Correlations of consumption patterns in social-economic networks

Yannick Leo[1], Márton Karsai[1,], Carlos Sarraute[2] and Eric Fleury[1]*

[1] *Univ Lyon, ENS de Lyon, Inria, CNRS, UCB Lyon 1, LIP UMR 5668, IXXI, F-69342, Lyon, France*
[2] *Grandata Labs, Bartolome Cruz 1818 V. Lopez. Buenos Aires, Argentina*

[*] *Corresponding author: marton.karsai@ens-lyon.fr*

Abstract

We analyze a coupled dataset collecting the mobile phone communication and bank transactions history of a large number of individuals living in Mexico. After mapping the social structure and introducing indicators of socioeconomic status, demographic features, and purchasing habits of individuals we show that typical consumption patterns are strongly correlated with identified socioeconomic classes leading to patterns of stratification in the social structure. In addition we measure correlations between merchant categories and introduce a correlation network, which emerges with a meaningful community structure. We detect multivariate relations between merchant categories and show correlations in purchasing habits of individuals. Our work provides novel and detailed insight into the relations between social and consuming behaviour with potential applications in recommendation system design.

1 Introduction

The consumption of goods and services is a crucial element of human welfare. The uneven distribution of consumption power among individuals goes hand in hand with the emergence and reservation of socioeconomic inequalities in general. Individual financial capacities restrict personal consumer behaviour, arguably correlate with one's purchasing preferences, and play indisputable roles in determining the socioeconomic position of an ego in the larger society [1, 2, 3, 4, 5]. Investigation of relations between these characters carries a great potential in understanding better rational social-economic behaviour [6], and project to direct applications in personal marketing, recommendation, and advertising.

Social Network Analysis (SNA) provides one promising direction to explore such problems [7], due to its enormous benefit from the massive flow of human behavioural data provided by the digital data revolution [8]. The advent of this era was propagated by some new data collection techniques, which allowed the recording of the digital footprints and interaction dynamics of millions of individuals [9, 10]. On the other hand, although social behavioural data brought us detailed knowledge about the structure and dynamics of social interactions, it commonly failed to uncover the relationship between social and economic positions of individuals. Nevertheless, such correlations play important roles in determining one's socioeconomic status (SES) [11], social tie formation preferences due to status homophily [12, 13], and in turn potentially stand behind the emergent stratified structure and segregation on the society level [4, 14]. However until now, the coupled investigation of individual social and economic status remained a great challenge due to lack of appropriate data recording such details simultaneously.

As individual economic status restricts one's capacity in purchasing goods and services, it induces divergent consumption patterns between people at different socioeconomic positions [6, 1, 2]. This is reflected by sets of commonly purchased products, which are further associated to one's social status [15]. Consumption behaviour has been addressed from various angles considering e.g. environmental effects, socioeconomic position, or social influence coming from connected peers [1]. However, large data-driven studies combining information about individual purchasing and interaction patterns in a society large population is still rare, even questions addressing correlation between consumption and social behaviour are at utmost interest.

In this study we address these crucial problems via the analysis of a dataset, which simultaneously records the mobile-phone communication, bank transaction history, and purchase sequences of millions of inhabitants of Mexico over several months. This corpus, one among the firsts at this scale and details, allows us to infer the socioeconomic status, consumption habits, and the underlying social structure of millions of connected individuals. Using this information our overall goal is to identify people with certain financial capacities, and to understand *how much money they spend, on what they spend, and whether they spend like their friends?* More precisely, we formulate our study around two research questions:

- Can one associate typical consumption patterns to people and to their peers belonging to the same or different socioeconomic classes, and if yes how much such patterns vary between individuals or different classes?
- Can one draw relations between commonly purchased goods or services in order to understand better individual consumption behaviour?

After reviewing the related literature in Section 2, we describe our dataset in Section 3, and introduce individual socioeconomic indicators to define socioeconomic classes in Section 4. In Section 5 we show how typical consumption patterns vary among classes and relate them to structural correlations in the social network. In Section 6 we draw a correlation network between consumption categories to detect patterns of commonly purchased goods and services. Finally we present some concluding remarks and future research ideas.

2 Related work

Earlier hypothesis on the relation between consumption patterns and socioeconomic inequalities, and their correlations with demographic features such as age, gender, or social status were drawn from specific sociological studies [16] and from cross-national social surveys [17]. However, recently available large datasets help us to effectively validate and draw new hypotheses as population-large individual level observations and detailed analysis of human behavioural data became possible. These studies shown that personal social interactions, social influence [1], or homophily [22] in terms of age or gender [20] have strong effects on purchase behaviour, knowledge which led to the emergent domain of online social marketing [21]. Yet it is challenging to measure correlations between individual social status, social network, and purchase patterns simultaneously. Although socioeconomic parameters can be estimated from communication networks [18] or from external aggregate data [19] usually they do not come together with individual purchase records. In this paper we propose to explore this question through the analysis of a combined dataset proposing simultaneous observations of social structure, economic status and purchase habits of millions of individuals.

3 Data description

In the following we are going to introduce two datasets extracted from a corpus combining the mobile phone interactions with purchase history of individuals.

DS1: Ego social-economic data with purchase distributions

Communication data used in our study records the temporal sequence of 7,945,240,548 call and SMS interactions of 111,719,360 anonymized mobile phone users for 21 months in Mexico. Each call detailed record (CDR) contains the time, unique caller and callee IDs, the direction (who initiate the call/SMS), and the duration of the interaction. At least one participant of each interaction is a client of a single mobile phone operator in Mexico, but other mobile phone users who are not clients of the actual provider also appear in the dataset with unique IDs. All unique IDs are anonymized as explained below, thus individual identification of any person is impossible from the data. Using this dataset we constructed a large social network where nodes are users (whether clients or not of the actual provider), while links are drawn between any two users if they interacted (via call or SMS) at least once during the observation period. We filtered out call services, companies, and other non-human actors from the social network by removing all nodes (and connected links) who appeared with either in-degree $k_{in} = 0$ or out-degree $k_{out} = 0$. We repeated this procedure recursively until we received a network where each user had $k_{in}, k_{out} > 0$, i. e. made at least one out-going and received at least one in-coming communication event during the nearly two years of observation. After construction and filtering the network remained with

82,453,814 users connected by 1,002,833,289 links, which were considered to be undirected after this point.

