Analysis of Households' Financial Diaries by Using Network Clustering

By Vahid Khatami

Introduction

A high percentage of population, specifically in developing and underdeveloped countries, are categorized as unbanked people, who do not have access to any formal financial account. Lack of an identification document, unstable income, collateral, and geographical restrictions are among the barriers which keep them away from access to financial tools such as loans, saving accounts, or insurance. On the other side, banks heavily rely on credit scores of their clients to find their associated risk. But low-income people have a lack of certified records showing their credit history and so are unable to access the banking system despite their higher demand for finance.

Uncertain sources of income and vulnerability to unpredicted disasters in case of low-income people result in using a wide variety of informal financial tools. Member-owned financial intermediaries such as financial cooperatives, self-help groups & rotating savings and credit association (ROSCAs) are among the major informal financial providers. Although their services are usually more flexible than formal sector, they provide less protected lending options. Low quality of monitoring and regulatory policies also heighten the probability of defaults in such markets.

Innovative solutions such as facilitating information flow to lessen financial risk may improve the functionality of financial markets and increase financial inclusion in poor neighborhoods. Since many context related variables are involved in identifying financial status of low-income households, more costumer-based approaches are necessary to address the problem. Economic activities, uncontrolled externalities, and family members' demands are highly context specific that motivate financial institutes to deliver more customized products.

This study is focused on providing a few solutions that can improve the quality of financial services for poor households mostly relying on agricultural activities. Based on FAO's (Food and Agriculture Organization of the United Nations) 2014 report, "there are an estimated 450 million smallholder farming households (representing 2 billion people) relying to various degrees on agricultural production for their livelihoods. They represent the largest client segment by livelihood of those living on less than $2 a day." Households constitutes a significant portion of agricultural activities in rural areas. For example, "85 percent of households in Tanzania are considered agricultural, meaning that they cultivated land, reared livestock, or managed fisheries (Derksen-Schrock et al. 2012)".

Research Questions

Based on above-mentioned comments, this study tries to address financial inclusion by categorizing smallholder households based on demographic features and financial behaviors. The following questions are among the related concerns:

- How many distinguished clusters can be identified through target households?
- To what extent can households' demographic features determine their financial pattern?
- What are the bundled item categories in households' portfolio?
- What are the common financial tools used by different clusters?
- What are the opportunities to customize financial products based on demographic features?

Literature Review

There are a few research studies about social network aspects of microfinance products. The common input information in many of these researches are survey data collected over specific time intervals. In some cases, studies use lab experiments in the field to identify particular patterns within networks. Evaluating network indicators such as centrality, diffusion, and mobilization are the common outcomes of such network studies. For example, Benerjee et al. (2013) in "the diffusion of microfinance" developed a model of information diffusion based on the census of households in 75 villages of India. They have tried to determine how individuals get knowledge about new products by focusing on the centrality of first-informed individuals. Another study on the same dataset by Davidson et al. (2017) has focused on self-help groups and the consequences of women's participation. They concluded that engagement of rural females in self-help groups enhances their social capital and expands their social network beyond the groups.

Many enterprises have recently involved in network analysis of low-income population. Their primary motivation is to address the asymmetric information problem in the financial market caused by a lack of credit history. They take advantage of network analysis to propose non-traditional credit scoring methods. Microfinance institutes can rely on network measurements to find the potential risk attached to each household or individual in poor regions. Furthermore, the progress of digital finance and the entrance of telecommunication companies to the financial market, have provided new opportunities based on high volume of transaction data.[1] For example, clients with highly volatile expenditures or low variety of financial resources in their portfolio are expected to impose higher risks to financial providers.

[1] https://cfi-blog.org/tag/credit-scoring/

Looking at the social network connections based on mobile phone metadata is another noble approach to estimate the credit scores of end users. Reconstructing social networks based on demographic features allows financial providers to cluster clients in terms of risk exposure. The proposed estimation method in this study rely on network clustering based on similarities in households' portfolio and demographic variables. [2]

Data

CGAP (Consultative Group to Assist the Poor) has published a survey dataset conducted between June 2014 and June 2015. This dataset, which is used as the input information in my research, captures the financial and in-kind transactions of about 270 households in impoverished northern Mozambique, the fertile farmlands of western Tanzania, and Punjab province, the breadbasket of Pakistan. CGAP retained the services of Bankable Frontier Associates (BFA) to manage the project. For in-country data collection, BFA worked with International Capital Corporation in Mozambique, Digital Divide Data in Tanzania, and RCons in Pakistan.[3]

The Diaries methodology combines in-depth quantitative and qualitative research. Research teams met participating families about every two weeks to collect granular data on cash flows in and out of the household, their financial tools, assets, major life events, and attitudes toward agriculture and financial services. The size of target respondents are shown in the following table.

Country	Number of sites (villages)	Number of households in the final sample
Mozambique	3	93
Tanzania	2	86
Pakistan	2	94

Although the dataset shows a close picture of a households' inflow and outflows, there are some limitations that make it difficult to develop general claims. The selection of households was purposeful rather than random, without focus on representativeness. Also, the concentration of sample households in two or three neighboring villages does not reflect the

[2] https://www.devex.com/news/how-alternative-credit-scoring-is-transforming-lending-in-the-developing-world-88487

[3] http://www.cgap.org/data/data-financial-diaries-smallholder-families#microdata

average picture in each country. Furthermore, the respondents may purposefully hide certain information from interviewers, or other family members. They might also forget their detailed transactions at the time of interview.

Methodology

I use clustering algorithm as the main method to identify major partitions. The partitions are the combination of households' demographic features and their item categories/economic activities, which are revealed through financial diaries. The most homogenous partitions represent those households, who have similar habits in their economic and financial status. Similar households are expected to have some common demographic features that I will explore as independent variables. Determining the bounded activities or financial tools is another outcome of the clustering method in this study, which helps to predict household's portfolio based on partial information.

Initial Observations

Looking at the visualized raw data enables some hypothesizing about clustering applications given this dataset. There are some common patterns in financial diaries when controlling demographic features. More details are explained bellow.

Seasonality:

Seasonality effects are captured in the following graph (Figure 1), which shows similar trend in Mozambique and Pakistan. Looking at item categories in terms of economy sector (Figure 2) reveals the reasons for seasonal trends in those countries. In Pakistan, a large portion of inflows to household diaries is specified in agricultural activities, which have seasonal patterns. Inflows from employment has the same importance in Mozambique, which seem to represent the positions related to or dependent on agricultural activities. As the financial sector has comprises the highest percentage in household diaries in Tanzania, there is no obvious seasonal pattern in wholesome diaries.

