.0001	0.0001
	(1.34)	(1.33)
$ExpRatio_{t-1}$	-0.004***	-0.003***
	(-6.29)	(-6.20)
$Turnover_{t-1}$	0.000	0.000
	(-0.15)	(-0.16)
$FundAge_{t-1}$	0.000	0.000
	(-0.93)	(-0.55)
$gsum > 20$	0.0043**	0.0041**
	(2.05)	(2.01)
FinCenter	0.0006***	0.0011***
	(2.71)	(2.64)
$\log(a_i)$	-0.0004	-0.0004
	(-0.75)	(-0.62)
ICI		0.0071
		(1.46)

Table 10: Summary statistics, Herfindahl index of stock picks of fund managers

The table shows the summary statistics of the Herfindahl index of fund managers' stock picks across the 20 cities over time. We first compute the time-series average of the Herfindahl index values for each individual fund i across all quarters, then we report the summary statistics of these time-series averages. s.d. denotes standard deviation and med denotes median.

	mean	s.d.	med	min	max
Herfindahl index	0.108	0.042	0.100	0.067	0.583

Table 11: Summary statistics of correlation between RPC score *gsum* and Herfindahl index of stocks picks for mutual funds

The table shows the summary statistics of the correlation between our RPC score *gsum* of mutual funds and the Herfindahl index of mutual fund managers' stock picks over time. We first compute this correlation for all funds in each quarter, then we report the summary statistics of the time series of those correlation values. s.d. denotes standard deviation and med denotes median.

	mean	s.d.	med	min	max
Correlation value	0.466	0.126	0.470	0.213	0.689

Table 12: Herfindahl index of stock picks of fund managers and mutual fund performances

This table reports the Fama-MacBeth estimates of the regression coefficients in the specification $pfm_{i,t} = \alpha + \beta HerfBig_{i,t-1} + x'_{i,t-1}\gamma + \varepsilon_{i,t}$, with t-statistics based on Newey-West HAC standard errors (of lag order 12) shown in parentheses. *, ** and *** denote statistical significance at the 10%, 5% and 1% levels respectively. The dependent variable is fund i's net return at quarter t. $HerfBig$ is a dummy variable that equals 1 if a fund's Herfindahl index of stock picks is larger than the upper quartile of the Herfindahl indices of all funds at quarter t. All other explanatory variables are the same as those in Table 9. The non-constant regressors are all measured at quarter $t-1$.

	(1)	(2)	(3)
$const$	0.023***	0.022***	0.020***
	(4.12)	(3.71)	(4.41)
$FundReturn_{t-1}$	0.128***	0.118***	0.110***
	(2.84)	(2.69)	(3.35)
$logTNA_{t-1}$	-0.0011**	-0.0010*	-0.0009***
	(-2.00)	(-1.84)	(-3.43)
$logFamSize_{t-1}$	0.0001	0.0001	0.0001
	(0.49)	(0.30)	(0.98)
$ExpRatio_{t-1}$	-0.003***	-0.003***	-0.003***
	(-2.91)	(-3.29)	(-3.93)
$Turnover_{t-1}$	0.000	0.000	0.000
	(0.19)	(-0.19)	(0.07)
$FundAge_{t-1}$	0.000	0.000	0.000
	(-0.07)	(-0.32)	(-1.74)
FinCenter	0.0005**	0.0005**	0.0005**
	(2.33)	(2.46)	(2.19)
$\log(a_i)$	0.0028	0.0032	0.0014
	(0.19)	(0.20)	(0.12)
ICI		0.0048	0.0065
		(0.47)	(0.90)
$HerfBig$	0.0032**	0.0030**	0.0015**
	(2.48)	(2.39)	(2.21)
$gsum > 20$			0.0032**
			(2.08)

Table 13: Proportion of managers who attended undergraduate schools in a particular city

The table reports the summary statistics of proportions of managers who attended undergraduate schools in a particular city among the 20 cities that we consider, using all available data on manager demographics. s.d. stands for standard deviation and med stands for median. For an explanation to the abbreviated city names, please refer to the note under Table 3.

