Advice and Influence:

The Flow of Advice and the Diffusion of Innovation

Juan Carlos Barahona
Massachusetts Institute of Technology
20 Ames St, Cambridge
Massachusetts, 02139-4307
barahona@media.mit.edu

Alex (Sandy) Pentland
Massachusetts Institute of Technology
20 Ames St, Cambridge
Massachusetts, 02139-4307
pentland@media.mit.edu

ABSTRACT

Finding the influential people in a community is key to diffusion process of technological innovations, as well as other kinds of products. The ability to recognize who are the influential members of a community is important for diffusion policy makers and managers. This information is traditionally obtained through costly ethnographic studies which are not necessarily efficient. In certain endeavors the use of socioeconomic and demographic measures characteristic of those ethnographic studies is not effective, because the target population is very homogeneous. In the specific case of diffusion of advanced digital technologies in underserved communities or rural areas the challenge of economic sustainability becomes an issue and the cost of traditional methods to find who are the influential members becomes prohibitive.

We explore the use of sociometric information as a supplement to socioeconomic and demographic variables to determine the influential members of a community, under conditions where conventional methods may fail. We believe that identifying the structural characteristics of the flow of advice plays a key role in this space. We explore the theoretical possibilities of different possible graph-theoretic measures given data about networks.

An empirical study of these ideas using data on a community of Costa Rican coffee growers is reported. We collected sociometric data from 122 producers and compare our results with an independent ethnographic study of the same population. It turns out that the flow of advice captured by a generalized measure of eigenvector centrality, controlling for age and innovativeness using a logistic regression method, produced a good predictor of the influential members of the community. In terms of the positive predicting value our results suggest that we can double the precision (for this particular data set we got 91.66% vs. 45% obtained by the conventional methods).

Sociometric data is expected to become more available and easier to record and process, as mobile phones, computers of all sizes and Internet become ubiquitous and better algorithms for data mining from those devices evolve. This work is part of a larger research agenda aimed at designing methods and applications informed by the structural properties of human dynamics to improve the flow of ideas and innovations.

Keywords. Advice Networks, Measurement, Performance, Design, Economics, Influence, Innovation, Reliability, Human Dynamics, Social Networks, Diffusion, Centrality, Rural Development.

1. INTRODUCTION

Classical writers such as John Stuart Mill and Karl Marx speculated that the standard of living could not rise indefinitely unless advances in technology increased the yield of the means of production. Neoclassical growth theory, based on capital accumulation, supports this intuition [1]. For this reason, it is clear why technology is increasingly stressed as a key element to help underdeveloped communities around the world.

Digitalization and the new communication technologies are the drivers of an exponential increase in the amount of information available and the velocity at which it can be shared, all at ever lower costs and through a widening variety of media. Economic globalization and record levels of productivity are driven in part by the ability to link applications, devices and people as nodes of highly distributed networks that can interact using the common language of 1s and 0s [2]. It is no surprise that ICT (Information and Communication Technology) for Development efforts around the world are primarily focused on computers and Internet access. They usually come in the form of school computer labs, "telecenters" or information kiosks. Capacity (readiness) to use digital technologies and access dominate most of the debate on the "Digital Divide".

However, Dutta and Jain [3] suggest that readiness to use technology and actual usage do not necessarily go hand in hand. There might be a readiness threshold. A country or community might need a certain level of readiness before effective usage of ICT can be achieved. In other words, there is a delay in the expected benefits of technology as a result of a low starting point in regards to ICT, suggesting a non-linear relationship between both variables, adding complexity to the discussion on diffusion of ICT in underserved communities. If there is a threshold to be exceeded before getting any impact, speed of diffusion and sustainability become crucial, as it drains the social and financial capital of the entrepreneur or organization promoting the use of ICT.

Unfortunately, most discussions on ICT for development are either technocentric, to the detriment of the analysis of the surrounding society, or are focused on the political economy and public policy aspects of national reforms. They lack careful exploration of the nature of different technologies, how they may interact with the local culture at community level, and the impact they have in the diffusion process and the overall success of the project or investment. If there is a better understanding of the dynamics of technology and information diffusion in underserved or rural communities then, it may be possible to get quicker and

higher returns on investments made by individuals, businesses, and governments by way of better diffusion of ideas and skills.

Diffusion of innovations depends on time, communication channels, and a social structure to support it [4]. Most studies on innovation have been retrospective; they lack information on interpersonal communication networks, and more important, few have attempted to use the lessons from diffusion research to accelerate the diffusion of innovations [5]. Valente and Davis' work [6] suggests, through simulation, the possibility of achieving a critical mass in a much shorter time by carefully selecting the opinion leaders of a social network. In general, identifying who are the influential members improves the design of diffusion strategies, regardless of what is being diffused through the network. In practice, the selection of influentials is usually accomplished by using conventional wisdom and traditional sociological theory, e.g. by looking for those with higher social and economic status and leaders of formal and informal organizations within the community. Selection is usually done after the definition of general criteria to select participants or "beneficiaries", ignoring the underlying network structure. In other words, many projects by design define a profile that usually tends to make the population of interest very homogeneous (e.g. programs designed to reach the poorest of the poor, or a specific gender within an income bracket) without consideration of the social network.

It is in this particular context that we explore how sociometric measures can provide useful information to determine who are the influential actors. Specifically, we look at different structural patterns and compare them with conventional socioeconomic variables in their ability to provide useful predictions of influence. These sociometric measures are expected to be more cost-efficient and less troubling than a conventional socio-economic survey, as it is well documented how troublesome it is to collect and use income related questions [7]. This paper develops a model that uses these sociometric measures to identify the key social members through the dynamics of the flow of advice and their use of media technologies.

2. STRUCTURAL PERSPECTIVES ON DIFFUSION OF INNOVATIONS

Most empirical research on diffusion of innovations confirms the premise that new ideas and practices spread through interpersonal communications. However, most foundational studies have focused on the spread of relatively simple and "static" technologies, such as weed spray in Iowa [8], hybrid seed corn [9] or tetracycline [10], as opposed to ever evolving modern technologies and their myriad of versions and the potential difficulties and complexities intrinsic to them.

The key to transfer those simple technologies is awareness and imitation. In other words, P gets the idea through personal communication with O (awareness) and P decides to imitate O (adoption), later P passes the information to Q and so on. This approach leads to the interest in parameters such as the rate of diffusion and how it correlates with proximity, communication or influence. Valente et al. [11] studied and confirmed the association between friendship ties and the adoption of contraceptive choices in Cameroon women. Their model defined network exposure as

$$E_i = \frac{\sum \omega_{ij} y_j}{\sum \omega_i} \quad (1)$$

Where ω is the social network weight matrix and y is the vector of adoptions. The network exposure is measured on direct contacts. ω can be transformed to reflect other social influence process through a family of relational, positional and centrality measures.

