ENDOGENOUS NETWORK FORMATION: EXPERIMENTS AND METHODS

by

Rong Rong
A Dissertation
Submitted to the
Graduate Faculty
of
George Mason University
in Partial Fulfillment of
The Requirements for the Degree
of
Doctor of Philosophy
Economics

Committee:

_____ Director

_____

_____

_____

_____ Department Chairperson

_____ Program Director

_____ Dean, College of Humanities and Social Sciences

Date: _____ Spring Semester 2013
George Mason University
Fairfax, VA

Endogenous Network Formation: Experiments and Methods

A Dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at George Mason University

by

Rong Rong
PhD of Economics
George Mason University, 2013

Director: Daniel Houser, Professor
Department of Economics

Spring Semester 2013
George Mason University
Fairfax, VA

DEDICATION

This is dedicated to my loving parents, Fang Liu and Wenming Rong.

ACKNOWLEDGEMENTS

I would like to thank the many friends, relatives, and supporters who have made this happen. Thank my parents who always support what I pursue unconditionally. My loving boyfriend, Chris, assisted me in editing my research and making sure that my body and mind is in good condition during this difficult time of writing and job searching. Dr. Houser, and the other members of my committee were of invaluable help on my research and writing. Finally, thanks go out to all my colleagues in the Interdisciplinary Center for Economic Science (ICES) for providing a thought provoking work environment.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

ENDOGENOUS NETWORK FORMATION: EXPERIMENTS AND METHODS

Rong Rong, Ph.D.

George Mason University, 2013

Dissertation Director: Dr. Daniel Houser

This dissertation develops both substantive and methodological themes on the topic of social networks. Substantively, I conduct experimental studies based on the game theoretical models that describe network formation in various settings. Methodologically, I review the procedure of cluster analysis that could be used to discover the nature and the number of behavioral rules used by individuals in network environments.

Chapter 1: Growing Stars: A Laboratory Analysis of Network Formation

The acquisition and dispersion of information often occurs through social networks (Jackson, 2009). Empirical and theoretical findings suggest that efficient information dispersion networks take the form of a star: small numbers of agents gather information for distribution to larger groups. Controlled randomized tests, however, have typically found little evidence of star network emergence. An exception is Goeree et al (2009), which reports reliable star network formation in an environment that includes ex

ante heterogeneous agents. While heterogeneity may explain network formation in some environments, in others it may play a smaller role. Here we show that specific institutional environments promote star network formation in the presence of ex ante homogeneous agents. Especially effective institutions include investment limits and the "right-of-first-refusal," both of which add stability to the decision environment. At the level of individual behavior, we find these institutions to encourage rational decision making and positive habit formation.

Chapter 2: Money or Friends: Social Identity and Truth Telling in Networks

Communication between departments within a firm may include deception. Theory suggests that small difference in monetary incentives explains why lying to outgroup members may be strategically optimal (Crawford and Sobel, 1982; Galeotti et al, 2012). In natural environments, however, social incentives also play an important role in determining the information people choose to share or to withhold. Unfortunately, little is known about how monetary and social incentives interact to determine truth-telling. We design a laboratory experiment to address this question. We found that absent social identity, players' choices are mostly consistent with the theoretical predictions. Interestingly, the effect of identity is asymmetric: sharing the same identity does not promote truth-telling but holding different identities reduces truthfulness. We find that identity has an overall detrimental impact on truthfulness. These results have important implication for intra-organizational conflict management, suggesting that only by

strengthening identity at the level of the organization can one create a positive impact on communication among different departments.

Chapter 3: Exploring Network Behavior using Cluster Analysis

Cluster analysis organizes a complicated data set into small number of groups based on patterns of similarity. It can be used to discover data structures without requiring strong ex ante assumptions about the properties of the data. Decision data from laboratory experiments are often generated by complex behavioral rules that can be difficult to specify a priori. These data may particularly benefit from clustering methods. This paper reviews key procedures and algorithms related to cluster analysis and discusses how to choose among clustering methods to analyze experimental data.