To calculate individual economic estimators we used a dataset provided by a single Bank in Mexico. This data records financial details of 6,002,192 people assigned with unique anonymized identifiers over 8 months from November 2014 to June 2015. The data provides time varying customer variables as the amount of their debit card purchases, their monthly loans, and static user attributes such as their billing postal code (zip code), their age and their gender.

A subset of IDs of the anonymized bank and mobile phone costumers were matched[1]. This way of combining the datasets allowed us to simultaneously observe the social structure and estimate economic status (for definition see Section 4) of the connected individuals. This combined dataset contained 999,456 IDs, which appeared in both corpuses. However, for the purpose of our study we considered only the largest connected component of this graph. This way we operate with a connected social graph of 992,538 people connected by 1,960,242 links, for all of them with communication events and detailed bank records available.

To study consumption behaviour we used purchase sequences recording the time, amount, merchant category code of each purchase event of each individual during the observation period of 8 months. Purchase events are linked to one of the 281 merchant category codes (mcc) indicating the type of the actual purchase, like fast food restaurants, airlines, gas stations, etc. Due to the large number of categories in this case we decided to group mccs by their types into 28 purchase category groups (PCGs) using the categorization proposed in [23]. After analyzing each purchase groups 11 of them appeared with extremely low activity representing less than 0.3% (combined) of the total amount of purchases, thus we decided to remove them from our analysis and use only the remaining K_{17} set of 17 groups (for a complete list see Fig.2a). Note that the group named *Service Providers* (k_1 with mcc 24) plays a particular role as it corresponds to cash retrievals and money transfers and it represents around 70% of the total amount of purchases. As this group dominates over other ones, and since we have no further information how the withdrawn cash was spent, we analyze this group k_1 separately from the other $K_{2\text{-}17} = K_{17} \backslash \{k_1\}$ set of groups.

This way we obtained DS1, which collects the social ties, economic status, and coarse grained purchase habit informations of ~ 1 million people connected together into a large social network.

DS2: Detailed ego purchase distributions with age and gender

From the same bank transaction trace of 6,002,192 users, we build a second data set DS2. This dataset collects data about the age and gender of individuals together with their purchase sequence recording the time, amount, and mcc of each debit card purchase of each ego. To receive a set of active users we extracted a corpus of 4,784,745 people that were active at least two months during the observation period. Then for each ego, we assigned a feature set $PV(u) : \{age_u, gender_u, SEG_u, r(c_i, u)\}$ where SEG assigns a socioeconomic group (for definition see Section 4) and $r(c_i, u)$ is an ego purchase distribution vector defined as

$$r(c_i, u) = \frac{m_u^{c_i}}{\sum_{c_i} m_u^{c_i}}. \qquad (1)$$

This vector assigns the fraction of $m_u^{c_i}$ money spent by user u on a merchant category c_i during the observation period. We excluded purchases corresponding to cash retrievals and money transfers, which would dominate our measures otherwise. A minor fraction of purchases are not linked to valid mccs, thus we excluded them from our calculations.

This way DS2 collects 3,680,652 individuals, without information about their underlying social network, but all assigned with a $PV(u)$ vector describing their personal demographic and purchasing features in details.

4 Measures of socioeconomic position

To estimate the personal economic status we used a simple measure reflecting the consumption power of each individual. Starting from the raw data of DS2, which collects the amount and type of debit card purchases, we estimated the economic position of individuals as their average monthly purchase (AMP). More precisely, in case of an ego u who spent $m_u(t)$ amount in month t we calculated the AMP as

$$P_u = \frac{\sum_{t \in T} m_u(t)}{|T|_u} \qquad (2)$$

[1] Note, that the matching, data hashing, and anonymization procedure was carried out through direct communication between the two providers (bank and mobile provider) without the involvement of the scientific partner. After this procedure only anonymized hashed IDs were shared disallowing the direct identification of individuals in any of the datasets.

where $|T|_u$ corresponds to the number of active months of user u (with at least one purchase in each month). After sorting people by their AMP values we computed the normalized cumulative distribution function of P_u as

$$C(f) = \frac{\sum_{f'=0}^{f} P_u(f')}{\sum_u P_u} \tag{3}$$

as a function of f fraction of people. This function (Fig.1a) appears with high variance and suggests large imbalances in terms of the distribution of economic capacities among individuals in agreement with earlier social theory [27].

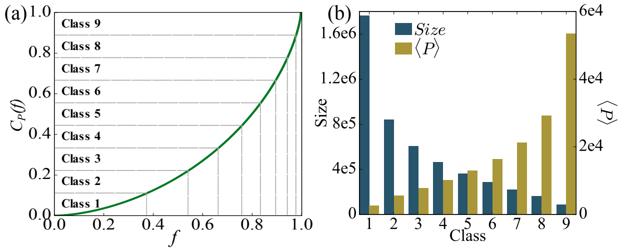

Fig. 1: **Social class characteristics (a)** Schematic demonstration of user partitions into 9 socioeconomic classes by using the cumulative AMP function $C(f)$. Fraction of egos belonging to a given class (x axis) have the same sum of AMP $(\sum_u P_u)/n$ (y axis) for each class. **(b)** Number of egos (green) and the average AMP $\langle P\rangle$ (in USD[2]) per individual (yellow) in different classes.

Subsequently we used the $C(f)$ function to assign egos into 9 economic classes (also called socioeconomic classes with smaller numbers assigning lower classes) such that the sum of AMP in each class s_j was the same equal to $(\sum_u P_u)/n$ (Fig.1). We decided to use 9 distinct classes based on the common three-stratum model [25], which identifies three main social classes (lower, middle, and upper), and for each of them three sub-classes [26]. There are several advantages of this classification: (a) it relies merely on individual economic estimators, P_u, (b) naturally partition egos into classes with decreasing sizes for richer groups and (c) increasing $\langle P\rangle$ average AMP values per egos (Fig.1b)[2].

5 Socioeconomic correlations in purchasing patterns

In order to address our first research question we were looking for correlations between individuals in different socioeconomic classes in terms of their consumption behaviour on the level of purchase category groups. We analyzed the purchasing behaviour of people in DS1 after categorizing them into socioeconomic classes as explained in Section 4.

