Count Models of Social Networks in Finance*

Harrison Hong† Jiangmin Xu‡

First Draft: November 29, 2013
This Draft: April 21, 2014

ABSTRACT

We use overdispersed Poisson regression models to study social networks in finance. We count an investor's social connections in different cities as proportional to the number of stocks held by this investor that are headquartered in those cities. When connections are formed in an i.i.d. manner, the count of such connections in any city follows a Poisson distribution. Using data from institutional investors' holdings, we find instead overdispersion for a number of cities like San Jose and San Diego, which suggests that investors have non-i.i.d. propensities to be connected to these cities. Overdispersed cities have a large number of graduates from local universities who work in the fund industry. Managers with relatively high non-i.i.d. propensities to pick stocks from overdispersed cities significantly outperform other managers.

*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 at NYU Stern School of Business, Hong Kong Science and Technology and Bank of International Settlements for helpful comments.

†Princeton University, NBER, CAFR, and Visiting Professor at NYU Stern School of Business, (e-mail: hhong@princeton.edu),

‡Princeton University (e-mail: jiangmin@princeton.edu)

1 Introduction

There is a growing use of social networks to model phenomena from all corners of financial economics. Models of social interactions, epidemics and network effects are thought to be the leading explanations for dramatic changes in stock market participation rates during the Internet Bubble Period or housing ownership rates during the Housing Bubble Period (see, e.g., Shiller and Pound (1989), Glaeser and Scheinkman (2001), Hong, Kubik, and Stein (2004), Glaeser, Gottlieb, and Tobio (2012), Han and Hirshleifer (2013)). Investor social networks, such as being a Boston money manager (Hong, Kubik, and Stein (2005)), being a Harvard Alumni (Cohen, Frazzini, and Malloy (2008)) or going to the same brokerage house in China (Feng and Seasholes (2008)), are associated with information sharing among group members which influence what stocks are held and how the stocks perform. Moreover, 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)).

A question fundamental and common to all these endeavors is how to measure the presence and value of social networks. Yet, no systematic approach has emerged. Instead, studies typically attack this challenge by being creative in utilizing special data and exploiting unique situations to identify network effects. This approach is largely necessitated by a lack of comprehensive information about social networks. While this approach has been highly effective in generating insights, the cost is that it is difficult to generalize results from one setting to another. And in many important settings, such detailed network data might simply not be available.

In this paper, we show that overdispersed Poisson regression models, relying mostly on holdings or trade data that are typically available in most finance settings, can be used to study social networks in finance. These models were originally developed by statisticians Zheng, Salganik, and Gelman (2006) to analyze 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.

Importantly, they distinguish between being gregarious and being part of a network. Gregariousness is defined as people who differ in the expected number of social connections. However, their connections are formed randomly or independent and identically distributed (i.i.d.) as in the random networks model literature following Erdös and Rényi (1959). In contrast, being part of a network means that people from certain groups have non-i.i.d. propensities to form ties with each other. These models 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 follows a Poisson distribution.

While the Poisson distribution fits well the count of friends in certain groups, like people named Nicole or postal workers, Zheng, Salganik, and Gelman (2006) find that it does not fit well the count of friends in other groups, like prisoners. For instance, the count of friends in the prison population is highly overdispersed, in that most people surveyed know zero but some know many prisoners. That is, the variance of this count distribution significantly exceeds the mean of the count distribution, in contrast to what one would find with a Poisson distribution in which the variance equals the mean. Overdispersion then captures social connections to the prison population that are formed in a non-i.i.d. manner as some people have a non-i.i.d. propensity to know prisoners. This is presumably because the prison population constitutes a network while people named Nicole do not.

Although such survey data are rare in financial markets, we show that these models can be extended to study social networks in finance by using plentiful data on the actions of agents in financial markets such as their investment holdings or trades. For concreteness, we study investors' social networks by modeling the count of acquaintances in different cities, as proportional to the number of firms or stocks an investor holds that are headquartered in a given city. The idea is that since the stocks an investor chooses is a function of his network, we can infer that an individual who owns a "disproportionate" (in a sense that we will make precise shortly) number of stocks that are located in a certain city is more likely to have contacts in these cities.

Our extension of the network model in Zheng, Salganik, and Gelman (2006) can be easily applied to many other contexts in finance, such as 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.

Using panel data on the holdings of institutional investors in different cities, our dependent variable of interest is a monotonic transformation of the count of the number of stocks in those groups that are held by an investor. We estimate this model while allowing for heterogeneity in a number of important dimensions.

First, we allow for different gregariousness across managers—more gregarious managers have in expectation more stocks. We view this set of estimates as akin to investor fixed effects that allow some investors to hold more stocks than others. But it does not affect our inference of whether an investor belongs to a network. This inference is instead made controlling for this heterogeneity similar to the aforementioned statistics literature on social networks. Having a lot of friends is not the same as being part of a network. One could simply have equal numbers of friends and hold a lot of stocks in every group by chance.

Second, we allow different cities to have different numbers of potential connections based on how many stocks are headquartered in the city. A city like New York, which has many firms headquartered there, will have many connections attached to it. Again, this is a control or adjustment as we want to keep city or industry sizes roughly similar.

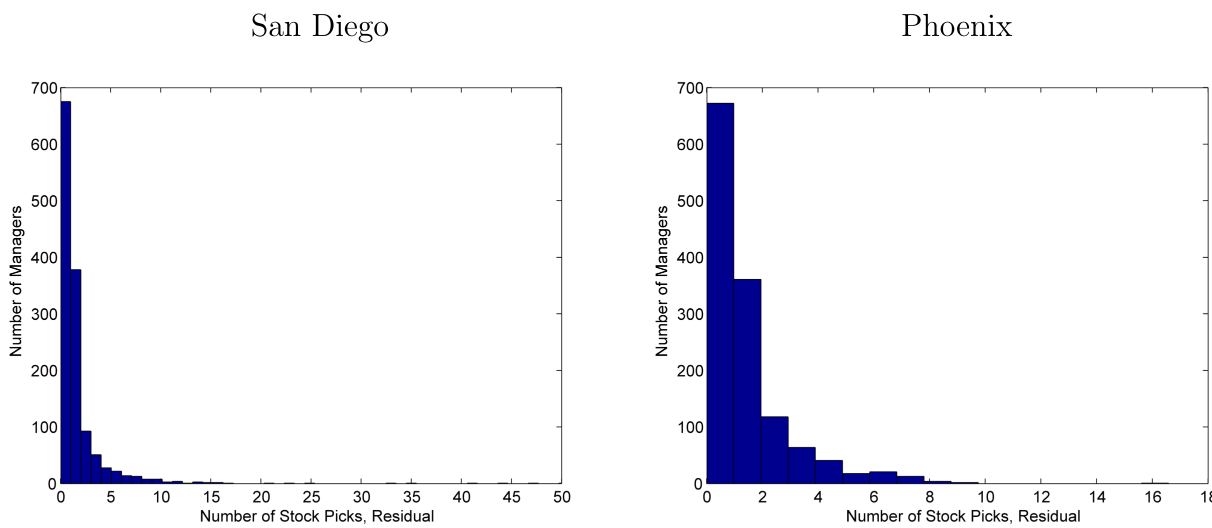
Controlling for these two factors, we can then use our model to estimate the degree of overdispersion of the cross-sectional distribution of the count of stocks in any given city held by investors. We allow the degree of overdispersion to vary across groups. That is, we can estimate a different overdispersion parameter for each city. If we have N investors, K groups, we end up estimating $N + 2K + J$ parameters with $N \times K$ number of observations reflecting the number of stock holdings in different cities. In addition, J is the degree of freedom needed to estimate semi-parametrically the transformation of the number of stocks into the number of social connections.

In our empirical analysis of the mutual fund holdings data from 1993 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 groupings, we find some overdispersion 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) or controlling for fund style of the managers.

To get an intuition for our set-up, we plot in Figure 1 a histogram of the count of stocks headquartered in San Diego (left-panel) and Phoenix (right-panel) using the holdings of 1315 mutual fund managers reported in the fourth quarter of 2005. The x-axis is the number of stocks held by a manager. The y-axis is the frequency of managers. The counts are residuals after controlling for manager fixed effects and city size as described above. For San Diego, the mean of the count of stocks held by the mutual fund manager population is 1.83 with a variance of 12.7. For Phoenix, the mean is 1.43 and the variance is 2.4.

San Diego is highly overdispersed while Phoenix is near Poisson. This indicates that there are some managers who 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 Phoenix. That is, managers who invest in San Diego are more likely to be part of a network that guides them toward San Diego stocks than Phoenix.

Figure 1: Histogram of residual count of stocks, San Diego versus Phoenix



Our major concern is that overdispersion might simply capture data errors or some outliers but that are otherwise uninformative. Indeed, overdispersion is often treated as a nuisance rather than something fundamentally informative as pointed out by Zheng, Salganik, and Gelman (2006). However, in the context of social networks, the Poisson null model has a very natural interpretation a la Erdös and Rényi (1959), which is what makes this overdispersed Poisson regression model informative about networks.

We then relate our model's outputs to demographic information about the investors as well as the performance of their investments. First, we try to understand what makes different cities overdispersed. We find that overdispersed cities tend to have a large number of graduates from local universities who work in the fund industry. We control for other city attributes such as the city's real GDP, whether the city leans Democratic or Republican and the number of stocks headquartered in that city. Only the local university representation variable comes in significantly. This is consistent with the notion of university being an important source of networks in Cohen, Frazzini, and Malloy (2008).

Second, we use our model to calculate for each investor his relative propensity to have contacts in a city (RPC) and relate these managerial RPC scores to the manager's demographic information. Our model gives a prediction for the expected number of stocks any investor should hold in a given city. An investor 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 investor's scores across all the cities. We find that managers with an advanced degree, of the female gender, who younger and Republican-minded have higher RPC scores. Having an advanced degree and being younger play by far the biggest roles, consistent with our above finding that local university representation in the fund industry explains the overdispersion of different cities.

Third, we then regress fund performance on these managerial RPC scores, while controlling for a host of the usual explanatory variables for fund performance. We find that managers with higher RPC city scores outperform those with lower RPC scores by around 2.5% 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. Moreover, our findings are also not driven by familiarity bias (Pool, Stoffman, and Yonker (2012)). Namely, we show that managers do not hold concentrated positions in the city where they go to college. Our findings that social networks are valuable echoes prior studies, which document the value of investor and CEO networks such as Cohen, Frazzini, and Malloy (2008) and Engelberg, Gao, and Parsons (2012) using unique data on these networks.

In the Internet Appendix, we consider a couple of additional sets of analyses. First, we replicate our findings using industry groupings as opposed to city groupings and find overdispersion for some industries. However, the performance results using industry groupings are weaker and the interpretation is potentially more problematic. Second, In addition to using institutional investor holdings data, we also use the Barber and Odean (2001) brokerage house retail investor holdings data and perform the same set of analyses. The problem with this dataset is that it is not representative of the retail investor universe in contrast to our mutual fund data which has the entire mutual fund universe. As a result, we generally find weaker social network effects for retail investors, but all the results are qualitatively similar in that we detect overdispersion in certain cities and industries and also find outperformance using our RPC scores. It is interesting to contrast the institutional and retail results as a statement about the prevalence of investor social networks across these two types of investors. The retail investor results also reassure us that our institutional investor results are not driven by unique mutual fund industry considerations. They seem more universal to investor networks.