Their approach implies at least four different levels of decision to design a study of the network effect on diffusion. The first one is the election of the type of network to observe and register. It could be a network of friendship, advice or any other convenient type. Second, if influence or other behavior determines P's probability of adoption, what set of structural features of networks capture such behaviors (relational, positional or centrality)? Third, within each set, which measures should be used? (There are probably more than a dozen different types of centrality measures). And once the above decisions are made, still there is an issue of fine tuning to decide the weight attached to each factor, generally based on social distance. For example, if O influence P and P influences Q. Should the influence of P and Q reflect the fact that O may or not be connected to a highly central or an isolated N?

ICT for development projects usually come in the form of computers for schools, community centers or other public or quasi-public spaces. In rural areas, probably more often, they come in the form of telecenters that embody a variety of different media that offer a wide range of potential solutions for community problems, all the way from telemedicine to e-commerce. In terms of ICT for development public policies, most discussions revolve around Internet access issues.

Those types of innovations are substantially different from the technologies mentioned above. They are knowledge intensive and for their adoption to be sustained over time there needs to be a continuous flow of information and support to keep up with the pace of new versions or even just to keep it functional. Voice over IP and wireless Internet solutions are frequently praised for their promising potential to serve isolated communities. But, updating to a newer version of hardware or software may cause operative systems to crash. In that moment, what may seem a simple operation (update a driver for instance) can become a real problem. It may come from previous experience (knowledge), advice (another villager has the knowledge and the villager has direct or indirect access to him or her) or from specialized technical assistance which depending on the type of source and the relative isolation could be scarce and expensive to acquire. In this particular setting exposure to the friendship network is probably not enough.

We claim that the relevant unit of analysis for the diffusion of advanced technologies is the community's social networks of advice, and that flow of ideas within these networks can be used to identify the influentials in order to better promote rapid diffusion of ideas. Rapid diffusion should be an objective to increase the collective knowledge base of the new technologies and surpass the critical threshold of adoption. Valente and others have shown that centrality is key to accelerate diffusion. The most influential problem solvers in the community should be the "entry" or starting points of the diffusion process as well.

2.1 Centrality and Advice

Since research on the idea of centrality applied to human communications was introduced in the late 40's by Bavelas at the Group Networks Laboratory, M.I.T., centrality has been related to reputations of power and influence over a community [12].

The most frequent form of organization of a social structure is the center-periphery pattern. It consists of a) a subgroup of relatively central prestigious actors who are connected by direct or short indirect ties and b) a subgroup of peripheral actors who are directly connected to the central actors rather than to other peripheral actors. In this form of organization, central actors tend to be resourceful and cohesively joined to other actors [13].

Within the family of centrality measures, there are four prominent ones due to their strong and distinct qualities[14]. They are also foundational in the field of social network analysis: degree, betweenness, closeness and eigenvector centralities.

Degree Centrality

The most simple and natural way of describing the concept of centrality is the star configuration. The center in this structure possesses 3 unique properties: it has the maximum degree [15-17]; it falls on the geodesics (shortest path[1] linking a given pair of points) between the largest possible number of other points and, since it is located at the minimum distance from all other points, it is maximally close to them (Freeman, 1978/79). Mathematically it is defined by equation (2).

$$C_D(p_k) = \sum_{i=1}^{n} a(p_i p_k) \tag{2}$$

Where $a(p_i p_k)=1$ if and only if p_i and p_k are connected by a line, otherwise it is 0.

Betweeness Centrality

Betweeness [12] usually indicates a node that can control the flow of information bridging disparate regions of the network. Because of its reliance on non-directed paths and geodesics, betweeness cannot be easily estimated for directed data [18].

Its assumptions are that the traffic will choose the shortest path, and if confronted with equally short paths, it will randomly choose only one. Traffic moves one to one instead of copying itself or being broadcast from a node. A second assumption is that it is not diffusing randomly. Since it is taking only the shortest path, then it "knows" its target from the origin [19] These last assumptions make Freeman's betweeness centrality measure unsuitable to be used in contexts where these assumptions do not hold, like the spread of computer viruses, diseases and other infections, or information movement in most cases. The characteristics of our latent variable and of the observed advice network fall outside of these assumptions and therefore we did not evaluate this particular measure.

Closeness centrality

Closeness is the theoretic distance of a given node to all other nodes and it is commonly used in the study of diffusion. As opposed to degree centrality, this measure takes into account indirect connections. In a directed graph the outgoing arcs will be related to the amount of steps one actor needed to reach the other actors. In terms of flow it is ordinarily interpreted as an index of the expected time until arrival of something flowing within the network [20].

The critical assumption of this measure is that information is following the shortest path or parallel duplication –where all paths are followed simultaneously, including the shortest path as well. It only works on connected or strongly connected graphs.

In our study, the networks of advice found and registered are not well connected. This limitation impedes the use of this measure in the current analysis.

Eigenvector centrality

It is the property of a node that has a high eigenvector score and that is connected to others who are also high scorers. This is measured by the principal eigenvector of the adjacency matrix of a network. It was designed to work with valued data but works on binary information as well. The use of eigenvector centrality is convenient when the status of an actor is a function of the status of those with who he is in contact (Bonacich, 1972). Given an adjacency matrix A, the eigenvector centrality of node i is:

$$c_i = \alpha \sum A_{ij} c_j \tag{3}$$

Where α is a required parameter to give the equations a non-trivial solution ($\alpha = 1/\lambda$, i.e. the reciprocal of the eigenvalue) and has no substantive interpretation.

It is usually interpreted as a measure of influence. It assumes that traffic moves via unrestricted walks and does not assume that things flowing will be transferred or copied to one neighbor at a time, so this measure is ideal for influence type processes [20].

2.2 Generalized eigenvector measure of the flow of advice

Among the measures of centrality the eigenvector method seems appropriate because one should expect that receiving advice/information from someone who is more central should add more to one's centrality than being advised by an isolated member of the community. Since our advise data is asymmetric we used "power" which is a generalized eigenvector measure of centrality, also known as Bonacich Power Centrality or Alpha Centrality [21]. It is represented by the following equation:

$$c_i(\alpha, \beta) = \sum_{j=1}^{n} A_{ij}(\alpha + \beta c_j) \tag{4}$$

The value of α is used to Normalize the measure and has no substantive interpretation. We use UCINET [22] to estimate Bonacich Power Centrality and in their solution the normalization

[1] A path is defined as a sequence of adjacent nodes in which no node is visited more than once

parameter is automatically selected so that the sum of squares of the node centralities is the size of the network [23]. The parameter β is an attenuation factor which gives the amount of dependence of each node's centrality on the centralities of the nodes it is adjacent to. It can be interpreted as the degree to which an individual's status is a function of the statuses of those to whom he is connected. Beta is an adjustable weight that can take positive and negative values, depending on the specific phenomena under analysis. There are cases like bargain where the advantage comes from being connected to less powerful individuals. In communication networks Beta should be positive, as one benefit from the information available to one's alters. Bonacich [24] suggests that in a communication network, a low positive value of Beta would be appropriate if most communications were local and not transmitted beyond the dyad. Since the nature of personal advice implies information on specific personal concerns, it seems reasonable to expect that most interaction happens at dyad level. But who advices the advisor should also be relevant. Thus, in the case of technical/business as well as personal advice it seems reasonable to choose the maximum value for β (note that if α and $\beta = 0$ then equation 4 is equal to equation 2). It was done using $\beta =$ very close to the absolute value of the reciprocal of the largest eigenvalue of both the adjacency matrices.