First for each class s_j we take every users $u \in s_j$ and calculate the m_u^k total amount of purchases they spent on a purchase category group $k \in K_{17}$. Then we measure a fractional distribution of spending for each PCGs as:

$$r(k, s_j) = \frac{\sum_{u\in s_j} m_u^k}{\sum_{u\in s} m_u^k}, \tag{4}$$

where $s = \bigcup_j s_j$ assigns the complete set of users. In Fig.2a each line shows the $r(k, s_j)$ distributions for a PCG as the function of s_j social classes, and lines are sorted (from top to bottom) by the total amount of money spent on the actual PCG[3]. Interestingly, people from lower socioeconomic classes spend more on PCGs associated to essential needs, such as *Retail Stores (St.)*, *Gas Stations*, *Service Providers* (cash) and *Telecom*, while in the contrary, other categories associated to extra needs such as *High Risk Personal Retail* (Jewelry, Beauty), *Mail Phone Order*, *Automobiles*, *Professional Services (Serv.)* (extra health services), *Whole Trade* (auxiliary goods), *Clothing St.*, *Hotels* and *Airlines* are dominated by people from higher socioeconomic classes. Also note that concerning *Education* most of the money is spent by the lower middle classes, while *Miscellaneous St.* (gift, merchandise, pet St.) and more apparently *Entertainment* are categories where the lowest and highest classes are spending the most.

From this first analysis we can already identify large differences in the spending behaviour of people from lower and upper classes. To further investigate these dissimilarities on the individual level, we consider the $K_{2\text{-}17}$ category set as defined in section 3 (category k_1 excluded) and build a spending vector $SV(u) = [SV_2(u), ..., SV_{17}(u)]$ for each ego u.

[2] To assign purchase values in USD we used the daily average currency rate (17.90 MXN/USD) on the 2nd March 2016.

[3] Note that in our social class definition the cumulative AMP is equal for each group and this way each group represents the same economic potential as a whole. Values shown in Fig.2a assign the total purchase of classes. Another strategy would be to calculate per capita measures, which in turn would be strongly dominated by values associated to the richest class, hiding any meaningful information about other classes.

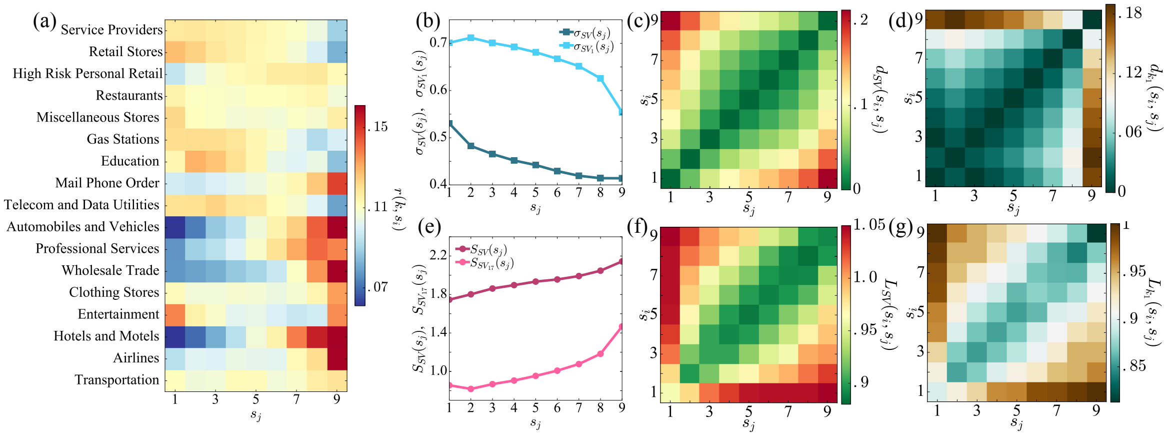

Fig. 2: **Consumption correlations in the socioeconomic network (a)** $r(k, s_i)$ distribution of spending in a given purchase category group $k \in K_{17}$ by different classes s_j. Distributions are normalised as in Eq.4, i.e. sums up to 1 for each category. **(b)** Dispersion $\sigma_{SV}(s_j)$ for different socioeconomic classes considering PCGs in $K_{2\text{-}17}$ (dark blue) and the single category k_1 (light blue). **(c)** (resp. **(d)**) Heat-map matrix representation of $d_{SV}(s_i, s_j)$ (resp. $d_{k_1}(s_i, s_j)$) distances between the average spending vectors of pairs of socioeconomic classes considering PCGs in $K_{2\text{-}17}$ (resp. k_1). **(e)** Shannon entropy measures for different socioeconomic classes considering PCGs in $K_{2\text{-}17}$ (dark pink) and in k_{17} (light pink). **(f)** (resp. **(g)**) Heat-map matrix representation of the average $L_{SV}(s_i, s_j)$ (resp. $L_{k_1}(s_i, s_j)$) measure between pairs of socioeconomic classes considering PCGs in $K_{2\text{-}17}$ (resp. k_1).

Here each item $SV_k(u)$ assigns the fraction of money m_u^k/m_u what user u spent on a category $k \in K_{2\text{-}17}$ out of his/her $m_u = \sum_{k \in K} m_u^k$ total amount of purchases. Using these individual spending vectors we calculate the average spending vector of a given socioeconomic class as $\overline{SV}(s_j) = \langle SV(u) \rangle_{u \in s_j}$. We associate $\overline{SV}(s_j)$ to a representative consumer of class s_j and use this average vector to quantify differences between distinct socioeconomic classes as follows.

The euclidean metric between average spending vectors is:

$$d_{SV}(s_i, s_j) = \|\overline{SV}_k(s_i) - \overline{SV}_k(s_j)\|_2, \qquad (5)$$

where $\|\vec{v}\|_2 = \sqrt{\sum_k v_k^2}$ assigns the L^2 norm of a vector $\vec{v}$. Note that the diagonal elements of $d_{SV}(s_i, s_i)$ are equal to zero by definition. However, in Fig.2c the off-diagonal green component around the diagonal indicates that the average spending behaviour of a given class is the most similar to neighboring classes, while dissimilarities increase with the gap between socioeconomic classes. We repeated the same measurement separately for the single category of cash purchases (PCG k_1). In this case euclidean distance is defined between average scalar measures as $d_{k_1}(s_i, s_j) = \|\langle SV_1 \rangle(s_i) - \langle SV_1 \rangle(s_j)\|_2$. Interestingly, results shown in Fig.2d. indicates that here the richest social classes appear with a very different behaviour. This is due to their relative underspending in cash, which can be also concluded from Fig.2a (first row). On the other hand as going towards lower classes such differences decrease as cash usage starts to dominate.