Our contributions are two-fold. First, we introduce a new approach to the modeling of investors' social networks by extending statistical models of surveys on acquaintances in groups. 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 (Glaeser and Scheinkman (2002), Hong, Kubik, and Stein (2005)). In other words, the null is that under non-social interaction there is no reason for the actions to be correlated after controlling for public signals. The challenge with the excess correlation approach is controlling for a rich enough set of public signals. This new approach differs in counting stock picks across different groups and making inferences on which groups are networks. Our new approach has a different null hypothesis premised on random social connections leading to a Poisson distribution of the count of investors' holdings in any given group. The challenge here is to rule out that overdispersion is simply due to some outliers.

Second, and at the same time, our study contributes to the literature in statistics on the sociology of networks pioneered in Zheng, Salganik, and Gelman (2006). These papers have developed rich statistical models to study answers to surveys of questions on acquaintanceship networks. However, these studies are limited in terms of demographic or other information about the respondents. In applying these models to investor networks, we tap into a vast and rich database of such information about investor respondents including their investment performance. As a result, we can ask and answer many more questions about the determinants of the structure of social networks and the value of such networks for those who have them.

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 and 5. We discuss extensions of our methodology in Section 6. We conclude in Section 7. The Internet Appendix materials are in Section 8.

2 The Model

Our model follows Zheng, Salganik, and Gelman (2006)'s analysis of social networks. The key difference is that we do not observe answers to connections in different groups. We will instead use the number of holdings an investors has across different groups to proxy for their social connections. We will allow for a general flexible monotonic transformation of the number of connections in different groups to the number of holdings. This flexible transformation will be estimated along with the parameters describing the social networks.

2.1 Notation

Following Zheng, Salganik, and Gelman (2006), we use the following notation for the social networks between investors and their acquaintances in different cities. There is a total population of N investors, with friends residing in a total of K groups. Here, "group" is used interchangeably with "city".

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

The parameters $\{a_i\}$ may be viewed as controls for investor fixed effects, while the parameters $\{b_k\}$ can be thought of as controls for group sizes.

We also model the count of acquaintances y_{ik} in different groups as an increasing transformation of the number of stocks z_{ik} an investor holds that belong to a given group.[1] Therefore, we have $y_{ik} = h(z_{ik})$, where h is the increasing transformation. In our baseline setup, we assume y_{ik} is proportional to z_{ik}, thus $y_{ik} = z_{ik}/c$ where c is the transformation parameter.

2.2 The Null 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 its mean $\lambda_{ik} = ab_k$ 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 individ-

[1] By "belong to", we mean a company is headquartered in a specific city (group).

ual investors. Hence y_{ik} follows a Poisson distribution with a mean $\lambda_{ik} = a_i b_k$, but relative propensities g_{ik} are still all equal to one. We call this our null model.

2.3 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]

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

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 $(\omega > 1)$. 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.

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 $\{a_i\}$, representing the expected

[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.

number of acquaintances know by investor i, and the group size parameters $\{b_k\}$ that gauge the proportion of social connections involving group k.

2.4 Likelihood Function

Following from the density expression in (1), the likelihood function of $y = \{y_{ik}\}$ in our overdispersed model is

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

and the log-likelihood

$$\mathcal{L} = \sum_{i=1}^{N}\sum_{k=1}^{K}\Big(LG(y_{ik}+\zeta_{ik}) - LG(\zeta_{ik}) - LG(y_{ik}+1) - \zeta_{ik}\log(\omega_k) + y_{ik}\big[\log(\omega_k - 1) - \log(\omega_k)\big]\Big),$$

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. Since we observe the stock holdings $\{z_{ik}\}$ of investors in different groups and $y_{ik} = h(z_{ik})$ under the transformation h, the (log-)likelihood function in terms of $z = \{z_{ik}\}$ can then be expressed as

$$\mathcal{L} = \sum_{i=1}^{N}\sum_{k=1}^{K}\Big(LG\big(h(z_{ik})+\zeta_{ik}\big) - LG(\zeta_{ik}) - LG\big(h(z_{ik})+1\big) - \zeta_{ik}\log(\omega_k) + h(z_{ik})\big[\log(\omega_k - 1) - \log(\omega_k)\big]\Big). \tag{2}$$

The parameters of interest in our model are $\theta = \big(\{\omega_k\}_{k=1}^{K}, \{a_i\}_{i=1}^{N}, \{b_k\}_{k=1}^{K}\big)'$, a $(N+2K)\times 1$ vector. We also have $N \times K$ observations of z_{ik}. In addition, in our baseline scenario, $y_{ik} = z_{ik}/c$, so that the transformation parameter c will be estimated jointly with θ.

We shall estimate our model parameters using the method of maximum likelihood (MLE) based on (2), 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.[5] In the appendix, we will discuss

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

[5]The intuition behind our MLE is similar to the idea 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 interest y is a increasing transform of the observable variable z. It can be understood

the estimation mechanism and the associated results for the case where h is a general increasing transform. Additionally, we will also consider the estimation for another interesting scenario where z is a censored version of y.

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 1990–2011. 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 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 1680 unique actively managed, non-sector domestic equity funds, or 111144 fund-quarter observations on stock holdings.[6]

Besides the institutional investor holdings data, we will also employ the retail investor hold-

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) attains *the* maximum with the associated θ estimate.

[6] On average, approximately 1263 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.

ings data from Barber and Odean (2001) in the Appendix. Their dataset contains the monthly investments of 78,000 households between January 1991 and December 1996 from a large discount brokerage firm. It includes all investment accounts opened by each household at this discount brokerage firm, thus we aggregate the account information if a household had multiple accounts. Moreover, we focus on the the common stock investments of these households and exclude investments in mutual funds (both open- and closed-end), American depository receipts, warrants, and options. In addition, we only consider those households that had 10 or more stocks in their monthly portfolios on average. This is because the subsequent analyses performed on retail investors are only meaningful if their numbers of stock picks are not too small. Finally, we will use the demographic information contained in Barber and Odean (2001)'s dataset on age, gender and household income for these retail investors.[7] Overall, we have 1609 unique retail investors (households) with demographic information and monthly holdings of 10 or more stocks on average, or about 93600 household-month observations.[8]

Next, we categorize the stocks held by mutual funds or retail investors into city groups and industry groups. We use the information on companies' headquartered cities and their SIC industry codes that is 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).[9] On the other hand, to obtain industry groups, we utilize the industry definitions of Fama-French 30 industry portfolios and classify each stock into a particular industry.[10]

We shall only consider the largest 20 cites (MSAs) or largest 20 industries in terms of the number of located companies. The reason is because the 20 largest groups, either cities or industries, already cover approximately 80% of all the stocks held by mutual funds or retail investors 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.

---

[7] Please refer to Barber and Odean (2001) and also Barber and Odean (2000) for a detailed description on the dataset.

[8] At the monthly level, there are about 1300 households with a monthly average holdings of 10 or more stocks who had their holdings reported in the dataset.

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

[10] The information on Fama-French 30 industry portfolios is available at
http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/Data_Library/det_30_ind_port.html.

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 both the mutual fund and the retail investor holdings data. To be more precise, at each point in time (quarter for mutual funds and month for retail investors), we will use the past 12 quarters or months of holdings data as a rolling subsample to estimate θ and c based on the log-likelihood (2). The observations z_{ik} are then the number of unique stock picks from a group k made by an investor i during the past 12 quarters or months. Therefore, our rolling estimates start at 1993Q1 (resp. Jan 1992) and end at 2011Q4 (resp. Dec 1996) for mutual funds (resp. retail investors).

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.

4 Main Results

In this section, we report our main estimation results based on the mutual fund data. We shall concentrate on the results having cities as groups, and relegate the results with industry groups to the Appendix. We also perform the same set of analyses on retail investor data and the associated results will be shown in the Appendix as well.

4.1 Transformation Parameters

Table 1 presents the estimates (Fama-MacBeth means of the quarterly rolling estimates) and related summary statistics of the transformation parameter for city groups. It shows that the mean of the transformation parameter c is 1.37 with a standard deviation of .1 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 1.37.

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

	mean	s.d.	med	min	max
City groups	1.37	0.10	1.38	1.13	1.87

4.2 Gregariousness Parameters

Next, Table 2 shows the summary statistics of the estimated values of the gregarious parameters a_i and Figure 2 illustrates the histogram of their Fama-MacBeth averages. We observe that the mean of a_i is 105 using city groups. This estimate can be interpreted literally as the typical manager having around 100 friends in the mutual fund industry overall and just in our sample. But there is a fairly sizeable standard deviation of around 120 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.

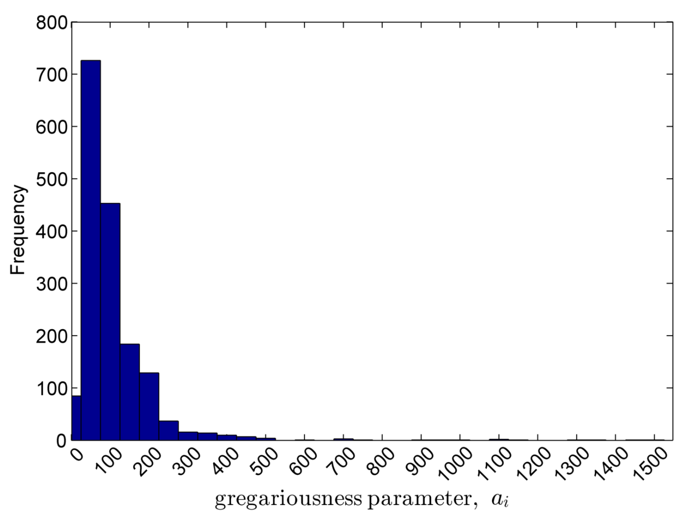
Nevertheless, we view the estimates of gregariousness parameters as more akin to investor fixed effects for some investors having more stocks than others. They are separate from and do not affect our inference on whether investors belong to a network. In other words, having a lot of friends is not the same as being part of a network since it could also be affected by other factors such as investment style.

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

	mean	s.d.	med	min	max
City groups	105.0	116.0	76.8	0.6	1497.0

Note: summary statistics are based on individual time-series averages of a_i.

Figure 2: Histogram of a_i estimates, city groups, mutual funds



4.3 Group Size Parameters

We then report the parameter estimates for b_k that gauge the relative sizes of cities. Table 3 and Figure 3 demonstrate the values of b_k for the 20 cities.

