

Inferring Personal Financial Status from Social Network Location

Shaojun Luo¹, Flaviano Morone¹, Carlos Sarraute² and Hernán A. Makse¹

¹*Levich Institute, City College of New York, New York, NY 10031*

²*Grandata Labs, (1638) Vicente López, Buenos Aires, Argentina*

Abstract

It has been theoretically established that the pattern of social ties affects people's financial status. Here we translate this concept into an operational definition at the network level, which allows us to infer the economic wellness of individuals through a measure of their location in the social network. We analyze two large-scale sources including telecommunication and financial data of a whole country population of 110 million people. Our results show that individual's location, measured as the optimal collective influence in the structural integrity of the social network, can be used to infer personal financial status. For pragmatic use and validation, we carry out a marketing campaign that shows a three-fold increase in response rate by targeting individuals identified by our social network metrics as compared to random targeting. Our strategy could also be useful in maximizing the effects of large-scale economic stimulus policies.

The problem of how the network of social contacts [1–3] impacts the financial situation of individuals has drawn tremendous attention due to its importance in a diversity of socio-economic issues ranging from policy to marketing [4–7]. Theoretical analyses have pointed to the importance of the social network in economic life [5] as a medium to diffuse ideas [8, 9] through the effects of “structural holes” [10] and “weak ties” in the network [4]. Likewise, research has recognized the positive economic effect of expanding individual’s contacts outside their own tightly connected friendship group [1, 11–13]. While previous work has established the importance of social network influence to economic status, the problem of how to quantify such correspondence via social network centralities or metrics [3, 14] remains open. Recently, a numerical study has tested the effect of network diversity on economic development [6]. This study analyzed economic development defined at the community level. However, the question of how social network metrics may be used to infer financial status at the individual level— necessary, for instance, for micro-target marketing or social intervention campaigns— still remains unanswered. The difficulty arises, in part, due to the lack of empirical data combining, simultaneously, individual’s financial information with the pattern of social ties at the large-scale network level.

In this work we address this problem directly by combining two datasets: a massively-large social network of the whole population of Mexico with financial banking data. The social network is constructed from mobile (calls and SMS) and residential communication data collected during the months of April to July, 2013 (SM I, aggregated data at `kcore-analytics.com`). The database contains 1.10×10^8 phone users. After filtering the non-human active nodes by a machine-learned model trained on human natural communication behavior (SM II), we construct a final network of 1.07×10^8 nodes in a giant connected component made of 2.46×10^8 links. The ties, or links, in the network correspond to phone call communication, since we expect that communication patterns are indicative of individual’s location in the social network [15–17] which in turn are a signature of financial well-being.

Financial status is obtained from the combined credit limit on credit cards assigned by banking institutions to each client. The credit limit is based on composite factors of income and credit history and therefore reflects the financial status of the individual. The credit limit is pulled from a well-encrypted bank database and identified by the encrypted clients’ phone number registered in the bank. Thus, we are able to precisely cross-correlate the

financial information of an individual with its social location in the phone call network at the country level. There are 5.02×10^5 bank clients who have been identified in the mobile network whose credit limit ranges from MXN $\$10^3$ to $\$6 \times 10^6$ (Mexican Peso = 0.057 US\$). Thus, the datasets are precisely connected providing an unprecedented opportunity to test the correlation between network location and financial status.

Figures 1A and 1B show the communication patterns geolocalized across the country of individuals in the top 10% and bottom 10% credit limit, respectively. The inequality in the patterns of communication between the top economic class and the lowest is striking. It is visually apparent that the top 10% (accounting for 45.2% of the total credit in the country) displays a completely different pattern of communication than the bottom 10%; the former is characterized by more active and diverse links, especially connecting remote locations and communicating with other equally affluent people. Particular examples of the extended ego-networks for two individuals (with same number of ties) ranking in the top 1% and bottom 20% provide a zoomed in picture of such differences (Figs. 1C and D, respectively). The wealthiest one-percenters have higher diversity in mobile contacts and are centrally located surrounded by other highly connected people (network hubs). On the other hand, the bottom poorest individuals have low contact diversity and are weakly connected to fewer hubs. The crux of the matter is to find a reliable social network centrality to quantify this notable evidence of network influence.

