



**HAL**  
open science

## Wisdom of the institutional crowd

Kevin Primicerio, Damien Challet, Stanislao Gualdi

► **To cite this version:**

Kevin Primicerio, Damien Challet, Stanislao Gualdi. Wisdom of the institutional crowd. 2017. hal-01484914

**HAL Id: hal-01484914**

**<https://hal.science/hal-01484914>**

Preprint submitted on 8 Mar 2017

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Wisdom of the institutional crowd

Kevin Primicerio<sup>1</sup>, Damien Challet<sup>1,2</sup>, and Stanislao Gualdi<sup>3</sup>

<sup>1</sup>Laboratory of Mathematics in Interaction with Computer Science, CentraleSupélec, Université Paris Saclay, Grande Voie des Vignes, 92290 Châtenay-Malabry, France

<sup>2</sup> Encelade Capital SA, Innovation Park, Building C, EPFL, 1015 Lausanne, Switzerland

<sup>3</sup>Capital Fund Management S.A., 23, rue de l'Université, 75007 Paris, France

March 8, 2017

## Abstract

The average portfolio structure of institutional investors is shown to have properties which account for transaction costs in an optimal way. This implies that financial institutions unknowingly display collective rationality, or Wisdom of the Crowd. Individual deviations from the rational benchmark are ample, which illustrates that system-wide rationality does not need nearly rational individuals. Finally we discuss the importance of accounting for constraints when assessing the presence of Wisdom of the Crowd.

## 1 Introduction

The collective ability of a crowd to accurately estimate an unknown quantity is known as the “Wisdom of the Crowd” [1] (WoC thereafter). In many situations, the median estimate of a group of unrelated individuals is surprisingly close to the true value, sometimes significantly better than those of experts [2, 3, 4, 5]. WoC may only hold under some conditions [1, 6]: for example social imitation is detrimental as herding may significantly bias the collective estimate [7, 8]. WoC is a reminiscent of collective rationality without explicit individual rationality: when it applies, it is a consistent aggregation of possibly inconsistent individual estimates [9]. This is to be contrasted with the economic paradigm that collective rationality reflects individual rationality, where only a “typical” decision maker – the representative agent – is considered [10] or team reasoning where the individual agents explicitly optimize the collective welfare [11].

Although almost all known examples of WoC are about a single number, there is no reason why it could not hold for whole functional relationships between several quantities. For example, Haerdle and Kirman analyze the prices and weights of many transactions in Marseille fish market: while the relationship between these two quantities is rather noisy, a local average does display the expected quantity discount [12]. More generically, many simple relationships found in Economic textbooks may only hold on average, but not for each agent or each transaction. Aggregation of quite diverse individual actions is still an open problem [13].

Asset price efficiency is an obvious instance of WoC in Finance: it states that current prices, determined by the actions of many traders, are the best possible estimates and fully reflect all available information [14, 15, 16]. Another WoC candidate in Finance is portfolios. The rationale is that many market participants, especially investment funds, strive to build optimal portfolios following their own criteria and constraints. If anything, the average portfolio structure of this population may be close to that of a rational benchmark.

In this paper we follow an approach similar to [17], where the benchmark is a simple mean-variance portfolio optimization with a given transaction cost structure. The relationship between the optimal number of assets in a portfolio and the portfolio value depends strongly on the transaction cost structure. Such dependence is found in the average portfolio structure of individual and professional investors over many orders of magnitude of portfolio value. In this article we focus on much larger investment funds facing additional constraints and show how the latter shape WoC.

## Results

The added value of an investment fund resides in how well the latter allocates capital to securities with respect to some predefined criterion (return on investment, risk, performance tracking, etc.). Many kinds of funds exist, each with its own objective and methods. Thus, understanding the minute structure of the portfolio of a given fund is both impractical and fortunately not necessary in our approach as we do not need to assume individual rationality. Therefore, we shall focus on simple properties of portfolios whose average can be related to a rational benchmark. Generally speaking, at time  $t$ ,

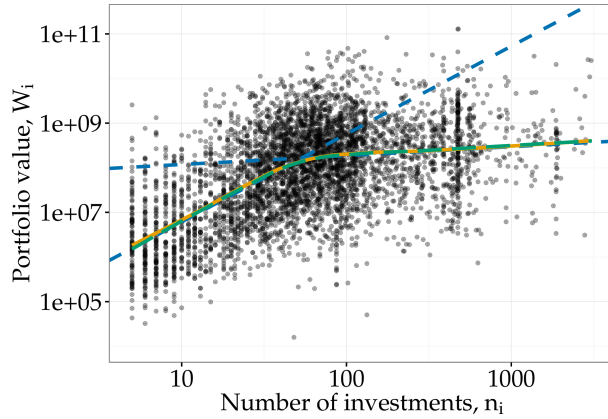


Figure 1: Total mark-to-market value  $W_i$  as a function of the number of investments  $n_i$ , with a robust locally weighted regression fit (yellow line) and two linear fits (blue dashed lines) for two different ranges of  $n$ . Robust locally weighted regression fit for the simulated data (in green).

fund  $i$  has capital  $W_i(t)$  and may invest it into  $N(t)$  existing securities. As a result, each security  $\alpha$  is found in  $m_\alpha(t)$  portfolios. The explicit time dependence is dropped hereafter for the sake of clarity. All the plots which are not time dependent refer to  $t = 2013Q1$  but the same qualitative behavior is observed for other dates.