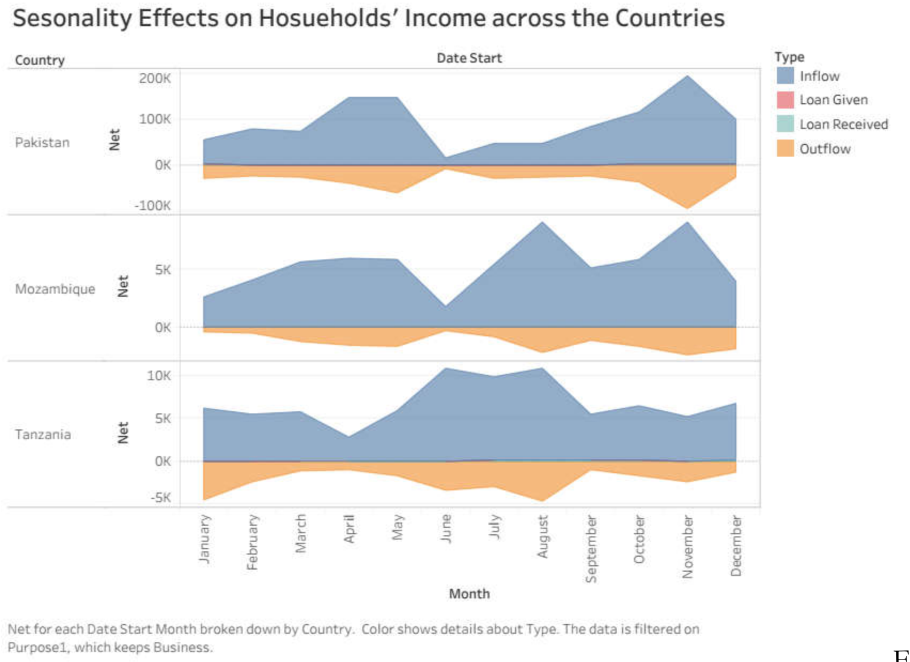


Figure 1

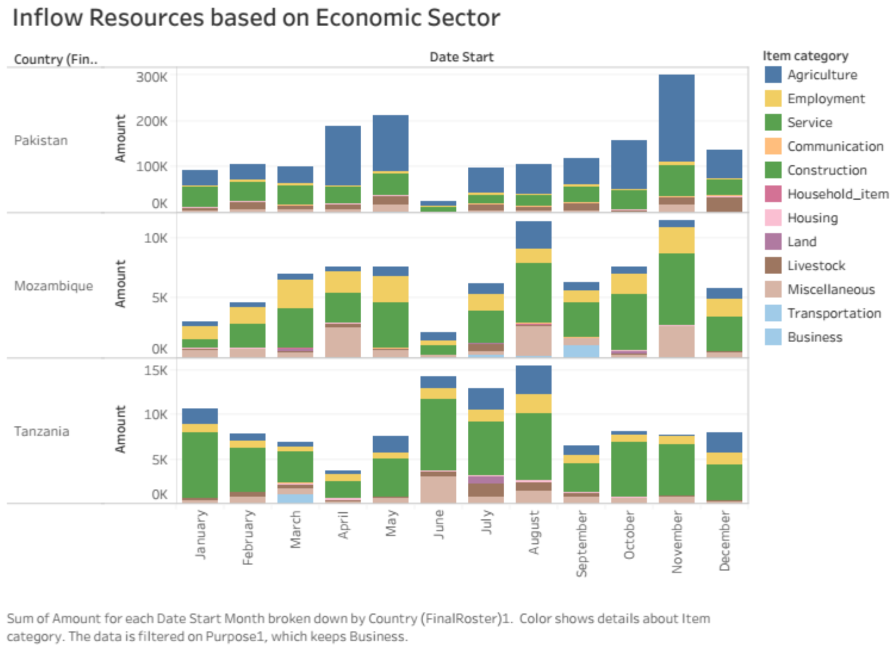
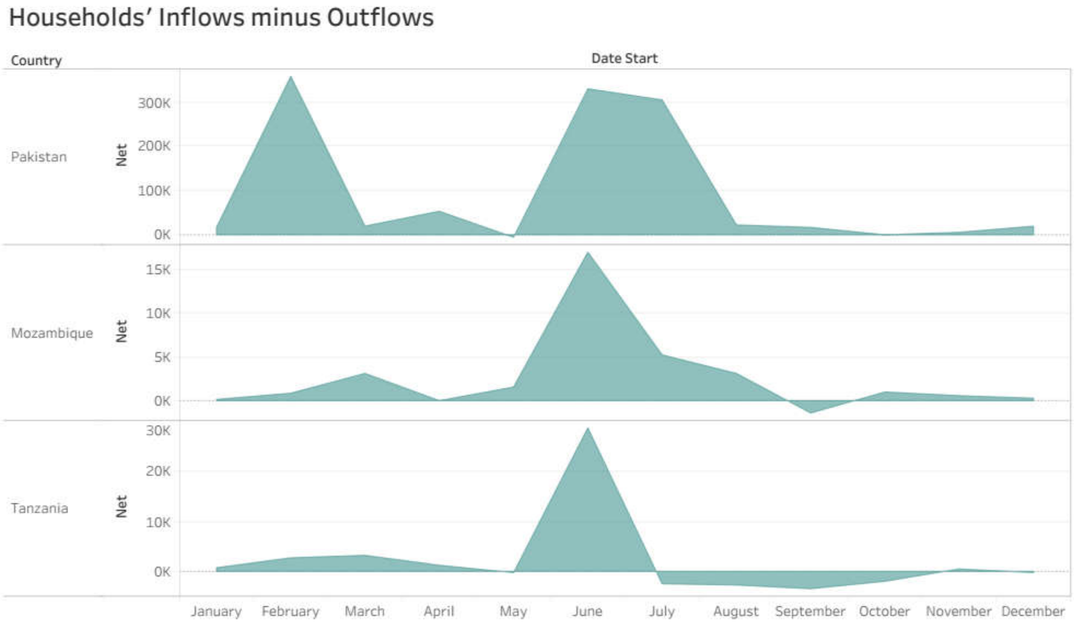


Figure 2

Inflows minus Outflows:

The difference between inflows/loans received and outflows/loans given for the families is shown in the following graph across different months and three countries (Figure 3). Although the magnitudes are not the same across countries, similar patterns are observed. An extra surplus in the beginning of summer and close to zero values for the winter periods are the main observed features.