Many metrics or centralities have been considered to characterize the influence or importance of nodes in a network [3, 14, 18]. We consider those centralities that can be scaled to the large network size considered here (SM III, Figs. 1E-F): (a) degree centrality k_i (number of ties of individual i) is one of the simplest [3], (b) PageRank, of Google fame [19], is an eigenvector centrality that includes the importance of not only the degree, but also the nearest neighbors, (c) the k-shell index k_s of a node (Fig. 1E), i.e., the location of the inner shell obtained by iteratively pruning all nodes with degree $k \leq k_s$ [20], and (d) the collective influence of a node with degree k_i in a region of size ℓ defined by the frontier of the influence ball $\partial\text{Ball}(i, \ell)$, and predicted to be $\text{CI} = (k_i - 1) \sum_{j \in \partial\text{Ball}(i, \ell)} (k_j - 1)$ (Fig. 1F) [21]. As opposed to the other heuristic centralities, CI is derived from theory of maximization of influence in the network [22]. The top CI nodes are thus identified as top influencers or superspreaders of information and they do so by position themselves at strategic locations at the center of spheres surrounded by hubs hierarchically placed at distances ℓ (Fig. 1F). These collective

influencers also constitute an optimal set that provide integrity to the social fabric: they are the smallest number of people that, upon leaving the network (a process mathematically known as optimal percolation [21]), would disintegrate the network into small disconnected pieces.

By definition, all metrics have similarities (e.g., they are proportional to k , and PageRank and CI are based on largest eigenvalues of the adjacency and non-backtracking matrix, respectively [21]), and indeed, we find that their values in the phone communication network are correlated (Supplementary Table II). More interestingly, Fig. 2 provides evidence of correlation between the four network metrics with financial status (ranked credit limit) when we control for age, indicating that the network location correlates with financials. In this figure, we plot the fraction of wealthy individuals (defined as top 4-quantile, equivalent to a credit limit greater than MXN \$70,000, see SM IV [17]) in a sampling grid for a given value of age and social metric as indicated.

While all social metrics show correlations with financial status when considered with age (Fig. 2) the question remains which metric is the most efficient predictor. Strong correlations with economic wellness are observed for the feature pairs (age, k-shell) ($R^2 = 0.96$, Fig. 2B) and (age, CI) ($R^2 = 0.93$, Fig. 2D). SM V provides further comparison when considering the metrics alone indicating the k-shell and CI capture better the correlation with credit line. Among these two metrics, CI guarantees both a requirement for strong correlation and sufficient resolution. k-shell could not capture further details due to its limitation of values (k-shell ranges from 1 to 23, a typical shell containing tens of million people, while CI spans over seven order of magnitude, Fig. S9D). This fact results in CI being the most efficient social signature for financial status of the individuals. According to its definition (Fig. 1F) a top CI node is a moderate to strong hub surrounded by other hubs hierarchically placed at distance ℓ . It is then no surprise that top CI predicts well the economic status of the top earners, since, this structure is similar to the ego-centric network evidenced by the top 1% earners in Fig. 1C.

To track the effect of CI independently of age we investigate the effects of CI inside two specific age groups in Figs. 3A and B. In both age groups, high CI is always accompanied by higher population of wealthy people. A relatively smaller slope in age group <30 suggests that the CI network effect is more sensitive for elder people with more mature and stable economic levels than younger people. When we combine age and CI quantile ranking into an

age-network composite: $ANC = \alpha \text{ Age} + (1 - \alpha) \text{ CI}$, with $\alpha = 0.5$ a remarkable correlation ($R^2 = 0.99$, Fig. 3C) is achieved. By combining network information with age, the probability to identify individuals with high credit limit reaches $\sim 70\%$ at the highest earner level. Such a level of accuracy renders the model practical to infer individual’s financial ability using network collective influence as we show next.