In the following we focus on the number of assets in portfolio of fund  $i$ , denoted by  $n_i$ , and on its relationship with the total assets under management  $W_i$ . We also investigate how the asset selection of funds shapes the relationship between the capitalization of asset  $\alpha$ , denoted by  $C_\alpha$ , and the number of funds which have invested in this security,  $m_\alpha$ . The key point is that only  $n_i$  depends on asset allocation strategies of fund  $i$  while  $W_i$ ,  $m_\alpha$  and  $C_\alpha$  are beyond its control. Given its scalar simplicity,  $n_i$  must be regarded as a convenient and concise proxy of possibly complex asset allocation algorithms, thus as a useful simplification. WoC, if any, is to be found in the relation between  $n_i$  and the three other quantities. Figure 1 plots  $W_i$  versus  $n_i$  in logarithmic scale: a cloud of point emerges, with a roughly increasing trend. The large amount of noise confirms the great diversity of fund allocation strategies. WoC may only appear in some average behavior. This is why we computed a locally weighted polynomial regression [18]. Remarkably, two distinct regions appear, each characterized by a roughly linear behavior. The axes of this figure being logarithmic,  $\log W_i = \mu_x \log n_i + \epsilon_i$ , where  $x \in \{<, >\}$  labels the region, and  $\epsilon_i$  is the deviation of  $W_i$  from its local average  $W$ . In other words, defining  $n$  as the local average of  $n_i$ , the average portfolio follows

$$W \propto n^{\mu_x}. \quad (1)$$

The boundary  $n^*$  between the two regions is algorithmically determined for each quarterly snapshot (see S.I.), which allows us to measure the two exponents  $\mu_{<}$  and  $\mu_{>}$ . The latter are quite stable as a function of time (see Fig. 11 in S.I.); their time-averages  $\overline{\mu_{<}} \simeq 2.1 \pm 0.2$  and  $\overline{\mu_{>}} \simeq 0.3 \pm 0.1$  are markedly different, which points to distinct collective ways of building portfolios in these two regions. The cross-over point  $n^*$  is also stable as time goes on (see Fig. 11 in S.I.).

Let us start with the small diversification region ( $n_i < n^*$ ). First, the time average of exponent  $\mu_{<}$  is close to 2. The same exponent was found in the portfolios of individual and professional investors, whose portfolio values were much smaller than investment funds [17]. This is not entirely surprising: the value  $\mu = 2$  corresponds to optimal mean-variance portfolios with constant cost per transaction [17] (see also S.I.), provided that the same amount is invested in all the chosen assets (equally-weighted portfolios). One checks that large institutional portfolios are indeed very close to being equally weighted in this region. Take fund  $i$ ; denoting by  $p_{i\alpha} = W_{i\alpha}/W_i$  its investment fraction in security  $\alpha$ , its diversification  $S_i$  may be measured with Shannon Entropy  $S_i = -\sum_{\alpha} p_{i\alpha} \log p_{i\alpha}$ , which equals 1 and is maximal when all the non-null  $p_{i\alpha}$  are equal. As shown in Fig. 2, it is the case for low-diversification funds. The small deviations from  $S_i = 1$  in this region are due at least in part to the fact that the values of assets evolve after the initial investment and that portfolios are not rebalanced just before the dates of our dataset.

Since one finds the same exponent  $\mu$  over many decades of portfolio values for a wide spectrum of market participants, and since  $\mu = 2$  corresponds to a realistic transaction cost per transaction, we argue that WoC is a plausible explanation of the average portfolio structure. Note that  $\mu = 2$  does not imply that funds really face constant transaction cost per transaction, only that their population acts as if it does. Finally, the fact that in this view WoC holds for a whole functional relationship, not only a single number, considerably extends its reach.

So far, bringing to light WoC only required to focus on the number of securities in a portfolio, not on how funds select securities. This implicitly assumed that funds could invest in all securities they wished, which is clearly not the case in the large diversification region: the fact that the exponent  $\mu$  is much smaller in this region implies that funds need on average to split their investments into many more securities. This is most likely due to liquidity constraints: large funds cannot

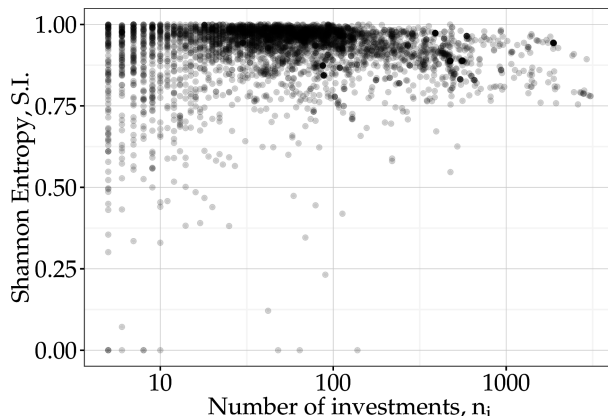


Figure 2: Shannon Entropy  $S_i$  as a function of the number of investments  $n_i$ , each dot represents a single fund.