An important property of this measure is that allows for negative values of β. In the case of advice it may have valuable applications since it is possible to think of negative values, for example, some actors may have conflict with a source of advice, potentially affecting the flow of information and ideas in the system. Therefore, even in the presence of small β, conceptually eigenvector centrality is the adequate point centrality measure for advice networks. This parallels the "graph-theoretic concept of "vulnerability". This is not to be developed in this paper. Cook et al. [25] among others have developed relevant work on the effect of negative edges in communication networks.

3. Empirical Analysis

We want to fit a model to predict which people are influential based on: conventional economic and demographic attributes, graphic-theoretical characteristics of the individuals and a measure that captures their use of advanced media.

We want to estimate the following logistic model to predict who is influential:

$$\log \frac{\Pr(y=1)}{(1-\Pr(y=1))} = \beta_0 + \beta_1 \mathrm{X} + \beta_2 C + \beta_3 M + \varepsilon \qquad (5)$$

Where y equals 1 if the respondent is influential, X is a set of socio-economic and/or demographic characteristics, C is a set of sociometric measures based on the eigenvector centralities and M their use of Media and ε the expected error.

The propositions and tests in the form of hypothesis are:

1. Sociometric measures are an important supplement to conventional social and economic status attainment measures in predicting who is influential.
2. Patterns of Advice received and given is a good predictor of who are the influential members of a community.
3. If Hypothesis 2 is true, there must be an important correlation with the early adoption of tools that are used to support and enhance communication, which leads to Hypothesis number 4.
4. If the use of media technology can be use as a predictor of influence, then a propensity to be an early adopter is correlated with patterns of advice and the use of media technology.

3.1 DATA

Sampling Region

We explored these ideas using data collected in 2003 from a community of coffee growers in the southern mountains of Costa Rica called Santa Maria de Dota. The community has roughly 4300 inhabitants; coffee production and exports represent about 80% of their income.

Santa Maria de Dota is a well-established and integrated rural community. An interesting characteristic of the region is the structure of land ownership, mostly very small producers with 1 or 2 acres, with not much land available to grow their crops. In being so small, coordination and diffusion of information is key to production, processing, and commercialization of their coffee beans. Entrepreneurship is in high demand, as they cannot divide their land among their offspring, thus forcing them to generate their own jobs or to migrate.

The homogeneous social and economic characteristics of this population are expected to produce a relatively small effect from the social and demographic characteristics. Most producers are organized in a local cooperative called COOPEDOTA. It is collectively owned by the coffee growers registered as members of the cooperative. COOPEDOTA receives their coffee cherries and processes them into coffee beans. Whole beans and ground coffee is commercialized by the cooperative on behalf of the producers.

Baseline

In order to establish a "ground truth" or baseline it is necessary to establish who are the influential members that the model is expected to capture in a more effective and efficient way.

During the summer of 2003, a team of two senior Costa Rican researchers, trained in social sciences, former professors at the university of Costa Rica and currently members of an NGO ("CEMEDCO"), volunteered to conduct an Ethnographic Diagnostic in Dota. They were familiar with the general ideas of the social networks approach but not with its methods. Their goal

was to identify key members in the community. Key members were understood to be people that influence the community's decisions and whose opinions and decisions have the potential to affect the socioeconomic development of the community as a whole. After 6 visits to the community and dozens of interviews they reported 53 influential members, among them 32 were coffee producers registered as members in the local cooperative.

The list with the 53 names was discussed for validation with a group of "community experts", identified by CEMEDCO's researchers based on the knowledge about the community they gained. The expert's validation reduced the list to 30 members. They were expected to be the most influential members. Among those 30 influential people, 19 were registered producers.

The group of 30 was invited and attended to a workshop sponsored by INCAE (an international business school in Latin America and research facility), where they completed a sociometric survey. A roster with their names was presented to them, and they were asked to provide information on friendship, advice and influence. This produced dyadic data. We used Freeman's in-degree centrality as a scale of influence. Only those that were considered influential by their peers were considered the "truly" influential people or baseline. Only 19 had an in-degree measure bigger than zero and among them 16 were registered producers. Table 1 summarizes the three different exercises that lead to the baseline estimation we described.

TABLE 1. The three exercises used to construct the baseline for this study.

	Community Members	*Subset of producers (among the Community Members)*
Ethnographic Diagnostic	53	32
Community Experts Validation	30	19
Sociometric Survey	19	16

Data Collection

All active producers have to personally approach the mill office to collect either a check or an equivalent form of payment for their processed crop. Usually, most of them arrive during the first three to five days. The producers were interviewed as they approached the mill during the peak four days. By the end of the fourth day 84.72% of all payments had being collected according to the administration files. Their arrival seems to follow an apparent log-normal distribution. One hundred and twenty three surveys were collected through a short interview (see Figure 1).

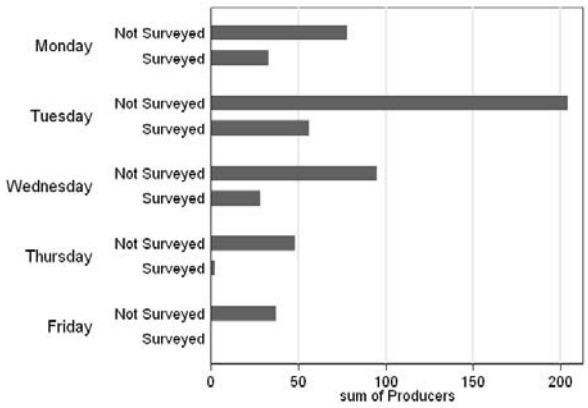


Figure 1. Distribution of Respondents as they approached the Coffee Mill to collect their payments during the week of data collection.

There were 123 respondents. (One was dropped because he was the son of a producer and his main occupation was not his family farm.) The respondents were asked to mention the names of the people who provide them with advice, using the *free recall* method (no roster of names were showed), they also were asked from the names they provided which were cooperative members. The names were validated later against with the cooperative membership records. There was no limit to the number of names that could be recorded. In social network surveys it is especially problematic because each missing answer becomes a gap in the social network under study [26]. Advice is not a troublesome dimension of social relations, and in general, subjects where not uncomfortable answering questions about it. The interview was done in the premises of the cooperative, which probably legitimate the willingness of the cooperative management to support the study. Also, an interviewer fill out the questionnaire and the interview was kept as short as possible. All of these factors contributed to an unexpectedly high response rate for the overall questionnaire and a 100% response rate for the questions related to advice. We know that a sample is often not representative of a network because the structure of a random sample seldom matches the structure of the overall network. Therefore, we must be careful about generalizations about the social structure of the population, but accepting the limitations of our data set, we do believe that it is large enough to capture the main patterns of the flow of advice.