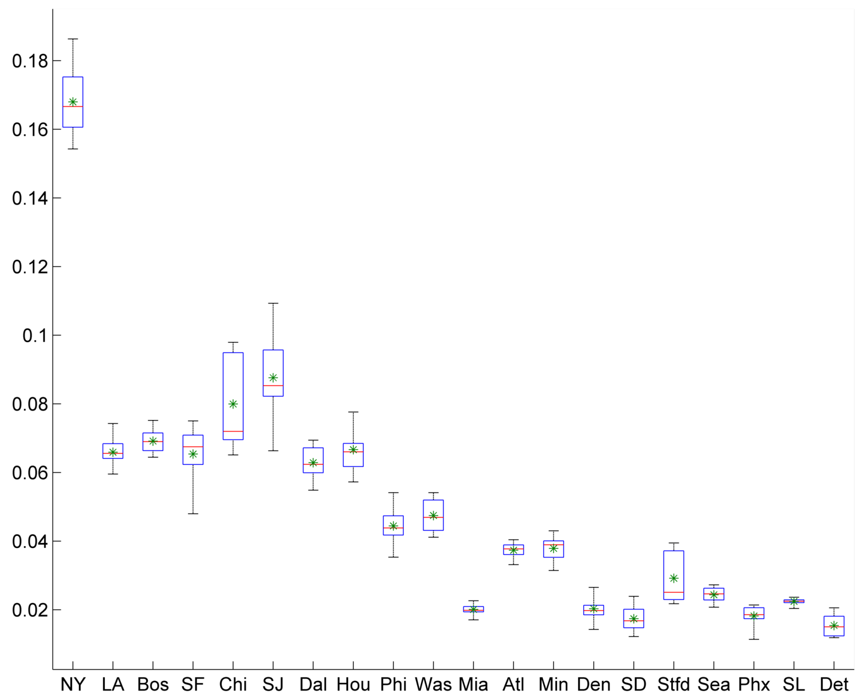
Table 3: Estimates of b_k, city groups, mutual funds

	mean	s.d.	med	min	max
NY	0.168	0.009	0.167	0.154	0.186
LA	0.066	0.003	0.066	0.060	0.074
Bos	0.069	0.003	0.069	0.064	0.075
SF	0.065	0.007	0.067	0.048	0.075
Chi	0.080	0.012	0.072	0.065	0.098
SJ	0.088	0.012	0.085	0.066	0.109
Dal	0.063	0.004	0.062	0.055	0.069
Hou	0.067	0.006	0.066	0.057	0.078
Phi	0.044	0.005	0.044	0.035	0.054
Was	0.047	0.004	0.047	0.041	0.054
Mia	0.020	0.001	0.020	0.017	0.023
Atl	0.037	0.002	0.038	0.033	0.040
Min	0.038	0.003	0.039	0.031	0.043
Den	0.020	0.003	0.020	0.014	0.027
SD	0.017	0.003	0.017	0.012	0.024
Stfd	0.029	0.007	0.025	0.022	0.039
Sea	0.024	0.002	0.025	0.021	0.027
Phx	0.018	0.003	0.019	0.011	0.021
SL	0.022	0.001	0.022	0.020	0.024
Det	0.015	0.003	0.015	0.012	0.021

Note: 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.

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 Chicago. 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.

Figure 3: Boxplot of b_k estimates, city groups, mutual funds



Note: 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.

4.4 Overdispersion Parameters

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, it is quite possible that there is no much variation in these relative propensities. However, larger values of ω_k would imply dissimilarities in individuals' relative propensities to make connections.

Table 4 and Figure 4 display the estimated overdispersions ω_k for city groups. There are three evident features. Firstly, 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. Secondly, 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 0.37, and the rank correlation between them is merely about 0.23. Thirdly, 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. Hence it implies that some investors do belong to certain integrated social networks even in the smaller cities (in terms of b_k) such as Miami or Minnesota.

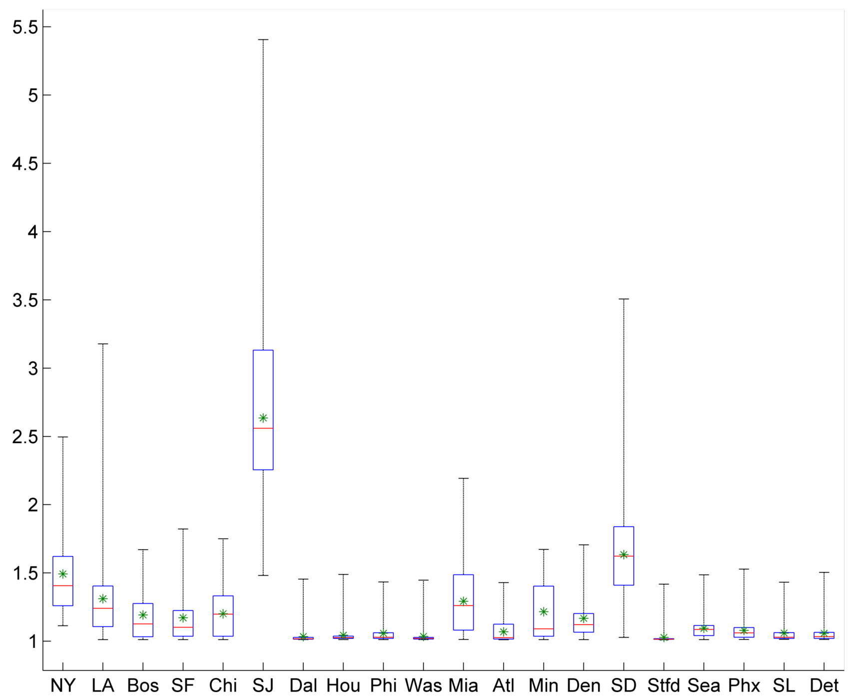
The overdispersion estimates therefore signify that a number of investors live among some intricate social networks. They do not have to be the most gregarious investors, nor are they necessarily tied to the largest cities or industries.

Table 4: Estimates of ω_k, city groups, mutual funds

	mean	s.d.	med	min	max	t-stat
NY	1.502	0.302	1.396	1.013	2.386	5.621
LA	1.411	0.313	1.331	1.012	3.167	4.820
Bos	1.182	0.191	1.117	1.002	1.660	4.488
SF	1.161	0.187	1.092	1.002	1.812	3.985
Chi	1.191	0.163	1.188	1.001	1.740	5.223
SJ	2.625	0.663	2.550	1.471	5.396	12.059
Dal	1.024	0.064	1.007	1.001	1.445	3.351
Hou	1.034	0.071	1.016	1.002	1.479	4.475
Phi	1.048	0.076	1.018	1.001	1.424	3.996
Was	1.025	0.067	1.010	1.001	1.437	3.494
Mia	1.283	0.233	1.251	1.003	2.182	5.269
Atl	1.059	0.082	1.016	1.001	1.419	4.175
Min	1.206	0.203	1.081	1.002	1.662	4.304
Den	1.158	0.156	1.111	1.001	1.696	4.927
SD	1.624	0.355	1.613	1.018	3.496	9.343
Stfd	1.018	0.059	1.005	1.001	1.407	2.828
Sea	1.084	0.078	1.076	1.002	1.476	7.071
Phx	1.069	0.085	1.051	1.002	1.518	5.118
SL	1.050	0.077	1.018	1.004	1.422	3.989
Det	1.048	0.075	1.023	1.003	1.494	4.769

Note: for an explanation to the abbreviated city names, please refer to the note under Table 3. 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).

Figure 4: Boxplot of ω_k estimates, city groups, mutual funds



Note: 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.

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.

Table 5: Robustness checks on ω_k, city groups, mutual funds

	No Local Response		No Growth Fund	
	mean	t-stat	mean	t-stat
NY	1.548	5.768	1.479	5.647
LA	1.473	5.287	1.427	5.221
Bos	1.160	5.351	1.214	4.203
SF	1.141	3.082	1.138	4.304
Chi	1.170	5.731	1.164	4.375
SJ	2.560	11.413	2.578	12.439
Dal	1.027	4.012	1.016	3.230
Hou	1.064	4.819	1.016	4.682
Phi	1.038	4.128	1.033	4.259
Was	1.027	3.293	1.017	3.478
Mia	1.237	5.302	1.356	5.380
Atl	1.041	3.776	1.050	4.751
Min	1.165	3.978	1.189	4.332
Den	1.171	4.661	1.169	3.948
SD	1.600	9.337	1.745	9.486
Stfd	1.016	3.473	1.017	2.652
Sea	1.072	6.459	1.068	7.410
Phx	1.063	4.467	1.072	4.273
SL	1.034	4.476	1.039	3.981
Det	1.060	3.473	1.031	5.366

Note: this table demonstrates the results of two robustness checks on the overdispersion estimates, using mutual fund data with city groups. "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.

5 Networks, Demographics and Performances

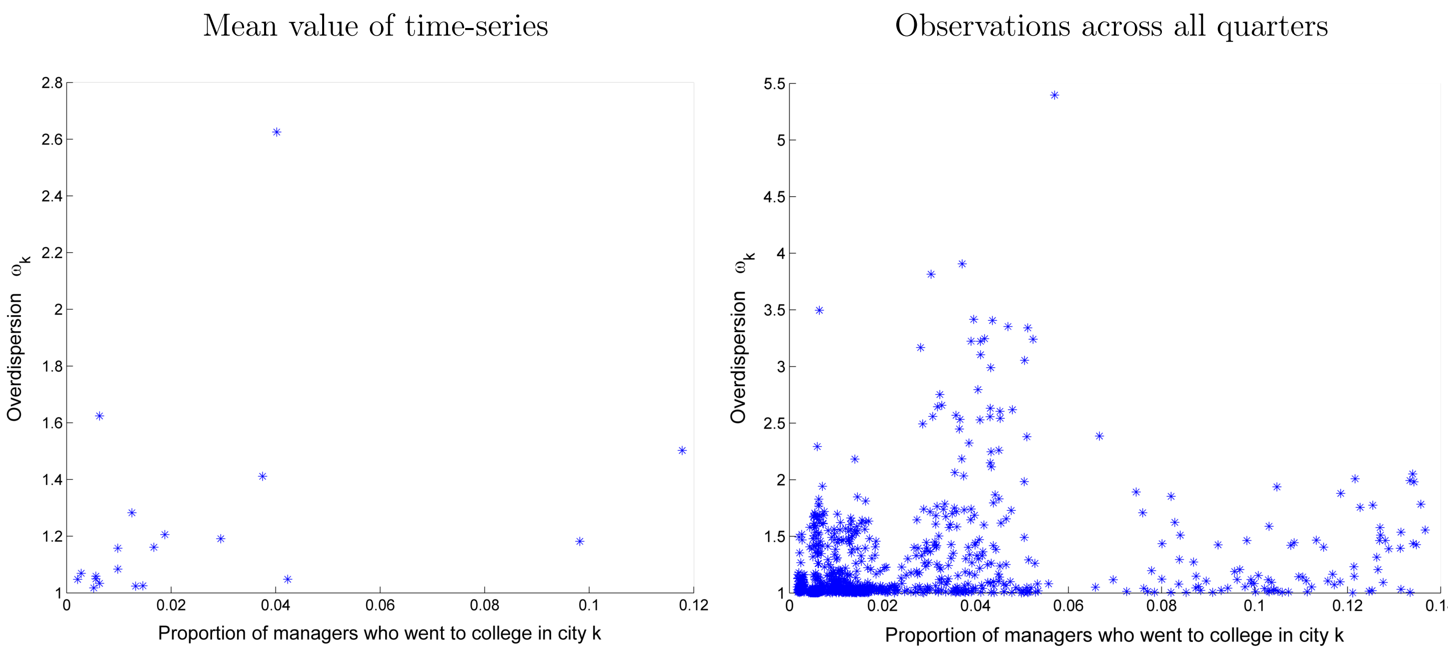
5.1 What Explains ω_k?