invest as much as they wish in some assets because there are simply not enough shares to build a position larger than a certain size without impacting too much their prices. Each fund has its own way to determine the maximal amount to invest in a given security  $\alpha$ ; a common criterion is to limit the fraction  $W_{i\alpha}/C_\alpha$ . Fig. 8 in S.I. strongly suggests that each fund fixes its upper bound

$$f_i^{(\max)} \geq \max_\alpha f_{i\alpha} \quad \text{where } f_{i\alpha} = \frac{W_{i\alpha}}{C_\alpha}. \quad (2)$$

It turns out that  $f_i^{(\max)}$  is highly heterogeneous among funds  $\log_{10}(f_i^{(\max)}) \simeq -3.0 \pm 1.0$  (see Fig. 9). The existence of such limits implies that portfolios are less likely to be equally weighted in the large diversification region, which is confirmed by the decrease of the typical portfolio weights entropy (see Fig. 2).

Funds, however, do not invest in a randomly chosen security, even in the low diversification region. Figure 3 displays a scatter plot of the capitalization  $C_\alpha$  of each security  $\alpha$  versus  $n_\alpha$ , the number of funds which have invested in this security, together with a local non-linear fit. As for  $W$  vs  $n$ , one finds a power-law relationship

$$\log C_\alpha = \gamma \log m_\alpha + \epsilon_\alpha \quad (3)$$

for large enough  $m$  (see S.I.). Hence  $C \propto m^\gamma$ , where  $m$  is the local average of  $n_\alpha$ . Exponent  $\gamma$  is stable during the period 2007-2014 (see Fig. 11 in S.I.) and its average  $\gamma \simeq 2.2 \pm 0.1$ .

In short, one needs to introduce a model of how funds choose to invest in securities both to bring WoC to light in the large diversification region and to reproduce the average behavior of both Eqs (3) and 1. Since one sees a cross-over between two types of behavior rather than an abrupt change, we create logarithmic bins of the axis  $n_i$  and denote the bin number of fund  $i$  by  $[n_i]$ . Two mechanisms must be specified: how a fund selects security  $\alpha$  and how much it invests in it. The latter point is dictated by Fig. 8 in the large  $n_i$  region where fund  $i$  invests  $W_{i\alpha} = f_i^{(\max)} C_\alpha$ ; for the sake of simplicity, we approximate  $f_i^{(\max)}$  by the median value of  $f_i^{(\max)}$  in the bin  $[n_i]$ , denoted by  $f_{[n_i]}^{(\max)}$ . In the small diversification region, we assume that  $n_i = n_i^{\text{opt}}$ , thus  $W_{i\alpha} = W_i/n_i^{\text{opt}}$  to be consistent with our previous results. We choose a security selection mechanism that rests on the market capitalization  $C_\alpha$  of a security  $\alpha$  (see S.I.) which is a good proxy of the liquidity (Fig. 10). We perform Monte-Carlo simulations from the empirical selection probabilities and  $f_{[n_i]}^{(\max)}$  and display the resulting  $W$  vs  $n$  and  $C$  vs  $m$  in Figs 1 and 3 (continuous green lines), in good agreement with the local averages (continuous orange lines). One notices a discrepancy in the relationship  $C$  vs  $m$  for large  $n$ , which mainly comes from funds in the large diversification region. (See Fig 12 S.I).

The large diversification region illustrates how constraints may considerably modify the rational benchmark. While the above mechanism of security selection is able to reproduce adequately the behavior of well diversified funds, we could not find a rational benchmark for the dependence of  $f^{\max}$  and  $n_i$ . Thus, the case for WoC in the large diversification region is not entirely closed.

## Data

Our dataset consists of an aggregation of the following publicly available reports (in order of reliability): the SEC Form 13F, the SEC's EDGAR system forms N-Q and N-CSR and (occasionally) the form 485BPOS. Our work focuses on the period starting from the first quarter of 2005 to the last quarter of 2013.

These forms are filled manually and are thus error prone. We partially solve this issue by cross-checking different sources (which often contains overlapping information) and by filtering data before processing (see details in S.I.).

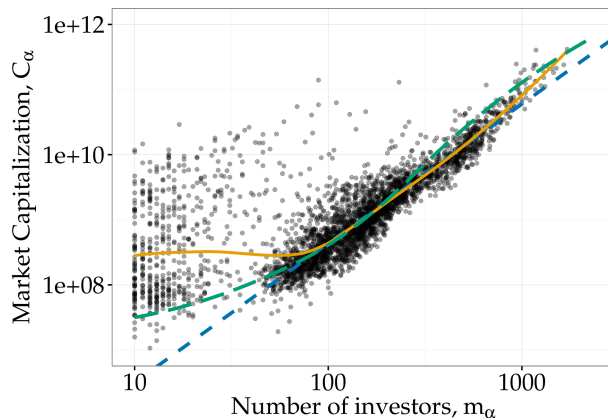


Figure 3: Market capitalization of securities as a function of the number of investors in logarithmic scale. From the local non-linear robust fit (yellow line) we observe a linear relationship for assets with more than about 100 investors. The blue dashed line corresponds to a linear fit on that group of asset. Hence  $W_\alpha \propto m_\alpha^\gamma$ , with  $\gamma \simeq 2.1$ . Robust locally weighted regression fit for the simulated data (in green).