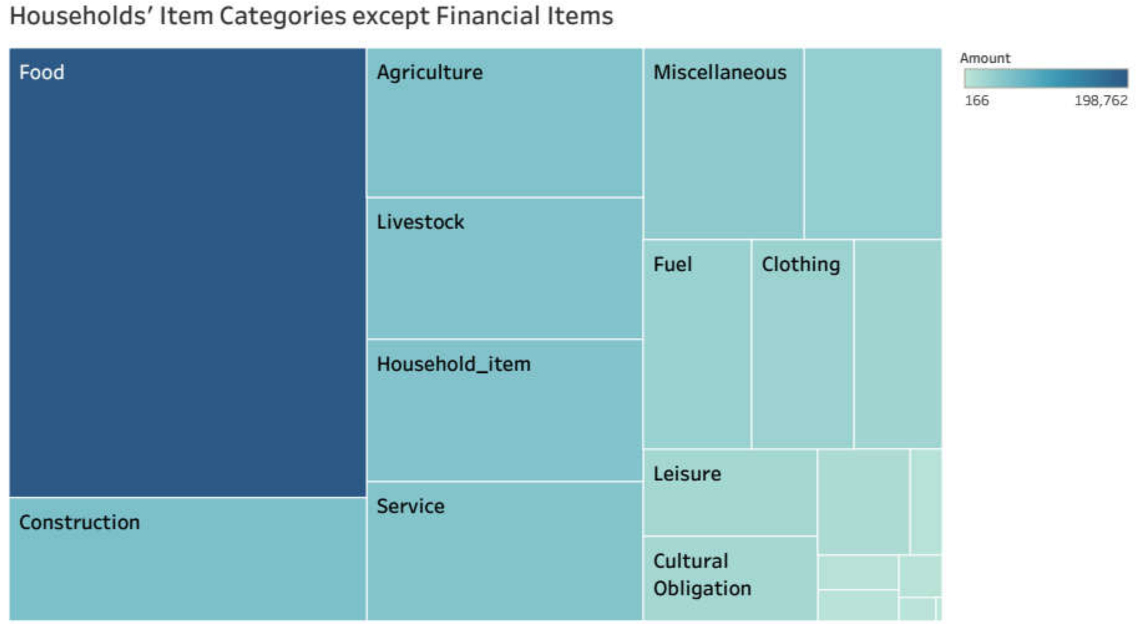

Households' Inflows minus Outflows

Net for each Date Start Month broken down by Country. The data is filtered on Type and Purpose1. The Type filter keeps Inflow, Loan Given, Loan Received, NA and Outflow. The Purpose1 filter keeps Business and Household.

Figure 3

Item Categories in Home-Based Diaries:

Financial item categories create the biggest portion of home-based transactions. The graph below (Figure 4) shows other major items of home-based transactions after excluding the financial ones.

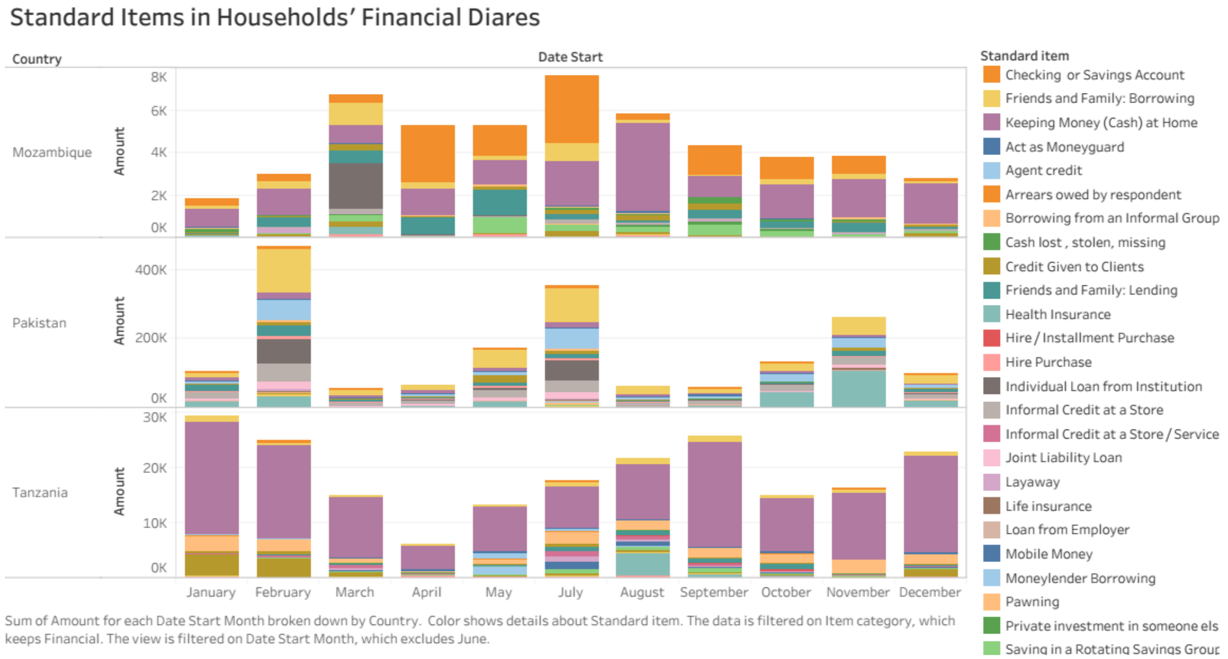

Households' Item Categories except Financial Items

Item category. Color shows sum of Amount. Size shows sum of Amount. The marks are labeled by Item category. The data is filtered on Purpose, which keeps H. The view is filtered on Item category, which has multiple members selected.

Figure 4

Financial Tools:

Friends and families are the major financial sources for Pakistani households in comparison to the Tanzanian, who are more willing to keep their cash at home, and Mozambican, who also rely on their checking and saving accounts as one of the main financial sources.