To explain better the differences between socioeconomic classes in terms of purchasing patterns, we introduce two additional scalar measures. First, we introduce the dispersion of individual spending vectors as compared to their class average as

$$\sigma_{SV}(s_j) = \langle \|\overline{SV}_k(s_j) - SV_k(u)\|_2 \rangle_{u \in s_j}, \qquad (6)$$

which appears with larger values if people in a given class allocate their spending very differently. Second, we also calculate the Shannon entropy of spending patterns as

$$S_{SV}(s_j) = \sum_{k \in K_{2\text{-}17}} -\overline{SV}_k(s_j) \log(\overline{SV}_k(s_j)) \qquad (7)$$

to quantify the variability of the average spending vector for each class. This measure is minimal if each ego of a class s_j spends exclusively on the same single PCG, while it is maximal if they equally spend on each PCG. As it is shown in Fig.2b (light blue line with square symbols) dispersion decreases rapidly as going towards higher socioeconomic classes. This assigns that richer people tends to be more similar in terms of their purchase behaviour. On the other hand, surprisingly, in Fig.2e (dark pink line with square symbols) the increasing trend of the corresponding entropy measure suggests that even richer people behave more similar in terms of spending behaviour they used to allocate their purchases in more PCGs. These trends are consistent even in case of k_1 cash purchase category (see $\sigma_{SV_1}(s_j)$ function depicted with dark blue line in in Fig.2b) or once we include category k_1 into the entropy measure $S_{SV_{17}}(s_j)$ (shown in Fig.2b with light pink line).

To complete our investigation we characterize the effects of social relationships on the purchase habits of individuals. We address this problem through an overall measure quantifying differences between individual purchase vectors of connected egos positioned in the same or different socioeconomic classes. More precisely, we consider each social tie $(u, v) \in E$ connecting individuals $u \in s_i$ and $v \in s_j$, and for each purchase category k we calculate the average absolute difference of their purchase vector items as

$$d^k(s_i, s_j) = \langle |SV_k(u) - SV_k(v)| \rangle_{u \in s_i, v \in s_j}. \tag{8}$$

Following that, as a reference system we generate a corresponding configuration network by taking randomly selected edge pairs from the underlying social structure and swap them without allowing multiple links and self loops. In order to vanish any residual correlations we repeated this procedure in $5 \times |E|$ times. This randomization keeps the degree, individual economic estimators P_u, the purchase vector $SV(u)$, and the assigned class of each people unchanged, but destroys any structural correlations between egos in the social network, consequently between socioeconomic classes as well. After generating a reference structure we computed an equivalent measure $d^k_{rn}(s_i, s_j)$ but now using links $(u, v) \in E_{rn}$ of the randomized network. We repeated this procedure 100 times and calculated an average $\langle d^k_{rn} \rangle (s_i, s_j)$. In order to quantify the effect of the social network we simply take the ratio

$$L_k(s_i, s_j) = \frac{d^k(s_i, s_j)}{\langle d^k_{rn} \rangle (s_i, s_j)} \tag{9}$$

and calculate its average $L_{SV}(s_i, s_j) = \langle L_k(s_i, s_j) \rangle_k$ over each category group $k \in K_{2\text{-}17}$ or respectively k_1. This measure shows whether connected people have more similar purchasing patterns than one would expect by chance without considering any effect of homophily, social influence or structural correlations. Results depicted in Fig.2f and 2g for $L_{SV}(s_i, s_j)$ (and $L_{k_1}(s_i, s_j)$ respectively) indicates that the purchasing patterns of individuals connected in the original structure are actually more similar than expected by chance (diagonal component). On the other hand people from remote socioeconomic classes appear to be less similar than one would expect from the uncorrelated case (indicated by the $L_{SV}(s_i, s_j) > 1$ values typical for upper classes in Fig.2f). Note that we found the same correlation trends in cash purchase patterns as shown in Fig.2g. These observations do not clearly assign whether homophily [12, 13] or social influence [1] induce the observed similarities in purchasing habits but undoubtedly clarifies that social ties (i.e. the neighbors of an ego) and socioeconomic status play deterministic roles in the emerging similarities in consumption behaviour.

6 Purchase category correlations

To study consumption patterns of single purchase categories PCGs provides a too coarse grained level of description. Hence, to address our second question we use DS2 and we downscale from the category group level to the level of single merchant categories. We are dealing with 271 categories after excluding some with less than 100 purchases and the categories linked to money transfer and cash retrieval (for a complete list of IDs and name of the purchase categories considered see Table 1). As in Section 3 we assign to each ego u a personal vector $PV(u)$ of four socioeconomic features: the age, the gender, the social economic group, and the distribution $r(c_i, u)$ of purchases in different merchant categories made by the central ego. Our aim here is to obtain an overall picture of the consumption structure at the level of merchant categories and to understand precisely how personal and socioeconomical features correlate with the spending behaviour of individuals and with the overall consumption structure.

As we noted in section 5, the purchase spending vector $r(c_i, u)$ of an ego quantifies the fraction of money spent on a category c_i. Using the spending vectors of n number of individuals we define an overall correlation measure between categories as

This symmetric formulae quantifies how much people spend on a category c_i if they spend on an other c_j category or vice versa. Therefore, if $\rho(c_i, c_j) > 1$,