The main limitation of this dataset is that it provides accurate figures for long positions only. The other positions (short, bonds, ...) are most of the time only partially known. The frequency of the dataset is also inhomogeneous: data for most of the funds are quarterly updated (depending on regulations), hence we decided to restrict ourselves to 4 points in a year only. Such frequency is probably too low for investigating the dynamics of individual behavior but is not a problem for we focus on an aggregate and static representation of the investment structure.

## Discussion and conclusion

While WoC is commonly applied to a population collectively guessing a single number, we investigate here a fundamentally different situation and provide evidence for a collective functional optimization of the asset ownership structure. What the reference function should be is dictated by optimality arguments. In the case of financial markets, the rational benchmark was not related to the efficient market hypothesis, but to the way a large population of professional fund managers build their portfolios. Whereas each fund has its own benchmark with respect to which the fund performance may be assessed, this, fortunately, has no discernible influence on the average structure of their portfolio. In addition, WoC is often meant as a collective guessing of non-experts; one thus may conclude that the population investigated here has decidedly more expertise than the subjects of other WoC studies. What kind of expertise the typical fund manager has is not obvious, at least when one looks at their pure performance (see e.g. [19]). In addition, the optimal relationship between the number of assets in a portfolio and the value of the latter is clearly not broadly known in these circles, as shown by the very large deviations from the ideal case in Fig. 1, and the collective expertise only appears when their decisions are suitably averaged. The presence of WoC when the subjects face strong constraints, as those of highly diversified funds, is more conjectural, and more work will be needed in that respect.

At a higher level, our results suggest that, while individuals may deviate much from the rational expectation theory, standard economic theory may hold at a collective level, without need for micro-founded individual decisions: the average decision may in some cases be approximated by a rational, representative agent. Our results however only hold on a snapshot of the system, for which individual fluctuations may be averaged out. In a dynamic setting, the very large deviations from the rational benchmark may not be neglected in the presence of feedback loops [20]. In other words, the dynamics of these fluctuations are worth investigating in their own right.

## References

- [1] J. Surowiecki, *The wisdom of crowds*. Anchor, 2005.
- [2] F. Galton, “Vox populi (the wisdom of crowds),” *Nature*, vol. 75, pp. 450–51, 1907.
- [3] S. Hill and N. Ready-Campbell, “Expert stock picker: the wisdom of (experts in) crowds,” *International Journal of Electronic Commerce*, vol. 15, no. 3, pp. 73–102, 2011.
- [4] H. E. Landemore, “Why the many are smarter than the few and why it matters,” *Journal of public deliberation*, vol. 8, no. 1, 2012.

- [5] M. Nofer and O. Hinz, “Are crowds on the internet wiser than experts? the case of a stock prediction community,” *Journal of Business Economics*, vol. 84, no. 3, pp. 303–338, 2014.
- [6] C. P. Davis-Stober, D. V. Budescu, J. Dana, and S. B. Broomell, “When is a crowd wise?,” *Decision*, vol. 1, no. 2, p. 79, 2014.
- [7] J. Lorenz, H. Rauhut, F. Schweitzer, and D. Helbing, “How social influence can undermine the wisdom of crowd effect,” *Proceedings of the National Academy of Sciences*, vol. 108, no. 22, pp. 9020–9025, 2011.
- [8] L. Muchnik, S. Aral, and S. J. Taylor, “Social influence bias: A randomized experiment,” *Science*, vol. 341, no. 6146, pp. 647–651, 2013.
- [9] R. M. Hogarth, “A note on aggregating opinions,” *Organizational Behavior and Human Performance*, vol. 21, no. 1, pp. 40–46, 1978.
- [10] J. E. Hartley and J. E. Hartley, *The representative agent in macroeconomics*. Routledge, 2002.
- [11] A. M. Colman, B. D. Pulford, and J. Rose, “Collective rationality in interactive decisions: Evidence for team reasoning,” *Acta psychologica*, vol. 128, no. 2, pp. 387–397, 2008.
- [12] W. Härdle and A. Kirman, “Nonclassical demand: A model-free examination of price-quantity relations in the Marseille fish market,” *Journal of Econometrics*, vol. 67, pp. 227–257, 1995.
- [13] A. P. Kirman, “Whom or what does the representative individual represent?,” *The Journal of Economic Perspectives*, vol. 6, no. 2, pp. 117–136, 1992.
- [14] B. G. Malkiel and E. F. Fama, “Efficient capital markets: A review of theory and empirical work,” *The journal of Finance*, vol. 25, no. 2, pp. 383–417, 1970.
- [15] B. G. Malkiel, “The efficient market hypothesis and its critics,” *The Journal of Economic Perspectives*, vol. 17, no. 1, pp. 59–82, 2003.
- [16] E. F. Fama, “Market efficiency, long-term returns, and behavioral finance,” *Journal of financial economics*, vol. 49, no. 3, pp. 283–306, 1998.
- [17] D. M. de Lachapelle and D. Challet, “Turnover, account value and diversification of real traders: evidence of collective portfolio optimizing behavior,” *New Journal of Physics*, vol. 12, no. 7, p. 075039, 2010.
- [18] W. S. Cleveland, E. Grosse, and W. M. Shyu, “Local regression models,” *Statistical models in S*, vol. 2, pp. 309–376, 1992.
- [19] L. Barras, O. Scaillet, and R. Wermers, “False discoveries in mutual fund performance: Measuring luck in estimated alphas,” *The Journal of Finance*, vol. 65, no. 1, pp. 179–216, 2010.
- [20] S. Gualdi, M. Tarzia, F. Zamponi, and J.-P. Bouchaud, “Tipping points in macroeconomic agent-based models,” *Journal of Economic Dynamics and Control*, vol. 50, pp. 29–61, 2015.
- [21] V. M. Muggeo, “Estimating regression models with unknown break-points,” *Statistics in medicine*, vol. 22, no. 19, pp. 3055–3071, 2003.