To validate our strategy we perform a social marketing campaign, whose objective is the acquisition of new credit card clients, by sending messages to affluent individuals identified by their CI values inviting the recipient to initiate the product request. We notice that in this experiment we use an independent dataset at a different time frame, and we use only the CI values extracted from the network to classify the targeted people. Specifically, we used the communication network resulting from the aggregation of calls and SMS exchanged between users over a period of 3 months from December 2015 to February 2016. The resulting social network contains 7.19×10^7 people and 3.51×10^8 links. The campaign was conducted on a total of 656,944 people who were targeted by a SMS message offering the product according to their CI values in the social network. We also sent messages to a control group of 48,000 nodes, chosen randomly. To evaluate the campaign, we measure the response rate, i.e. the number of recipients who requested the product divided by the number of targeted people as a function of CI. In the control group, the response rate to the messages was 0.331%. Our results show that groups of increasing CI show an increase in their response rate, with a sound three-fold gain in the rate of response of the top influencers identified by top CI values when compared to the random case. The results of the experiment are summarized in Table I and in Fig. 4.

Our combined datasets offer also the possibility to test at the level of single individuals the importance of diversity of links as measured by the ties to distant communities in the network not directly connected to the individual’s own community [4–6]. To this end, we first detect the communities in the social network by applying fast fold modularity detection algorithms (SM VI) [23, 24]. The diversity of an individual can be quantified through the diversity ratio $DR = W_{\text{out}}/W_{\text{in}}$ [10], defined as the ratio of total communication events with people outside its own community, W_{out} , and inside its own community, W_{in} . This ratio is weakly correlated to CI ($R = 0.4$) suggesting that it captures a different feature of network influence. We implement the same statistics of composite ranking as before resulting in an age-diversity-composite $ADC = \alpha \text{ Age} + (1 - \alpha) \text{ DR}$, with weight $\alpha = 0.5$. The result

CI range	Count	Quantile	Answered Yes	Response Rate
[0,48]	66495	0.1	170	0.26%
(48, 246]	65164	0.2	218	0.33%
(246, 600]	65961	0.3	316	0.48%
(600, 1144]	65376	0.4	332	0.51%
(1144, 1992]	65477	0.5	363	0.55%
(1992, 3408]	65477	0.6	458	0.70%
(3408, 6032]	65736	0.7	493	0.75%
(6032, 11772]	65641	0.8	555	0.8%
(11772, 28740]	65683	0.9	657	1.0%
(28740, 2719354]	65683	1.0	573	0.87%

TABLE I. Results of the “real-life” marketing campaign where individuals (Count) were targeted according to their quantile CI ranking in the whole social network obtained from the phone communication activity. The response to the campaign (Answered Yes) was computed to calculate the Response Rate.

(Fig. 3D) shows that ADC correlates with individual financial well-being generalizing the aggregated results in [6] to the individuals. Thus, the considered social metrics, DR and CI, express the fact that higher economic levels are correlated (we notice that no causal inference can be established with the present data) with the ability to communicate with individuals outside our local tightly-knit friendship community [4], and to position at particular network locations of high CI that are optimal for information spreading and structural stability of the social network.

This result highlights the possibility of predicting financial situation as well as the benefits of social target policies from network metrics leading to tangible improvements in social targeting campaigns. This has an immediate impact in designing optimal marketing campaigns by identifying the affluent targets based on their influential position in the social network. This finding may be also raised at the level of a principle, a fact that would explain the emergence of the phenomenon of collective influence itself as the result of the local optimization of socio-economic interactions.