CHAPTER ONE: GROWING STARS: A LABORATORY ANALYSIS OF NETWORK FORMATION

I. Introduction

How information is acquired and subsequently dispersed among people is widely studied in economics (Rogers, 1995). In many relevant contexts it occurs through networks of agents (Jackson, 2009). Empirical and theoretical findings suggest that efficient information networks take the form of a star: small numbers of agents gather information and then distribute it to a larger group (Weimann, 1994; Bala and Goyal, 2000; Galeotti and Goyal, 2010). Despite the theoretical advances, one persistent challenge has been to discover conditions under which star networks emerge within controlled laboratory environments. One way to generate star networks reliably is to incorporate ex ante agent heterogeneity (Goeree, et al, 2009). Although able to explain the formation of star networks in many cases, ex-ante heterogeneity is perhaps less important in other naturally occurring network environments (Feick and Price, 1987; Conley and Udry, 2010). Our paper addresses network emergence from an alternative perspective. We build features of naturally occurring institutions into our experiment design, and find them successful at promoting star networks with ex-ante homogeneous agents.

Early work on star networks dates to the 1950s. In their pioneering paper, Katz and Lazersfeld (1955) coined the term "opinion leaders" to describe a small subset of

highly connected people[1]. Half a century later, studies continue to provide empirical support for the existence of opinion leaders in politics and marketing (Weimann, 1994; Katz and Lazersfeld, 2006). Opinion leaders clearly make a difference. For instance, empirical evidence has shown that words from opinion leaders boost sales of consumer products (Godes and Mayzlin, 2009), contribute to the prevention of AIDS (Kelly et al, 1992), and transmit political thought and ideas (Roch, 2007). Given the importance of the opinion leaders in disseminating information, people in both the private and public sectors are eager to identify them (Iyengar et al, 2008). A better understanding of the formation and growth of star networks, which can be thought of as stylized opinion leader networks, would facilitate such efforts[2].

Theoretical studies of star networks have shown that under certain conditions star networks emerge as efficient and stable equilibria (Bala and Goyal, 2000; Galeotti and Goyal, 2010[3]). Requirements include that network goods are non-rival and that agents are able to form links unilaterally[4]. These conditions are easily implementable in laboratory

[1] The concept of "influentials" can largely be used interchangeably (Merton, 1968; Gladwell, 2000).
[2] Research on star networks also connects to the empirical literature on so-called scale-free networks, where the majority of nodes have only a small number of links while a small number of nodes are highly linked. Examples of scale-free network include scientific citation networks, coauthor networks, internet Pagerank networks, among others. The star network we study is a special case of a scale-free network.
[3] Some non-game-theoretical models of star network formation build upon preferential attachment and study the behavior of large networks (Barabasi and Albert, 1999; Jackson and Rogers, 2007). A direct test of those models needs validation of its behavioral assumptions, which is difficult to achieve with a lab experiment.
[4] Jackson and Wolinsky (1996) discussed two cases in which those two conditions are lacking. They found that network efficiency and stability is hard to achieve under regular payoff functions.

studies and are attractive in the sense that they are features of many environments where information dispersion is important[5].

To investigate star-network formation, economists have collected laboratory data in various network environments (Callander and Plott, 2005; Falk and Kosfeld, 2003; Goeree et al, 2009). Loosely speaking, in a typical network formation experiment, players decide how to form "links" with other players in light of the benefits those links confer. These studies generally, however, did not succeed in discovering star networks[6]. To our knowledge, the single exception is Goeree et al (2009), which finds star networks to emerge reliably in the presence of ex ante heterogeneous agents.

Ex ante heterogeneity may help to explain the emergence of networks in many environments (e.g., co-authorship), but may not be the entire solution in other naturally occurring network environments (Feick and Price, 1987; Conley and Udry, 2010). For this reason, we thought it important to investigate institutional effects on star network formation with ex ante homogeneous agents.

Our focus is institutions that add "stability" to the decision environment, in the sense of reducing period-to-period changes in one's choice. Note that agent heterogeneity can accomplish this. The reason is that people who have an advantage in investing or linking may be more likely to do so repeatedly, and others may be more able to form accurate expectations about their play. Institutions can also play this role in principle, and

[5] Knowledge generated in academia is one example. Open source software is another example. This excludes cases where information is protected by IPR and dispersion of that information requires bilateral agreements such as those found in some enforceable contracts.

[6] Falk and Kosfeld (2010) found equilibrium "wheel" networks to emerge, but were not able to observe the formation of equilibrium star networks.

thus promote the emergence of star networks. To test this possibility, we collect data from laboratory experiments under the following institutions:

(1) Sequential decisions. Sequencing is a feature of invitations on Facebook or Twitter, among others. They regularly ask current users to invite their friends to join the site, and then those friends ask their friends. It is plausible that sequential moves can mitigate unnecessary "trial and error" and therefore stabilize period-to-period decisions.

