Wang et al. *EPJ Data Science* (2020) 9:11
https://doi.org/10.1140/epjds/s13688-020-00230-1

EPJ Data Science
a SpringerOpen Journal

REGULAR ARTICLE Open Access

Comparative analysis of layered structures in empirical investor networks and cellphone communication networks

Peng Wang[1], Jun-Chao Ma[1], Zhi-Qiang Jiang[1*], Wei-Xing Zhou[1] and Didier Sornette[2,3]

*Correspondence:
zqjiang@ecust.edu.cn
[1]School of Business and Research Center for Econophysics, East China University of Science and Technology, Shanghai, China
Full list of author information is available at the end of the article

Abstract

Empirical investor networks (EIN) proposed by Ozsoylev et al. are assumed to capture the information spreading path among investors. Here, we perform a comparative analysis between the EIN and the cellphone communication networks (CN) to test whether EIN is an information exchanging network from the perspective of the layer structures of ego networks. We employ two clustering algorithms (*k*-means algorithm and *H/T* break algorithm) to detect the layer structures for each node in both networks. We find that the nodes in both networks can be clustered into two groups, that one has a layer structure similar to the theoretical Dunbar Circle that alters in ego networks exhibit a four-layer hierarchical structure with the cumulative number of 5, 15, 50 and 150 from the inner layer to the outer layer, and the other one has an additional inner layer with about 2 alters compared with the Dunbar Circle. We also find that the scale ratios, which are estimated based on the unique parameters in the theoretical model of layer structures, conform to a log-normal distribution for both networks. Our results not only deepen our understanding on the topological structures of EIN, but also provide empirical evidence of the channels of information diffusion among investors.

Keywords: Empirical investor networks; Cellphone communication networks; Layered structure; Cluster analysis

1 Introduction

Due to the lack of data recording the information communication between investors, uncovering the information spreading path in investors is a great challenge. Recently, Ozsoylev et al. [22] first proposed the empirical investor network (EIN) as a novel representation of the information diffusion network, based on their order placements: two investors are said to be connected if they placed the same type (ask or bid) of orders within a short time window (usually 30 seconds). The underlying hypothesis behind the EIN is that, when new information comes, it spreads from source nodes to peripheral nodes in investor social networks and the time lags with which the information reaches different investors determine the lags between their order placements. However, no direct evidence is found to support the assumption of EIN, which is crucial for applying EIN on determining asset

pricing dynamics and understanding the trading behaviors and profitability of investors. As social network is the backbone of information diffusion, EIN can be regarded as a proxy of investor social network. We thus propose to check the validity of the EIN construction by studying the specific structures in social networks, such that the degree distribution of entire network and the layer structures of ego network. As a reference and for comparison, we also test the hierarchical structures present in cellphone communication networks (CN), which are usually considered as information spreading network. It is found that EIN and CN share very similar network structures, characterized by the same distribution of (weighted) degree, the same layer structure of ego network, and the same distribution of scale ratio, giving credence to the hypothesis that EIN uncovers a significant part of the information spreading path between investors.

The contributions of this paper are listed as follows. First, differing from the strand of literatures which focus on the layered structure in western social networks, like cellphone communication network in Europe, Christmas card exchange network in UK, and online social networks in Facebook and Twitter [7, 13, 20], we empirically uncover the layered structure in the Chinese social networks, including EIN and CN. Second, by investigating the subtle structures in EIN and CN, we initially find that there are great similarities in the layered structures between EIN and CN, which complements the existing studies on Dunbar's number and EIN. Third, Ozsoylev et al. [22] conclude that EIN captures information diffusion between investors by the evidence that central investors in EIN trade earlier and gain higher returns than their peripheral neighbors. We complementally contribute the evidence supporting that EIN reveals the information spreading path between investors from the perspective of social network structures.

The present work is related to the research on Dunbar's number and its generalised discrete hierarchical structure in social networks. Recall that Dunbar's number of about 150 represents the average size of the personal ego network, i.e., the group of people one can typically maintain stable social relationships with due to cognitive limits [3, 4]. Furthermore, the social relations in human and animal network have been found to form layer structures, each layer representing different emotional closeness [5, 6]. And layer structures have approximately the configuration of 3–5, 10–15, 30–50, and 100–200 alters from the inner layer to outer layer [30]. Many empirical ego networks are found to exhibit such layer structures, including the network abstracted from the exchange of Christmas cards [13], the hunter-gatherer social networks [11, 30], and online societies in virtual world [8].

Another strand of literature relevant to our work is the use of cellphone and internet communication data that enable one to test the classical social theories empirically in large scale individuals. For example, the weak tie theory [9] has been validated for cellphone communication networks [17, 21]. Such data have also been used to verify the hierarchical layer structures in social networks [20]. Arnaboldi et al. [1] found that the co-author networks in academic fields also have discrete hierarchical structures. By scanning the online social network from Facebook and Twitter, Dunbar et al. [7] found that the ego networks exhibit limited size and hierarchical structures. More importantly, such layer structure can be considered as a "social fingerprint" for a specific individual, because it is stable and not affected by the change of friends [24].

This paper is organized as follows. Data and methods are given in Sect. 2. Section 3 presents the results on the degree distribution, clustering, and theoretical model fits. Section 4 concludes.

2 Data and methods

2.1 Empirical investor networks

Our empirical investor networks (EIN) are constructed from the order flows of 100 stocks included in the Shenzhen 100 index (399,004). The order flow data span the whole year of 2013. Following Ozsoylev et al. [22], on each trading day, the EIN is obtained by connecting investors if they submit at least 3 buy (or sell) orders for the same stocks within 30 seconds. By aggregating the EIN on each trading day together, we obtain the annual EIN, which contains 381,345 nodes and 8,143,541 links. Ozsoylev et al. [22] argue that the links in EIN may reflect the potential channels of information diffusion among investors, indicating the existence of localized structures in social networks formed by investors. Thus, the larger the occurrence of links between two investors, the higher the probability for the existence of social connections between them. We further employ a statistical validated method [2, 10, 12, 19, 25, 27] to check whether two investors are occasionally connected, which provides us with the statistical validated empirical investor networks, abbreviated as SVEIN.