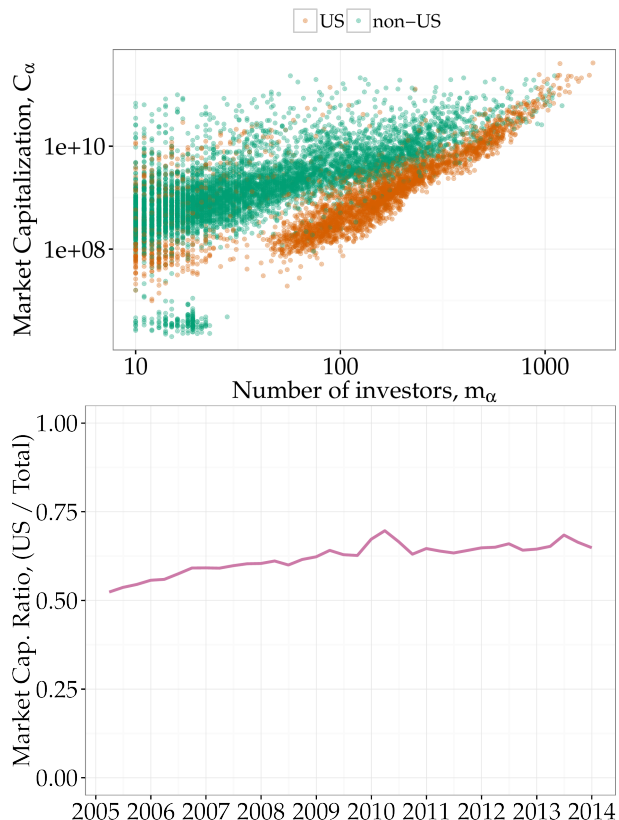


Figure 4: Top: Market capitalization as a function of the number of investors for all securities. Bottom: Temporal evolution of the aggregated market capitalization of US over the total market capitalization.

## Supporting Information (SI)

### 2 Filtering

In order to remove inconsistencies in the dataset, we applied the following filters

#### 2.1 Country of origin

Our dataset is sparse and heterogeneous. Indeed, the quality of the sources of data is directly related to each country’s disclosure regulations. For these reasons we decided to keep only the entities which use an US based mail address.

About 60% of the total market capitalization of the dataset is concentrated in US based securities. Figure 4 shows two large clouds of dots, each of them correspond to a different region of origin: green (resp. orange) cloud corresponds to non-US (resp. US) based securities. The origin of this large difference between these two regions are not clear: it could for example come from differences in regulations in non-US countries. It turns out that the ratio of the investment values in US and non-US assets varies little as a function of time (see Fig. 4), which does not affect the exponent  $\mu$  in Eq. 1. As a consequence we focused on US securities.

#### 2.2 Frequency

Large funds are requested to report their positions at a frequency which depends on the applicable regulation. As a result, reporting frequency ranges from monthly to yearly, most funds filing quarterly reports. We therefore focused of the latter.

#### 2.3 Penny Stocks

The “penny stocks”, i.e., usually securities which trade below \$5 per share in the USA, are not listed on a national exchange. Since they are considered highly speculative investments and are subject to different regulations, we filtered them out.

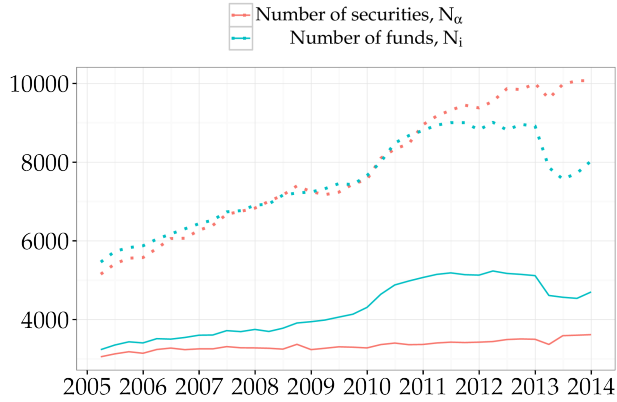


Figure 5: Temporal evolution of the number of funds  $N_i$  and securities  $N_\alpha$  in the data base. Unfiltered in dashed lines and US based only in solid lines.