-
- [1] M. Newman, *Networks: An Introduction* (Oxford University Press, Oxford, 2010).
- [2] G. Caldarelli, A. Vespignani, *Large Scale Structure and Dynamics of Complex Networks: From Information Technology to Finance and Natural Science* (World Scientific, Singapore, 2007).
- [3] S. Wasserman, K. Faust, *Social Network Analysis* (Cambridge University Press, Cambridge, 1994).
- [4] M. S. Granovetter, *Am. J. Sociol.*, **78**, 360–1380 (1973).
- [5] M. S. Granovetter, *J. Eco. Persp.*, **19**, 33–50 (2005).
- [6] N. Eagle, M. Macy, R. Claxton, *Science* **328**, 1029–1031 (2010).
- [7] V. K. Singh, L. Freeman, B. Lepri, A. S. Pentland, *2013 IEEE International Conference on Social Computing (SocialCom)*, pp. 174–179 (2013).
- [8] L. Smith-Doerr, W. W. Powell, in *The Handbook of Economic Sociology* (Princeton University Press, 2005) Vol. 2, pp. 379–402.
- [9] D. Strang, S. A. Soule, *Annu. Rev. Sociol.* 265–290 (1998).
- [10] R. S. Burt, *Structural holes: The social structure of competition* (Harvard University Press, 2009).
- [11] S. Page, *The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies* (Princeton University Press, 2007).
- [12] R. Fernandez, N. Weinberg, *Stanford GSB Research Paper Series*, no. **1382**, 1 (1994).
- [13] C. Zimmer, *The Art and Science of Entrepreneurship* (Ballinger, Cambridge, MA, 1986), pp. 3–23.
- [14] L. C. Freeman, *Soc. Net.* **1**, 215–239 (1979).
- [15] J. P. Onnela, *et al.*, *Proc. Natl. Acad. Sci. USA* **104**, 7332–7336 (2007).
- [16] M. C. Gonzalez, C. A. Hidalgo, A.-L. Barabasi, *Nature* **453**, 779–782 (2008).
- [17] N. Eagle, A. S. Pentland, D. Lazer, *Proc. Natl. Acad. Sci. USA* **106**, 15274–15278 (2009).
- [18] S. Pei, H. A. Makse, *J. Stat. Mech.*, P12002 (2013).
- [19] L. Page, S. Brin, R. Motwani, T. Winograd, *The PageRank citation ranking: bringing order to the web* (Stanford InfoLab, 1999).
- [20] M. Kitsak, *et al.*, *Nat. Phys.* **6**, 888–893 (2010).
- [21] F. Morone, H. A. Makse, *Nature* **6**, 65–68 (2015).

- [22] D. Kempe, J. Kleinberg, É. Tardos, *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 137–146 (2003).
- [23] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, E. Lefebvre, *J. Stat. Mech.*, P10008 (2008).
- [24] M. E. Newman, *Phys. Rev. E* **70**, 056131 (2004).

Acknowledgments

Research was sponsored by the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-09-2-0053 (the ARL Network Science CTA). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

We thank B. Min and M. Travizano for discussions.

FIG. 1. Visualization of communication activity of population in the top 10% **(A)** and bottom 10% **(B)** total credit limit classes. **(C)** An example of the ego-network (extended to two layers) for an individual in the top 1% wealthy class with credit limit larger than \sim MXN \$430,000 and **(D)** the bottom 20% class with credit limit smaller than MXN \$10,000. Both individuals have the same degree $k = 30$. Different colors represent the communities clustered by modularity (SM VI). The top one-percenter has more diverse and rich ego-network structure connecting to 25 different communities and \sim 5,000 nodes, many of them are also hubs. The bottom 20% individual connects to only 9 communities and \sim 300 nodes. **(E)** Schematic representation of a network under k-shell decomposition. **(F)** Example of the calculation of CI. The collective influence $\text{Ball}(i, \ell)$ of radius $\ell = 3$ around node i is the set of nodes contained inside the sphere and ∂Ball is the set of nodes on the boundary (brown). CI is the degree minus one of the central node times the sum of the degree minus one of the nodes at the boundary of the ball.

FIG. 2. Correlation between the fraction of wealthy individuals vs. age and **(A)** k ($R^2 = 0.92$), **(B)** k-shell ($R^2 = 0.96$), **(C)** PageRank ($R^2 = 0.96$), and **(D)** $\log_{10}\text{CI}$ ($R^2 = 0.93$). Only the groups with population larger than 20 are shown in the plot. The four metrics correlate well with financial status when considered with age. Further correlations are studied in SM V indicating that CI could be considered as an efficient metric out of the four of them.