2.2 Cellphone communication network

The cellphone call records obtained from one Chinese cellphone operator cover periods from June 28th to July 24th and October 1st to December 31st in 2010. By excluding the days October 12th, November 5th, 6th, 13th, 21st and 27th, and December 6th, 8th, 21st and 22nd on which the data were missing, we have a total of 109 days. In the data, there are 91,911,735 cell phone users and 4,599,472,652 calls. As we cannot access the call records from the other cellphone operators, only the call records in which both mobile phone subscribers belong to the data provider are included in our analysis, which leads to 1,173,501,607 records. As it is known that the frequency of calls may represent the intimacy between friends, the higher the communication frequency between two cellphone users, the stronger their assumed intimacy. We exclude the users who are identified as robots, telecom frauds and telephone sales [15]. Finally, we build cellphone communication networks based on the reciprocal calls between normal users. The statistical validated method mentioned above is also employed to remove the random calls, thus providing us with the statistical validated cellphone communication networks, abbreviated as SVCN.

2.3 Statistical validated method

As is well known, EIN and CN contain a great deal of noise: for instance, two investors may submit orders at the same time by pure coincidence and callers may make wrong calls to callees. This suggests to remove such irrelevant signals by testing whether two nodes are randomly connected. For this, we employ a statistically validated method, proposed by Tumminello et al. [25] and used in different systems [2, 10, 12, 19, 27] to extract the links that are not randomly generated.

For two given nodes i and j, the purpose of the statistical validation is to check whether i preferentially connects to j. The EIN is taken as an example to illustrate the statistical validation method. Let us denote by N is the total number of transactions between investors in EIN, by N_{ic} the number of transactions initiated by investor i, by N_{jr} the number of transactions matched by investor j, and by $X = N_{icjr}$ the number of transactions initiated by investor i and matched by investor j. We can then calculate the probability of observing

X co-occurrences via the following equation [25, 26],

$$H(X|N, N_{ic}, N_{jr}) = \frac{C_{N_{ic}}^{X} C_{N-N_{ic}}^{N_{jr}-X}}{C_{N}^{N_{jr}}}, \tag{1}$$

where $C_{N_{ic}}^{X}$ is a binomial coefficient. We can also estimate the p-value associated with the observed N_{icjr} as follows:

$$p(N_{icjr}) = 1 - \sum_{X=0}^{N_{icjr}-1} H(X|N, N_{ic}, N_{jr}). \tag{2}$$

For the EIN, we need to perform $2 \times 8{,}143{,}541 = 16{,}287{,}082$ tests. The corresponding Bonferroni correction of our multiple testing hypothesis is $p_b = 0.01/N_E$ where $N_E = N(N-1)/2$ is the maximal possible number of edges. If the estimated $p(N_{icjr})$ is less than p_b, we can infer that investor i preferentially connects to investor j. Otherwise, we conclude that the edge pointed from i to j is randomly generated.

For a given edge between node i and node j in the CN, we are able to estimate the p-value for the number of calls N_{jcir} initiated by j and received by i in a similar way. We need to conduct $2 \times 296{,}928{,}030 = 593{,}856{,}060$ tests. And the Bonferroni correction is set as $p_b = 0.01/N_E$. When $p(N_{icjr})$ is less than p_b, this suggests that individual i preferentially calls individual j. Only when the two conditions that (1) i preferentially calls j and (2) j preferentially calls i are simultaneously satisfied, we conclude that the edge between i and j is significant.

2.4 Clustering method

Figure 1 illustrates the layer structure of a typical ego network. The ego in the center are surrounded by the alters, who have direct connections with the ego. The alters usually form a layer structure, in which their emotional closeness decrease from the inner layer to the outer layer. The theoretical Dunbar Circle corresponds to a four-layer hierarchical structure with the cumulative number of 5, 15, 50, and 150 from inside to outside. We employ two clustering algorithms, including the k-means algorithm and the head-to-tail (H/T) break algorithm [14], to detect the layer structures of the ego network in the SVEIN and SVCN based on the activity frequencies on links. The k-means algorithm is implemented with the *R* package *CKmeans.1d.dp* [28]. The optimized number of clusters are determined according to the BIC. In the H/T break algorithm, the data is split into two parts according to the data mean m_1, and the head part in which all values are larger than m_1 is further separated into two parts according to the head mean m_2. Such process iterates until the head is not heavy-tailed distributed. The H/T break algorithm is proposed to cluster the data with a heavy-tailed distribution, corresponding to the case of link weights in the SVEIN and SVCN.

3 Result

3.1 Degree distribution

We first report the descriptive statistics of both filtered networks. As reported in Panel A of Table 1, in the SVEIN we find that there are 2.23%, 6.39%, and 91.37% of the total number of users (21,806 users) whose degrees are in the range of $k > 100$, $50 < k \leq 100$, and

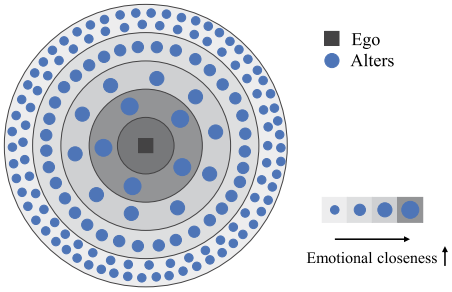

Figure 1 Illustration of the theoretical Dunbar Circle in ego networks. The square in the center represent the ego and the circles around are the alters, who have direct connection with the ego. The circle size is proportional to the emotional closeness between the alters and the ego. According to the emotional closeness, the alters form a hierarchical structure with different layers in which their closeness to the ego decrease from inner layer to the outer layer. The theoretical Dunbar Circle corresponds to a four-layer hierarchical structure with the cumulative number of 5, 15, 50, and 150 from inside to outside

Table 1 Statistical descriptions of SVEIN and SVCN. k denotes the degree of users in the network

			Degree			Weighted degree		
	N	f	mean	std	std/mean	mean	std	std/mean
Panel A: SVEIN								
$k > 100$	487	2.23%	142.9	38.5	26.95%	18,487.1	10,984.6	59.42%
$50 < k \leq 100$	1394	6.39%	68.8	13.9	20.22%	5504.3	2935.4	53.33%
$k \leq 50$	19,925	91.37%	10.0	11.8	117.95%	477.0	1134.0	237.73%
Panel B: SVCN								
$k > 100$	60,748	1.46%	142.2	45.8	32.23%	1544.7	775.0	50.17%
$50 < k \leq 100$	177,076	4.25%	69.4	13.7	19.79%	780.3	410.9	52.66%
$k \leq 50$	3,930,604	94.29%	8.1	10.0	124.08%	92.1	161.7	175.68%

$k < 50$, respectively. And their average degree and standard deviation are 142.9 and 38.5, 68.8 and 13.9, and 10.0 and 11.8, leading to a coefficient of variation of 26.95%, 20.22%, and 117.95% (standard deviation/mean). Their average weighted degree and standard deviation are 18,487.1 and 10,984.6, 5504.3 and 2935.4, and 477.0 and 1134.