## 2.4 Size

We also filtered out small funds and securities :  $W_i > 10^5$  USD,  $C_\alpha > 10^5$  USD,  $n_i \geq 5$ ,  $m_\alpha \geq 10$ .

## 2.5 Output

We restricted our study to 36 quarterly snapshots starting from the first quarter of 2005 and ending with the last quarter of 2013. Figure 5 reports the evolution of the number of securities and funds in the database before and after filtering.

# 3 Asset selection

The framework we introduce in this paper follows a series of a few elementary steps described below. The aim is for the model to be sensitive to the different constraints which dominates the portfolio selection of a fund.

## 3.1 Finding $n^*$

For date  $t$ , we define the cross-over point  $n^*$  between the two regions which appear in the local polynomial regression. We determine this point value with a likelihood maximization of the model

$$W = \mu_{<}n + (\mu_{>} - \mu_{<})(n - n^*)\theta(n - n^*), \quad (4)$$

where  $\theta(x)$  is the Heaviside function. We use a recursive method to find parameters  $\mu_{<}$ ,  $\mu_{>}$  and  $n^*$  [21]. Figure 11 shows that  $n^*$  is stable as a function of time.

## 3.2 Selection algorithm

### 3.2.1 Small diversification region $n_i < n^*$

In this region, we consider the equally weighted portfolio hypothesis to be true. Each position has a size  $\frac{W_i}{n_i^{\text{opt}}}$ , where  $n_i^{\text{opt}}$  is the optimal number of position computed with eq 1. The funds select their asset randomly with a probability proportional to  $C_\alpha$ . Also, in order to construct an equally-weighted portfolio, a position is valid only if it is of size  $\frac{W_i}{n_i^{\text{opt}}}$ .

### 3.2.2 Large diversification region $n_i \geq n^*$

In this region, the liquidity constraints make it harder for funds to keep an equally weighted portfolio and portfolio values are thus spread on a larger number of assets. We model this process by introducing both a maximum fraction  $f$  which materializes the lack of liquidity and a probabilistic asset selection mechanism based on the rescaled rank of the security in terms of capitalization. In particular, we extract the rescaled rank from a beta distribution and take the corresponding security. The parameters of the distribution are fitted from the empirical density function (see later) Fig. 6. <sup>1</sup>

<sup>1</sup>We do not use the same rank-based selection mechanism in the low-diversification region because in this case it is harder to have a good fit with the beta distribution. This is however only a minor point since the capitalization is approximately power-law distributed and the two selection mechanisms are basically equivalent (the rank is proportional to a power of the capitalization). Results are indeed very similar in both cases. We chose a security selection mechanism which rests on the scaled rank of capitalization of security  $\alpha$ , defined as  $\rho_\alpha = \frac{r_\alpha}{N}$  where  $r_\alpha$  is



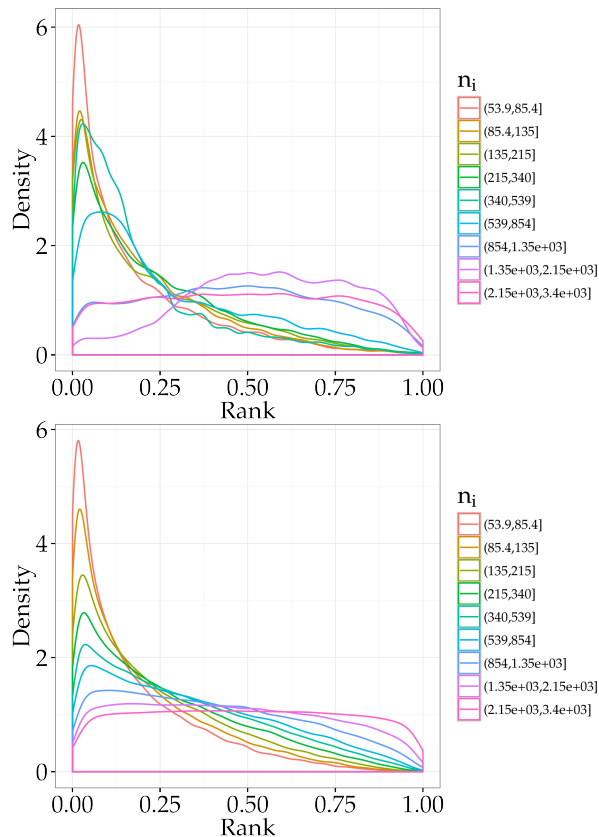


Figure 6: Top: Empirical probability density function of investing in a security of scaled rank  $\rho$  given the diversification  $n_i$  of the fund. Bottom: Probability density function of investing in a security of scaled rank  $\rho$  given the diversification  $n_i$  of the fund, given by the model.