Graph-theoretic Data Sets

There are three generic social boundary specification strategies [27] : formal membership criteria based on node's attributes; an event based approach and a relational approach based on social connectedness. In this paper we are using each of these methods to set the boundaries of three possible data sets. The overall criterion to select the interviewees was membership to the coffee cooperative. Those that actually had a chance to participate in the study were selected upon the event that they show up during the week of data collection. The open question (with no roster) on who you look for advice… generated names of producers as well as names of other members in the community. The total list of names presented the possibility to define two different data sets based on a relational criterion mixed with an attribute criterion:

a) Those mentioned (connected) by the interviewees, who are registered producers and had been interviewed (n=122).

b) Those mentioned (connected) by the interviewees who are registered producers (n=169)

c) Those mentioned (connected) by the interviewees either registered producers or not (n=298).

In each case two n x n matrices were created. All of them are a one-mode matrix $\mathbf{A}$ where the (i,j) entry in the matrix is denoted by X_{ij} and represents the value of the tie from actor i to j. In this case this is a dichotomous relation where

$X_{ij} = 1$ if $i \rightarrow j$, $X_{ij} = 0$ otherwise.

We chose to use a) as the data set to work with. We can treat it like a whole network, since all the respondents sending nominations will have an equivalent likelihood of being nominated by his or her peers[2].

Attribute Data

Since recollection of sociometric data using a paper survey places a burden in the respondent and the interviewer, as much as possible attribute data has been collected from different secondary public or semi-public sources. The master database has 1296 records corresponding to community members. In the case of the producers non-sensitive data was provided by the cooperative, other sources as phone books and qualified informants have being used as well. Some data collection on attributes was done specifically for this study other than the surveys as is the case of neighborhood status that we will describe later.

3.2 Explanatory Variables

Control Variables: Demographic and socioeconomic individual characteristics.

Lipset, cited by Blau and Duncan ([15-17]) says that "position in the social structure is usually associated with a certain level of income, education, family structure, community reputation and so forth". This paper tried to follow as much as possible Lipset's intuition to construct an equation that predicts a person's influence using the socio economic and demographic variables used in most theories about influence.

Income. The INCAE survey did not ask for income, nor did the cooperative had this data available. However, we had access to the amount of coffee beans they brought in 2003 to the cooperative to be processed. Since coffee is the main source of income for the vast majority it should be a good proxy for income and the records of coffee processed were reliable since no other company nearby was offering a better price than them, not to mention a legal obligation of exclusivity and a natural restriction associate with costs of transportation from farm to mill.

[2] Dataset are available at www.media.mit.edu/~barahona/datasets

TABLE 2 Description of selected variables in their original dimensions, some usual transformations and interaction variables that were tested

Variable	Mean	Standard Deviation	Min	Max	Partial Correlation	p-value
Influential	0.13	0.34	0	1		
Age in years	48.80	13.50	23	87	-0.0408	0.6820
Squared Age	2560.01	1463.30	529	7569	0.0536	0.5910
Mature (range 35-70)	0.80	0.40	0	1	0.2037	0.0390
Education in Years	8.70	4.30	3	18	0.1553	0.1170
Respondent has Secondary Education	0.14	0.35	0	1	-0.0080	0.9360
Gender (Male)	0.80	0.40	0	1	-0.0912	0.3590
Perceived socio-economic status of Neighborhood	3.31	0.63	2	4	-0.0520	0.6020
Volume of Coffee Crop	3112.30	3020.60	83.5	12681.3	-0.0897	0.3680
Log of the Volume of Coffee Crop	7.50	1.10	4.4	9.4	0.0840	0.3990
Freeman Indegree for the personal advise network	0.16	0.39	0	2	0.0921	0.3550
Freeman Indegree for the economic advise network	0.74	7.70	0	85	-0.0377	0.7050
Freeman Outdegree for the personal Advice Network	0.16	0.39	0	2	-0.0528	0.5960
Freeman Outdegree for the economic Advice Network	0.74	0.49	0	2	-0.1054	0.2890
Advice Centrality Index	0.23	0.54	0	2	0.2911	0.0030
Innovation (is early adopter of e-mail, fax and mobile)	0.36	0.63	0	3	0.2213	0.0250
Respondent has 3 or more channels of Communication	0.08	0.28	0	1	-0.0669	0.5020
Interaction of Mature x Years of Education	6.93	5.19	0	18	-0.1676	0.0910
Interaction of Mature x secondary education	0.10	0.30	0	1	0.0789	0.4280
Interaction of Mature x ACI	0.20	0.51	0	2	0.0642	0.5200
Interaction of ACI x Innov/comm	0.28	1.04	0	6	0.0118	0.9060

Since we had access to the exact home address of every producer, we created a supplementary "social status" variable based on the local perception of the social status of the producer's neighborhood. A list of all neighborhoods was produced and provided to a young local health professional, a local taxi driver, and to a business man who is in the construction business. They were asked independently to assign a value from 1 to 5, according to their perception, of the socio-economic status of each neighborhood. When there was no consensus, two votes decided the assigned status. There was no case in which all three answers were different.

Age, education and gender were provided by the cooperative. A dichotomous variable call "Mature" was created to capture this age range, from 35 to 70 years old, reflecting what a producer described as "the age when you and society know who you really are". Gender has no significant correlation with being influential which is not a surprise in this community[3]. Education data is consistent with this observation. When comparing the level of education of all male and female producers they share the same average amount of years ($\overline{x}$ =8.7 years, $p = 0.0332$).

Graph-theoretic variables

The correlation of power-centrality with the response variable is higher for the personal advice network than the one corresponding to the technical/business advice network. This difference across domains may suggest that the influential's advice is most sought after in interpersonal issues. This is consistent with the results of a study conducted by the Allensbach Institute on a German national sample (n=3843) reported by Weimann [28]. They found that in the financial and political domain the influentials had clear dominance, but compared with these and 16 other domains in their study, the influentials advice is most sought after in "dealing with others" and "recreation".

A paired correlation of the power centrality measures also shows this relationship (pair wise correlation=.43, p<.001). This is strong evidence that there is a correlation between both matrices of advice. When structural autocorrelation is present, Krackhart [29] recommends the use of Quadratic Assignment Procedure (QAP) to test the independence of the coefficients, since OLS can become severely biased under this condition. This is because the assumption of zero covariance between any two errors [30] is not met. Each person in a dyad will contribute to (N-1) dyads, and hence there is a high likelihood that the error that characterizes one dyad involving ego is similar to the error characterizing another dyad involving ego, or that the errors are "auto correlated". QAP attempts to solve this problem [31]. In this procedure the relation matrices are permuted to examine whether the results are artifacts of the structure of the network rather than genuine relations among the actors. A hypothesis

[3] Costa Rica is known in Latin America as a pioneer in women rights and as an international advocate of women and children rights (e.g. the country adopted the Law of Responsible Fatherhood, which gave women the legal right to name and receive support from fathers who did not recognize their children when born out of wedlock, leaving the alleged father to bear with a legal process and the use of genetic evidence to proof her wrong in court). People in Los Santos are familiar with women as members or president of the board of businesses and civic organizations, driving a 4x4 taxi or running a mechanical workshop.

test using QAP effectively suggests the existence of a correlation between both advice matrices (Pearson Correlation= 0.062, p=0.005).