	mean	s.d.	med	min	max
NY	11.78%	2.36%	11.67%	8.25%	15.08%
LA	3.75%	0.63%	3.62%	2.81%	4.69%
Bos	9.82%	1.96%	9.73%	6.87%	12.57%
SF	1.67%	0.24%	1.54%	1.32%	2.02%
Chi	2.95%	0.49%	2.82%	2.21%	3.69%
SJ	4.02%	0.67%	3.97%	3.02%	5.03%
Dal	1.32%	0.17%	1.17%	1.07%	1.57%
Hou	0.63%	0.07%	0.59%	0.53%	0.73%
Phi	4.23%	0.85%	4.11%	2.96%	5.50%
Was	1.46%	0.18%	1.40%	1.19%	1.73%
Mia	1.25%	0.16%	1.19%	1.02%	1.48%
Atl	0.56%	0.06%	0.53%	0.48%	0.64%
Min	1.88%	0.37%	1.84%	1.33%	2.43%
Den	0.98%	0.11%	0.87%	0.83%	1.13%
SD	0.63%	0.07%	0.63%	0.53%	0.73%
Stfd	0.52%	0.05%	0.50%	0.45%	0.59%
Sea	0.98%	0.11%	0.87%	0.83%	1.13%
Phx	0.28%	0.03%	0.27%	0.24%	0.32%
SL	0.56%	0.06%	0.56%	0.48%	0.64%
Det	0.21%	0.02%	0.19%	0.18%	0.24%

Table 14: Using demographic *gsum* score to predict manager demographic attribute

This table reports the Fama-MacBeth estimates of the regression coefficients in the logit regression models where the dependent variable is the demographic attribute (being a female ($FEMALE$), being younger than 45 years old ($YOUNG$), being a republican (REP) or having attended college in a "university" city ($SCHOOL$)) of a manager, with t-statistics based on Newey-West HAC standard errors shown in parentheses. *, ** and *** denote statistical significance at the 10%, 5% and 1% levels respectively. The main explanatory variable is the *log* of the demographic *gsum* score of a manager, with $gsumFEM$ being constructed based on cities having a relatively high female-to-male sex ratio, $gsumYNG$ on cities having a relatively high proportion of adults between the age of 25 and 45, $gsumREP$ on cities being republican-affiliated and $gsumSCH$ on cities having a relatively large proportion of managers who went to college there. The cities based on which the three different demographic scores are calculated have been detailed in Footnote 10. *const* denotes the constant term, SAT is the log of the median SAT score of the undergraduate school that a manager attended, and ADV, $FEMALE$, $YOUNG$, REP and $SCHOOL$ are all dummy variables that equal 1 if a manager holds a graduate degree, is a female, is less than 45 years old, has a Republican political affiliation, and received an undergraduate degree in a "university" city respectively. For our logit regression in the last column where the dependent variable is REP, we exclude managers that are neither a republican nor a democrat.

	$FEMALE$	$YOUNG$	REP	$SCHOOL$
const	9.82***	-3.17	21.5***	-33.6***
	(4.92)	(-1.63)	(4.55)	(-37.3)
SAT	-1.86***	0.20	-3.05***	4.59***
	(-5.87)	(0.75)	(-4.47)	(36.9)
ADV	-0.40***	-0.02	0.87***	-0.15
	(-3.08)	(-0.08)	(14.79)	(-1.48)
$FEMALE$		0.20	-0.28*	1.07***
		(1.10)	(-1.92)	(8.39)
$YOUNG$	0.24		-0.50***	-0.04
	(1.38)		(-3.41)	(-0.88)
REP	0.03	-0.92***		0.01
	(0.15)	(-9.71)		(0.23)
$SCHOOL$	1.04***	-0.03	-0.30**	
	(5.67)	(-0.63)	(-2.43)	
$gsumFEM$	0.43***			
	(3.03)			
$gsumYNG$		0.79***		
		(4.33)		
$gsumREP$			0.37***	
			(5.63)	
$gsumSCH$				0.30**
				(2.16)

Figure 1: Histogram of count of stocks, Seattle versus San Diego

This figure shows the histogram of the count of stocks that are held by actively managed mutual funds and are headquartered in Seattle (left-panel) and San Diego (right-panel), using the holdings of 1066 mutual fund managers reported in the fourth quarter of 2004. The x-axis is the number of stocks held by a manager. The y-axis is the frequency of managers.