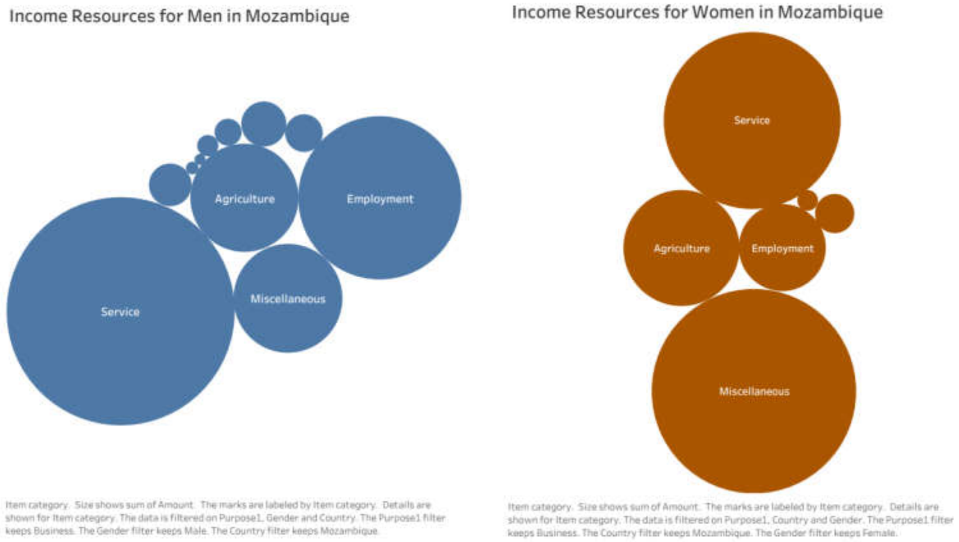


Sum of Amount for each Date Start Month broken down by Country. Color shows details about Standard item. The data is filtered on Item category, which keeps Financial. The view is filtered on Date Start Month, which excludes June.

Figure 5

Economic Activities Based on Gender:

Men and women in the studied villages have similar involvement in economic sectors except in Mozambique, which is shown in the following graph. (Figure 6)



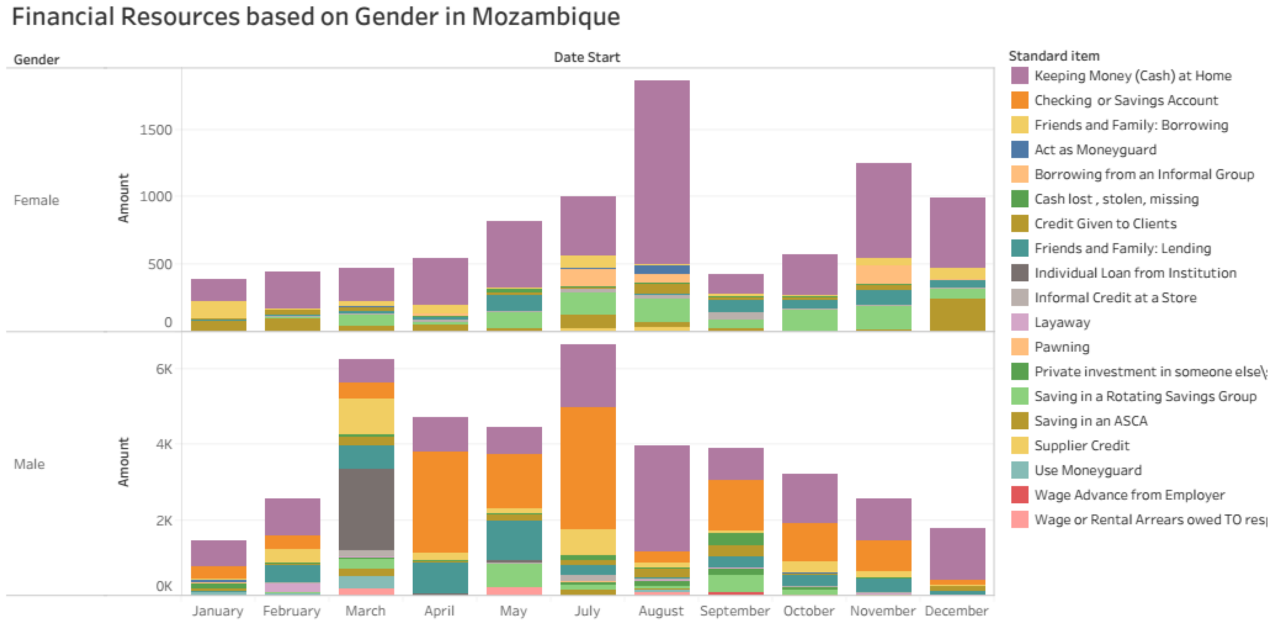
Item category. Size shows sum of Amount. The marks are labeled by Item category. Details are shown for Item category. The data is filtered on Purpose1, Gender and Country. The Purpose1 filter keeps Business. The Gender filter keeps Male. The Country filter keeps Mozambique.

Item category. Size shows sum of Amount. The marks are labeled by Item category. Details are shown for Item category. The data is filtered on Purpose1, Country and Gender. The Purpose1 filter keeps Business. The Country filter keeps Mozambique. The Gender filter keeps Female.

Figure 6

Financial Tools Based on Gender:

Women in Mozambique also have different patterns in their financial diaries, as shown in the graph below (Figure 7). They use less checking or savings accounts and keep more portion of their money at home relative to men.

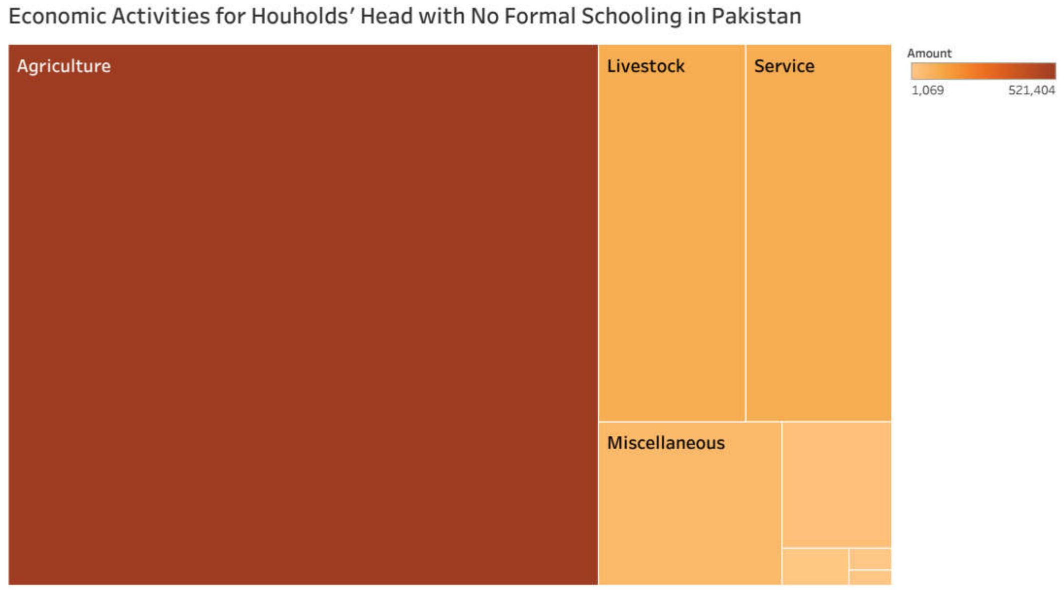


Sum of Amount for each Date Start Month broken down by Gender. Color shows details about Standard item. The data is filtered on Item category and Country. The Item category filter keeps Financial. The Country filter keeps Mozambique. The view is filtered on Date Start Month, which excludes June.