To avoid the problem of confounded variables we constructed a new variable ACI (Advice Centrality Index) to reflect the combined effect of both the personal and business/technical advice domains. We first dichotomized each power centrality variable using a 2.5 cut-off after inspecting the data (see Figures 2 and 3). Then the new variable resulted from summing up the "power advisors" of each network. Therefore the values for the new variable are 0 for non advisors, 1 for those who are power advisors in one of the networks and 2 for those powerful advisors in both networks.

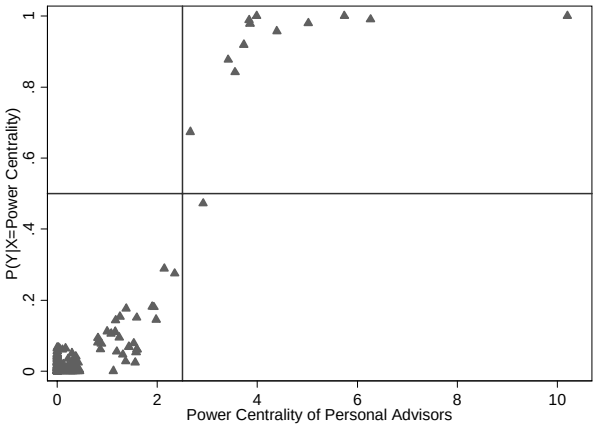


Figure 2. Probability of being an influential producer conditioned on Power Centrality of Personal Advice

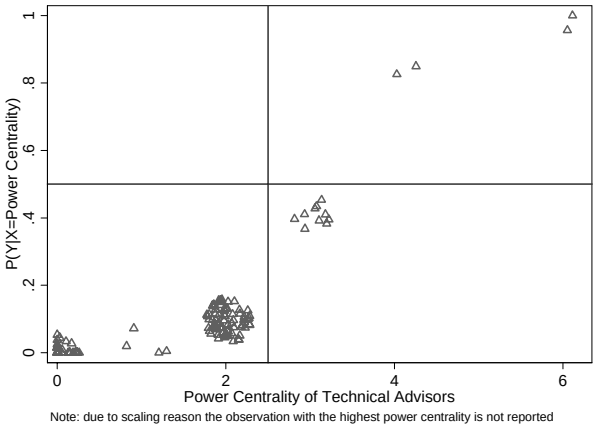


Figure 3 Probability of being an influential producer conditioned on Power Centrality of business/economical Advice

Since media technology plays an important role in the flow of ideas, availability and use of communication tools should also play a role in the community member's capacity to influence. Research shows that the strength of ties between nodes is associated with multiple relationships and the use of more media to communicate [32]. In the process, communicators will reach a common understanding of the media and work together to a joint communications solution [33]. So, we explored the use of communication technologies in the community.

Most of the producers have access to land phone, fax, mobile phone and email. As one might expect, having access to the latter is more difficult do to infrastructure limitations. So we created a simple ordinal variable called channel that adds up the number of channels a subject employs. By observing the data (mode 1, average 1.54, s. d. 0.85) we chose having 3 or more as the cutoff value (91.8% had 2 or less) to create a new binary variable to distinguish those having an exceptional number of communication channels.

One fourth of the respondents had a computer at home but only 5.7% of all respondents used e-mail, and the correlation between having a computer and using e-mail was rather weak ($\chi^2 = 3.95$, ρ=0.047). Thus, independently of having a computer at home or not, it seems fair to expect that the few using e-mail are early adopters. The second and third least popular channels were faxes and mobile phones (8% and 22%). To capture the propensity to be early adopters and the use of multiple channels for communication we used the presence of e-mail, fax and mobile phone as a proxy for the pattern of adoption of new communication channels. We called the variable Innovativeness.

4. RESULTS

All variables, transformations, and interactions presented in Table 1 were divided into three subsets: socio-economic conventional, sociometric, and Media/Innovation. Stepwise regression was used only for the subset of socioeconomic and demographic variables. Stepwise regression [34] was used first to discriminate among the subset of socio economic and demographic variables and then to compare our variables for centrality and for innovativeness, and to screen possible interaction effects among the variables. No significant interaction effects were found. We run the hierarchical stepwise regression using ρ=0.25 in the forward steps and ρ=0.10 in the backward steps.

For this particular data set (n=122), we found only the variable "Mature" (being within the age range 35 to 70 years) being significant among the socio economic and demographic variables. This should come to no surprise, remember that this is a particularly homogeneous group of people. The Alpha Centrality Index was used as our sociometric measure, as discussed above.

We then tested "Mature", Innovativeness and Alpha Centrality Index against the null hypothesis of being simultaneously zero. We conducted a Wald test after running a logistic regression against the binary response variable (isInfluential). We obtained

strong evidence to reject the hypothesis that the effects of these variables are simultaneously equal to zero (($\chi^2 = 11.30$, df =3, ρ=0.0102). Table 2 describes the equation of the logit regression model. The second, third and fourth columns present the results of running a logistic regression independently for each variable against the response variable. Column 4 is the full model.

TABLE 2 Logistic Regression Results for the components and the final model. (n=122)

Variable	X (Age)	M (Innov.)	C *(Alpha)*	$X + M + C$ Full Model
LR chi2 (a)	2.61	29.65	48.07	58.40
D. of Freedom	1	1	1	3
Prob > chi2	0.1061	0.0000	0.0000	0.0000
Pseudo $-R^2$ (b)	0.0275	0.3127	0.5070	0.6160
Log likelihood	-46.0992	-32.5814	-23.3687	-18.2056

(a) The likelihood-ratio chi-square is defined as 2(L1 - L0), where L0 represents the log likelihood for the "constant-only" model and L1 is the log likelihood for the full model with constant and predictors.

(b) Technically, R2 cannot be computed the same way in logistic regression as it is in OLS regression. The pseudo-R2, in logistic regression, is defined as (1 - L1)/L0, where L0 represents the log likelihood for the "constant-only" model and L1 is the log likelihood for the full model with constant and predictors. This statistic will equal zero if all coefficients are zero. It will come close to 1 if the model is very good

Table 3 presents two nested models and the full model. Model 1 stands for the sociometric and demographic variables, in this case age, which was not significant by itself. Model 2 combines Innovativeness with Mature and was significant at 1%. The full model adds the centrality measure. For the combined model the strongest association is for the sociometric variable, and the weakest is age.

TABLE 3 Odds Ratios and p-values of the Main Effects Model

	Model (1)	Model (2)	Model (3)
Mature	4.157	5.008	14.916
	(0.179)	(0.146)	(0.072)*
Innovation in Comm. Channels		13.534	10.101
		(0.000)***	(0.014)**
Alpha Centrality Index			35.586
			(0.000)***
Observations	122	122	122
Pseudo R-squared	0.028	0.343	0.616
Log Lik Intercept Only	-47.405		
Log Lik Full Mod	-46.099	-31.121	-18.206
Likelihood Ratio LR		32.567	58.398

p values in parentheses

* Significant at 10%; ** significant at 5%; *** significant at 1%

Table 4 describes the estimated unstandardized coefficients for the full model.