### 3.2.3 Importance of capitalization ranking

Figure 6 shows that the distribution of investment of a fund is sensitive to its diversification  $n_i$ . The Beta distribution, which is limited to a  $[0, 1]$  interval, is flexible enough to describe the asset selection mechanism of a fund.

$$f(x; a, b) = \frac{1}{B(a, b)} x^{a-1} (1-x)^{b-1}, \quad (5)$$

where  $a$  and  $b$  are the shape parameters of the distribution, and  $B$  is a normalization constant.

### 3.2.4 Maximum investment ratio

The funds limit their investment in a given asset. They seem to follow a simple rule: defining the investment ratio  $f_{i,\alpha} = \frac{W_{i\alpha}}{C_\alpha}$ , one easily sees in Fig. 8 that each fund has a maximum investment ratio

$$f_i^{\max} = \max_\alpha \left( \frac{W_{i\alpha}}{C_\alpha} \right) \quad (6)$$

Since the average exchanged dollar-volume of an asset is proportional to its capitalization (Fig. 10), the existence of  $f_i^{\max}$  is a way to account for the available liquidity.

Although that limit is clear for an individual fund, there is a large range of empirical values  $f_i^{\max}$  Fig. 9.

## 4 Simulation

The simulation is done in a few simple steps:

1. Compute  $n^*$  using the segmented model Eq. 4.

the rank of capitalization  $C_\alpha$  at a given time. The selection probability  $P(W_{i\alpha} > 0 | \rho_\alpha)$  is then obtained by parametric fit to a beta distribution in each logarithmic bin.

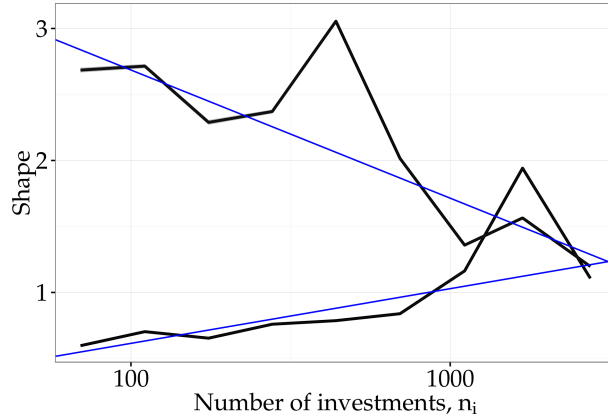


Figure 7: Coefficients  $a$  and  $b$  of the Beta Distribution 5 as a function of  $n_i$

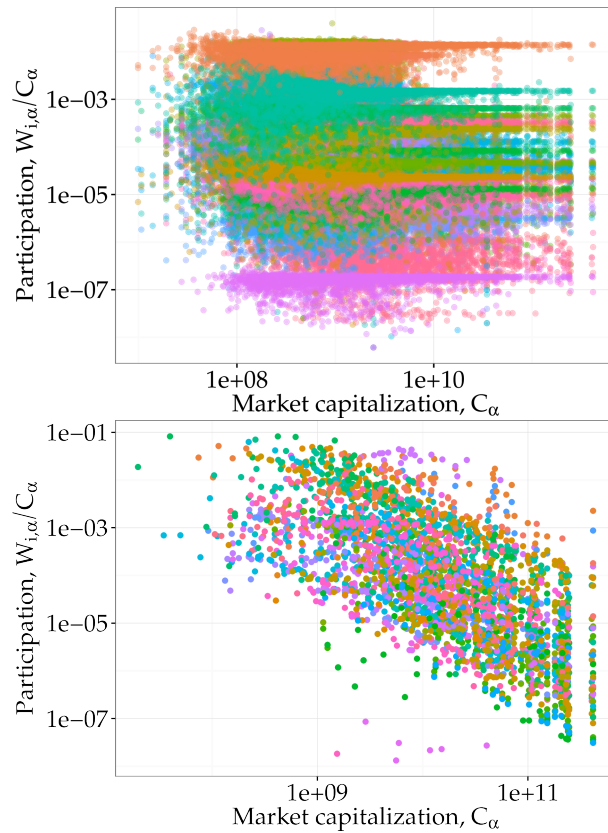


Figure 8: Fraction of the market capitalization of a security held by a fund. Each color represent a different fund. Top: Funds with a large diversification ( $n_i > 800$ ). We can clearly see a delimitation for most of the funds, which correspond to the maximum fraction  $f_i^{\max}$ . The value of  $f_i^{\max}$  widely differs from one fund to another. Bottom: Funds with a low diversification ( $n_i < 60$ ),  $f_i^{\max}$  doesn't appear.

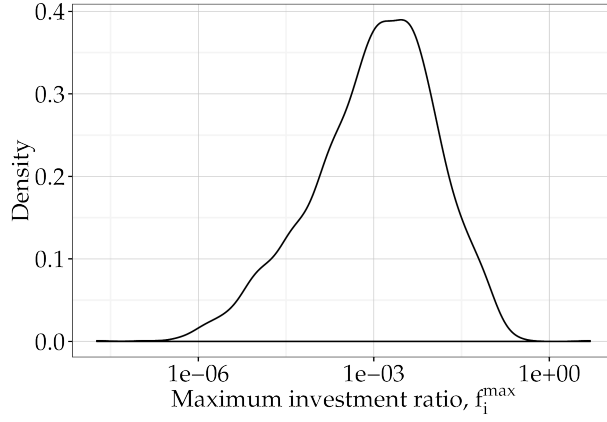


Figure 9: Empirical probability density function of  $f_i^{\max}$  for all the funds.