Overdispersion is often treated as a nuisance rather than something fundamentally informative. But in the context of social networks, the Poisson null model has a very natural interpretation a la Erdös and Rényi (1959) of random social connections, which is what makes this overdispersed Poisson regression model informative about networks.

Our major concern is that overdispersion might simply capture data error or some outliers but that are otherwise uninformative. To show that this is not the case, we relate our model's output to demographic information about the investors as well as the performance of their investments.

We first conjecture that managers who went to colleges in a particular city are more likely to form social connections within that city via their college-friend networks.[11] To give an idea about the local university representation of managers, Table 6 tabulates the summary statistics of proportions of managers who attended undergraduate schools in different cities and Figure 5 gives the scatter plot of ω_k's against these proportions.

Figure 5: Scatter plot of ω_k against local university representation



[11] This is related to the idea of fund manager-corporate director college links in Cohen, Frazzini, and Malloy (2008).

Table 6: Proportion of managers who attended undergraduate schools in a city

	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%

To examine our conjecture, we first analyze how the overdispersion parameter ω_k of each city depends on the number of managers who attended colleges in that city. We employ the following regression specification:

$$\log(\omega_{k,t}) = \alpha + \beta\, prop_{k,t} + \gamma' Ctrl_{k,t} + \eta\, DotCom_t + \varepsilon_{k,t},$$

where $prop_{k,t}$ is the proportion of managers who attended undergraduate schools in city k at quarter t, $Ctrl_{k,t}$ is a vector of relevant control variables, and $DotCom_t$ is a time dummy variable that equals one if quarter t belong to the dotcom bubble period of 1997Q1 to 2001Q4. We restrict the coefficients in front of $DotCom_t$'s to be all equal to keep it parsimonious. This is then performed as a panel regression with city fixed effects, and the estimation results are displayed in Table 7.

It is clear from Table 7 that the number of managers who went to college in city k has a positive and statistically significant effect on the overdispersion ω_k of that city. When only $prop_{k,t}$ is included (column (1)), the coefficient estimate of β equals 2.06 with a t-statistic equal

Table 7: Relationship between ω_k and Number of Managers Graduating from City k

	(1)	(2)	(3)	(4)	(5)	(6)
$prop$	2.061**	1.818**	1.198**	1.799**	1.310**	1.714**
	(2.03)	(2.45)	(1.98)	(2.25)	(1.96)	(2.36)
$StkNums$			0.044		0.030	0.047
			(1.51)		(1.09)	(0.96)
$RGDP$				0.219	0.201*	0.358**
				(1.46)	(1.87)	(2.05)
$RepCity$			0.025	0.043	0.040	0.041
			(0.84)	(1.28)	(1.28)	(0.77)
$DotCom$		0.029**	0.066**	0.077**	0.094**	0.053**
		(2.27)	(2.44)	(2.51)	(2.35)	(2.30)
$MktCap$						-0.364
						(-1.604)

Note: this table reports the estimation results of various forms of the regression $\log(\omega_{k,t}) = \alpha + \beta\, prop_{k,t} + \gamma' Ctrl_{k,t} + \varepsilon_{k,t}$, with t-statistics shown in parentheses. *, ** and *** denote statistical significance at the 10%, 5% and 1% levels respectively. StkNums denotes the number of stocks headquartered in a city. RGDP is the real GDP per capita of a city. RepCity is a dummy variable that equals one if the political affiliation of a city is Republican (based on the most recent election results). MktCap denotes the market-cap Herfindahl index of a city, which is a Herfindahl index computed based on the market capitalization of each individual stock located in that given city. The vector of controls $Ctrl_{k,t}$ includes either some or all of the variables among StkNums, RGDP, RepCity and MktCap, depending on the specifications. City-level (or MSA-level to be more precise) real GDP data are obtainable from Bureau of Economic Analysis. Political affiliations of different counties where cities are located are obtained from the website of David Leip's atlas of U.S. presidential elections at *http://uselectionatlas.org/*.

to 2.03 (hence significant at the 5% level). It indicates that if the proportion of managers who attended colleges in city k increases by 1 percentage point, ω_k would rise by 2.06%. The effect does not diminish if we include other city-level control variables such as the number of stocks headquartered in a city, the political affiliation of a city, and/or the real GDP per capita of a city (column (2) to (4)). The estimate of β is in the range of 1.2 to 2.0 and is always significant at the usual significance levels. This result thus supports our conjecture since if more managers ended up going to schools in city k, more of them were likely to skew their social networks into that city, resulting in a higher overdispersion of acquaintance connections over there.

Furthermore, in column (6) of Table 7, we incorporate the market-cap Herfindahl index of a city ($MktCap$) as an extra control. The market-cap Herfindahl index is a Herfindahl index calculated based on the market capitalization of each individual stock located in a given city. The purpose to include this particular variable is to make sure that the overdispersion of a city is not *mechanically* driven by the presence of a few dominant, large-sized companies with many other small companies, since it might be the case that most managers hold the stock of one large, well-known company while there are some managers holding stocks in a lot of the small companies. The displayed result shows that the presence of some large companies in a city does not have any positive effect on the city's overdispersion, thus overdispersions could not be mechanically caused by such presence.

In addition, Table 7 also tells us that our overdispersion parameters $\{\omega_k\}$ indeed capture social network effects, since their subtlety are not simply explained by other standard economic variables such as a city's real GDP level or the number of stocks it has. As discussed before, overdispersion does not have to arise just because cities are large (in terms of number of stocks), and we can see here that cities being richer does not necessarily lead to more profound overdispersions either.

5.2 Managerial RPC Measure and Managerial Demographics

We next use our model to generate for each manager his propensity to be connected in a non-i.i.d. way to groups in these cities and relate these managerial RPC scores to managerial demographics. Recall that in our model, investors' *expected* relative propensities to know a member in group 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 group. The RPC *measure* for any investor in a particular group k is computed as $g_{ik} = y_{ik}/(a_i b_k)$.[12] Our model predicts that an investor should have an expected number of $a_i b_k$ connections in a given group, 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 group is more likely to be part of and has $g_{ik} > 1$ in that group, i.e. being part of that network.

Then we sum up investors' RPC measures across all the groups, 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 $K = 20$ groups. However, if there are structured social networks among various groups, 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.

Table 8 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.

Table 8: Correlations between g_{ik} and a_i, mutual funds, city groups

NY	LA	Bos	SF	Chi	SJ	Dal	Hou	Phi	Was
-0.043	0.040	0.027	0.147	-0.022	0.058	0.013	0.004	0.041	0.006
Mia	Atl	Min	Den	SD	Stfd	Sea	Phx	SL	Det
0.185	-0.017	-0.010	0.012	0.220	-0.030	-0.027	0.202	0.033	0.110

Note: correlations are based on the Fama-MacBeth time-series means of g_{ik} (for each k) and a_i. For explanations on abbreviated city group names, please refer to the notes under Table 3.

The summary statistics for our RPC scores $gsum_i$ are demonstrated in Table 9 as well as in Figure 6. We notice that the mean of RPC scores are close to 20 with city groupings, yet the standard deviation (around 5) is sizeable. Once more, this is another piece of evidence showing that certain investors have non-i.i.d. propensities to form ties with members from different

[12]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.

cities. Furthermore, we find in Table 10 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 nine. 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.

Table 9: Summary statistics of $gsum_i$, mutual funds

	mean	s.d.	med	min	max
City groups	19.20	5.18	18.91	2.07	359.9

Note: summary statistics are based on individual time-series averages of $gsum_i$.

Figure 6: Histogram of $gsum_i$, city groups, mutual funds

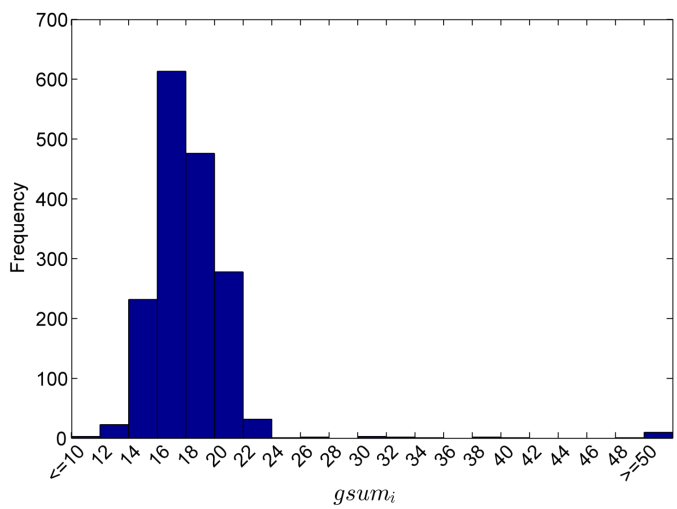


Table 10: Statistics on numbers of cities with $g_{ik} > 1$ for mutual funds having $gsum > 20$

	mean	s.d.	med	min	max
City groups	8.96	0.43	9.07	7.81	9.76

Note: the table contains the summary statistics on the average of the number of groups where $g_{ik} > 1$, for mutual funds that have $gsum > 20$.

Having calculated our RPC measure and score for each manager, we are now in a position to study how they are related to demographic information about the mutual fund managers.

To assess what type of managers are more likely to have higher RPC scores, we regress manager RPC score $gsum$ on a vector of manager demographic characteristics with the following regression equation:

$$\log(gsum)_i = \alpha + \beta_1 MedSAT_i + \beta_2 Adv_i + \beta_3 Female_i + \beta_4 Old_i + \beta_5 Rep_i + \varepsilon_i,$$

where MedSAT is the median SAT score of the undergraduate school that a manager went to, Adv is a dummy variable that equals 1 if a manager attended gruadate school, Female is a dummy for being a female, Old is a dummy for being over 45 years old, and Rep is a dummy for being a Republican. We estimate this for each quarter and then take the Fama-MacBeth time-series means and Newey-West standard errors of the resulting quarterly estimates. The estimation result is shown in Table 11.

Table 11: RPC Score and Manager Demographic Characteristics

Dependent variable: $log(gsum)$					
const	MedSAT	Adv	Female	Old	Rep
2.917***	-0.004	0.032***	0.022*	-0.029***	0.016***
(45.03)	(-0.48)	(10.12)	(3.38)	(-9.59)	(4.13)

Note: this table reports the Fama-MacBeth estimation results of the regression $\log(gsum)_i = \alpha + \beta_1 MedSAT_i + \beta_2 Adv_i + \beta_3 Female_i + \beta_4 Old_i + \beta_5 Rep_i + \varepsilon_i$, 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.

It can be seen from Table 11 that managers who went to graduate school have $gsum$'s that are 3.2% higher on average, and the effect is statistically significant. This suggests that managers with a graduate degree tend to be more socially connected, which is essentially consistent with our earlier conjecture of the college-friend networks. Furthermore, we find that female as well as Republican managers are likely to have more social ties than others, while older managers seem to have fewer. Having an advanced degree and being young are by far the strongest variables, which is consistent with the role of the influence of a city's local universities in the mutual fund universe in explaining ω's.