TABLE 4 Estimated Coefficients, Standard Errors, z-Scores, Two-Tailed p-Values and 95% Confidence Intervals for the Final Logistic Regression Model (n=122)

	Coeff.	*Std.Err.*	*Z*	*P>|z|*	*95% Conf. Interval*	
Mature	2.7024	1.5030	1.80	0.072	-.2434	5.6484
Innov.	2.3127	.9435	2.45	0.014	.4634	4.1619
ACI	3.5720	.8932	4.00	0.000	1.8213	5.3227
_cons	-7.1570	1.8903	-3.79	0.000	-10.8619	-3.4522

The following histogram illustrates show how many "influential producers" are predicted by each component of the model. The histograms of Figures 3 and 4 show that even though a majority of early adopters are not influential, amongst the influential the majority are innovators.

Among influentials and non-influentials, producers of mature age are the dominant ones, but the ratio of mature that respects those outside of the 40-70 range is much higher for the influentials.

Figure 3. Distribution of Influential and non influential Producers According to Age (n=122).

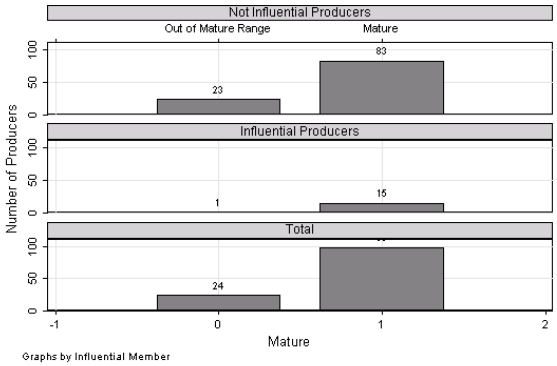
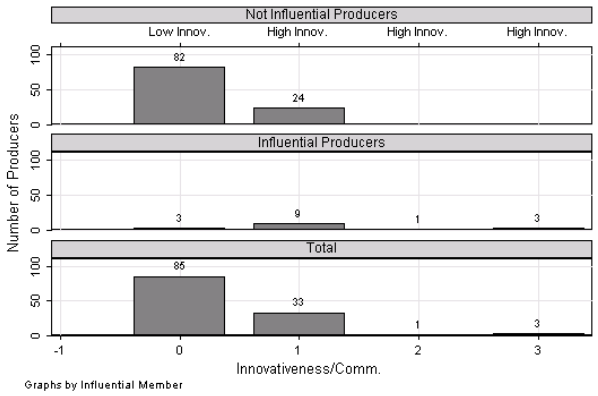


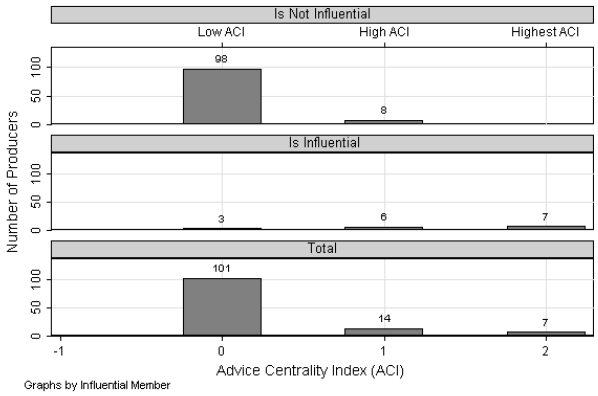
Figure 5 shows the power of the advice centrality index to predict influentials. 100% of producers with ACI of 2 are influentials, 45% of those with an ACI of 1 are influentials, and only 4% of those with an ACI of 0 are influential.

Figure 4. Distritution of Producers according to Innovativeness between influentials and non-influentials (n=122)



If all variables are held equal to zero, the probability of being influential is close to zero ($Pr(y|x) = 0.0015$) and someone meeting the three criteria has a probability of 0.9216 being influential. This can be appreciated best in a graphical form. Figure 5 combines the three effects and presents the predicted values for the sample. Note how the expected "S-shape" curve shapes.

Figure 5. Distribution of Producers according to Advice Centrality coefficient clearly discriminating influentials from non-influentials (n=122).



In the next figure age is represented by the size of the marker and a diamond shape means that the subject is an innovator in the terms defined in this paper. Although in the left hand side of the Figure there are mature aged people and innovators, they as a group are dominant within those with high alpha centrality.

5. DISCUSSION

This paper suggests that patterns of advice captured by sociometric measures are a powerful predictor of influence. The model is effective for classification of who the influential producers, according with the success and failures in the result from the model. In terms of accuracy (total correctly classified divided by total population), our classifier was 95.08% accurate and the ethnographic study has 85.42% accuracy. But accuracy is not the right metric, since it implies that all errors are equal [35]. We argue that in this context there are much higher costs associated with type I errors (false positive) than with type II (false negative).

In the context were it is desirable to tell apart who belongs to the group of influential members and who do not, with the purpose of working with them to foster an optimized diffusion process, both errors have very different consequences. For example, imagine someone gathers 11 influential members of the same community and none of the non influential members is present. They will recognize each other as influential and they will easily recognize what other influential people should be there, in case they are missing. It is so because core people tend to have a dense collection of relationships among themselves [36]. This structure has being recognized and documented in community influence systems [37]. Thus, missing a few will tend to be autocorrected by the knowledge and well established relationships of the core group. Now imagine the scenario were they are together, but share the room with other people that are not influential. It may be confusing to recognize what the group is about for them. The rules of engagement will be somehow different about the members of the two different groups and the effectiveness will suffer rising the organizational cost. To correct this, then they or someone else would have to ask the "false influential members" to leave, which would imply a social and emotional cost. To use a measure that is adequate to compare the conventional and our methodology in these terms, let us introduce the corresponding confusion matrices.

TABLE 6. Confusion Matrices

		Predicted by Model	
		Negative	Positive
Actual	Negative	105	1
	Positive	5	11

		Predicted by Conventional Methods	
		Negative	Positive
Actual	Negative	*87*	0
	Positive	19	16

While the conventional way of classifying the influential is extremely efficient with zero type II errors, it produces a false positive rate (type I) equal to 17.92%. These values for our model are 31.25% and 0.94% respectively. It is an important difference that is blurred by the accuracy measure. Instead, we should use the proportion of the predicted positive cases that were correct. This ratio is called in the machine learning literature the precision of the classifier, also known as the positive predicting value. In these terms our results suggest that we can get a 91.66% precision as opposed to 45% estimated for the ethnographic study.

Being an "established" member of the community and being an innovator plays a significant but much less important role. The findings are consistent with our intuition: influence follows the

flow of advice and information. The ability to capture the dynamics of diffusion of ideas has the potential to have a very positive impact in the way ideas are promoted and especially in the way that technology is deployed in underserved communities, by making interventions more effective and efficient by nurturing the flow of advice.