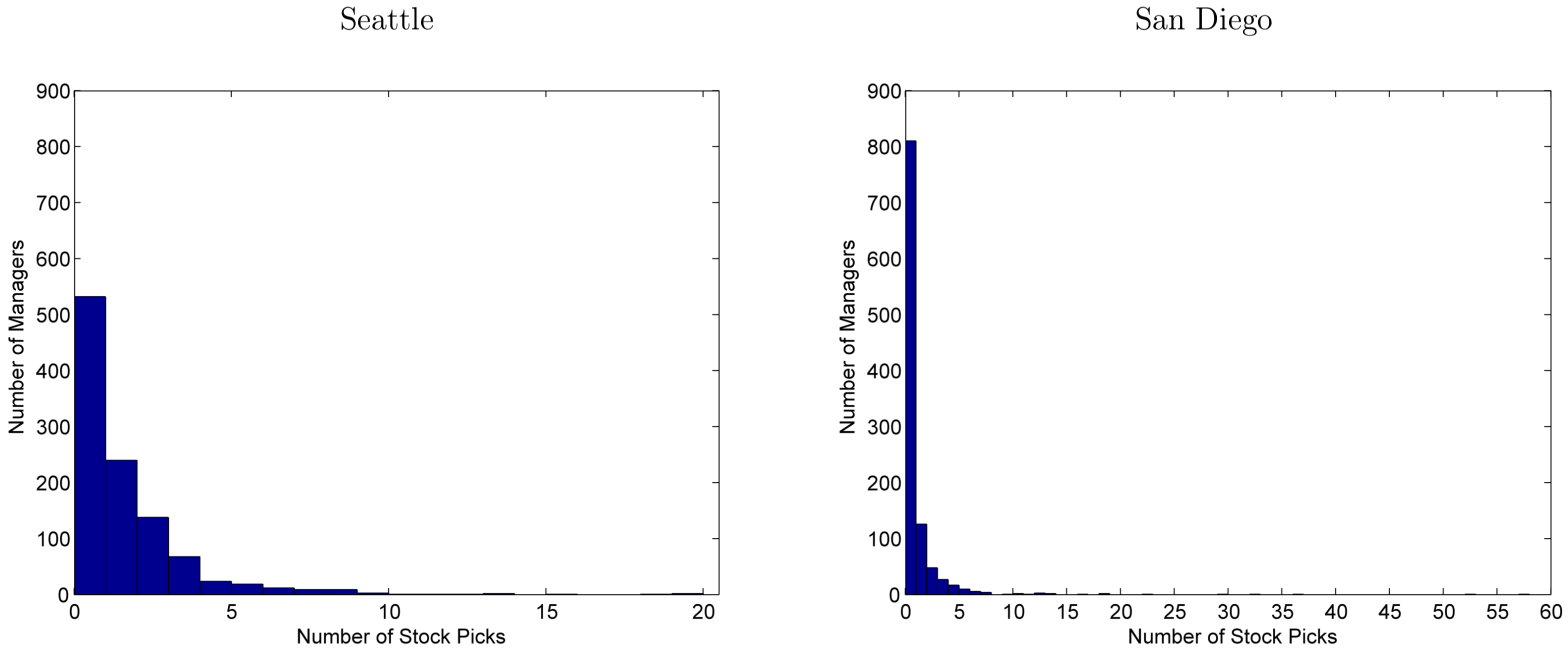
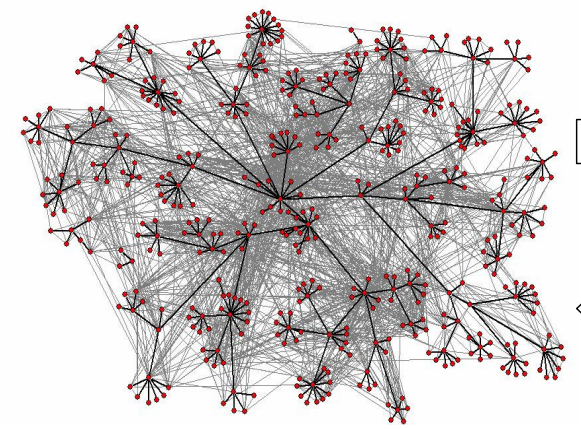