In Panel B of Table 1, we find that the number of users in the SVCN with degree $k > 100$, $50 < k \leq 100$, and $k < 50$ are 60,748, 177,076, and 3,930,604, accounting for 1.46%, 4.25%, and 94.29% of the users, respectively. The corresponding average degree and standard deviation are 142.2 and 45.8, 69.4 and 13.7, and 8.1 and 10, resulting in a coefficient of variation of 32.23%, 19.79%, 124.08%. And their average weighted degree and standard deviation are 1544.7 and 775, 780.3 and 410.9, and 92.1 and 161.7. The absolute number of nodes with $k > 100$ in the SVEIN is much smaller than those in the SVCN, and the relative numbers are very close to each other. According to the descriptive statistics, both filtered networks exhibit great similarities in their degree distributions.

We further fit the empirical degree and weighted degree distributions of the SVEIN and SVCN with the following four distributions, including the power-law, the normal, the exponential, and the log-normal distribution,

$$f_P(x) = \frac{\alpha - 1}{x_{\min}} \left(\frac{x}{x_{\min}} \right)^{-\alpha}, \quad \alpha > 1, \tag{3}$$

$$f_N(x) = \frac{1}{\sqrt{2\pi}\sigma_N} \exp\left[-\frac{(x - \mu_N)^2}{2\sigma_N^2} \right], \tag{4}$$

$$f_E(x) = \lambda e^{-\lambda x}, \quad x > 0 \tag{5}$$

$$f_L(x) = \frac{1}{\sqrt{2\pi}\sigma_L x} \exp\left[-\frac{(\ln x - \mu_L)^2}{2\sigma_L^2} \right]. \tag{6}$$

Table 2 Results of fitting the (weighted) degrees to the power-law, normal, exponential, and log-normal distribution for the SVEIN and SVCN and statistically testing on whether the (weighted) degrees are drawn from the four distributions. The symbols *, **, and *** indicate the significant levels of 5%, 1%, and 0.1%, respectively

	SVEIN				SVCN			
	Degree		Weighted degree		Degree		Weighted degree	
	$k > 0$	$k > 100$	$k > 0$	$k > 100$	$k > 0$	$k > 100$	$k > 0$	$k > 100$
Panel A: Fits to the power-law distribution.								
α	1.50	3.50	1.50	1.84	1.50	3.50	1.50	1.50
KS	0.19	0.11	0.33	0.26	0.16	0.09	0.42	0.42
p-value	0.00***	0.13	0.00***	0.00***	0.00***	0.00***	0.00***	0.00***
AIC	161,511.77	4693.74	325,265.95	10,568.02	28,130,126.25	580,173.13	49,774,220.93	1,084,909.45
Panel B: Fits to the normal distribution								
μ_N	−1031.56	−93.68	1200.60	18,487.09	12.61	−2292.22	142.47	1422.38
σ_N	131.70	107.86	3570.82	10,973.34	23.12	327.80	298.20	881.68
KS	0.26	0.03	0.37	0.15	0.31	0.03	0.32	0.09
p-value	0.00***	0.77	0.00***	0.00***	0.00***	0.40	0.00***	0.00***
AIC	166,010.36	**4632.85**	418,656.91	10,447.39	38,012,462.68	576,466.02	59,330,827.17	975,316.25
Panel C: Fits to the exponential distribution								
α	16.71	42.91	1200.60	18,487.09	12.62	42.23	142.47	1446.69
KS	0.24	0.05	0.41	0.28	0.24	0.02	0.30	0.28
p-value	0.00***	0.46	0.00***	0.00***	0.00***	0.63	0.00***	0.00***
AIC	166,432.62	4637.31	352,848.35	10,540.11	29,472,982.01	576,270.11	49,680,143.77	1005,464.72
Panel D: Fits to the log-normal distribution								
μ_L	1.83	4.65	5.08	9.67	1.61	4.20	3.34	7.24
σ_L	1.43	0.39	2.04	0.54	1.30	0.54	2.04	0.45
KS	0.13	0.04	0.06	0.08	0.11	0.02	0.07	0.01
p-value	0.00***	0.71	0.00***	0.00**	0.00***	0.61	0.00***	0.17
AIC	**157,319.09**	4634.67	**314,478.41**	**10,208.38**	**27,447,309.97**	**575,902.17**	**45,639,797.91**	**955,503.08**

The parameters of these distributions are obtained by Maximum Likelihood Estimation (MLE). The results are listed in Table 2. Kolmogorov–Smirnov (KS) tests are also conducted to check whether the (weighted) degrees are drawn from the four distributions. The null hypothesis is that the data set follows one of the four distributions. One find that, for both networks, the samples of the degree with $k > 0$ and the weighted degree with $k > 0$ and $k > 100$ conform precisely to none of the four distributions. This is not surprising, given the large sizes of our data sets, which will thus reject null hypotheses on the basis of even slight deviations. However, we can still compare the goodness of the fits by the four distributions using the Akaike information criterion (AIC) listed in Table 2. Except for the sample with $k > 100$ in the SVEIN, the log-normal distribution has the smallest AIC value. Thus, among the four distributions, the log-normal distribution fits the empirical degree distributions best.

The results of Table 2 strongly suggest that the correct distribution of degrees is a mixture of at least two log-normal distribution, one for small k and one for large k. Roughly, we can find a threshold k_H, the degrees less than k_H are fitted by the left truncated log-normal distributions and the degrees greater than k_H are fitted by the right truncated log-normal distribution. Following Refs. [16, 29], the threshold k_H can be estimated by minimizing the following residual,

$$R = \frac{\left\{\sum_i^{n_s}\left[\frac{K^s_{i,\mathrm{fit}}-K^s_{i,\mathrm{emp}}}{K^s_{i,\mathrm{fit}}+K^s_{i,\mathrm{emp}}}\right]+\sum_j^{n_l}\left[\frac{K^l_{j,\mathrm{fit}}-K^l_{j,\mathrm{emp}}}{K^l_{j,\mathrm{fit}}+K^l_{j,\mathrm{emp}}}\right]\right\}^{\frac{1}{2}}}{\sqrt{n_s+n_l}}, \tag{7}$$