FIG. 3. Correlation between the fraction of wealthy individuals as given by the top 25% credit limit and CI in different age groups of **(A)** 18-30, **(B)** >45 . Correlations between top economic status and large collective influence as determined by CI values in different ages are significant in all age groups while the slope of the linear regression is larger in elder group (0.053 compared to 0.037). **(C)** Age-network composite ranking $\text{ANC} = 1/2 \text{ Age} + 1/2 \text{ CI}$, and **(D)** age-diversity composite ranking $\text{ADC} = 1/2 \text{ Age} + 1/2 \text{ DR}$. By combining the network metrics with age into a composite index, the chances to identify people of high financial status reaches up to $\sim 70\%$ for high values of the composite and both R^2 show high level of correlation ($R^2 = 0.99$ and 0.96 for ANC and ADC, respectively) making both composites good predictor of wealth in practical applications.

FIG. 4. Response rate of acquisition vs. CI quantile in the real-life CI-targeted marketing campaign. The response rate increases approximately linear with CI ranking. The CI-targeted campaign shows a three-fold gain for the top influencers with high CI as compared

with a campaign targeting a randomized control group.

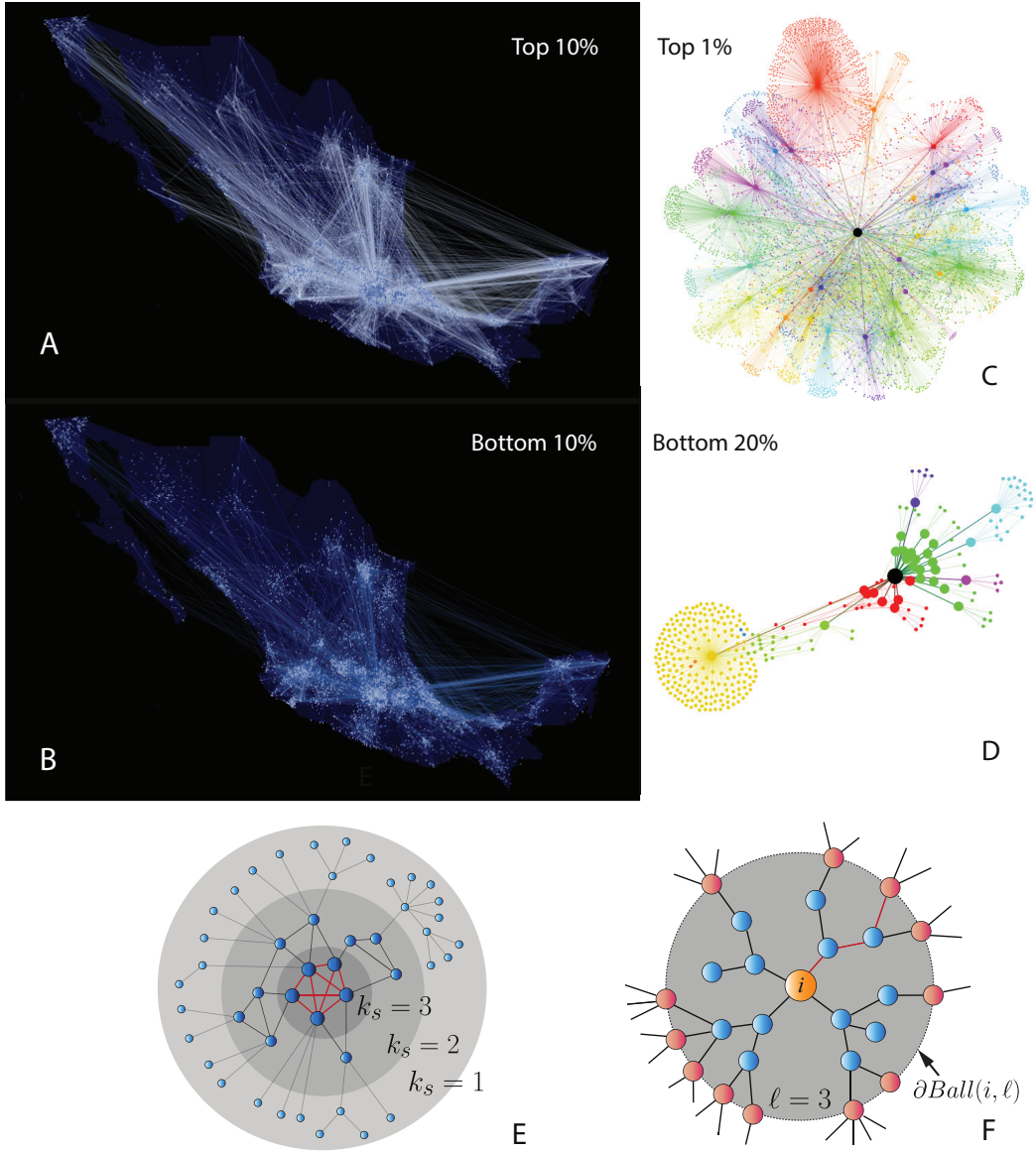


FIG. 1.