In the same vein as addressing what explains $\{\omega_k\}$, we next analyze whether the RPC measures $\{g_{i,k}\}$ of managers are higher in those cities where they attended college. For this purpose, we use the following regression

$$g_{i,k} = \alpha + \beta D_{i,k} + \gamma SameStyle_{i,k} + \eta Local_{i,k} + \varepsilon_{i,k},$$

where $D_{i,k}$ is a dummy variable that equals one if manager i went to college in city k, $SameStyle_{i,k}$ is the average $g_{i,k}$ in city k of all funds that have the same CRSP fund style categorization as fund i, and $Local_{i,k}$ is a dummy variable that equals one if fund i is also headquartered in city k. If fund managers tilt their social connections towards their college cities, then we should find that β is positive and statistically significant.

We estimate the above regression quarter by quarter and then take the Fama-MacBeth time-series means and Newey-West standard errors of the quarterly estimates. The estimated value of β equals 0.17 and is significant at the 1% level (with a t-statistic equal to 6.93). Hence this tells us that the RPC measure of a typical fund is 0.17 point higher in the city where the manager attended college than in the rest of the cities on average. Therefore, it confirms our conjecture that managers tend to have more network ties from cities where they received college education and as a result, prefer to have disproportionally more stock *picks* in these cities as well.

5.3 Overdispersion in Counts vs. Portfolio Weights

One natural question that arises is whether fund managers also tilt their portfolio weights towards cities where they attended undergraduate schools, since the result we just obtained focuses on their RPC measures $\{g_{i,k}\}$ (and hence stock picks). To this end, we estimate a similar regression of the form

$$w_{i,k} = \alpha + \beta D_{i,k} + \gamma SameStyle_{i,k} + \eta Local_{i,k} + \varepsilon_{i,k},$$

but with the dependent variable changed to portfolio weight $w_{i,k}$ (measured in percentages) and the regressor $SameStyle_{i,k}$ to the average $w_{i,k}$ in city k of all funds that have the same CRSP fund style categorization as fund i. What we find this time is that the estimated β equals 0.96, yet the t-statistic is only 0.982. This indicates that although managers overweight stocks to some extent in cities where they went to college, the overweighting effect is noisy and insignificant.

So why the social network effect shows up in our RPC measures $\{g_{i,k}\}$ and hence stock picks, but not in portfolio weights? We consider the possible reason to be as follows. Social connections are defined in terms of *counts*, and our RPC measures (and also stock picks) are criteria that aim at capturing the *numbers* of such network ties. They are likely to be a cleaner and a more direct gauge of the links between people in networks within any particular city than portfolio weights. This is because portfolio weights might as well be affected by other factors such as the need for benchmarking that are not immediately related to network connections. To put it in another way, a manager may have quite a few stock picks in a city based on his social networks, yet he might not allocate a large portfolio weight to each of these stocks because his preference for benchmarking.

5.4 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 performance:

$$pfm_{i,t} = \alpha + \beta RPCdummy_{i,t-1} + x'_{i,t-1}\gamma + \varepsilon_{i,t} \ , \qquad (4)$$

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 $t-1$. They include: (1) fund i's lagged pfm at $t-1$, (2) log of total net asset of fund i, (3) log of one plus the total net asset of other funds in fund i's family, (4) expense ratio of fund i, (5) 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). They are contained in the regressor x as well. Finally, α is a constant term and $\varepsilon_{i,t}$ is a generic error term uncorrelated with all other explanatory variables in 4. We will carry out the regression 4 quarter by quarter and then take the Fama-MacBeth time-series means and Newey-West standard errors of the quarterly estimates.

Table 12 depicts our fund performance regression results with cities as groups. Most of the 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 the 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 about 2.5% a year using city groups. However, we notice that being gregariousness does not necessarily lead to outperformance, as the coefficient on $\log(a_i)$ is close 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 12 where ICI is included as an extra explanatory variable in the regression specification of (4). In addition, our result that social networks are valuable to the tune of 2.5% 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).

6 Extensions

In this section, we detail the estimation process for the general model in which (1) the transform $h : h(z) = y$ is of a general increasing form and (2) z is a censored version of y.

6.1 A General Increasing Transform

When h is a general increasing transform, we follow the approach of Murphy (1994) and Murphy (1995). We denote the unique realizations of z_{ik} by z^u_s, where $s = 1, 2, \ldots$ are respectively the first, second, … unique counts of z. A normalization is applied so that $h(0) = 0$, i.e. if z_{ik} is zero, so is y_{ik}. Then we let the step sizes for the transformation h be $\Delta_s := h(z^u_{s+1}) - h(z^u_s),\ s = 1, 2, \ldots$. Next, we define a matching function

Table 12: RPC scores and mutual fund performances, city groups

	(1)	(2)
$const$	0.012*	0.008
	(1.65)	(1.06)
$FundReturn_{t-1}$	0.063**	0.064*
	(2.34)	(1.95)
$logTNA_{t-1}$	-0.0010**	-0.0008*
	(-2.23)	(-1.94)
$logFamSize_{t-1}$	0.0001	0.0001
	(0.72)	(0.65)
$ExpRatio_{t-1}$	-0.003***	-0.002**
	(-4.28)	(-2.15)
$Turnover_{t-1}$	-0.001	-0.001
	(-0.59)	(-0.64)
$FundAge_{t-1}$	-0.000	-0.000
	(-1.24)	(-1.38)
$gsum > 20$	0.0057**	0.0050***
	(2.15)	(2.51)
FinCenter	0.0007***	0.0012***
	(2.74)	(3.31)
$\log(a_i)$	-0.0006	-0.0007
	(-0.77)	(-0.86)
ICI		0.0078
		(0.93)

Note: this table reports the Fama-MacBeth estimates of the regression coefficients in specification 4, 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's net return at quarter t. $gsum > 20$ 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). ICI denotes the Industry Concentration Index (ICI), which is constructed in a similar manner as in Kacperczyk, Sialm, and Zheng (2005). But for simplicity, we use an equally weighted index instead.

$$H(z_{ik}, \Delta) = \begin{cases} \displaystyle\sum_{s=1}^{m \mid z_{m+1}^u = z_{ik}} \Delta_s & \text{if } m > 0 \\ h(0) & \text{if } m = 0. \end{cases}$$

The matching function transforms the value of z_{ik} by summing up all the Δ_s's up to the index m where z_{ik} equals the unique count number z_{m+1}^u. Thus we could use the matching function to map each z_{ik} back to its corresponding value of y_{ik}. In other words, if we know all the $\{\Delta_s\}$, then we know the increasing transform h (hence y_{ik}) for each unique z_{ik} that we observe. Therefore, the step sizes $\{\Delta_s\}$ are the additional parameters that we would like to estimate.

In this situation, let us denote the dimensionality of $\Delta = \{\Delta_s\}$ by J, where J equals the number of unique realizations of z_{ik} minus one. Our main parameters of interest are $\theta = (\{\omega_k\}, \{a_i\}, \{b_k\})'$, an $N + 2K$ vector. Hence we will use an $N \times 2K$ number of observations $\{z_{ik}\}$ to estimate a $N + 2K + J$ number of parameters θ and Δ. The log-likelihood function in terms of $z = \{z_{ik}\}$ then becomes

$$\mathcal{L} = \sum_{i=1}^{N} \sum_{k=1}^{K} \Big(LG\big(H(z_{ik}, \Delta) + \zeta_{ik}\big) - LG(\zeta_{ik}) - LG\big(H(z_{ik}, \Delta) + 1\big) - \zeta_{ik} \log(\omega_k) + H(z_{ik}, \Delta)\big[\log(\omega_k - 1) - \log(\omega_k)\big]\Big) ,$$

where $H(z_{ik}, \Delta)$ is defined as above. To estimate the parameters under a general transform h, we will adopt a technique similar to the profile maximum likelihood (see, e.g. Murphy, Rossini, and van der Vaart (1997), Murphy and van der Vaart (2000)). This method has been used in transformation models where the underlying interest y is a increasing transform of the observable variable z. The intuition has been discussed in the main body of this article and can be understood as follows. For each possible values of $\{\Delta_s\}$, we first compute the maximum likelihood estimate of θ and the corresponding maximal value of the log-likelihood, then we find the values of $\{\Delta_s\}$ such that the log-likelihood attains *the* maximum with the associated θ estimate.

We estimated our model under a general transform h using the mutual fund data with city groups, and we illustrate the results here. Since most of the results are similar to the ones from our baseline case, we will keep the illustration parsimonious and concise.

To start with, Table 13 show the estimates for the gregariousness parameters. We can notice that compared to the baseline result in Table 2, the estimates are somewhat larger, but the differences are quite small.[13]

Table 13: Estimates of a_i under a general h, mutual funds with city groups

mean	s.d.	med	min	max
131.0	142.5	104.8	0.6	1571.6

Next, the overdispersion parameter estimates are reported in Table 14. We observe that the overdispersions have increased in magnitude. However, New York, Los Angeles, San Jose and San Diego remain as the cities that appear to have substantial overdispersions, which is in line with our baseline results.

Table 14: Estimates of ω_k under a general h, mutual funds with city groups

	mean	s.d.	med	min	max	t-stat
NY	1.81	0.36	1.79	1.03	2.51	6.06
LA	1.77	0.38	1.52	1.03	3.51	4.93
Bos	1.62	0.23	1.40	1.01	1.94	4.68
SF	1.55	0.26	1.37	1.01	2.35	4.38
Chi	1.54	0.20	1.57	1.01	2.19	5.39
SJ	2.93	0.71	2.68	1.48	5.58	12.22
Dal	1.18	0.09	1.15	1.01	1.69	3.44
Hou	1.19	0.12	1.16	1.02	1.88	4.76
Phi	1.22	0.14	1.26	1.02	1.70	4.45
Was	1.17	0.12	1.12	1.01	1.93	3.87
Mia	1.48	0.28	1.58	1.01	2.35	5.56
Atl	1.28	0.14	1.24	1.01	1.60	4.60
Min	1.56	0.25	1.45	1.01	1.83	4.53
Den	1.33	0.20	1.27	1.01	2.04	5.28
SD	1.97	0.40	1.68	1.03	3.86	9.64
Stfd	1.18	0.10	1.14	1.03	1.63	3.14
Sea	1.31	0.14	1.28	1.02	1.82	7.19
Phx	1.28	0.15	1.19	1.02	1.74	5.54
SL	1.27	0.13	1.11	1.02	1.80	4.48
Det	1.26	0.14	1.16	1.01	1.89	5.18

Lastly, we can see from Table 15 and Table 16 that the RPC scores based on the current estimates under a general transform h are similarly related to managers' demographic characteristics as in the baseline case, and these RPC scores have a positive impact on mutual fund returns too. In particular, the outperformance number from the RPC score is 2.54% a year, consistent with what we found earlier in our baseline model.

[13]The estimates of b_k are very close to the baseline estimates and are hence not shown.