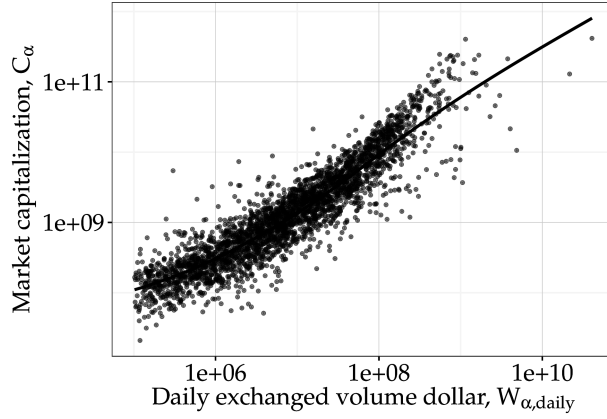


Figure 10: Market capitalization as a function of the daily exchange volume dollar. We find a slope close to 1 for all the dates in our database, confirming the hypothesis that the daily exchange volume dollar of an asset is proportional to its market capitalization.

2. Select a fund  $i$ , with a number of assets  $n_i$ .
3. If  $n_i < n^*$ :
  - (a) Compute its optimal portfolio value using Eq. 1. The fund will invest  $\frac{W_i^{\text{opt}}}{n_i}$  for every position.
  - (b) Select assets randomly with a probability proportional to  $C_\alpha$ .
4. Else if  $n_i \geq n^*$ :
  - (a) Compute its  $f_i^{\max}$ , so that the fund  $i$  will invest  $f_i^{\max}$  in  $n_i$  assets.
  - (b) Select assets randomly following a Beta probability distribution Fig. 6 with the parameters found in Fig. 7.

By iterating those steps we obtain Fig. 1

Since the simulation outputs a portfolio for every fund, we can directly infer the number of investors  $m_\alpha$  of every security.

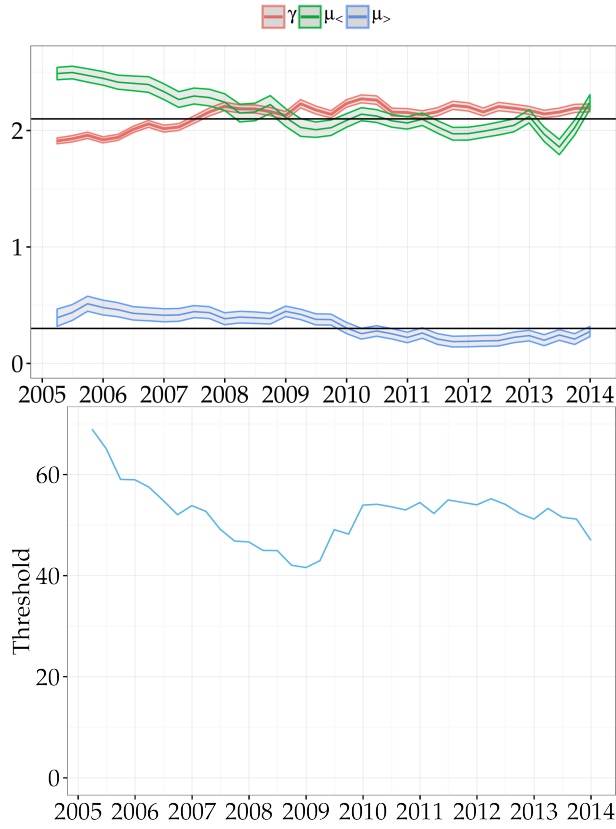


Figure 11: Top: Temporal evolution of the coefficients  $\mu_<$ ,  $\mu_>$  and  $\gamma$ . Bottom: Temporal evolution of the value of the cross-over point  $n^*$  between the two regions as a function of time. It reaches a global minimum after the 2008 crisis, and it also seems to decrease again after 2013.

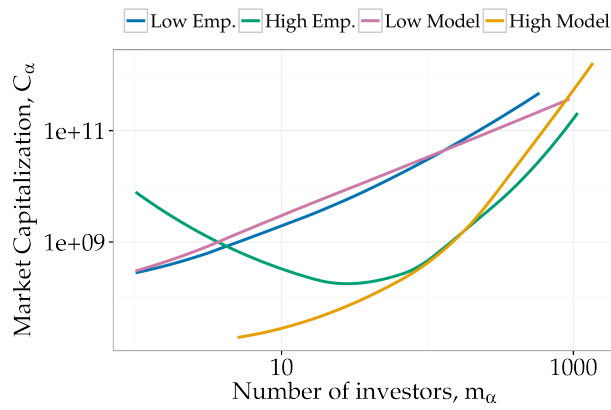


Figure 12: We separate the contribution from the low and highly diversified region. The origin of the discrepancy observed in Fig. 3 appears to be mainly due to the highly diversified region (Green dots for the empirical data, and orange dots for the model).