(2) Investment limits (or budget constraints). Budget constraints are present in all natural environments where information dispersion is important. There are the usual expense constraints in R&D projects (Dimasi et al, 2003; Dimasi and Grabowski, 2007), or sometimes government policies simply rule out multiple investments in the same area (Tran, 2009). Moreover, when personal relationships are involved, investments may face natural constraints with regard to time or distance (Marsden and Campbell, 1984).

(3) "Right of first refusal" (which we often denote by RFR). In contracts, RFR ensures that investors are able, should they desire to do so, to continue their investments. This stability allows for long-term planning by both the investor as well as others in the environment who are impacted by the presence of such investments. In addition, RFR emerges regularly whenever economic outcomes favor persistent investments in one agent rather than a spreading of resources among multiple agents. For instance, families in developing countries may offer higher education only to one child while keeping others siblings with minimum mandatory education. The predictability of an investor's identity may promote coordination and facilitate star network formation.

The key finding of our paper is that, in the presence of these environment-stabilizing institutions, stars reliably emerge in the presence of ex ante homogeneous agents. In particular, in both simultaneous and sequential decision environments, we find that combining investment limits with RFR generates robust star networks. As noted above, these findings complement those of Goeree et al (2009) by helping to explain the emergence of star networks in environments where agent heterogeneity may play a smaller role.

The remainder of the paper is organized as follows: The next section briefly reviews the theoretical and experimental literature on network formation. Section 3 lays out the theoretical background of the study. Section 4 presents the experimental design and procedure, and sets up the hypothesis. Section 5 reports experimental results. Section 6 explores decision making patterns at the individual level, and how those patterns are impacted by the institutional environment. Section 7 concludes.

II. Literature Review

II.1 Theoretical work on star network formation

Many theoretical studies have attempted to shed light on the process of network formation in general (Jackson, 2003), and recently specific theoretical progress has been made on understanding the conditions under which star networks can form (Bala and Goyal, 2000; Bramoulle et al, 2004; Galeotti and Goyal, 2010). For all the cases that we study in this paper, equilibrium star networks are also efficient. Star networks feature asymmetry in equilibrium actions by participants, because it pays to send links when others invest and vise versa. Note that this environment is characterized by "strategic

substitutes", and includes in general both anti-coordination games and games related to public goods provision[7].

An early paper by Bala and Goyal (2000) studied an environment with non-rival network goods and the possibility of forming links unilaterally. They found that star networks emerge in equilibrium only when the benefit of information flows between two agents regardless of who sends the link[8]. Their study was followed by Bramoulle et al (2004), who examined network formation in an anti-coordination game. They found that the shape of the equilibrium network need not be a star; with the exact network shape depending on the cost of link formation. More recently, a study by Galeotti and Goyal (2010) extended the model of Bala and Goyal (2000) by endogenizing the choice to invest. Their study showed that star networks emerge in equilibrium as well[9]. These advances of course leave open the question of whether the conditions required by theory are sufficient to generate star networks reliably in a controlled laboratory environment.

II.2 Experiments on star network formation

Despite the abundance of empirical evidence describing the existence of star networks[10], we are aware of only four experimental studies on star network formation (Callander and Plott, 2005; Falk and Kosfeld, 2003; Berninghaus et al, 2007; Goeree et al, 2009[11]). Falk and Kosfeld (2003) tested the theory of Bala and Goyal (2000). In particular, they studied whether and how equilibrium networks can form under "one-

[7] Bramoulle and Kranton (2007) discussed public goods provision in exogenerous networks extensively.
[8] If a link sender receives information from a link receiver, then the equilibrium network is a wheel.
[9] Galeotti and Goyal (2010) predicts peripheral-sponsored stars, while center-sponsored stars are predicted in the decay free model of Bala and Goyal (2000) and Bramoulle et at (2004).
[10] See Katz and Lazarsfeld (1955), Rogers (1995), Valente (1995)
[11] Experimental studies on endogenous networks other than stars include Deck and Johnson (2004), Ule (2005), Corbae & Duffy (2008), Knigge & Buskens (2010), Berninghaus et al (2011).

way" and "two-way" information flows. In contrast with theoretical predictions, they found that when information flows two ways the network fails to converge to a star. They concluded that the need for asymmetric strategies combined with inequality aversion might contribute to the difficulty in realizing star networks.