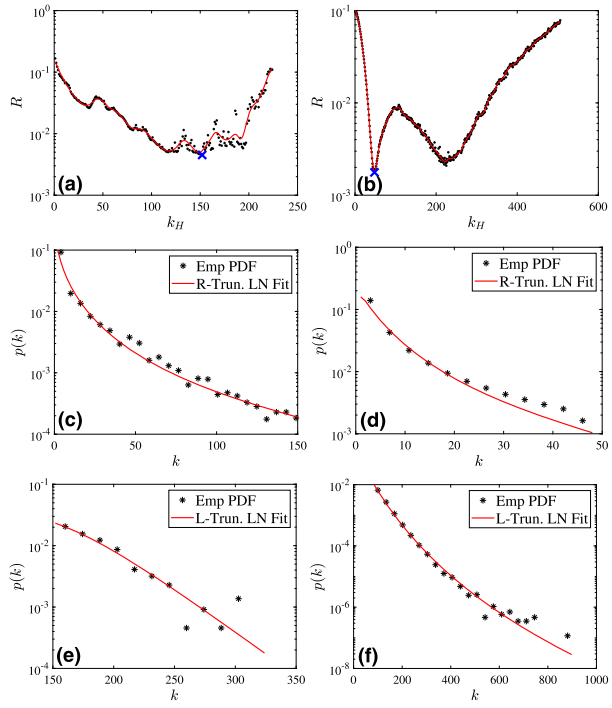

Figure 2 Results of the optimal truncated distribution of degrees for SVEIN (**a**), (**c**), (**e**) and SVCN (**b**), (**d**), (**f**). (**a**), (**b**) Plots of the fitting residual (Eq. (7)) as a function of threshold k_H. (**c**), (**d**) Plots of the right-truncated degree distributions. (**e**), (**f**) Plots of the left-truncated degree distributions

where K_{fit} and K_{emp} represent the fitting distribution and empirical distribution, the superscripts s and l stand for the sample less and greater than the threshold k_H, and n is the sample size. The parameters of both truncated distributions are determined through the Maximum Likelihood Estimation (MLE). Figure 2(a) and (b) illustrate the fitting residuals as a function of the possible thresholds for the degrees of SVEIN and SVCN. Thus, we can find that the optimal threshold are 152 and 48 for SVEIN and SVCN, respectively. The corresponding right-truncated and left-truncated degree distributions are plotted in Fig. 2(c)–(f) for SVEIN and SVCN. The solid lines in each panel represent the best fits to the truncated log-normal distributions. For the weighted degrees of both networks, we perform the same analysis and illustrate the results in Fig. 3. The optimal thresholds are 374 and 653 for the weighted degrees of SVEIN and SVCN, respectively. One can see that the empirical distributions agree well with the fitted distributions in Figs. 2 and 3, which support that the (weighted) degrees of both network conform to a mixed log-normal distribution.

As is well known, the log-normal distribution plays an important role in describing natural phenomena in which growth processes are driven by the accumulation of many small percentage changes (growth rates), which is additive on the logarithmic scale. If each percentage change is small enough, the summation on the logarithmic scale tends to be normally distributed according to the central limit theorem, which means that the percentage change follows a log-normal distribution in the linear scale. One intriguing feature of the log-normal distribution is that the growth rate is independent of its size. According to

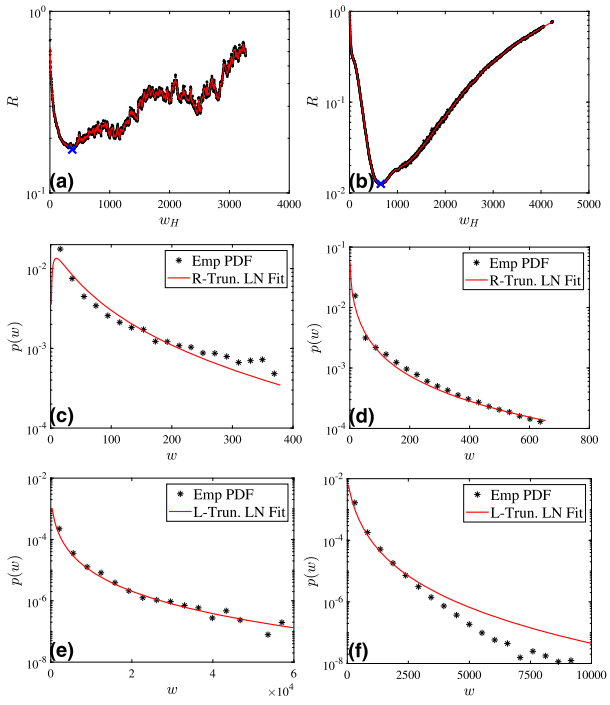

Figure 3 Results of the optimal truncated distribution of weighted degrees for SVEIN (**a**), (**c**), (**e**) and SVCN (**b**), (**d**), (**f**). (**a**), (**b**) Plots of the fitting residual (Eq. (7)) as a function of threshold k_H. (**c**), (**d**) Plots of the right-truncated degree distributions. (**e**), (**f**) Plots of the left-truncated degree distributions

the log-normal degree distributions, one can infer that the growth rate of one's "friends" should be independent of one's current number of "friends" in the SVEIN and SVCN.

3.2 Clusters

The layer structures in ego networks are usually determined based on the emotional closeness on links [23]. Here, we cannot measure the emotional closeness directly. As an alternative, we employ the number of order placements in the EIN and the number of calls in the CN as a proxy for the emotional closeness on links. For a given node with n links, we first normalize the number of order placements (respectively, the number of calls) W_i ($i = 1, 2, 3, \ldots, n$) on each link via the following equation,

$$\hat{W}_i = \frac{W_i - W_{\min}}{W_{\max} - W_{\min}}, \tag{8}$$

where $W_{\min} = \min(\{W_i\})$ and $W_{\max} = \max(\{W_i\})$. Equation (8) insures $0 \leq \hat{W}_i \leq 1$. The presence of natural breaks (associated with network layers) should then be reflected in the existence of sharp peaks in the distributions of $\hat{W}_i$. We thus plot the distribution of the normalized weights $\hat{W}_i$ in Fig. 4 for both networks. As shown in Fig. 4(a), no break can be observed for the SVEIN. A possible explanation is that the data sample of SVEIN is too small. In contrast, there is a significant peak at around 0.1 for the SVCN, as illustrated in Fig. 4(b), which corresponds to the natural break $w_i \approx 0.1 = 15/150$, i.e. the second layer at 15 of Dunbar's discrete hierarchy. In the following, we use the clustering algorithms