There are several different reasons to consider these results useful and worth of more testing. From an empirical point of view, it shows that sociometric information could have a significant role in helping identify influential members of a community, especially under conditions where the population of interest is highly homogeneous. Many settlements, housing projects, or communities are very homogeneous in their attribute values, giving more importance to relational sociometric measures.

The "advice centrality index" also has advantages in terms of efficiency. It is well known that traditional socioeconomic surveys have serious problems. Many people don't like to answer income or social status related questions. As a result data quality is poor and large survey samples required. However, this research suggests that a light and neutral question like "Who do you look for when you need technical or business information" or "who do you look to for personal advice", can provide enough information to recognize the influential members of the group, those who are key for the diffusion of ideas and innovations. It is important to note that satisfactory results were obtained working with a partial data network.

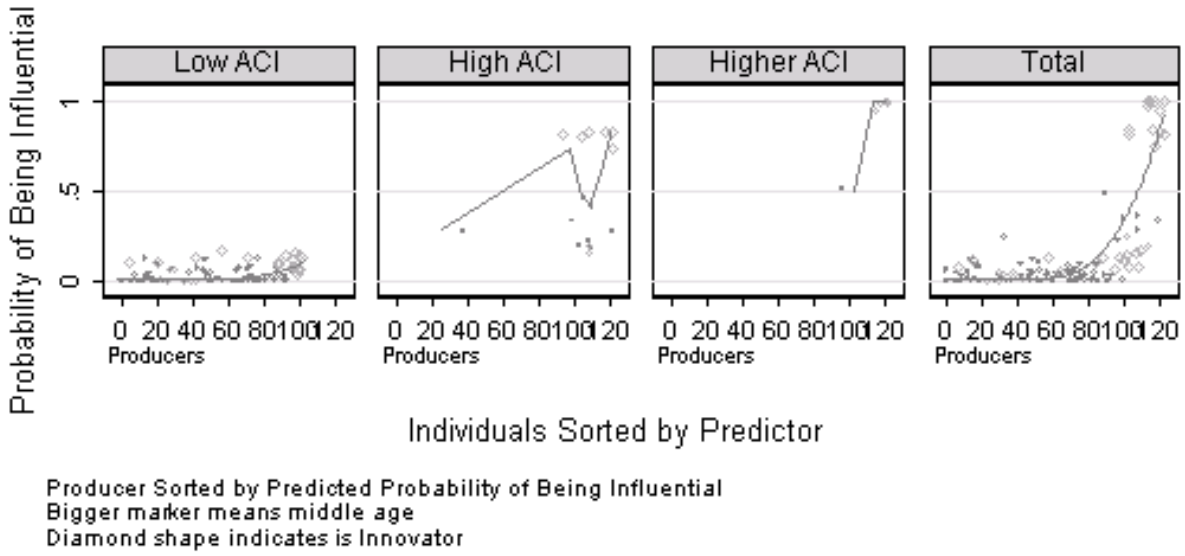
Improved precision through the use of our proposed sociometric method can have a major effect, particularly with costly interventions. For example, the diffusion of technological innovations with a high learning curve, where almost personal support and follow up is needed for long periods of time, is difficult and expensive, but crucial to pass certain threshold. It can also be effectively used as the first step to develop cognitive social structure studies [38].

6. Future Direction

Sociologists and more recently economists have devoted considerable attention to the impact of social structure and networks on the economy [39]. However these have been few attempts to translate this work into practical field methods. This work is one of the first of its kind.

There are still theoretical and empiric problems to solve before practical use of the abundant information about social networks can be used by communities. We foresee a role for machine-learning tools that can be used to develop stochastic models and methods to reconstruct whole networks out of partial and incomplete information. A future direction for this research is to test the model under conditions where the boundaries of the network are more diffuse and replication of the study with a different sample or different population will be sought.

Figure 6 Associated probability of being an influential producers according to the combination of Advice Centrality Index, Age and Pattern of Adoption of New Communication Channels.



ACKNOWLEDGMENTS

We thank INCAE for facilitating the data collection process, COOPEDOTA and COOPESANTOS for sharing data and providing enthusiastic support to our research. Rodrigo Jimenez and Roberto Mata provided valuable insights on the community dynamics. CEMEDCO provided time and effort to collect most of the reputational data through non-sociometric measures. Prof. Peter V. Marsden provided us with useful advice on an early draft of this paper.

REFERENCES

1. Grossman, G.M., *Innovation and growth in the global economy*. 2000, Cambridge, MA: MIT Press.
2. Wilson, E.J., *The information revolution and developing countries*. 2004, Cambridge, Mass.: MIT Press. xiv, 431 p.
3. Dutta, S. and A. Jain, *The Networked Readiness Index 2003-2004: Overview and Analysis Framework*, in *The Global Information Technology Report: Towards an Equitable Information Society 2003-2004*, S. Dutta, B. Lanvin, and F. Paua, Editors. 2004, Oxford University Press: New York.
4. Rogers, E.M., *Diffusion of innovations*. 5th ed. 2003, New York: Free Press. xxi, 551 p.
5. Valente, T.W., *Network models of the diffusion of innovations*. Quantitative methods in communication. 1995, Cresskill, N.J.: Hampton Press. xiii, 171 p.
6. Valente, T.W. and R.L. Davis, *Accelerating the Diffusion of Innovations Using Opinion Leaders.* The ANNALS of the American Academy of Political and Social Science, 1999. **566**(1): p. 55-67.
7. Korinek, A., J.A. Mistiaen, and M. Ravallion, *Survey Nonresponse and the Distribution of Income*. World Bank Policy Research Working Paper Series, March 2005.
8. Rogers, E., *Diffusion of Innovations*. 2003, New York.
9. Ryan, R. and N. Gross, *The Diffusion of Hybrid Seed Corn in Two Iowa Communities.* Rural Sociology, 1943. **8**(1): p. 15-24.
10. Coleman, J.S., E. Katz, and H. Menzel, *Medical Innovation: A Diffusion Study. New York: Bobbs Merrill.* 1966.
11. Valente, T.W., et al., *Social Network Associations with Contraceptive Use Among Cameroonian Women in Voluntary Associations.* Social Science and Medicine, 1997(45): p. 677-687.
12. Freeman, L.C., *Centrality in Social Networks Conceptual Clarification.* Social Networks, 1979. **1**: p. 215-239.
13. Friedkin, N.E., *Structural Bases of Interpersonal Influence in Groups: A Longitudinal Case Study.* Americal Sociological Review, 1993. **58**(December): p. 861-872.