Figure 7

Education and Economics Activities:

The education level of a household's head seems to effect the type of economic activities across countries. For example, in Pakistan, heads of households with secondary education have a higher percentage of involvement in the service sector than others.

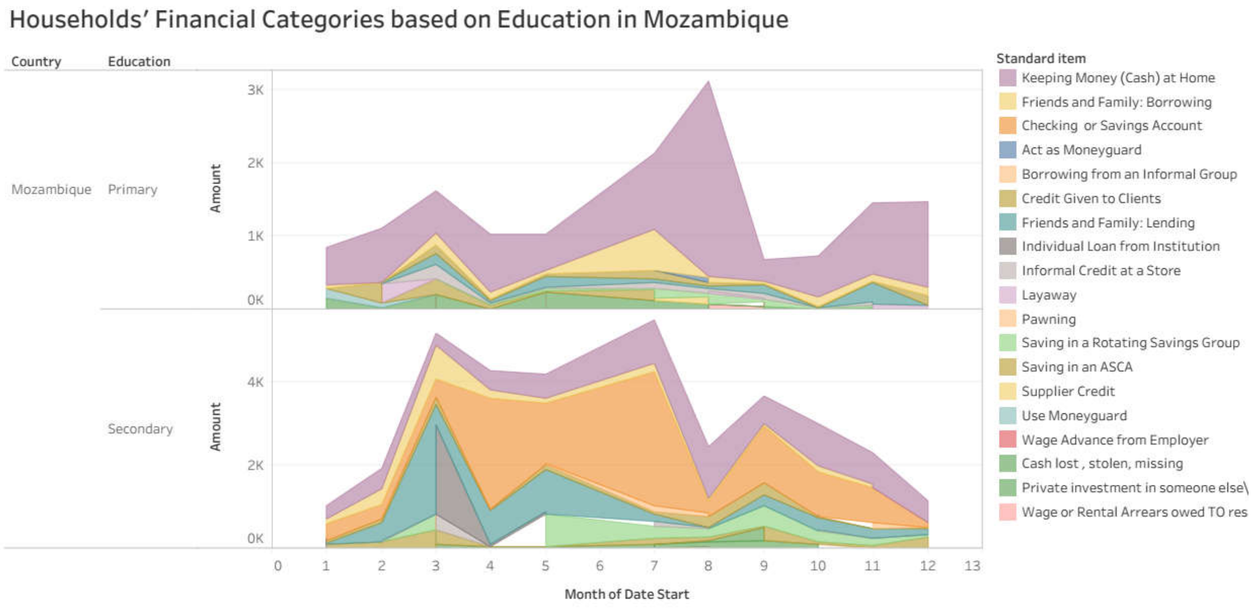


Item category. Color shows sum of Amount. Size shows sum of Amount. The marks are labeled by Item category. The data is filtered on Education, Purpose1 and Country. The Education filter keeps No formal schooling. The Purpose1 filter keeps Business. The Country filter keeps Pakistan.

Figure 8

Education and Financial Diaries:

Major differences in financial transactions based on education level is observed in Mozambique. It seems that head of households with secondary education have more access to checking or serving accounts than others with primary education, who keep more portion of their money at home.



The plot of sum of Amount for Date Start Month broken down by Country and Education. Color shows details about Standard item. The data is filtered on Item category and Date Start Month. The Item category filter keeps Financial. The Date Start Month filter excludes June. The view is filtered on Education and Country. The Education filter keeps Primary and Secondary. The Country filter keeps Mozambique.

Figure 9

Network Graph

Using network visualization provides a better understanding of common financial patterns among households. To conduct this analysis, I have built two datasets of links and nodes by merging transaction data and households' demographics (with households' ID as the mutual column). Three sample graphs have been selected for this analysis. They include the home-purposed item categories, business-purposed categories, and financial standard items through household's portfolios.

For each of the mentioned graphs, I have set up two different datasets indicating the nodes and links. The nodes' dataset includes the ID of each household and the name of corresponding transaction category. The links' dataset indicates the connection between each specific household and transaction category if the household spends more than a specific percentage of her total transaction over the specific category. The weight of each link indicates the total amount of transactions. For the next step, I use forced-directed graph drawing algorithm through "networkD3" package in R for visualization. This algorithm tries to relocate the nodes and edges by assigning forces among them in order to have the least number of cross edges as possible.[4]

As it can be seen in the following graphs (Figure 10-12), dark and light blue nodes represent households and transaction categories respectively. Not all households are included in the graph since some did not possess enough transactions on the filtered categories. Some of the households, which are usually centered in the middle, have a higher portfolio in terms of variety because of their numerous connections with the light blue nodes. The same observation for the transaction categories are also valid. Some of them have more attached links because of their popularity on the households' portfolio.