Table 15: RPC scores and demographic characteristics of mutual fund managers, general h

Dependent variable: $log(gsum)$					
const	MedSAT	Adv	Female	Old	Rep
2.904***	-0.006	0.014***	0.014*	-0.010***	0.011***
(51.74)	(-0.68)	(4.64)	(1.84)	(-3.13)	(2.82)

Table 16: RPC scores and mutual fund performances, general h

$const$	0.011
	(1.49)
$FundReturn_{t-1}$	0.064**
	(2.38)
$logTNA_{t-1}$	-0.0011**
	(-2.28)
$logFamSize_{t-1}$	0.0000
	(0.69)
$ExpRatio_{t-1}$	-0.003***
	(-4.20)
$Turnover_{t-1}$	0.000
	(-0.49)
$FundAge_{t-1}$	0.000
	(-1.32)
$gsum > 20$	0.0063**
	(2.29)
FinCenter	0.0008***
	(2.82)
$\log(a_i)$	-0.0003
	(-0.67)

6.2 A Censored Model

In the second part of our extensions, we discuss the scenario where our observable $\{z_{ik}\}$ is a censored version of the underlying $\{y_{ik}\}$. To be more specific, we consider a censoring threshold U, such that

$$\begin{cases} y_{ik} = z_{ik} & \text{if we observe } z_{ik} \text{ is strictly less that } U \\ y_{ik} \geq U & \text{if we observe } z_{ik} \text{ is greater than or equal to } U. \end{cases}$$

Therefore, with such a censoring, the log-likelihood becomes

$$\mathcal{L} = \sum_{i=1}^{N} \sum_{k=1}^{K} \Big[d_{ik} \Big(LG(z_{ik} + \zeta_{ik}) - LG(\zeta_{ik}) - LG(z_{ik} + 1) - \zeta_{ik} \log(\omega_k) + z_{ik} \big[\log(\omega_k - 1) - \log(\omega_k) \big] \Big) + (1 - d_{ik}) \log \big(1 - P(U|\zeta_{ik}, \omega_k) \big) \Big]$$

where d_{ik} is an indicator variable such that $d_{ik} = 1$ if $z_{ik} < U$, and $P(U|\zeta_{ik}, \omega_k)$ is the negative binomial cumulative distribution function

$$P(U|\zeta_{ik}, \omega_k) = \sum_{x=0}^{U} \frac{\Gamma(x + \zeta_{ik})}{\Gamma(\zeta_{ik})\Gamma(x + 1)} \left(\frac{1}{\omega_k} \right)^{\zeta_{ik}} \left(\frac{\omega_k - 1}{\omega_k} \right)^{x}.$$

This model can then be estimated using the usual maximum likelihood method. Additionally, one could also adapt it to the more complex case where y is an increasing transform of z if $z < U$.

We estimated the censored version of our model as illustrated above, with a range of censoring levels $U = 50, 75, 100, 125$. In general, we find that the estimation results under censoring are all qualitatively similar to the results from our baseline model. The main noticeable difference is that the estimates for the overdispersion parameters $\{\omega_k\}$ become larger due to the effect of censoring.

7 Conclusion

In this paper, we extend the overdispersed Poisson regression models used in statistics and sociology to study social networks in finance. Even though detailed network data is not typically available in finance settings, we show that we can model the count of an investor's social connections in different groups, such as cities or industries, as proportional to the number of stocks an investor holds that are headquartered in these cities or part of these industries. When connections are formed randomly, the count of these connections in any group follows a Poisson distribution. When connections are formed in a non-i.i.d. manner, the count of these connections in any group follows an overdispersed Poisson. Using data from institutional and retail investors' holdings, we estimate the degree of overdispersion for different groups. We find substantial overdispersion for some city groups such as San Diego, Los Angeles and San Jose, and for some industry groups such as Finance and Utilities. Our model also allows us to predict the relative propensity of any investor to be connected to a group. We show that these propensities are tied to investor demographics and are highly correlated with superior investor performance, suggesting that such networks are valuable.

These models can be used to study any financial network where investment data are available. Our set-up can be easily applied to many other contexts in finance such as 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. In short, the value of our set-up is that it connects the study of social networks in finance, which is hampered by limited data on social connections, to the study of social networks in statistics and sociology, which is hampered by data on performance. We hope this application of count models of social networks in financial markets might be useful for many different endeavors.

8 Internet Appendix

In the appendix, we first display the results for industry groups using mutual fund holdings data. Then we depict our findings using retail investor data with both city and industry groups.

8.1 Mutual Fund Results for Industry Groups

Table 17 presents the estimates of the transformation parameter for mutual funds with industry groupings, while Table 18 and Figure 7 show the estimates of the gregariousness parameters a_i. Compared to the city-group results in the main paper, we can see that the transformation parameter and the gregariousness parameters are larger when industry groups are considered for stocks. However, the differences between the estimates in the two group classifications are not substantial.

Table 17: Summary statistics, estimates of transformation parameter c, mutual funds

	mean	s.d.	med	min	max
Industry groups	1.53	0.16	1.55	1.18	1.79

Table 18: Summary statistics, estimates of a_i, mutual funds

	mean	s.d.	med	min	max
Industry groups	124.7	134.6	87.7	0.2	1220.9

Note: summary statistics are based on individual time-series averages of a_i.

Figure 7: Histogram of a_i estimates, industry groups, mutual funds

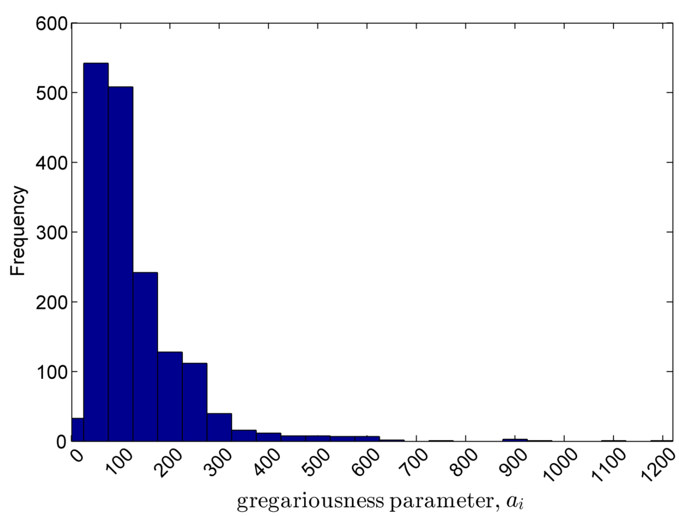


Table 19 and Figure 8 show the estimated values of b_k for the 20 industries. Similar to the city results, there are a few groups that have a much larger number of potential social connections attached to them comparing to the rest, and the sizes of various groups are stable across time.

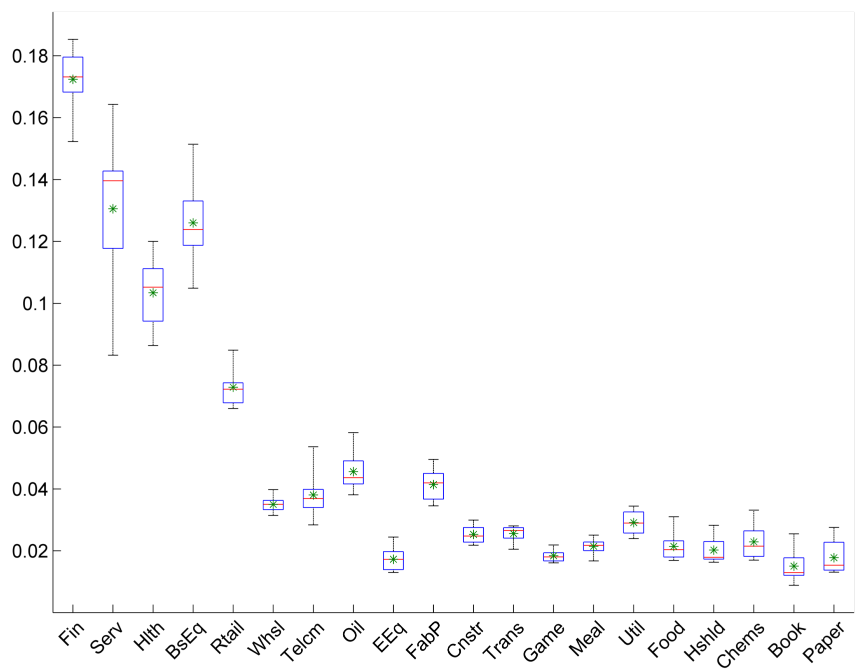
Table 19: Estimates of b_k, industry groups, mutual funds

	mean	s.d.	med	min	max
Fin	0.172	0.009	0.173	0.152	0.185
Serv	0.131	0.024	0.140	0.083	0.164
Hlth	0.103	0.010	0.105	0.086	0.120
BsEq	0.126	0.012	0.124	0.105	0.151
Rtail	0.073	0.006	0.072	0.066	0.085
Whsl	0.035	0.002	0.035	0.031	0.040
Telcm	0.038	0.007	0.037	0.028	0.054
Oil	0.046	0.006	0.044	0.038	0.058
EEq	0.017	0.004	0.017	0.013	0.024
FabP	0.041	0.005	0.042	0.035	0.050
Cnstr	0.025	0.003	0.025	0.022	0.030
Trans	0.026	0.002	0.027	0.021	0.028
Game	0.018	0.002	0.018	0.016	0.022
Meal	0.021	0.002	0.022	0.017	0.025
Util	0.029	0.003	0.029	0.024	0.034
Food	0.021	0.004	0.020	0.017	0.031
Hshld	0.020	0.004	0.018	0.016	0.028
Chems	0.023	0.005	0.022	0.017	0.033
Book	0.015	0.004	0.013	0.009	0.025
Paper	0.018	0.005	0.015	0.013	0.028

Note: the full names for the industry abbreviations are as follows. Fin: Finance, Serv: Service, Hlth: Health Care, BsEq: Business Equipment, Rtail: Retail, Whsl: Wholesale, Telcm: Telecommunication, Oil: Oil, EEq: Electrical Equipment, FabP: Fabricated Product, Cnstr: Construction, Trans: Transportation, Game: Recreation, Meal: Restaurant and Hotel, Util: Utility, Food: Food Products, Hshld: Consumer Products, Chems: Chemical, Book: Printing and Publishing, Paper: Paper Supplies.

Table 20 and Figure 9 present the estimated overdispersions ω_k for industry groups. Compared to the ω_k estimates from city groups, we notice a couple of key similarities. One is that there are some industries that show up as being much more overdispersed, such as Finance, Service and Utility. The other is that the t-statistics of testing the null Poisson distribution of $\omega = 1$ are also all significant at the 5% level, indicating the existence of delicate networks within each of the industry groups. On the other hand, it is observable that there is more overdispersion along industry classifications than city categories, which suggests that network

Figure 8: Boxplot of b_k estimates, industry groups, mutual funds



Note: 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 industry names, please refer to the note under Table 19.

connections are more prevalent along industry lines than city lines. Nonetheless, overdispersions are present independent of the type of group categorizations under consideration.