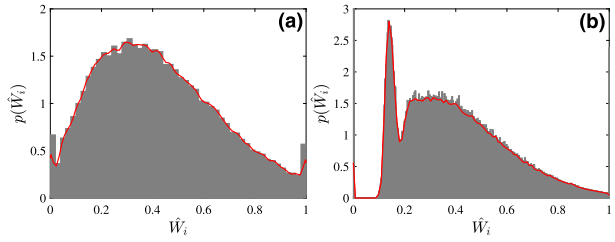

Figure 4 Probability distribution of the normalized weights $\hat{W}_i$. (**a**) SVEIN. (**b**) SVCN

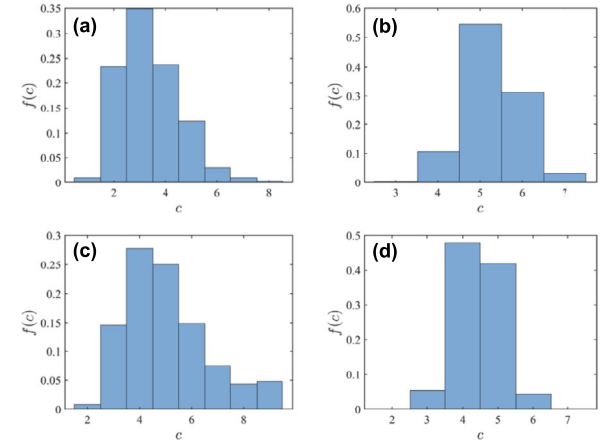

Figure 5 Plots of the percentage of the users who have the same number of layers in the SVEIN (**a**), (**b**) and SVCN (**c**), (**d**) based on the *k*-means (**a**), (**c**) and *H*/*T* (**b**), (**d**) break algorithm

(*k*-means and H/T break) to uncover the discrete hierarchical structure of the node with $k > 100$ based on the normalized weights $\hat{W}_i$.

Figure 5 shows the percentage of users who have the same number of layers according to the clustering algorithm of *k*-means and H/T break. As shown in Fig. 5(a) and (b), the alters belonging to investors with degree $k > 100$ in the SVEIN are mainly divided into 2–4 classes and 4–6 classes according to the *k*-means and H/T Break algorithm, respectively. And we also find 56.9% of the investors whose alters can be grouped into 5 layers. In order to measure the similarity and robustness of the clustering results, we further estimate the Jaccard coefficient between the clustering results of the two algorithms for the same user. The average Jaccard coefficient of all users is 0.11. As illustrated in Fig. 5(c) and (d), we find that in the SVCN the alters of the users with degree $k > 100$ are mainly divided into 3–6 classes and 4–5 classes based on the *k*-means algorithm and the H/T Break algorithm. And the average Jaccard coefficient of the clustering results is 0.23. Our results indicate that the overlapping degree of the clusters from both algorithms is low.

Table 3 shows the comparison of the clustering results for the users with degree $k > 100$ in both networks based on the *k*-means and H/T break algorithms. The results of the two clustering algorithms for the SVEIN are reported in panel A of Table 3. We find that 43% of users with degree $k > 100$ in SVEIN are grouped into 3 layers and the average cumulative number of alters in layers is 10.9, 45.8 and 141.7, in which the last two layers correspond to the middle two layers of the empirical discrete hierarchical structure and the first layer

Table 3 Comparison of the clustering results for the users with degree $k > 100$ based on the k-mean and H/T break algorithm for the SVEIN and SVCN. N and f represents the total number and the percentage of users. n_k stands for the cumulative number of users in the k-th layer. $\langle r \rangle$ is the average scale ratio

	N	f	n_1	n_2	n_3	n_4	n_5	n_6	$\langle r \rangle$
Panel A: Clustering results of SVEIN									
k-means									
$c = 2$	114	27.9%	27.8	121.7					3.84
$c = 3$	176	43.0%	10.9	45.8	141.7				3.04
$c = 4$	119	29.1%	5.4	20.8	57.5	151.4			2.64
H/T break									
$c = 4$	54	11.3%	2.9	11.0	37.1	133.1			3.45
$c = 5$	273	56.9%	1.6	5.3	15.0	42.8	133.0		3.00
$c = 6$	153	31.9%	1.2	3.2	7.5	18.5	51.3	156.0	2.88
Panel B: Clustering results of SVCN									
k-means									
$c = 4$	16,918	41.1%	3.0	12.8	42.8	132.0			3.22
$c = 5$	15,209	36.9%	2.1	7.3	20.4	54.2	141.4		2.66
$c = 6$	9049	22.0%	1.6	5.1	12.5	28.9	66.5	154.0	2.33
H/T break									
$c = 3$	3308	5.7%	5.0	27.1	126.7				4.71
$c = 4$	29,125	50.2%	2.1	8.7	33.4	133.9			3.97
$c = 5$	25,539	44.1%	1.2	3.8	11.7	39.5	147.6		3.61
Zhou			5	15	50	150			3.00

seems to correspond to the amalgamation of the first two layers of the empirical structure reported in Refs. [13, 30]. The H/T Break algorithm reveals that about 90% of the investors whose alters exhibit a configuration with 5 and 6 layers. One can observe that the number of alters in the outer four layers are very close to the theoretical Dunbar Circle 5, 15, 50, and 150. The number of alters in the inner or two layers is only 1–3.

Panel B of Table 3 lists the cumulative number of friends in each layer for the SVCN. For the k-means algorithm, we find that 16,918 (a fraction of 41.1%) users have a four-layer structure. The average cumulative number of alters from inside to outside are 3.0, 12.8, 42.8 and 132.0, which is in agreement with the discrete hierarchical structure 3–5, 10–15, 30–50, and 100–200 reported in Refs. [13, 30]). The corresponding scale ratio is 3.22 which is near to the Dunbar number 3. We also find that there are 15,209 users have a five-layer structure with an average accumulative number of 2.1, 7.3, 20.4, 54. 2, and 141.4. Besides the inner layer $n_1 = 2.1$, the number of alters in the outside four layers are very close to the reported hierarchical structure in Refs. [13, 30]. For the H/T Break algorithm, 29,125 users (about 50.2%) exhibit a four-layer structure and the average cumulative number of alters are 2.1, 8.7, 33.4 and 133.9. There are 25,539 (about 44.1%) users whose alters can be classified into 5 layers and the average accumulative number of alters in successive layers are 1.2, 3.8, 11.7, 39.5 and 147.6.