[4] https://cran.r-project.org/web/packages/networkD3/networkD3.pdf

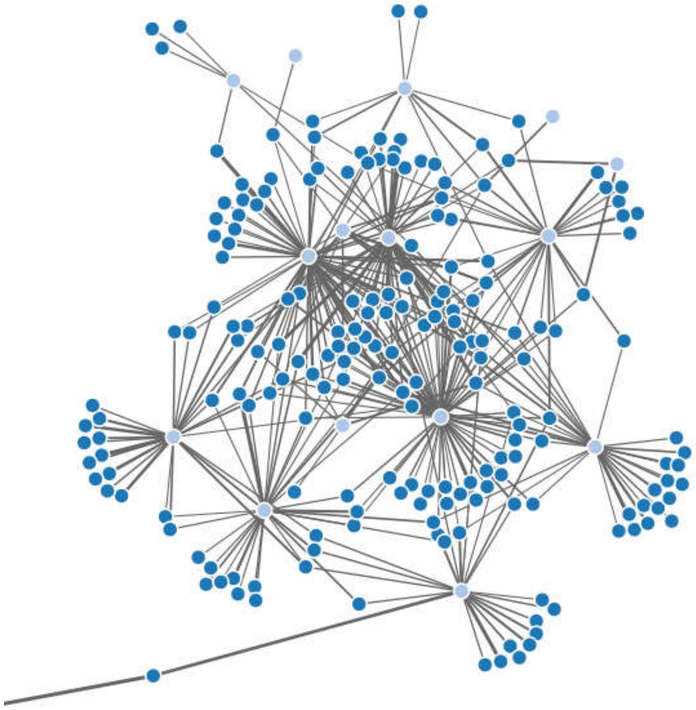


Figure 10 – The network graph of households (dark blues), and home-purposed item categories (light blues)

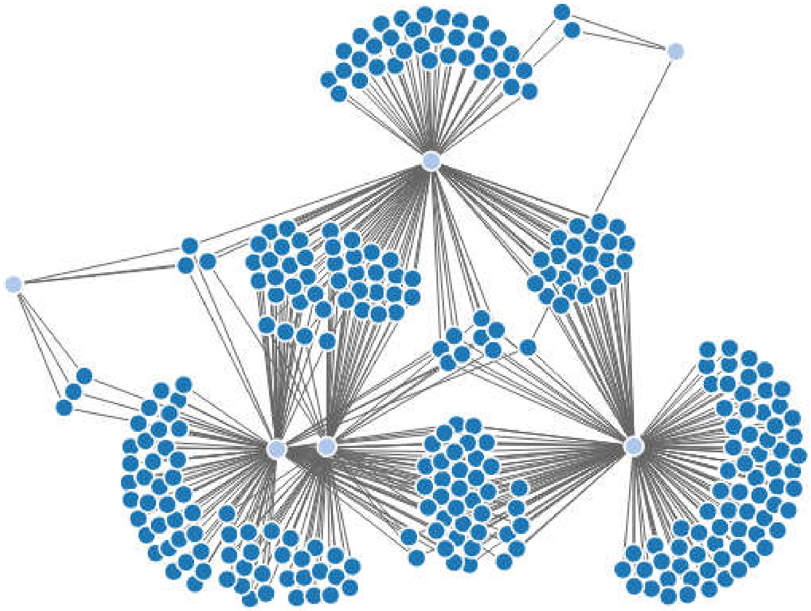


Figure 11– The network graph of households (dark blues), and business-purposed item categories (light blues)

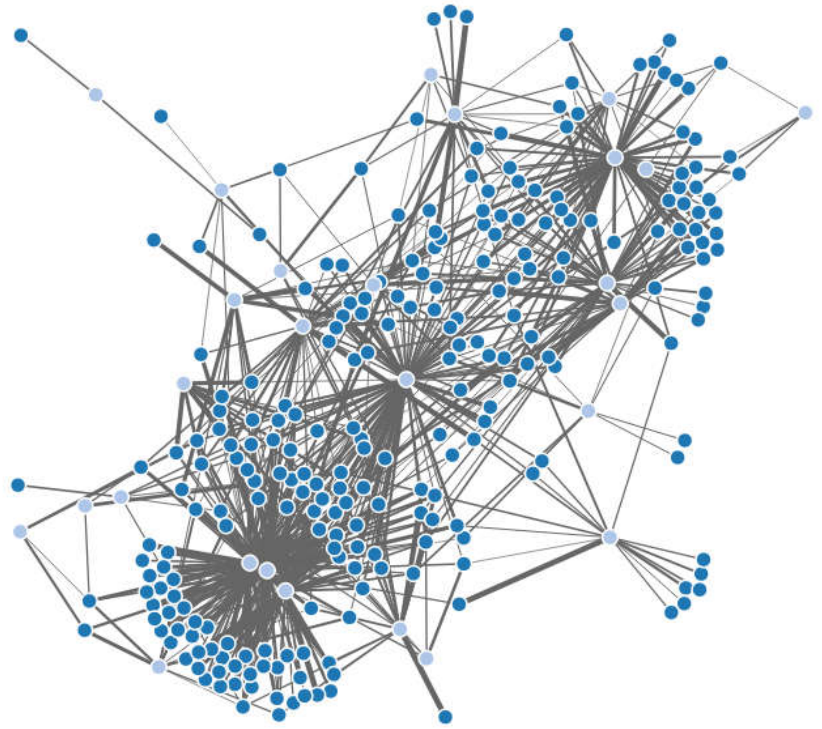


Figure 12– The network graph of households (dark blues), and financial standard items (light blues)

Two-Mode Blockmodeling

Since both initial observations and graph visualizations show the possibility of using clustering applications, I have developed a more in-depth analysis by using a two-mode block modeling algorithm. This method is one of the analytical tools in some of the social science applications, where there are two important distinct sets of objects. In many of these cases, the data are available in the form of a two-mode binary matrix, where elements of 1 indicate the presence of a bond (or tie) between the row and column object, and elements of 0 indicate the lack of a bond. (SAGE, Ch 29.) Examples of two-mode binary social network data include the attendance of women at various social events (Davis, Gardner, & Gardner, 1941), affiliation of executives with various clubs/boards (Faust, 1997), the participation of civic organizations in various community projects (Brusco & Steinley, 2006), and the voting patterns of Supreme Court justices on a set of cases (Doreian, Batagelj, & Ferlogoj, 2004). (Brusco, 2013)

One of the possible ways to think about the two-dimensional matrix in this case is to consider each row as a specific household and each column as a transaction category in the households' portfolio. I evaluate the matrix elements by setting a threshold as the minimum percentage of households' overall amount of transaction value over the specific category. If they spend more than the threshold, the elements' value will be 1 and 0 otherwise. With these assumptions, the goal of clustering optimization is to obtain a partition that includes the blocks with highest number of 1 values. (Brusco, 2013)

Within the context of two-mode block modeling, I have used K-means clustering method developed by Doreian et al. (2004) as "a relocation heuristic method that produces blockmodels that are locally optimal with respect to all possible transfers of row and column objects from their current cluster to one of the other clusters, as well as all possible exchanges of cluster memberships for pairs of objects not in the same cluster." (SAGE, Ch. 29)

The above-mentioned analysis has been applied in three different cases similar to those in the network graph section. For each case, the resulted binary matrix after clustering optimization is shown. Furthermore, the demographic summary of one of the bold clusters in each matrix has been calculated. The comparison of these summaries with the overall summary of households presents a few findings on the common features of specific clusters. Overall demographic features of target households can be summarized in the following table:

Age		Education		Marital Status		Gender		Country	
Min	17.0					Female	33%	Mozambique	32%
1st Qu	33.0	Beyond Secondary	0.3%	Divorced	8%				
Median	40.00	Missing	0.3%	Married	57%			Pakistan	34%
Mean	41.95	No formal schooling	30%	Missing	26%	Male	67%		
3rd Qu	49.00	Primary	55%	Single	4%			Tanzania	34%
Max	76.00	Secondary	14%	Widowed	4%				

As shown in the table, there are relatively equal numbers of households in each country. However the number of male participants is twice that of the female participants. More than half of respondents are married, and around the same percentage have primary education as the highest education level. The age of respondents covers a big interval between 17 and 76 years old, with an average of 42.