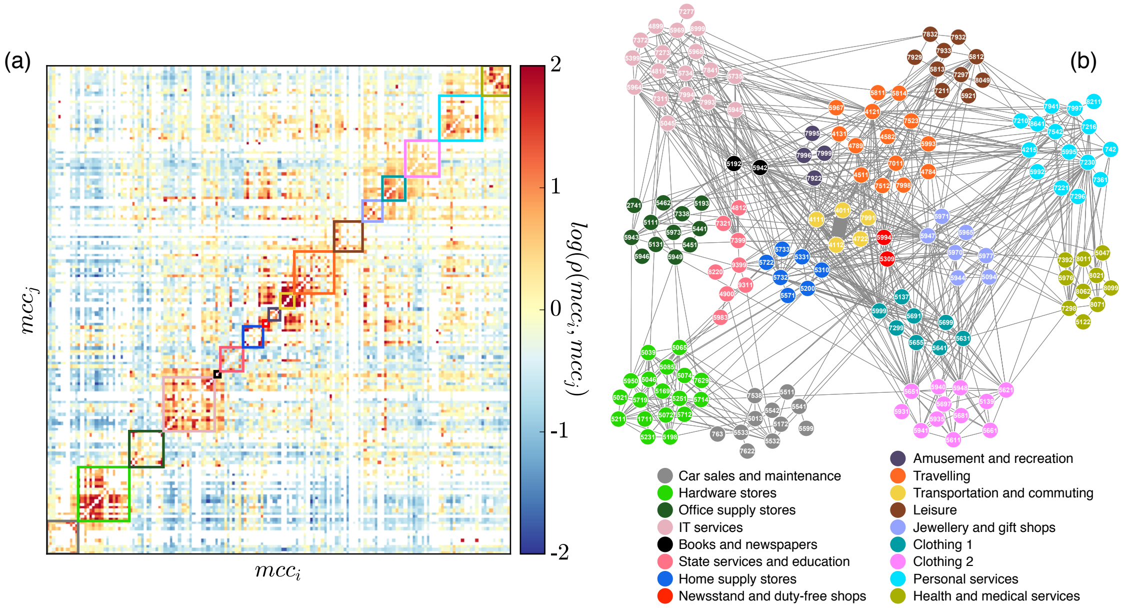

Fig. 3: **Merchant category correlation matrix and graph (a)** 163×163 matrix heatmap plot corresponding to $\rho(c_i, c_j)$ correlation values (see Eq. 10) between categories. Colors scale with the logarithm of correlation values. Positive (resp. negative) correlations are assigned by red (resp. blue) colors. Diagonal components represent communities with frames colored accordingly.**(b)** Weighted $G_\rho^>$ correlation graph with nodes annotated with MCCs (see Table 1). Colors assign 17 communities of merchant categories with representative names summarized in the figure legend.

the categories c_i and c_j are positively correlated and if $\rho(c_i, c_j) < 1$, categories are negatively correlated. Using $\rho(c_i, c_j)$ we can define a weighted correlation graph $G_\rho = (V_\rho, E_\rho, \rho)$ between categories $c_i \in V_\rho$, where links $(c_i, c_j) \in E_\rho$ are weighted by the $\rho(c_i, c_j)$ correlation values. The weighted adjacency matrix of G_ρ is shown in Fig.3a as a heat-map matrix with logarithmically scaling colors. Importantly, this matrix emerges with several block diagonal components suggesting present communities of strongly correlated categories in the graph.

To identify categories which were commonly purchased together we consider only links with positive correlations. Furthermore, to avoid false positive correlations, we consider a 10% error on r that can induce, in the worst case 50% overestimation of the correlation values. In addition, to consider only representative correlations we take into account category pairs which were commonly purchased by at least 1000 consumers. This way we receive a $G_\rho^>$ weighted sub-graph of G_ρ, shown in Fig.3b, with 163 nodes and 1664 edges with weights $\rho(c_i, c_j) > 1.5$.

To identify communities in $G_\rho^>$ indicated by the correlation matrix in Fig.3a we applied a graph partitioning method based on the Louvain algorithm [28]. We obtained 17 communities depicted with different colors in Fig.3b and as corresponding colored frames in Fig.3a. Interestingly, each of these communities group a homogeneous set of merchant categories, which could be assigned to similar types of purchasing activities (see legend of Fig.3b). In addition, this graph indicates how different communities are connected together. Some of them, like *Transportation, IT* or *Personal Serv.* playing a central role as connected to many other communities, while other components like *Car sales and maintenance* and *Hardware St.*, or *Personal* and *Health and medical Serv.* are more like pairwise connected. Some groups emerge as standalone communities like *Office Supp. St.*, while others like *Books and newspapers* or *Newsstands and duty-free Shops (Sh.)* appear as bridges despite their small sizes.

$$\rho(c_i, c_j) = \frac{n(\sum_u r(c_i, u) r(c_j, u))}{(\sum_u r(c_i, u))(\sum_u r(c_j, u))}. \tag{10}$$

Note that the main categories corresponding to everyday necessities related to food (*Supermarkets,*

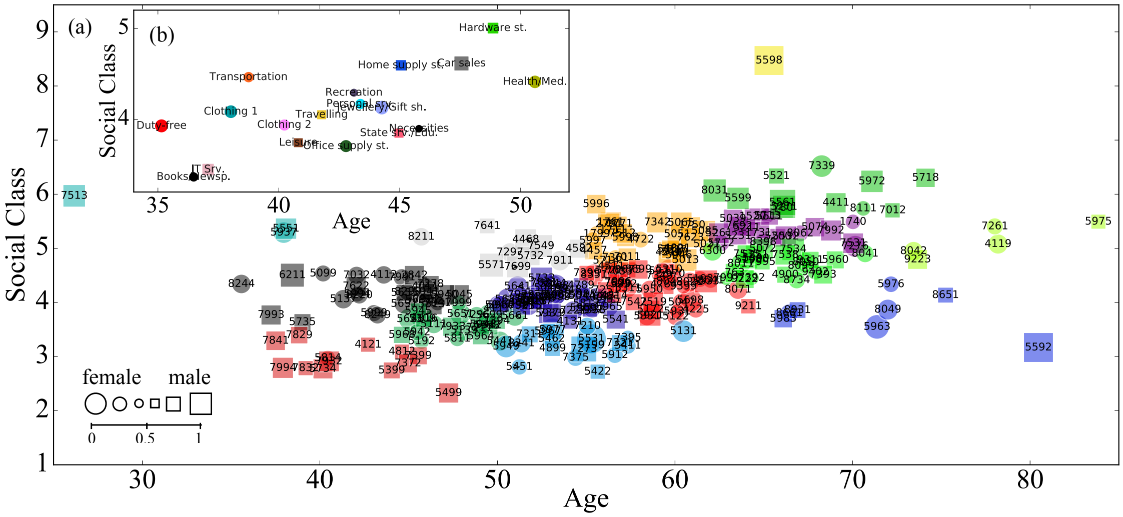

Fig. 4: **Socioeconomic parameters of merchant categories (a)** Scatter plot of $AFS(c_i)$ triplets (for definition see Eq. 11 and text) for 271 merchant categories summarized in Table 1. Axis assign average age and SEG of purchase categories, while gender information are assigned by symbols. The shape of symbols assigns the dominant gender (circle-female, square-male) and their size scales with average values. **(b)** Similar scatter plot computed for communities presented in Fig.3b. Labels and colors are explained in the legend of Fig.3a.