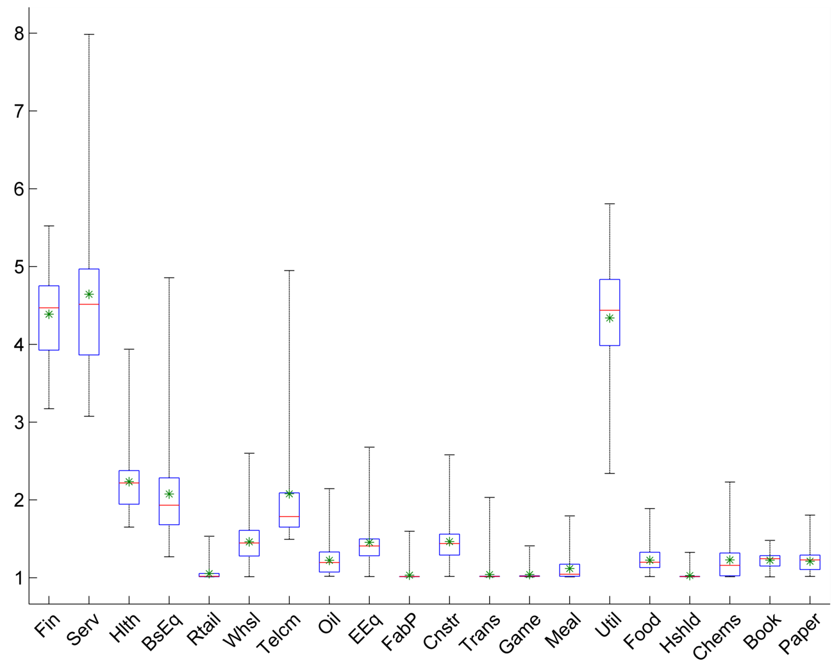
Table 20: Estimates of ω_k, industry groups, mutual funds

	mean	s.d.	med	min	max	t-stat
Fin	4.398	0.558	4.500	3.163	5.512	25.176
Serv	4.785	1.265	4.516	3.065	7.975	11.760
Hlth	2.225	0.427	2.209	1.641	3.927	14.404
BsEq	2.067	0.690	1.921	1.259	4.846	6.890
Rtail	1.043	0.093	1.009	1.001	1.524	3.342
Whsl	1.452	0.329	1.438	1.003	2.590	6.395
Telcm	2.069	0.718	1.777	1.484	4.939	6.222
Oil	1.215	0.181	1.185	1.008	2.135	6.073
EEq	1.445	0.353	1.399	1.005	2.669	5.700
FabP	1.024	0.087	1.004	1.001	1.587	1.873
Cnstr	1.455	0.326	1.429	1.006	2.570	6.133
Trans	1.031	0.126	1.007	1.002	2.023	1.809
Game	1.031	0.073	1.009	1.002	1.400	3.015
Meal	1.109	0.135	1.035	1.002	1.784	4.197
Util	4.330	0.748	4.428	2.331	5.796	20.664
Food	1.215	0.145	1.189	1.004	1.879	7.221
Hshld	1.020	0.055	1.006	1.002	1.316	2.418
Chems	1.219	0.262	1.149	1.003	2.221	3.670
Book	1.216	0.097	1.234	1.002	1.470	13.072
Paper	1.203	0.135	1.219	1.006	1.794	7.923

Note: for an explanation to the abbreviated industry names, please refer to the note under Table 19. 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).

Table 21 illustrates the correlations between the RPC measures g_{ik} and the gregariousness parameter estimates a_i of mutual funds for industry groups, while the summary statistics of the RPC scores $gsum_i$ are demonstrated in Table 22 as well as in Figure 10 and Table 23. They resemble the results in the main paper with cities as groups. In addition, we compute the correlation between fund managers' RPC scores using city groups and those using industry groups. Interestingly, managers' scores from city groups are not very correlated with their scores from industry groups and the correlation coefficient is approximately 0.2. Thus it suggests that city and industry networks can be dissimilar for investors.

Figure 9: Boxplot of ω_k estimates, industry groups, mutual funds



Note: 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 industry names, please refer to the note under Table 19.

Table 21: Correlations between g_{ik} and a_i, mutual funds, industry groups

Fin	Serv	Hlth	BsEq	Rtail	Whsl	Telcm	Oil	EEq	FabP
0.020	0.035	-0.006	0.016	0.090	0.328	-0.040	0.003	0.165	0.003
Cnstr	Trans	Game	Meal	Util	Food	Hshld	Chems	Book	Paper
0.291	0.200	0.148	0.222	0.003	-0.039	-0.024	0.016	0.057	0.036

Note: correlations are based on the Fama-MacBeth time-series means of g_{ik} (for each k) and a_i. For explanations on abbreviated industry group names, please refer to the notes under Table 19.

Table 22: Summary statistics of $gsum_i$, mutual funds

	mean	s.d.	med	min	max
Industry groups	19.61	4.79	19.28	0.30	167.5

Note: summary statistics are based on individual time-series averages of $gsum_i$.

Figure 10: Histogram of $gsum_i$, industry groups, mutual funds

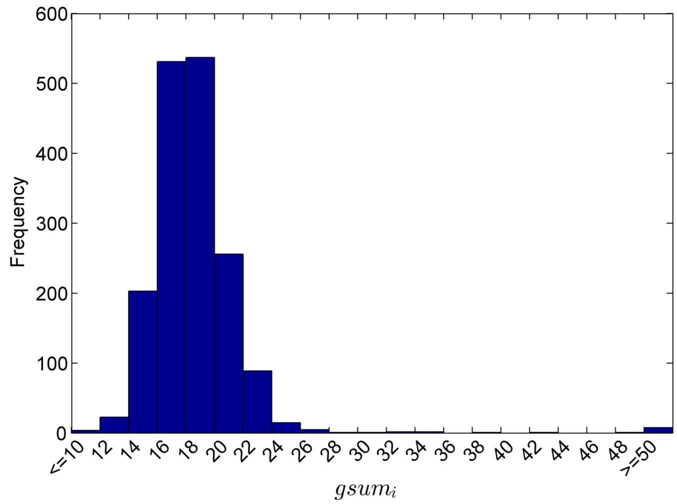


Table 23: Statistics on numbers of groups with $g_{ik} > 1$ for mutual funds having $gsum > 20$

	mean	s.d.	med	min	max
Industry groups	9.02	0.50	9.14	7.08	9.94

Note: the table contains the summary statistics on the average of the number of groups where $g_{ik} > 1$, for mutual funds that have $gsum > 20$.

Table 24 depicts our fund performance regression results using industry groups. Fund managers with higher RPC scores outperforms by about 2% per annum. As in the case of city groups, this result does not diminish even if ICI is included in the performance regression. Moreover, when we include both city RPC scores and industry RPC scores in the regression, both scores entail significant outperformance numbers. Nevertheless, the effect from city RPC scores is somewhat larger and more significant than the effect from industry RPC scores.

Table 24: RPC scores and mutual fund performances

	Industry			Both $gsum$'s
$const$	0.010	0.005	$const$	0.009
	(1.16)	(0.54)		(1.10)
$FundReturn_{t-1}$	0.061**	0.066**	$FundReturn_{t-1}$	0.061**
	(2.36)	(2.29)		(2.36)
$logTNA_{t-1}$	-0.0011**	-0.0009*	$logTNA_{t-1}$	-0.0011**
	(-2.05)	(-1.83)		(-2.16)
$logFamSize_{t-1}$	0.0000	0.0001	$logFamSize_{t-1}$	0.0001
	(0.36)	(0.50)		(0.51)
$ExpRatio_{t-1}$	-0.003***	-0.002*	$ExpRatio_{t-1}$	-0.003***
	(-4.43)	(-1.65)		(-4.33)
$Turnover_{t-1}$	-0.000	-0.000	$Turnover_{t-1}$	-0.000
	(-0.32)	(-0.13)		(-0.46)
$FundAge_{t-1}$	-0.000	-0.000	$FundAge_{t-1}$	-0.000
	(-1.25)	(-1.57)		(-1.06)
$gsum > 20$	0.0044**	0.0038**	$gsum > 20$, city	0.0040**
	(1.98)	(2.21)		(2.09)
FinCenter	0.0009***	0.0012***	$gsum > 20$, industry	0.0036*
	(3.64)	(3.81)		(1.92)
$\log(a_i)$	0.0017	0.0006	FinCenter	0.0010***
	(1.25)	(0.76)		(2.58)
ICI		0.0085		
		(1.01)		

Note: this table reports the Fama-MacBeth estimates of the regression coefficients in specification 4, 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's net return at quarter t. $gsum > 20$ 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). ICI denotes the Industry Concentration Index (ICI), which is constructed in a similar manner as in Kacperczyk, Sialm, and Zheng (2005). But for simplicity, we use an equally weighted index instead.

8.2 Retail Investor Results for City Groups

To ensure that our results on institutional investor networks are not driven by unique mutual fund industry considerations, we perform the same set of analyses on Barber and Odean (2001)'s retail investor stock holdings data. The relevant results with cities as groups are reported first, and the industry-group results are shown next. Because all the results on retail investor data are qualitatively similar to those on mutual fund data, we shall keep our discussions brief.

8.2.1 Transformation and Gregariousness Parameters

To start with, Table 25 and 26 illustrate respectively the transformation and the gregariousness parameter estimates with city groups for retail investors. In general, both of the two sets of estimates are smaller compared to those from mutual fund data. This is because retail investors hold a much smaller number of stocks in contrast to mutual fund managers.

Again, we are interpreting these coefficients as investor fixed effects. Though one could also reasonably conclude that retail investors are likely to have much smaller investor networks than mutual fund managers.

Table 25: Summary statistics, estimates of transformation parameter, retail investors

	mean	s.d.	med	min	max
City groups	1.06	0.03	1.06	1.02	1.12

Table 26: Summary statistics, estimates of a_i, retail investors

	mean	s.d.	med	min	max
City groups	14.1	8.0	12.5	3.5	194.9

8.2.2 Group Sizes and Overdispersion Parameters

Next, Table 27 shows the estimated values of relative city sizes b_k for retail investors. They are very close to the estimates from using mutual fund data. In particular, the correlation between the "mutual-fund" city group sizes and the "retail-investor" city group sizes is 0.96.

Table 28 then documents the overdispersion estimates with city group classifications for retail investors. In general, we find that the overdispersion parameters become smaller when retail investor data are used, suggesting that social networks effects are weaker for retail investors. However, qualitatively, the estimates are still similar to the ones from mutual fund data in that every city is overdispersed to some extent and that there are major overdispersions in certain cities (e.g. San Jose).