Case study 1:

In the following graph (Figure 13), you can find the result of block modeling on the matrix households versus home-purposed item categories. Two relatively large categories of food and financial are excluded from the horizontal items because of their presence in almost all the households' portfolios, which makes them as the less predictive variables.

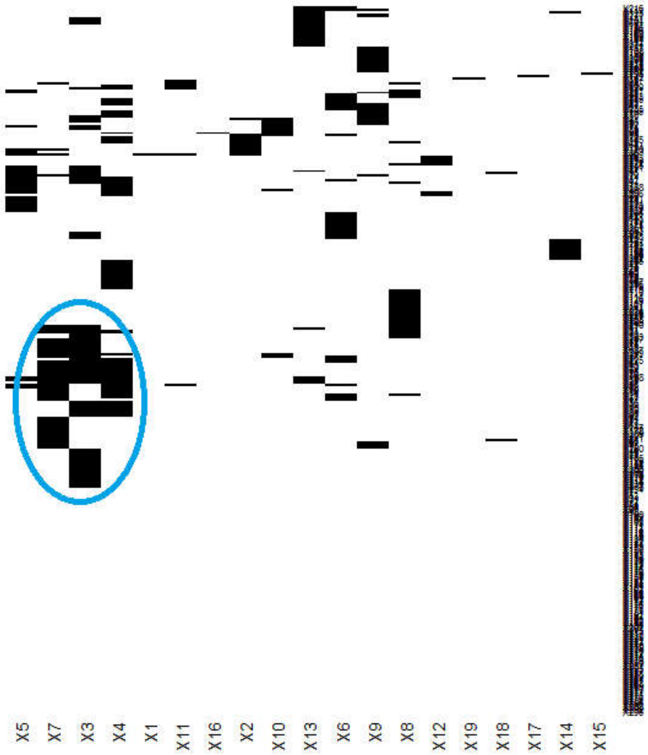


Figure 13– The block modeling of households (vertical), versus home-purposed item categories (horizontal)

Corresponding categories for the horizontal axis are as follows:

```
[1] Health          Fuel           Household_item     Clothing
[5] Leisure         Service        Transportation     Miscellaneous
[9] Construction      Employment        Communication        Education
[13] Livestock        Agriculture        Cultural Obligation   Funeral
[17] Housing          Land             Holiday or Celebration
```

By separating the households in the circled cluster, we can also study their demographic features as the following table:

```
    Age                Education            Marital Status    Gender          Country
Min.   :17.0                              : 0            : 0   Female:29  Mozambique:42
1st Qu.:34.0  Beyond Secondary            : 0  Divorced: 8     Male  :40  Pakistan  : 9
Median :40.0   Missing                    : 0  Married :19                Tanzania  :18
Mean   :41.8  No formal schooling         :10   Missing :34
3rd Qu.:49.0   Primary                    :41   Single   : 3
Max.   :74.0  Secondary                   :18   Widowed : 5
```

By comparing the statistical results of the mentioned circle with overall summary, we can see that they include more educated, less married, and more female participants than average. More households from Mozambique and less from Pakistan than the average distribution are included in the cluster.

Case Study 2:

In the following graph (Figure 14), you can find the result of block modeling on the matrix households versus business-purposed categories.

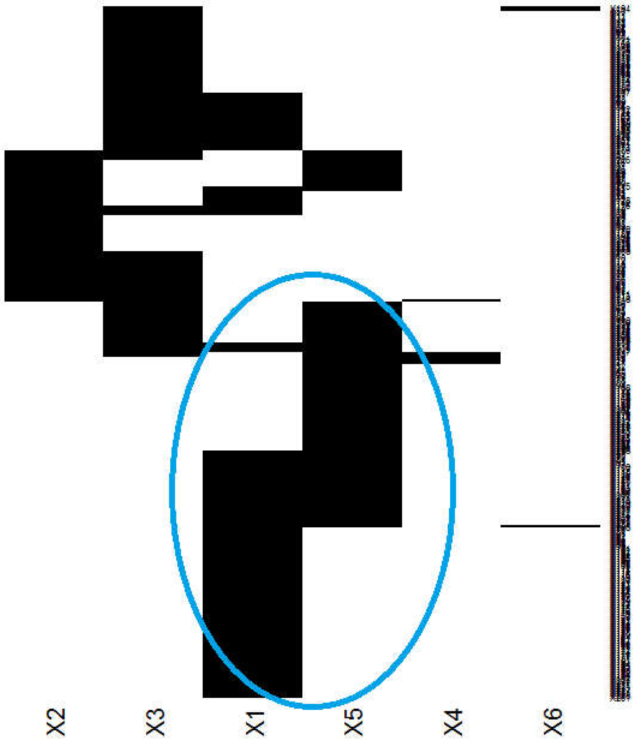


Figure 14– The block modeling of households (vertical), versus business-purposed item categories (horizontal)

Corresponding categories for the horizontal axis are as follows:

[1] Agriculture [2] Miscellaneous [3] Employment [4] Business [5] Service [6] Housing

By focusing on the households in the circled cluster, we can study their demographic features as the following table:

```
  Age              Education          Marital Status        Gender  Country
Min.   :17.00                        : 1             :  1   Female: 39  Mozambique:27
1st Qu.:34.00   Beyond Secondary   : 1   Divorced    : 12   Male  :126  Pakistan  :83
Median :40.00   Missing            : 0   Married     :112               Tanzania  :55
Mean   :42.32   No formal schooling:63   Missing     : 27
3rd Qu.:49.00   Primary            :83   Single      :  8
Max.   :76.00   Secondary          :17   Widowed     :  5
```

By comparing the statistical results of the mentioned circle with overall summary, we can see that they include more non-educated, more married, and less female participants than the average. More households from Pakistan and less from Mozambique than the average distribution are included in the cluster.