Food St.) and telecommunication (*Telecommunication Serv.*) do not appear in this graph. Since they are responsible for the majority of total spending, they are purchased necessarily by everyone without obviously enhancing the purchase in other categories, thus they do not appear with strong correlations.

Finally we turn to study possible correlations between purchase categories and personal features. An average feature set $AFS(c_i) = \{\langle age(c_i)\rangle, \langle gender(c_i)\rangle, \langle SEG(c_i\}\rangle\}$ is assigned to each of the 271 categories. The average $\langle v(c_i)\rangle$ of a feature $v \in \{age, gender, SEG\}$ assigns a weighted average value computed as:

$$\langle v(c_i)\rangle = \frac{\sum_{u\in\{u\}_i} \alpha_i(v_u) v_u}{\sum_{u\in\{u\}_u} \alpha_i(v)}, \qquad (11)$$

where v_u denotes a feature of a user u from the $\{u\}_i$ set of individuals who spent on category c_i. Here

$$\alpha_i(v_u) = \sum_{(u\in\{u\}_i | v_u = v)} \frac{r(c_i, u)}{n_i(v_u)} \qquad (12)$$

corresponds to the average spending on category c_i of the set of users from $\{u\}_i$ sharing the same value of the feature v. $n_i(v_u)$ denotes the number of such users. In other words, e.g. in case of $v = age$ and c_{742}, $\langle age(c_{742})\rangle$ assigns the average age of people spent on Veterinary Services ($mcc = 742$) weighted by the amount they spent on it. In case of $v = gender$ we assigned 0 to females and 1 to males, thus the average gender of a category can take any real value between $[0, 1]$, indicating more females if $\langle gender(c_i)\rangle \leq 0.5$ or more males otherwise.

We visualize this multi-modal data in Fig.4a as a scatter plot, where axes scale with average age and SEG, while the shape and size of symbols correspond to the average gender of each category. To further identify correlations we applied k-means clustering [29] using the $AFS(c_i)$ of each category. The ideal number of clusters was 15 according to several criteria: Davies-Bouldin Criterion, Calinski-Harabasz criterion (variance ratio criterion) and the Gap method [30]. Colors in Fig.4a assign the identified k-mean clusters.

The first thing to remark in Fig.4a is that the average age and SEG assigned to merchant categories are positively correlated with a Pearson correlation coefficient 0.42 ($p < 0.01$). In other words, elderly people used to purchase from more expensive categories, or alternatively, wealthier people tend to be older, in accordance with our intuition. At the same time, some signs of gender imbalances can be also