Table 27: Estimates of b_k, city groups, retail investors

	mean	s.d.	med	min	max
NY	0.201	0.013	0.208	0.173	0.214
LA	0.066	0.007	0.064	0.055	0.079
Bos	0.061	0.002	0.061	0.057	0.065
SF	0.071	0.005	0.070	0.064	0.079
Chi	0.113	0.003	0.114	0.107	0.117
SJ	0.086	0.018	0.079	0.069	0.126
Dal	0.061	0.002	0.061	0.055	0.065
Hou	0.045	0.006	0.044	0.036	0.056
Phi	0.034	0.001	0.034	0.031	0.037
Was	0.032	0.002	0.031	0.029	0.036
Mia	0.019	0.001	0.019	0.017	0.021
Atl	0.045	0.003	0.045	0.039	0.049
Min	0.028	0.001	0.028	0.026	0.029
Den	0.013	0.002	0.013	0.010	0.015
SD	0.018	0.002	0.018	0.012	0.022
Stfd	0.038	0.002	0.038	0.035	0.042
Sea	0.027	0.001	0.028	0.025	0.029
Phx	0.013	0.002	0.012	0.010	0.015
SL	0.014	0.001	0.014	0.013	0.015
Det	0.015	0.001	0.015	0.014	0.016

8.2.3 RPC scores and Portfolio Returns

Lastly, we depict for retail investors their RPC scores and how these RPC scores are tied to their investment portfolio returns. Importantly, Table 30 suggests that retail investors' RPC scores also lead to outperformance in their common-stock portfolios, similar to our institutional investor results. For retail investors, the outperformance is about 1.33% per year using city groups.[14]

Through comparing the institutional and retail investor results, we can see that social networks are more prevalent among mutual fund managers than among ordinary households. However, the qualitative resemblance between the two sets of results implies that the impact of

[14]Remember the holdings data of retail investors are at the monthly level. Hence the outperformance number — the coefficient in front of $gsum > 5$ in Table 30 is at the monthly level too.

Table 28: Estimates of ω_k, city groups, retail investors

	mean	s.d.	med	min	max	t-stat
NY	1.032	0.020	1.016	1.010	1.111	1.99
LA	1.110	0.067	1.096	1.009	1.275	6.91
Bos	1.175	0.116	1.099	1.053	1.390	4.98
SF	1.156	0.107	1.205	1.027	1.317	4.94
Chi	1.154	0.065	1.177	1.014	1.255	9.61
SJ	2.006	0.309	1.884	1.616	2.567	7.79
Dal	1.014	0.011	1.005	1.002	1.077	3.24
Hou	1.043	0.030	1.049	1.002	1.107	12.71
Phi	1.028	0.021	1.026	1.001	1.106	4.26
Was	1.013	0.008	1.012	1.001	1.063	3.86
Mia	1.244	0.041	1.252	1.169	1.343	21.86
Atl	1.006	0.005	1.004	1.001	1.017	1.87
Min	1.436	0.094	1.465	1.209	1.578	13.28
Den	1.087	0.040	1.082	1.004	1.153	11.26
SD	1.127	0.098	1.147	1.010	1.291	4.77
Stfd	1.004	0.003	1.004	1.001	1.010	1.91
Sea	1.341	0.076	1.358	1.181	1.454	13.24
Phx	1.048	0.034	1.032	1.010	1.158	5.92
SL	1.125	0.040	1.130	1.050	1.267	15.24
Det	1.045	0.016	1.043	1.012	1.074	7.02

social networks seems to be universal and is not confined to the particular system of the mutual fund industry.

Table 29: Summary statistics of $gsum_i$, retail investors

	mean	s.d.	med	min	max
City groups	18.57	7.00	17.62	4.82	161.9

Table 30: RPC scores and retail investor portfolio performances, city groups

Dep Var: $return_t$							
$const$	$return_{t-1}$	$PortVal$	$commission$	$turnover$	$HHequity$	$gsum > 5$	$\log(a_i)$
0.008**	0.031	0.0001	-0.0011**	0.001	0.0002**	0.0011**	-0.0007
(2.05)	(1.47)	(0.29)	(-1.97)	(0.81)	(2.36)	(2.13)	(-1.10)

Note: dependent variable is $return_t$, the monthly gross return on a investor's common-stock portfolio. Each column stands for one specific regressor. t-statistics based on Newey-West HAC standard errors (of lag order 12) are shown in parentheses. *, ** and *** denote statistical significance at the 10%, 5% and 1% levels respectively. $PortVal$ is the market value of an investor's stock portfolio measured in logs, $commission$ is the monthly commissions paid (from trades) as a percentage of the $PortVal$, and $HHequity$ is the total household equity value of an investor measured in logs. For retail investors, the RPC score dummy, $gsum > 5$, is over the largest 5 groups only where investors actually have a meaningful number of stock picks.

8.3 Retail Investor Results for Industry Groups

Table 31 and 32 illustrate the transformation and gregariousness parameter estimates respectively for retail investors, using industry as groups. These estimates are similar to the city-group estimates shown in the main paper.

Table 31: Summary statistics, estimates of transformation parameter, retail investors

	mean	s.d.	med	min	max
Industry groups	1.10	0.05	1.09	1.03	1.20

Table 33 shows the estimated values of b_k of industry groups for retail investors. They are very close to the estimates from using mutual fund data. In particular, the correlation between the "mutual-fund" industry group sizes and the "retail-investor" industry group sizes is 0.87.

Table 34 then documents the overdispersion estimates with industry group classifications for retail investors. Similar to the case of city groups, we find that the overdispersion parameters

Table 32: Summary statistics, estimates of a_i, retail investors

	mean	s.d.	med	min	max
Industry groups	16.7	10.0	14.7	1.2	270.1

become smaller when retail investor data are used instead of mutual fund data, but qualitatively the estimates are still similar to the ones from mutual fund data.

Table 33: Estimates of b_k, industry groups, retail investors

	mean	s.d.	med	min	max
Fin	0.119	0.006	0.119	0.104	0.128
Serv	0.080	0.011	0.078	0.067	0.101
Hlth	0.146	0.010	0.149	0.129	0.160
BsEq	0.145	0.012	0.139	0.131	0.173
Rtail	0.073	0.004	0.074	0.065	0.076
Whsl	0.020	0.002	0.021	0.017	0.023
Telcm	0.051	0.002	0.051	0.048	0.057
Oil	0.046	0.006	0.045	0.035	0.054
EEq	0.021	0.005	0.020	0.014	0.033
FabP	0.027	0.001	0.027	0.026	0.030
Cnstr	0.025	0.001	0.025	0.023	0.028
Trans	0.018	0.001	0.018	0.017	0.021
Game	0.016	0.004	0.016	0.011	0.024
Meal	0.020	0.001	0.020	0.019	0.022
Util	0.039	0.002	0.039	0.035	0.044
Food	0.050	0.004	0.050	0.043	0.056
Hshld	0.044	0.003	0.044	0.038	0.050
Chems	0.029	0.002	0.029	0.023	0.031
Book	0.016	0.001	0.016	0.014	0.017
Paper	0.015	0.001	0.015	0.013	0.018

Lastly, we depict for retail investors their RPC scores based on industry groups and how these RPC scores are related to their investment portfolio returns. Table 36 demonstrates that the industry RPC scores lead to an outperformance of about 1.45% per annum in retail investors' common-stock portfolios, which is very close to the number found in city RPC scores.

Table 34: Estimates of ω_k, industry groups, retail investors

	mean	s.d.	med	min	max	t-stat
Fin	1.575	0.087	1.546	1.454	1.723	17.11
Serv	1.364	0.204	1.281	1.110	1.995	5.44
Hlth	1.865	0.129	1.855	1.646	2.231	19.15
BsEq	1.939	0.469	1.704	1.427	2.832	4.71
Rtail	1.248	0.045	1.260	1.147	1.332	19.73
Whsl	1.137	0.127	1.084	1.005	1.363	2.98
Telcm	1.928	0.090	1.908	1.793	2.065	24.84
Oil	1.399	0.053	1.408	1.300	1.498	21.38
EEq	1.200	0.165	1.186	1.003	1.482	3.82
FabP	1.160	0.068	1.143	1.035	1.304	8.64
Cnstr	1.068	0.055	1.048	1.007	1.240	4.95
Trans	1.115	0.075	1.102	1.018	1.302	4.63
Game	1.068	0.064	1.079	1.002	1.184	6.37
Meal	1.059	0.030	1.060	1.011	1.125	6.42
Util	2.547	0.166	2.531	2.324	2.932	23.24
Food	1.313	0.028	1.319	1.237	1.374	43.39
Hshld	1.026	0.019	1.023	1.004	1.071	5.17
Chems	1.143	0.077	1.162	1.017	1.252	9.39
Book	1.109	0.032	1.106	1.037	1.178	19.27
Paper	1.043	0.033	1.022	1.003	1.243	2.76

Table 35: Summary statistics of $gsum_i$, retail investors

	mean	s.d.	med	min	max
Industry groups	18.60	6.72	17.87	8.70	196.2

Table 36: RPC scores and retail investor portfolio performances

	Industry		Both $gsum$'s
$const$	0.008**	$const$	0.007*
	(2.17)		(1.88)
$return_{t-1}$	0.030	$return_{t-1}$	0.030
	(1.41)		(1.47)
$PortVal$	0.0002	$PortVal$	0.0000
	(0.44)		(0.08)
$commission$	-0.0011**	$commission$	-0.0011**
	(-2.05)		(-2.26)
$turnover$	0.001	$turnover$	0.001
	(0.93)		(0.90)
$HHequity$	0.0002**	$HHequity$	0.0003**
	(2.52)		(2.18)
$gsum > 5$	0.0012**	$gsum > 5$, city	0.0006*
	(2.12)		(1.81)
$\log(a_i)$	-0.0006	$gsum > 5$, industry	0.0010**
	(-1.36)		(2.06)

Note: dependent variable is $return_t$, the monthly gross return on a investor's common-stock portfolio. t-statistics based on Newey-West HAC standard errors (of lag order 12) are shown in parentheses. *, ** and *** denote statistical significance at the 10%, 5% and 1% levels respectively. $PortVal$ is the market value of an investor's stock portfolio measured in logs, $commission$ is the monthly commissions paid (from trades) as a percentage of the $PortVal$, and $HHequity$ is the total household equity value of an investor measured in logs. For retail investors, the RPC score dummy, $gsum > 5$, is over the largest 5 groups only where investors actually have a meaningful number of stock picks.

References

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.

Barber, B. M., and T. Odean, 2000, "Trading is Hazardous to Your Wealth: The Common Stock Investment Performance of Individual Investors," *Journal of Finance*, 55, 773–806.

——— , 2001, "Boys will be Boys: Gender, Overconfidence, and Common Stock Investment," *Quarterly Journal of Economics*, 116, 261–292.

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.

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., J. D. Gottlieb, and K. Tobio, 2012, "Housing Booms and City Centers," NBER Working Paper 17914.

Glaeser, E. L., and J. A. Scheinkman, 2001, "Measuring Social Interactions," *Social Dynamics,*, S. Durlauf and P. Young (eds.), MIT Press.

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

Han, B., and D. Hirshleifer, 2013, "Self-Enhancing Transmission Bias and Active Investing," Working Paper, http://sites.uci.edu/dhirshle/files/2013/07/SET_6_8_2013_v1.pdf.

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., 1994, "Consistency in a Proportional Hazards Model Incorporating a Random Effect," *Annals of Statistics*, 22, 712–731.

——— , 1995, "Asymptotic Theory for the Frailty Model," *Annals of Statistics*, 23, 182–198.

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.

Pool, V. K., N. Stoffman, and S. E. Yonker, 2012, "No Place Like Home: Familiarity in Mutual Fund Manager Portfolio Choice," *Review of Financial Studies*, 25, 2563–2599.

Shiller, R. J., and J. Pound, 1989, "Survey evidence on diffusion of interest and information among investors," *Journal of Economic Behavior and Organization*, 12, 47–66.

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.