Case study 3:

In the following graph (Figure 15), you can find the result of block modeling on the matrix households versus financial options. Two relatively large categories of keeping money (cash) at the home and borrowing money from friends or family are excluded from the horizontal items because of their presence in almost all the households' portfolio, which makes them as weak predictive variables.

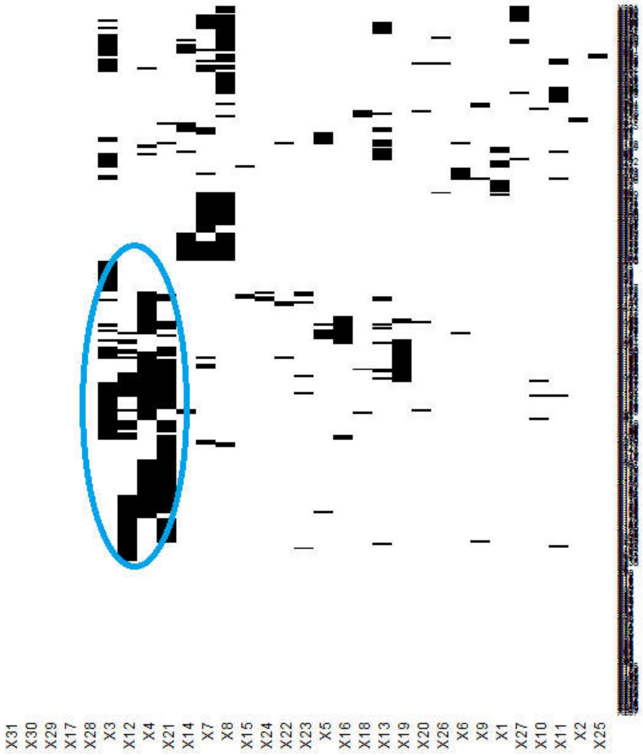


Figure 15– The block modeling of households (vertical), versus financial standard items (horizontal)

Corresponding categories for the horizontal axis are as follows:

Financial Tools:

[1] Cash lost , stolen, missing
[2] Private investment in someone else\\s business"
[3] Friends and Family: Lending
[4] Informal Credit at a Store
[5] Checking or Savings Account
[6] Layaway
[7] Saving in an ASCA
[8] Borrowing from an Informal Group
[9] Act as Moneyguard
[10] Pawning
[11] Wage or Rental Arrears owed TO respondents
[12] Use Moneyguard
[13] Credit Given to Clients
[14] Saving in a Rotating Savings Group
[15] Wage Advance from Employer
[16] Individual Loan from Institution
[17] Mortgage
[18] Supplier Credit
[19] Joint Liability Loan
[20] Moneylender Borrowing
[21] Agent credit
[22] Life insurance
[23] Hire Purchase
[24] Loan from Employer
[25] Arrears owed by respondent
[26] Mobile Money
[27] Informal Credit at a Store / Service Provider (e.g., boda boda)
[28] Health Insurance
[29] Hire / Installment Purchase
[30] Welfare Group
[31] Tafu airtime credit

By separating the households in the circled cluster, we can also study their demographic features as the following table:

```
     Age                  Education   Marital Status      Gender          Country
 Min.   :17.00                         : 1               :  2    Female: 44      Mozambique:31
 1st Qu.:33.00   Beyond Secondary   : 0   Divorced: 10    Male  :121      Pakistan  :81
 Median :40.00   Missing               : 0   Married  :112                    Tanzania  :53
 Mean   :41.69   No formal schooling:62   Missing : 26
 3rd Qu.:50.00   Primary               :84   Single    : 9
 Max.   :76.00   Secondary             :18   Widowed : 6
```

By comparing the statistical results of the mentioned circle with overall summary, we can see that they include slightly more non-educated, more married, and less female participants than the average. More households from Pakistan and less from Mozambique than the average distribution are included in the cluster. Demographic features of this group are pretty similar to the last case study, which shows a close connection between the financial options used by this type of households and their business-purposed transactions.

Conclusion

Addressing asymmetric information in case of relation between unbacked populations and financial market is a focal point to expand financial inclusion around the world. Network analysis brings new opportunities to find non-traditional credit scoring methods to address this problem. Because of the similarity between households' transaction diaries, clustering algorithms can be proposed as a solving algorithm. Reconstructing network graphs based on item categories in households' portfolio creates a tool to identify different kinds of clusters among households. Findings reveal that there is an opportunity to predict item categories and customize financial tools for each cluster based on partial information. More in-depth studies supported by high volume detailed financial diaries can quantify the level of risk associated with each household as an analytical tool for microfinance organizations.

References

- Banerjee, Abhijit, et al. "The diffusion of microfinance." *Science* 341.6144 (2013): 1236498.
- Brusco, Michael, and Douglas Steinley. "Inducing a blockmodel structure of two-mode binary data using seriation procedures." *Journal of Mathematical Psychology* 50.5 (2006): 468-477.
- Brusco, Michael, et al. "An exact algorithm for blockmodeling of two-mode network data." *The Journal of Mathematical Sociology* 37.2 (2013): 61-84.
- Davidson, Thomas, and Paromita Sanyal. "Associational Participation and Network Expansion: Microcredit Self-Help Groups and Poor Women's Social Ties in Rural India." *Social Forces*: 1-30.
- Davis, Allison, et al. *Deep South: A Sociological Anthropological Study of Caste and Class*. University of Chicago Press, 1941.
- Derksen-Schrock, K., A. Pennington, K. Stahley, A. Chew, R. Natali, M. K. Gugerty, and C. L. Anderson. "Tanzania National Panel Survey. LSMS—ISA: Highlights." Evans School Policy Analysis and Research (EPAR). EPAR Brief No. 184. Seattle: Evans School of Public Affairs, University of Washington.
- Doreian, Patrick, Vladimir Batagelj, and Anuška Ferligoj. "Generalized blockmodeling of two-mode network data." *Social networks* 26.1 (2004): 29-53.
- Faust, Katherine. "Centrality in affiliation networks." *Social networks* 19.2 (1997): 157-191.

- Scott, John & Carrington, Peter J. (2014). *The SAGE Handbook of Social Network Analysis*. London: SAGE.