742: Veterinary Serv.	5072: Hardware Supp.	5598: Snowmobile Dealers	5950: Glassware, Crystal St.	7296: Clothing Rental	7941: Sports Clubs
763: Agricultural Cooperative	5074: Plumbing, Heating Equip.	5599: Auto Dealers	5960: Dir Mark - Insurance	7297: Massage Parlors	7991: Tourist Attractions
780: Landscaping Serv.	5085: Industrial Supplies	5611: Men Cloth. St.	5962: Direct Marketing - Travel	7298: Health and Beauty Spas	7992: Golf Courses
1520: General Contr.	5094: Precious Objects/Stones	5621: Wom Cloth. St.	5963: Door-To-Door Sales	7299: General Serv.	7993: Video Game Supp.
1711: Heating, Plumbing	5099: Durable Goods	5631: Women?s Accessory Sh.	5964: Dir. Mark. Catalog	7311: Advertising Serv.	7994: Video Game Arcades
1731: Electrical Contr.	5111: Printing, Office Supp.	5641: Children?s Wear St.	5965: Dir. Mark. Retail Merchant	7321: Credit Reporting Agencies	7995: Gambling
1740: Masonry & Stonework	5122: Drug Proprietaries	5651: Family Cloth. St.	5966: Dir Mark - TV	7333: Graphic Design	7996: Amusement Parks
1750: Carpentry Contr.	5131: Notions Goods	5655: Sports & Riding St.	5967: Dir. Mark.	7338: Quick Copy	7997: Country Clubs
1761: Sheet Metal	5137: Uniforms Clothing	5661: Shoe St.	5968: Dir. Mark. Subscription	7339: Secretarial Support Serv.	7998: Aquariums
1771: Concrete Work Contr.	5139: Commercial Footwear	5681: Furriers Sh.	5969: Dir. Mark. Other	7342: Exterminating Services	7999: Recreation Serv.
1799: Special Trade Contr.	5169: Chemicals Products	5691: Cloth. Stores	5970: Artist?s Supp.	7349: Cleaning and Maintenance	8011: Doctors
2741: Publishing and Printing	5172: Petroleum Products	5697: Tailors	5971: Art Dealers & Galleries	7361: Employment Agencies	8021: Dentists, Orthodontists
2791: Typesetting Serv.	5192: Newspapers	5698: Wig and Toupee St.	5972: Stamp and Coin St.	7372: Computer Programming	8031: Osteopaths
2842: Specialty Cleaning	5193: Nursery & Flowers Supp.	5699: Apparel Accessory Sh.	5973: Religious St.	7375: Information Retrieval Serv.	8041: Chiropractors
4011: Railroads	5198: Paints	5712: Furniture	5975: Hearing Aids	7379: Computer Repair	8042: Optometrists
4111: Ferries	5199: Nondurable Goods	5713: Floor Covering St.	5976: Orthopedic Goods	7392: Consulting, Public Relations	8043: Opticians
4112: Passenger Railways	5200: Home Supply St.	5714: Window Covering St.	5977: Cosmetic St.	7393: Detective Agencies	8049: Chiropodists, Podiatrists
4119: Ambulance Serv.	5211: Materials St.	5718: Fire Accessories St.	5978: Typewriter St.	7394: Equipment Rental	8050: Nursing/Personal Care
4121: Taxicabs	5231: Glass & Paint St.	5719: Home Furnishing St.	5983: Fuel Dealers (Non Auto)	7395: Photo Developing	8062: Hospitals
4131: Bus Lines	5251: Hardware St.	5722: House St.	5992: Florists	7399: Business Serv.	8071: Medical Labs
4214: Motor Freight Carriers	5261: Nurseries & Garden St.	5732: Elec. St.	5993: Cigar St.	7512: Car Rental Agencies	8099: Medical Services
4215: Courier Serv.	5271: Mobile Home Dealers	5733: Music Intruments St.	5994: Newsstands	7513: Truck/Trailer Rentals	8111: Legal Services, Attorneys
4225: Public Storage	5300: Wholesale	5734: Comp.Soft. St.	5995: Pet Sh.	7519: Mobile Home Rentals	8211: Elem. Schools
4411: Cruise Lines	5309: Duty Free St.	5735: Record Stores	5996: Swimming Pools Sales	7523: Parking Lots, Garages	8220: Colleges Univ.
4457: Boat Rentals and Leases	5310: Discount Stores	5811: Caterers	5997: Electric Razor St.	7531: Auto Body Repair Sh.	8241: Correspondence Schools
4468: Marinas Serv. and Supp.	5311: Dep. St.	5812: Restaurants	5998: Tent and Awning Sh.	7534: Tire Retreading & Repair	8244: Business Schools
4511: Airlines	5331: Variety Stores	5813: Drinking Pl.	5999: Specialty Retail	7535: Auto Paint Sh.	8249: Training Schools
4582: Airports, Flying Fields	5399: General Merch.	5814: Fast Foods	6211: Security Brokers	7538: Auto Service Shops	8299: Educational Serv.
4722: Travel Agencies	5411: Supermarkets	5912: Drug St.	6300: Insurance	7542: Car Washes	8351: Child Care Serv.
4784: Tolls/Bridge Fees	5422: Meat Prov.	5921: Alcohol St.	7011: Hotels	7549: Towing Serv.	8398: Donation
4789: Transportation Serv.	5441: Candy St.	5931: Secondhand Stores	7012: Timeshares	7622: Electronics Repair Sh.	8641: Associations
4812: Phone St.	5451: Dairy Products St.	5932: Antique Sh.	7032: Sporting Camps	7623: Refrigeration Repair	8651: Political Org.
4814: Telecom.	5462: Bakeries	5933: Pawn Shops	7033: Trailer Parks, Camps	7629: Small Appliance Repair	8661: Religious Orga.
4816: Comp. Net. Serv.	5499: Food St.	5935: Wrecking Yards	7210: Laundry, Cleaning Serv.	7631: Watch/Jewelry Repair	8675: Automobile Associations
4821: Telegraph Serv.	5511: Cars Sales	5937: Antique Reproductions	7211: Laundries	7641: Furniture Repair	8699: Membership Org.
4899: Techno St.	5521: Car Repairs Sales	5940: Bicycle Sh.	7216: Dry Cleaners	7692: Welding Repair	8734: Testing Lab.
4900: Utilities	5531: Auto and Home Supp. St.	5941: Sporting St.	7217: Upholstery Cleaning	7699: Repair Sh.	8911: Architectural Serv.
5013: Motor Vehicle Supp.	5532: Auto St.	5942: Book St.	7221: Photographic Studios	7829: Picture/Video Production	8931: Accounting Serv.
5021: Commercial Furniture	5533: Auto Access.	5943: Stationery St.	7230: Beauty Sh.	7832: Cinema	8999: Professional Serv.
5039: Constr. Materials	5541: Gas Stations	5944: Jewelry St.	7251: Shoe Repair/Hat Cleaning	7841: Video Tape Rental St.	9211: Courts of Law
5044: Photographic Equip.	5542: Automated Fuel Dispensers	5945: Toy,-Game Sh.	7261: Funeral Serv.	7911: Dance Hall & Studios	9222: Government Fees
5045: Computer St.	5551: Boat Dealers	5946: Camera and Photo St.	7273: Dating/Escort Serv.	7922: Theater Ticket	9223: Bail and Bond Payments
5046: Commercial Equipment	5561: Motorcycle Sh.	5947: Gift Sh.	7276: Tax Preparation Serv.	7929: Bands, Orchestras	9311: Tax Payments
5047: Medical Equipment	5571: Motorcycle Sh.	5948: Luggage & Leather St.	7277: Counseling Services	7932: Billiard/Pool	9399: Government Serv.
5051: Metal Service Centers	5592: Motor Homes Dealers	5949: Fabric St.	7278: Buying/Shopping Serv.	7933: Bowling	9402: Postal Serv.
5065: Electrical St.					

Tab. 1: Codes and names of 271 merchant categories used in our study. MCCs were taken from the Merchant Category Codes and Groups Directory published by American Express [23]. Abbreviations correspond to: Serv. - Services, Contr. - Contractors, Supp. - Supplies, St. - Stores, Equip. - Equipment, Merch. - Merchandise, Prov. - Provisioners, Pl. - Places, Sh. - Shops, Mark. - Marketing, Univ. - Universities, Org. - Organizations, Lab. - Laboratories.

concluded from this plot. Wealthier people appear to be commonly males rather than females. A Pearson correlation measure between gender and SEG, which appears with a coefficient 0.29 ($p < 0.01$) confirmed it. On the other hand, no strong correlation was observed between age and gender from this analysis.

To have an intuitive insight about the distribution of merchant categories, we take a closer look at specific category codes (summarized in Table 1). As seen in Fig.4a elderly people tend to purchase in specific categories such as *Medical Serv.*, *Funeral Serv.*, *Religious Organisations*, *Motorhomes Dealers*, *Donation*, *Legal Serv.*. Whereas categories such as *Fast Foods*, *Video Game Arcades*, *Cinema*, *Record St.*, *Educational Serv.*, *Uniforms Clothing*, *Passenger Railways*, *Colleges-Universities* are associated to younger individuals on average. At the same time, wealthier people purchase more in categories as *Snowmobile Dealers*, *Secretarial Serv.*, *Swimming Pools Sales*, *Car Dealers Sales*, while poorer people tend to purchase more in categories related to everyday necessities like *Food St.*, *General Merch.*, *Dairy Products St.*, *Fast Foods* and *Phone St.*, or to entertainment as *Billiard* or *Video Game Arcades*. Typical purchase categories are also strongly correlated with gender as categories more associated to females are like *Beauty Sh.*, *Cosmetic St.*, *Health and Beauty Spas*, *Women Clothing St.* and *Child Care Serv.*, while others are preferred by males like *Motor Homes Dealers*, *Snowmobile Dealers*, *Dating/Escort Serv.*, *Osteopaths*, *Instruments St.*, *Electrical St.*, *Alcohol St.* and *Video Game Arcades*.