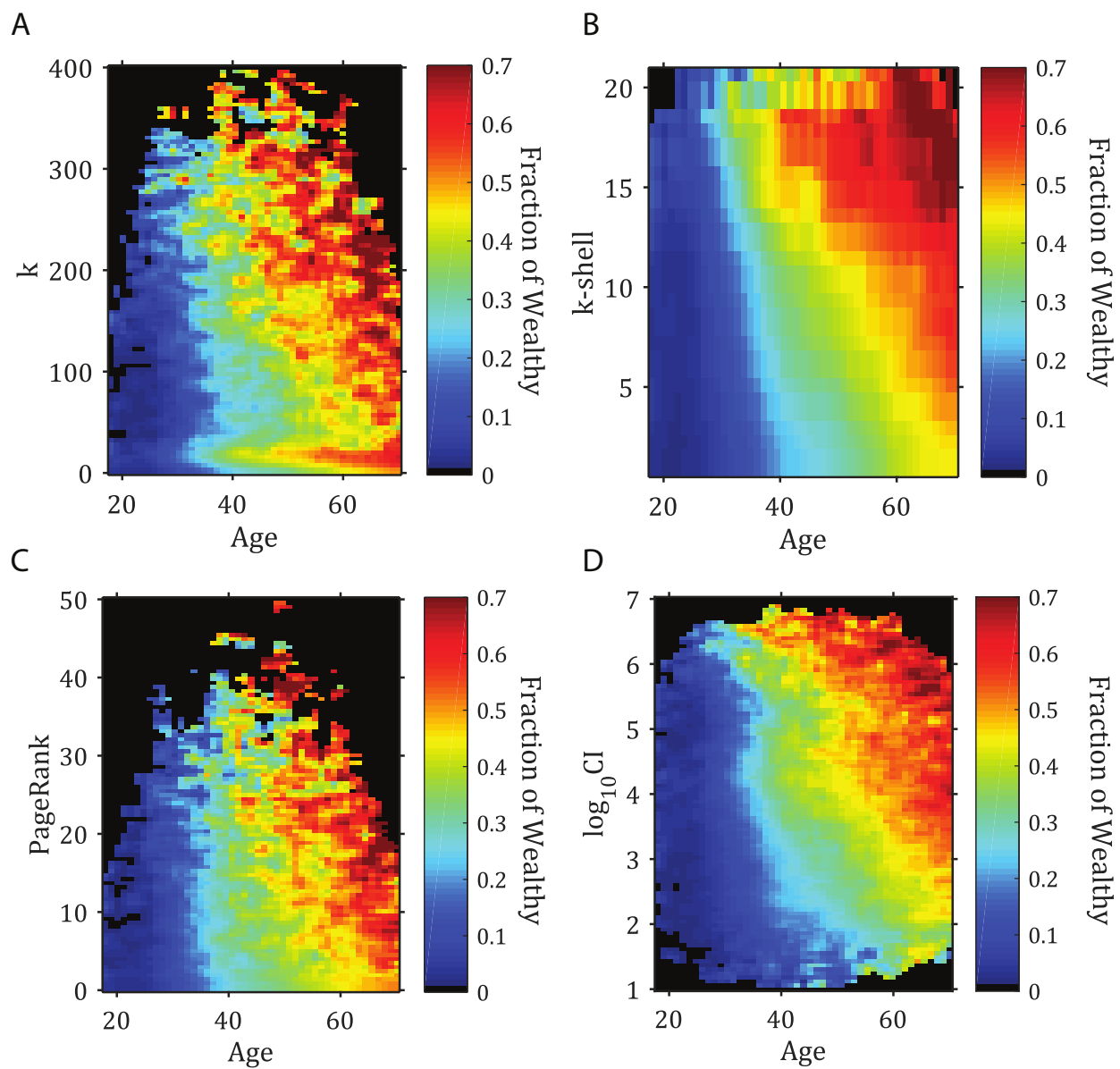


FIG. 2.