Figure 2: Social networks of fund managers and their stock picks in different cities

The figure on the left is a illustrative graph of the potential social networks of mutual fund managers, which is based on Figure 3 used in Adamic and Adar (2005) and is downloadable at `http://www-personal.umich.edu/~ladamic/img/hplabsemailhierarchy.jpg`. The table on the right is an example of managers' possible stock picks in different cities. The figure and the table are for illustrative purposes only and thus are not based on the data used in this paper.

Social Networks of Managers (Latent)



Stock Picks of Managers (Observed)

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

Figure 3: Histogram of a_i estimates for mutual funds

The figure shows the histogram of the estimated values of the gregarious parameters $\{a_i\}$ using mutual fund data with 20 city groups as described in the main article. As in Table 2, We compute the time-series average of the quarterly a_i estimates for each individual fund i and use these time-series averages as our estimated gregarious parameter values. The x-axis is the value of the a_i estimate and the y-axis is the frequency of managers.

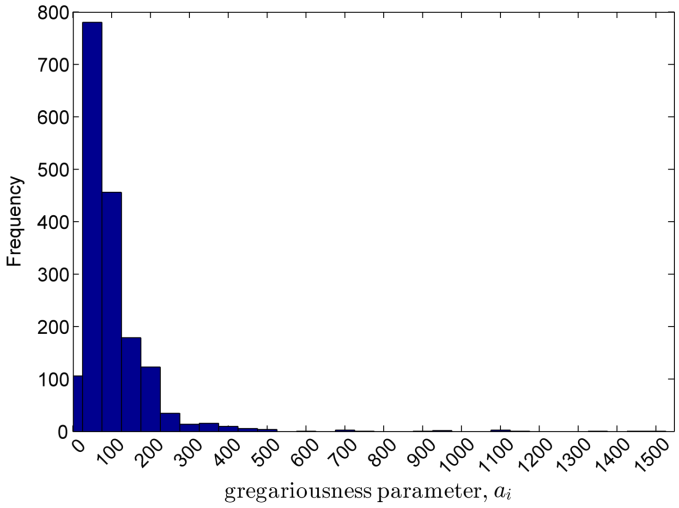


Figure 4: Boxplot of b_k estimates of 20 cities

This figure gives the boxplot of the quarterly estimates of the relative group size parameters $\{b_k\}$ for the 20 cities. The green marker is the mean, the red line is the median, the box is the interquartile range, and the tails extend to the min and the max. For an explanation to the abbreviated city names, please refer to the note under Table 3.