Both clustering algorithms reveal a similar discrete hierarchical structure in cellphone networks. We find that there is an extra innermost layer (1.2–2.1), with about 1–2 alters, for the users with four layers in their ego networks. We further fix the number of clusters to 4 for the k-means algorithm and estimate the cumulative numbers of in each layer, obtaining 2.5, 10.3, 36.8, and 142.2. In addition, we perform the clustering analysis on the link activities for each ego network, in which the ego investor with degrees $50 < k < 100$, by means of the k-means algorithm. We find that there are 621 investors (about 44.9%)

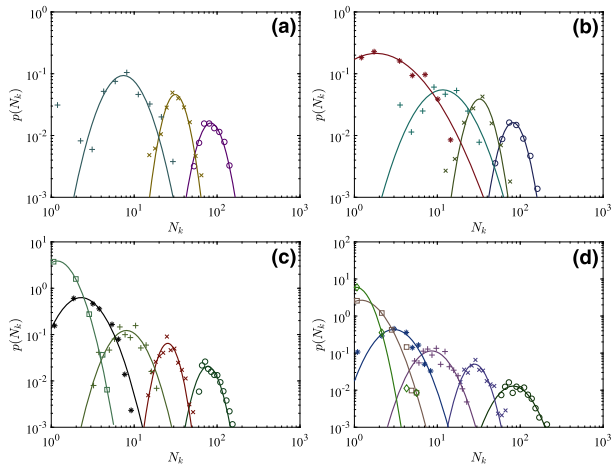

Figure 6 Probability distribution of the number of alters in different layers for the SVEIN. The solid curves represent the best log-normal fits to the empirical distribution. (**a**) Egos with three layers obtained from the *k*-means algorithm. (**b**) Egos with four layers obtained from the *k*-means algorithm. (**c**) Egos with five layers obtained from the *H/T* break algorithm. (**d**) Egos with six layers obtained from the *H/T* break algorithm

having a two-layer structure and the corresponding layer structure is 19.8 and 67.2, which is close to the middle two layers of the reported hierarchical structure [13, 30].

The empirical hierarchical structures of the personal ego networks in SVEIN and SVCN are compatible with the structure of 3–5, 10–15, 30–50, 100–200 from the inner to the outer layer, which is close to the theoretical Dunbar Circle. And the average empirical scaling ratio is close to the theoretically value 3 [18].

Figures 6 and 7 show the distributions of the numbers of alters in each layer for the egos having degree $k > 100$ in the SVEIN and SVCN. We only show the nodes whose personal ego networks having three-layer and four-layer (respectively, five-layer and six-layer) structures in the SVEIN (SVCN). For both networks, the clustering results of both algorithms are not in agreement with each other, as reflected by the low values of their Jaccard coefficients. An intriguing phenomenon is that the empirical distributions of the number of alters in each layer can be well fitted by the log-normal distributions, evidenced by the solid curves. Such log-normal distributions are robust when using different clustering algorithms, which are in agreement with the results of the online social network from Facebook and Twitter [7].

3.3 Fits to the theoretical model

We further fit the clustering results to the theoretical model of layer structures in personal social network [24]. According to this model, the probability, that the alters of an individual are divided into $\boldsymbol{\ell} = (\ell_1, \ell_2, \ldots, \ell_r)$, is calculated as follows

$$P(\boldsymbol{\ell}|\mathcal{L},\mu,N) = \mathcal{B}\left(L, \frac{\mathcal{L}}{N-1}, N-1\right)\left(\frac{e^{\mu}-1}{e^{\mu r}-1}\right)^{L}\binom{L}{\boldsymbol{\ell}} e^{\mu\sum_{k=0}^{r-1} k\ell_{k+1}}, \tag{9}$$

where $\boldsymbol{\ell} = (\ell_1, \ell_2, \ldots, \ell_r)$ represents the number of alters in each layer. $\mathcal{L}$ represents the sum of the alters expectation of each layer and is equal to the total number of alters L. N is the total number of individuals in the network. $\mathcal{B}(L,p,N) = \binom{N}{L}p^{L}(1-p)^{\mathrm{NL}}$ represents a binomial distribution. There is a unique parameter μ in the model, which is an indicator

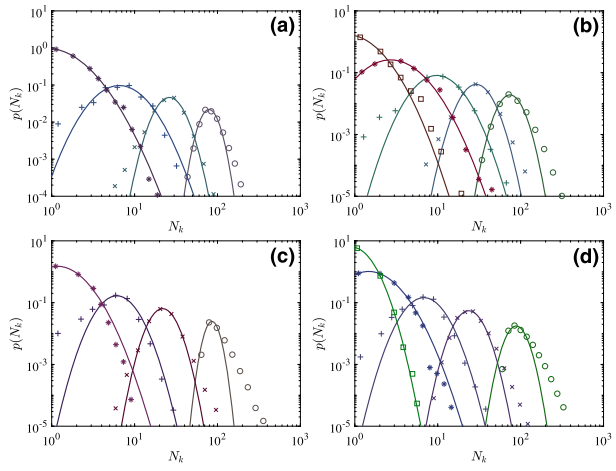

Figure 7 Probability distribution of the number of alters in different layers for the SVCN. The solid curves represent the best log-normal fits to the empirical distribution. (**a**) Egos with four layers obtained from the *k*-means algorithm. (**b**) Egos with five layers obtained from the *k*-means algorithm. (**c**) Egos with four layers obtained from the *H*/*T* break algorithm. (**d**) Egos with five layers obtained from the *H*/*T* break algorithm

of the discrete hierarchy for the ego network. The parameter μ is approximately equal to the logarithm of the scale ratio $\log(r)$ between the cumulative numbers of individuals in successive layers, if the personal investment (time and energy) decrease linearly with the layers [24].

Once the empirical hierarchical structure of egos is obtained, we calculate the average scale ratio $\langle r \rangle$ between adjacent cumulative layers based on the model proposed by Tamarit et al. [24]. The estimated theoretical scale ratios of both algorithms are listed in the last column of Table 3. For the SVEIN, the *k*-means algorithm indicates that the users are preferentially divided into the group having a three-layer structure while the H/T break algorithm uncovers that the ego networks exhibit a configuration of five layers. And their scale ratio are very close to the scaling ratio 3 discovered by Zhou et al. [30]. However, we find the existence of significant differences in the average scale ratio between the two clustering algorithms for the SVCN. On average, the average scale ratio of the H/T break algorithm is larger than 3.5 and the scale ratio obtained with the *k*-means algorithm is smaller than 3.5. Both clustering algorithms reveal that most of the users exhibit a four-layer structure in their ego networks, for which the scale ratio are respectively 3.2 and 4.0, which are roughly compatible with the scale ratio reported in Ref. [30].