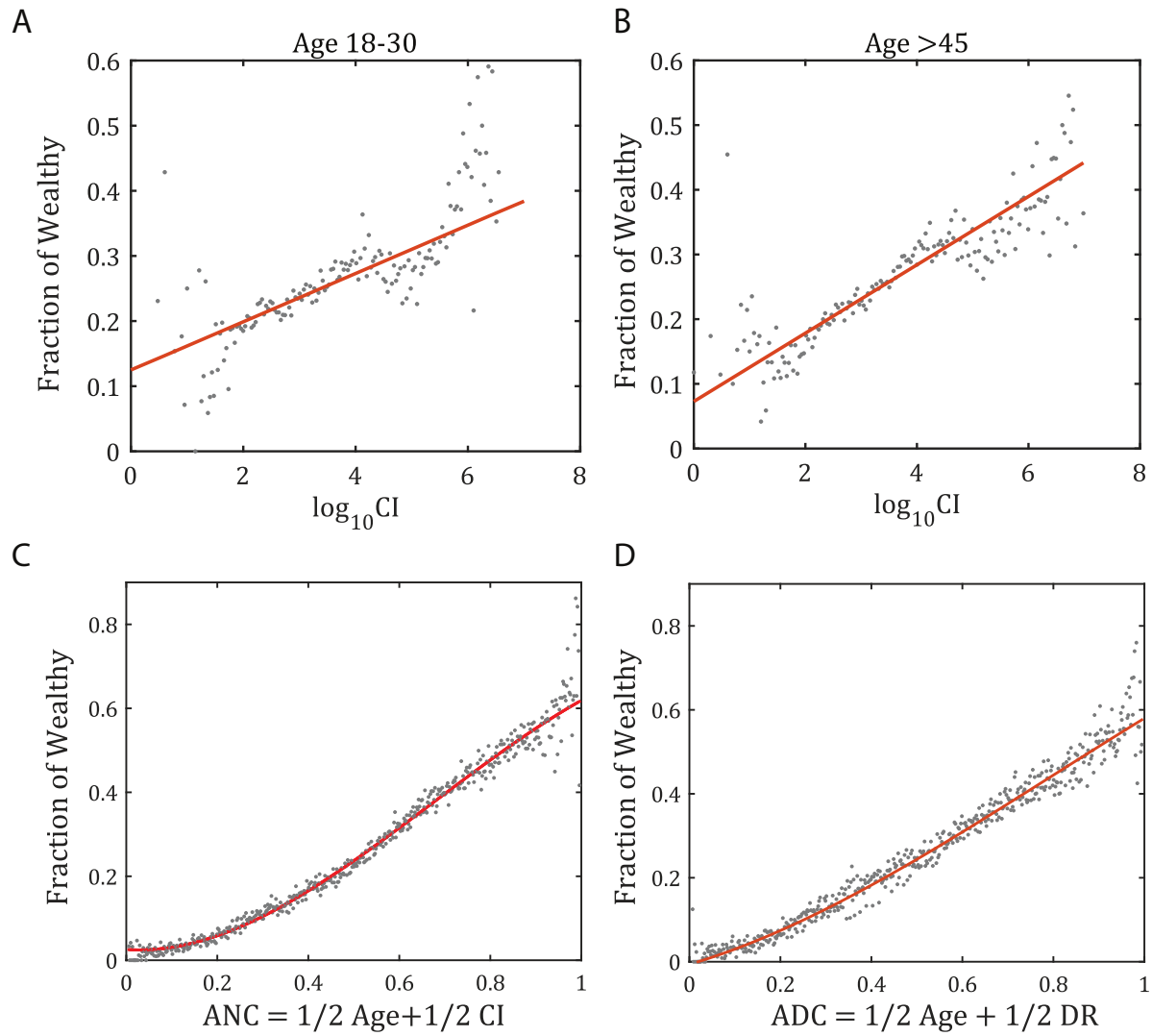


FIG. 3.