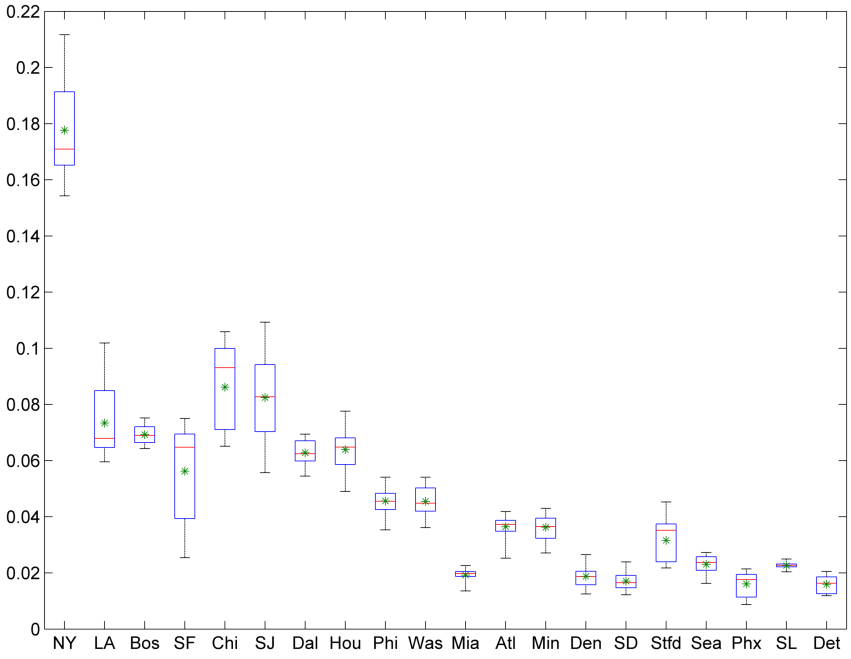


Figure 5: Boxplot of ω_k estimates of 20 cities

The figure gives the boxplot of the quarterly estimates of the overdispersion parameters $\{\omega_k\}$ for the 20 cities. The green marker is the mean, the red line is the median, the box is the interquartile range, and the tails extend to the min and the max. For an explanation to the abbreviated city names, please refer to the note under Table 3.

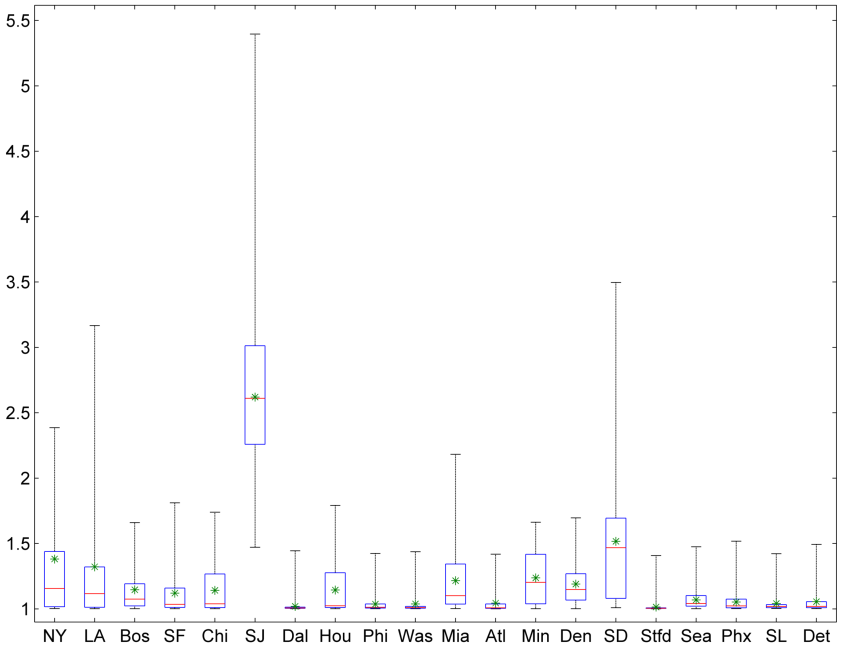


Figure 6: Histogram of $gsum_i$ of mutual funds

This figure shows the histogram of the values of our managerial RPC scores $\{gsum_i\}$ of mutual funds. As in Table 7, We compute the time-series average of the quarterly $gsum_i$ values for each individual fund i and plot the graph based on these time-series averages. The x-axis is the value of the RPC score $gsum_i$ and the y-axis is the frequency of managers.