Finally we repeated a similar analysis on communities shown in Fig.3b, but computing the AFS on a set of categories that belong to the same community. Results in Fig.4b disclose positive age-SEG correlations as observed in Fig.4a, together with somewhat intuitive distribution of the communities.

7 Conclusion

In this paper we analyzed a multi-modal dataset collecting the mobile phone communication and bank transactions of a large number of individuals living in Mexico. This corpus allowed for an innovative global analysis both in term of social network and its relation to the economical status and merchant habits of individuals. We introduced several measures to estimate the socioeconomic status of each individual together with their purchasing habits. Using these information we identified distinct socioeconomic classes, which reflected strongly imbalanced distribution of purchasing power in the population. After mapping the social network of egos from mobile phone interactions, we showed that typical consumption patterns are strongly correlated with the socioeconomic classes and the social network behind. We observed these correlations on the individual and social class level.

In the second half of our study we detected correlations between merchant categories commonly purchased together and introduced a correlation network which in turn emerged with communities grouping homogeneous sets of categories. We further analyzed some multivariate relations between merchant categories and average demographic and socioeconomic features, and found meaningful patterns of correlations giving insights into correlations in purchasing habits of individuals.

We identified several new directions to explore in the future. One possible track would be to better understand the role of the social structure and interpersonal influence on individual purchasing habits, while the exploration of correlated patterns between commonly purchased brands assigns another promising directions. Beyond our general goal to better understand the relation between social and consuming behaviour these results may enhance applications to better design marketing, advertising, and recommendation strategies, as they assign relations between copurchased product categories.

Acknowledgment

We thank M. Fixman for assistance and Grandata to share the data. We acknowledge the support from the SticAmSud UCOOL project, INRIA, and the SoSweet (ANR-15-CE38-0011-01) and CODDDE (ANR-13-CORD-0017-01) ANR projects.

References

[1] A. Deaton, Understanding Consumption. *Clarendon Press* (1992).

[2] A. Deaton and J. Muellbauer, Economics and Consumer Behavior. *Cambridge University Press* (1980).

[3] T. Piketti, Capital in the Twenty-First Century. (*Harvard University Press*, 2014).

[4] S. Sernau, Social Inequality in a Global Age. (*SAGE Publications*, 2013).

[5] C. E. Hurst, Social Inequality. 8th ed. (*Pearson Education*, 2015).

[6] J. E. Fisher, Social Class and Consumer Behavior: the Relevance of Class and Status", in Advances in Consumer Research Vol. 14, eds. M. Wallendorf and P. Anderson, Provo, UT : Association for Consumer Research, pp 492–496 (1987) .

[7] S. Wasserman, K. Faust, Social Network Analysis: Methods and Applications. (*Cambridge University Press*, 1994).

[8] S. Lohr, The age of big data. (*New York Times*, 2012).

[9] D. Lazer, et. al. Computational Social Science. *Science* **323**, 721–723 (2009)

[10] A. Abraham, A-E. Hassanien, V. Smasel (eds.), Computational Social Network Analysis: Trends, Tools and Research Advances. (*Springer-Verlag*, 2010).

[11] P. Bourdieu, Distinction: A Social Critique of the Judgement of Taste. *Harvard University Press* (Cambridge MA) (1984).

[12] M. McPherson, L. Smith-Lovin, J. M. Cook, Birds of a Feather: Homophily in Social Networks. *Ann. Rev. Sociol.* **27** 415–444 (2001).

[13] P. F. Lazarsfeld, R. K. Merton, Friendship as a Social Process: A Substantive and Methodological Analysis. In Freedom and Control in Modern Society. (*New York: Van Nostrand*, 1954) pp. 18–66.

[14] D. B. Grusky, Theories of Stratification and Inequality. In The Concise Encyclopedia of Sociology. pp. 622-624. (*Wiley-Blackwell*, 2011).

[15] P. West, Conspicuous Compassion: Why Sometimes It Really Is Cruel To Be Kind. *Civitas, Institute for the Study of Civil Society* (London) (2004).

[16] T. W. Chang, Social status and cultural consumption *Cambridge University Press* (2010)

[17] A. Deaton, The analysis of household surveys: a microeconometric approach to development policy. *World Bank Publications* (1997)

[18] Y. Dong, et. al., Inferring user demographics and social strategies in mobile social networks. *Proc. of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 15–24 (2014)

[19] N. Eagle, M. Macy, R. Claxton, Network diversity and economic development. *Science* **328**, 1029–1031 (2010)

[20] L. Kovanen, et. al., Temporal motifs reveal homophily, gender-specific patterns, and group talk in call sequences. *Proc. Nat. Acad. Sci.*, **110**, 18070–18075 (2013)

[21] R. Felix, P. A. Rauschnabel, C. Hinsch, Elements of strategic social media marketing: A holistic framework. *J. Business Res.* online 1st (2016)

[22] W. Wood, T. Hayes, Social Influence on consumer decisions: Motives, modes, and consequences. *J. Consumer Psych.* **22**, 324–328 (2012).

[23] Merchant Category Codes and Groups Directory. *American Express @ Work Reporting Reference* (http://tinyurl.com/hne9ct5) (2008) (date of access: 2/3/2016).

[24] P. Martineau, Social classes and spending behavior. *Journal of Marketing* 121–130 (1958).

[25] D.F. Brown, Social class and Status. In Mey, Jacob *Concise Encyclopedia of Pragmatics. Elsevier* p. 953 (2009).

[26] P. Saunders, Social Class and Stratification. (*Routledge*, 1990).

[27] V. Pareto, Manual of Political Economy. *Reprint (New English Trans) edition* (1971).

[28] V. Blondel, et. al., Fast unfolding of communities in large networks. *J. Stat.l Mech: theory and experiment* P10008 (2008).

[29] C. M. Bishop, Neural Networks for Pattern Recognition. (*Oxford University Press*, Oxford, England) (1995).

[30] R. Tibshirani, G. Walther, T. Hastie, Estimating the number of clusters in a data set via the gap statistic. *J. Roy. Stat. Soc. B* **63**, 411-423 (2001).