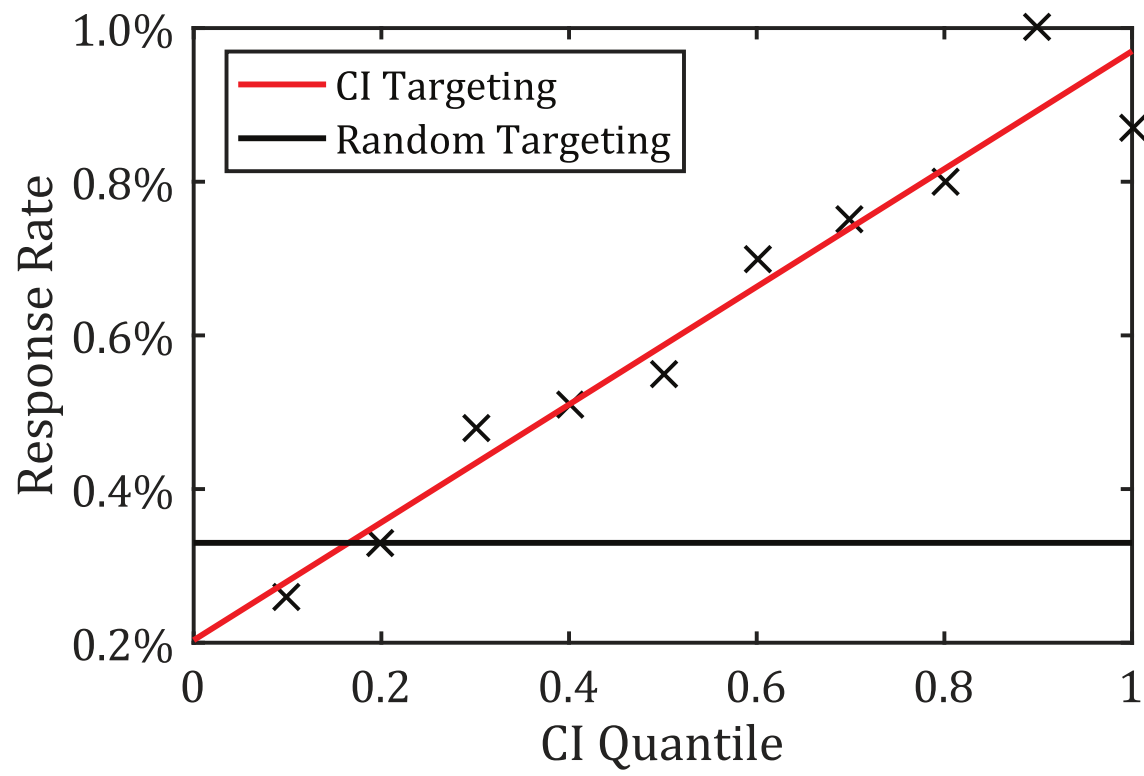


FIG. 4.