14. Everett, M.G. and S.P. Borgati, *Extending Centrality*, in *Models and Methods in Social Network Analysis*, P.J. Carrington, J. Scott, and S. Wasserman, Editors. 2005, Cambridge University Press: Cambridge.
15. Blau, P.M., et al., *Structures of power and constraint: papers in honor of Peter M. Blau.* 1990, Cambridge [England]; New York, NY, USA: Cambridge University Press. x, 495 p.
16. Blau, P.M. and O.D. Duncan, *The American occupational structure*. 1967, New York: Wiley. xvii, 520 p.
17. Blau, P.M., O.D. Duncan, and A. Tyree, *The American occupational structure*. 1978, New York: Free Press. xvii, 520 p.
18. Wasserman, S. and K. Faust, *Social network analysis: methods and applications*. Structural analysis in the social sciences; 8. 1994, Cambridge; New York: Cambridge University Press. xxxi, 825 p.
19. Borgatti, S.P., *Centrality and Nework Flow.* Social Networks. **Accepted**.
20. Borgatti, S.P., *Centrality and Aids.* Connections, 1995. **18**(1): p. 112-114.
21. Bonacich, P. and P. Lloyd, *Eigenvector-like measures of centrality for asymmetric relations.* Social Networks, 2001(23): p. 191-201.
22. Borgatti, S.P., M.G. Everett, and L.C. Freeman, *Ucinet 6 for Windows: Software for Social Network Analysis*, H.A. Technologies, Editor. 2002.
23. Borgatti, Everett, and Freeman, *UCINET 6 for Windows*. Reference Guide, ed. I. Analytic Technologies. 2002.
24. Bonacich, P., *Power and Centrality: A Family of Measures.* The American Journal of Sociology, March, 1987. **92**(5): p. 1170-1182.
25. Cook, K.S., R.M. Emerson, and M.R. Gillmore, *The Distribution of Power in Exchange Networks: Theory and Experimental Results.* The American Journal of Sociology, 1983. **89**(2): p. 275-305.
26. De-Lange, D., F. Agneessens, and H. Waege, *Asking Social Network Questions: A Quality Assessment of Different Measures.* Metodoloski zvezki, 2004. **1**(2): p. 351-378.
27. Marsden, P.V., *Network Data and Measurement.* Annual Review of Sociology, 1990. **16**: p. 435-463.
28. Weimann, G., *The influentials: people who influence people*. 1994, Albany: State University of New York Press. xiv, 370 p.
29. Krackhardt, D., *Predicting with Networks: Nonparametric Multiple Regression Analysis of Dyadic Data.* Social Networks, 1988(10): p. 359-381.
30. Wooldridge, J.M., *Introductory Econometrics: A Modern Approach*. 3RD ED. ed. 2006, Mason: SOUTH-WESTERN.
31. Martin, J.L., *A General Permutation-Based QAP Analysis Approach for Dyadic Data.* Connections 22, 1999. **2**: p. 50-60.
32. Haythornwaite, C. *Tie Strength and the Impact of New Media*. in *Hawaii International Conference on System Sciences*. 2001. Maui, Hawaii: IEEE.
33. DeSanctis, G. and M.S. Poole, *Capturing the Complexity in Advanced Technology Use: Adaptive Structuration Theory.* Organization Science, 1994. **5**(2): p. 121-147.
34. Berk, R.A., *Regression analysis: a constructive critique*. 2004, Thousand Oaks, Calif.: Sage Publications. xix, 259 p.
35. Provost, F., T. Fawcett, and R. Kohavi. *The Case Against Accuracy Estimation for Comparing Induction Algorithms*. in *15th international conference on machine learning*. 1998.
36. Borgatti, S.P. and M.G. Everett, *Models of core/periphery structures.* Social Networks, 2000. **21**(4): p. 375.
37. Laumann, E.O. and F.U. Pappi, *Networks of collective action: a perspective on community influence systems*. Quantitative studies in social relations. 1976, New York: Academic Press. xx, 329 p.
38. Krackhardt, *Cognitive social structures.* Social networks, 1987. **9**.
39. Granovetter, M., *The Impact of Social Structure on Economic Outcomes.* Journal of Economic Perspectives, 2005. **19**: p. 33-50.

Appendix A. Correlation Matrix

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
1 Influential	1.000																				
2 Age in years	0.021	1.000																			
3 Squared Age	-0.005	0.988	1.000																		
4 Mature (range 35-70)	0.131	-0.050	-0.180	1.000																	
5 Education in Years	0.236	-0.083	-0.086	-0.043	1.000																
6 Respondent has Secondary Education	0.054	0.052	0.058	-0.099	0.215	1.000															
7 Gender (Male)	0.041	-0.029	-0.035	0.059	-0.004	-0.014	1.000														
8 Perceived socio-economic status of Neighborhood	0.001	-0.005	0.009	-0.017	0.033	-0.049	-0.203	1.000													
9 Volume of Coffee Crop	0.000	-0.088	-0.085	0.082	0.063	-0.110	0.198	-0.018	1.000												
10 Log of the Volume of Coffee Crop	0.007	-0.190	-0.177	-0.021	0.060	-0.077	0.276	0.080	0.871	1.000											
11 Freeman Indegree for the personal advise network	0.209	-0.111	-0.101	-0.056	0.018	0.013	0.072	-0.041	0.058	0.121	1.000										
12 Freeman Indegree for the economic advise network	0.238	-0.025	-0.033	0.048	0.202	-0.036	0.043	0.097	-0.075	-0.106	-0.038	1.000									
13 Freeman Outdegree for the personal Advice Network	0.085	-0.041	-0.015	-0.056	0.155	-0.168	-0.041	-0.041	0.046	0.136	0.145	-0.038	1.000								
14 Freeman Outdegree for the economic Advice Network	0.158	0.059	0.040	0.155	-0.005	0.022	0.034	-0.054	-0.049	-0.082	-0.032	0.047	0.053	1.000							
15 Advice Centrality Index	0.734	-0.054	-0.070	0.058	0.316	-0.040	0.106	-0.042	0.096	0.082	0.248	0.310	0.210	0.257	1.000						
16 Innovation (is early adopter of e-mail, fax and mobile)	0.550	-0.020	-0.023	-0.011	0.191	-0.043	0.078	0.110	-0.119	-0.079	0.093	0.389	0.126	0.200	0.578	1.000					
17 Respondent has 3 or more channels of Communication	0.415	0.009	-0.004	0.073	0.132	-0.034	0.132	-0.053	-0.049	-0.015	0.104	0.310	0.180	0.098	0.482	0.637	1.000				
18 Interaction of Mature x Years of Education	0.259	-0.039	-0.129	0.664	0.646	0.065	0.015	0.020	0.110	0.022	0.026	0.201	0.176	0.115	0.294	0.183	0.161	1.000			
19 Interaction of Mature x secondary education	0.116	-0.056	-0.082	0.164	0.176	0.821	-0.077	0.012	-0.120	-0.114	0.073	-0.028	-0.138	0.008	0.013	-0.014	0.002	0.261	1.000		
20 Interaction of Mature x ACI	0.712	0.031	-0.003	0.192	0.321	-0.016	0.085	0.014	0.104	0.067	0.168	0.337	0.209	0.273	0.914	0.602	0.533	0.401	0.035	1.000	
21 Interaction of ACI x Innov/comm	0.623	-0.021	-0.038	0.094	0.270	-0.040	0.077	0.094	-0.039	-0.047	0.049	0.512	0.191	0.208	0.766	0.779	0.642	0.303	-0.009	0.803	1.000