Figures 8 and 9 show the distribution of the estimated average scale ratios for the egos having the same layer structure for both networks. We find that the scale ratio distributions given by the Tamarit's model conform to the log-normal distributions for both clustering algorithms. The χ^2 test, KS test and AD test cannot reject the null hypothesis, that the scale ratios are log-normal distributed, at the significant level of 5%. The solid curves in Figs. 8 and 9 are the best fits to the log-normal distributions. The scaling ratios given by $\exp(\hat{\mu})$ are located in the range of 2.5–3.3, which is compatible with the previous scaling ratio 3 discovered by Zhou et al. [30]. Our results reveal that the ego networks in SVEIN exhibit very similar layer structures to those in SVCN, confirming that the SVEIN captures the information spreading channels between investors.

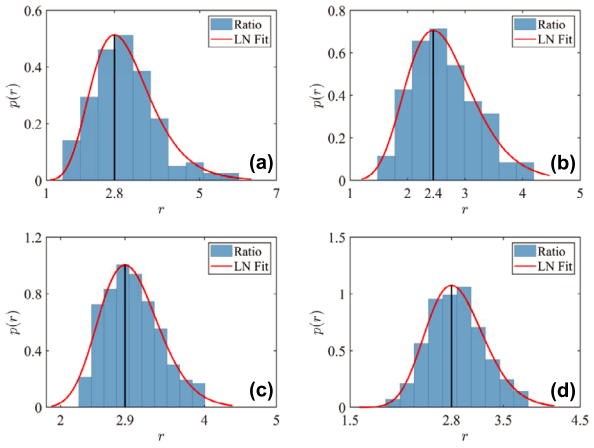

Figure 8 Empirical distribution of the scale ratios for the egos with different layers based on different cluster algorithms for the SVEIN. (**a**) Three layers and *k*-means algorithm. (**b**) Four layers and *k*-means algorithm. (**c**) Five layers and *H*/*T* break algorithm. (**d**) Six layers and *H*/*T* break algorithm

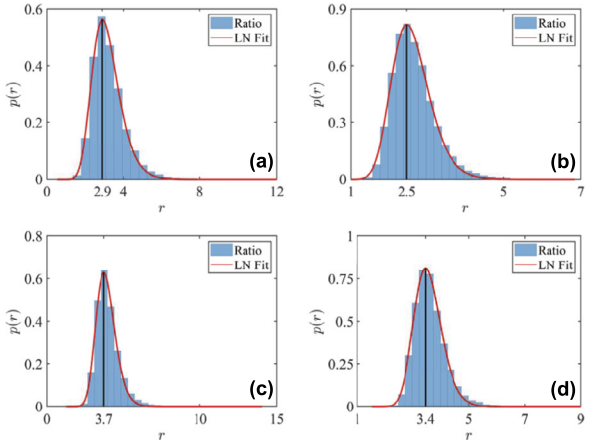

Figure 9 Empirical distribution of the scale ratios for the egos with different layers based on different cluster algorithms for the SVCN. (**a**) Four layers and *k*-means algorithm. (**b**) Five layers and *k*-means algorithm. (**c**) Four layers and *H*/*T* break algorithm. (**d**) Five layers and *H*/*T* break algorithm

4 Conclusion

We have performed a comparative analysis to detect the layer structures in Empirical Investor Networks and Cellphone Communication Networks. The layer structures have been quantified by two clustering algorithms, namely the *k*-means and H/T break algorithms. And both clustering algorithms reveal that there are two types of inner structure for both networks: one exhibits a layer structure similar to that of the theoretical Dunbar Circle, while the other has an additional inner layer, which is also found in Facebook and Twitter datasets [7]. Furthermore, we also find that both networks have a similar scale ratio (close to 3). And more interesting, these scale ratios remain stable even when old alters are replaced by new alters. By fitting our empirical clustering results to the theoretical model of layer structures [24], we confirm that the scale ratios of different egos follow a log-normal distribution for both networks. Our results suggest strong evidence that the structures of ego networks in EIN and CN exhibit great similarities, which captures the

information spreading routes between investors and validates the underlying assumption of EIN.

The Dunbar Circle referred to the layered structure of ego social network is ubiquitous in online and offline social networks [7, 13, 20]. In such layered structure, the size of each layer increases as the emotion closeness decrease, which can be attributed to the fact that individuals are restrained for maintaining more emotionally close social relationships due to the constrained cognitive capacity [4]. Our work demonstrates that EIN share very similar layered structure as CN, supporting that how many neighbors the investor having to exchange information is dominated by his/her cognitive ability. Furthermore, the Dunbar Circle also reveals that the ego shares information with the investors in inner structure more often than those in outer structure, indicating the existence of tight cliques between investors for exchanging trading information.

Acknowledgements
The authors would like to thank the anonymous reviewers for their valuable comments and suggestions to improve the quality of the paper.

Funding
This work was partially supported by the National Natural Science Foundation of China (U1811462, 91746108), the Shanghai Philosophy and Social Science Fund Project (2017BJB006), the Program of Shanghai Young Top-notch Talent (2018), the Shanghai Outstanding Academic Leaders Plan, and the Fundamental Research Funds for the Central Universities.

Abbreviations
EIN:, Empirical Investor Networks; CN:, Cellphone Communication Networks; SVEIN:, Statistical Validated Empirical Investor Networks; SVCN:, Statistical Validated Cellphone Communication Networks.

Availability of data and materials
The datasets used during the current study are available from the corresponding author on reasonable request.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
PW, JCM, ZQJ, WXZ, and DS designed research, performed research, analyzed data, and wrote the paper. All authors read and approved the final manuscript.

Author details
[1]School of Business and Research Center for Econophysics, East China University of Science and Technology, Shanghai, China. [2]Department of Management, Technology and Economics, ETH Zurich, Zurich, Switzerland. [3]Swiss Finance Institute, University of Geneva, Geneva, Switzerland.

Publisher's Note

Received: 20 July 2019 Accepted: 28 April 2020 Published online: 07 May 2020

References

1. Arnaboldi V, Dunbar RIM, Passarella A, Conti M (2016) Analysis of co-authorship ego networks. In: Wierzbicki A, Brandes U, Schweitzer F, Pedreschi D (eds) Advances in network science. Springer, Cham, pp 82–96
2. Curme C, Tumminello M, Mantegna RN, Stanley HE, Kenett DY (2015) Emergence of statistically validated financial intraday lead-lag relationships. Quant Finance 15(8):1375–1386. https://doi.org/10.1080/14697688.2015.1032545
3. Dunbar RIM (1992) Neocortex size as a constraint on group size in primates. J Hum Evol 22(6):469–493. https://doi.org/10.1016/0047-2484(92)90081-J
4. Dunbar RIM (1993) Coevolution of neocortical size, group size and language in humans. Behav Brain Sci 16(4):681–694. https://doi.org/10.1017/S0140525X00032325
5. Dunbar RIM (1998) The social brain hypothesis. Evol Anthropol 6(5):178–190. https://doi.org/10.1002/(SICI)1520-6505(1998)6:5<178::AID-EVAN5>3.0.CO;2-8
6. Dunbar RIM, Shultz S (2007) Evolution in the social brain. Science 317(5843):1344–1347. https://doi.org/10.1126/science.1145463
7. Dunbar RIM, Arnaboldi V, Conti M, Passarella A (2015) The structure of online social networks mirrors those in the offline world. Soc Netw 43:39–47. https://doi.org/10.1016/j.socnet.2015.04.005

8. Fuchs B, Sornette D, Thurner S (2014) Fractal multi-level organisation of human groups in a virtual world. Sci Rep 4:6526. https://doi.org/10.1038/srep06526
9. Granovetter MS (1973) The strength of weak ties. Am J Sociol 78(6):1360–1380
10. Gualdi S, Cimini G, Primicerio K, Di Clemente R, Challet D (2016) Statistically validated network of portfolio overlaps and systemic risk. Sci Rep 6):39467. https://doi.org/10.1038/srep39467
11. Hamilton MJ, Milne BT, Walker RS, Burger O, Brown JH (2007) The complex structure of hunter-gatherer social networks. Proc R Soc Lond B 274(1622):2195–2203. https://doi.org/10.1098/rspb.2007.0564
12. Hatzopoulos V, Iori G, Mantegna RN, Miccichè S, Tumminello M (2015) Quantifying preferential trading in the e-MID interbank market. Quant Finance 15(4):693–710. https://doi.org/10.1080/14697688.2014.969889
13. Hill RA, Dunbar RIM (2003) Social network size in humans. Hum Nat 14(1):53–72. https://doi.org/10.1007/s12110-003-1016-y
14. Jiang B (2013) Head/tail breaks: a new classification scheme for data with a heavy-tailed distribution. Prof Geogr 65(3):482–494. https://doi.org/10.1080/00330124.2012.700499
15. Jiang ZQ, Xie WJ, Li MX, Podobnik B, Zhou WX, Stanley HE (2013) Calling patterns in human communication dynamics. Proc Natl Acad Sci USA 110(5):1600–1605. https://doi.org/10.1073/pnas.1220433110
16. Jiang ZQ, Xie WJ, Li MX, Zhou WX, Sornette D (2016) Two-state Markov-chain Poisson nature of individual cellphone call statistics. J Stat Mech Theory Exp 073:210. https://doi.org/10.1088/1742-5468/2016/07/073210
17. Kovanen L, Kaski K, Kertész J, Saramäki J (2013) Temporal motifs reveal homophily, gender-specific patterns, and group talk in call sequences. Proc Natl Acad Sci USA 110(45):18.070–18.075. https://doi.org/10.1073/pnas.1307941110
18. Lera SC, Sornette D (2019) A theory of discrete hierarchies as optimal cost-adjusted productivity organisations. PLoS ONE 14(4):1–12. https://doi.org/10.1371/journal.pone.0214911
19. Li MX, Jiang ZQ, Xie WJ, Miccichè S, Tumminello M, Zhou WX, Mantegna RN (2014) A comparative analysis of the statistical properties of large mobile phone calling networks. Sci Rep 4:5132. https://doi.org/10.1038/srep05132
20. MacCarron P, Kaski K, Dunbar RIM (2016) Calling Dunbar's numbers. Soc Netw 47:151–155. https://doi.org/10.1016/j.socnet.2016.06.003
21. Onnela JP, Saramäki J, Hyvönen J, Szabó G, Lazer D, Kaski K, Kertész J, Barabási AL (2007) Structure and tie strengths in mobile communication networks. Proc Natl Acad Sci USA 104:7332–7336. https://doi.org/10.1073/pnas.0610245104
22. Ozsoylev HN, Walden J, Yavuz MD, Bildik R (2014) Investor networks in the stock market. Rev Financ Stud 27(5):1323–1366. https://doi.org/10.1093/rfs/hht065
23. Sutcliffe A, Dunbar RIM, Binder J, Arrow H (2012) Relationships and the social brain: integrating psychological and evolutionary perspectives. Br J Psychol 103(2):149–168. https://doi.org/10.1111/j.2044-8295.2011.02061.x
24. Tamarit I, Cuesta JA, Dunbar RIM, Sánchez A (2018) Cognitive resource allocation determines the organization of personal networks. Proc Natl Acad Sci USA 115(33):8316–8321. https://doi.org/10.1073/pnas.1719233115
25. Tumminello M, Miccichè S, Lillo F, Piilo J, Mantegna RN (2011) Statistically validated networks in bipartite complex systems. PLoS ONE 6:e17994. https://doi.org/10.1371/journal.pone.0017994
26. Tumminello M, Miccichè S, Lillo F, Varho J, Piilo J, Mantegna RN (2011) Community characterization of heterogeneous complex systems. J Stat Mech Theory Exp 2011:P01019. https://doi.org/10.1088/1742-5468/2011/01/P01019
27. Tumminello M, Lillo F, Piilo J, Mantegna RN (2012) Identification of clusters of investors from their real trading activity in a financial market. New J Phys 14:013041. https://doi.org/10.1088/1367-2630/14/1/013041
28. Wang HZ, Song MZ (2011) Ckmeans. 1d.dp: optimal *k*-means clustering in one dimension by dynamic programming. R J 3(2):29–33. https://doi.org/10.32614/RJ-2011-015
29. Wu Y, Zhou CS, Xiao JH, Kurths J, Schellnhuber HJ (2010) Evidence for a bimodal distribution in human communication. Proc Natl Acad Sci USA 107:18.803–18.808. https://doi.org/10.1073/pnas.1013140107
30. Zhou WX, Sornette D, Hill RA, Dunbar RIM (2005) Discrete hierarchical organization of social group sizes. Proc R Soc Lond B 272(1561):439–444. https://doi.org/10.1098/rspb.2004.2970