Callander and Plott (2005) also tested Bala and Goyal (2000) in the lab. They considered various conditions that differed in terms of the linking cost, as well as the value of information. They also examined the impact of having network agents with heterogeneous payoff structure, an issue unaddressed by the model. Their main finding was that star networks did not consistently emerge under theoretical conditions, and that even introducing payoff heterogeneity did not lead to systematic formation of star networks. Consequently, they report that "significant and persistent inefficiency" is a feature of all of their network environments.

Berninghaus et al (2007) provided yet another test of Bala and Goyal (2000) but focused on the comparison between discrete and continuous time environments. In the discrete environments, their results show that players have a tendency to reduce network distance over time. However, the overall average frequency of star networks found in their data (11.33%) is no greater than what we found in our baseline environment.[12]

In light of the complications with generating star networks, and following Callander and Plott (2005), Goeree et al (2009) explored whether common knowledge of agent heterogeneity combined with two-way information flows might promote star networks. They reported that: (1) compared to homogeneous agent treatments,

[12] Due to the design of the continuous environments in their study as well as Berninghaus et al (2006), the results cannot be compared to data from discrete environments.

significantly more stars are observed when agents' payoffs are heterogeneous[13]; and (2) perfect information about the nature of heterogeneity plays an important role in facilitating the coordination on star networks.

Like the above studies, we explore what conditions may facilitate the emergence of star networks. But in contrast, our study emphasizes the importance of homogeneous agent assumption and explores how institutional characteristics may impact network formation.

While Callander and Plott (2005) and Goeree et al (2009) demonstrated the importance of individual heterogeneity in network environments, there may be some environments where individual differences play a smaller role. For instance, information about heterogeneity may not always be easily available in natural environments, due to the fact that it goes unobserved. Indeed, substantial empirical research on market mavens has found no differences between the observable characteristics of agents who play different roles in the network (Feick and Price, 1987; Geissler and Edison, 2005; Wiedman et al, 2001; Williams and Slama, 1995). Others have pointed out that obtaining information about the costs and benefits of other network agents in agricultural environment may be difficult, given that people have an incentive to conceal their private information (Conley and Udry, 2010). Moreover, ex ante agent heterogeneity is not required by theory for star network emergence[14].

[13] According to their experimental data, heterogeneous link costs do not seem significantly to promote star network formation.

[14] Jackson and Lopez-Pintado (2011), Larrosa and Tahme (2011), and Vandenbossche and Demuynck (2010), developed models with heterogeneous agents. However, none of these relate to incentives associated with information acquisition or diffusion; therefore, the predictions generated from those models are not star-shaped networks. Galeotti et al (2006) developed models showing that star networks are an

In view of the fact that individual-level information is costly and sometimes infeasible to obtain, and to avoid introducing artificial focal points, we design our experiments to include ex ante homogeneity. We investigate conditions under which ex ante identical agents will take asymmetric equilibrium actions to establish efficient and stable star networks.

III. Theoretical Background

Our study is based on the model of network game in Galeotti and Goyal (2010). In their model, a group of identical rational agents face the choice of either investing in information or obtaining it less expensively by linking to another who currently invested in information. The level of investment by agent i is discrete $x_i \in \{0,1\}$. The set of links sent by agent i is denoted by a vector $g_i = (g_{i1},...,g_{ii-1},g_{ii+1},...,g_{in})$, where $g_{ij} = 1$ if player i sent a link to player j. Linking choices are then combined to determine the directed network structure $g = (g_1, g_2,..., g_n)$ 15. The key assumptions of the model are that information is non-rival and flows both ways across network links.16

The non-directed version of the network is denoted by $\bar{g}$, where $\bar{g}_{ij} = \max\{g_{ij}, g_{ji}\}$ for each agent i and j. Define $N(i;g) = \{j : g_{ij} = 1\}$ as the set of agents

equilibrium in environments where agents have heterogeneous benefits for information, while under heterogeneous costs stars are no longer equilibrium.

[15] A directed graph is a graph where the edges are associated with a direction.

[16] These assumptions are important due to the fact that they closely characterize certain situations of information dispersion in natural environments. For example, knowledge about agricultural technology is mostly non-rival, and could be shared between personal connections of farmers regardless of the linking direction.

to whom i has sent a link and $N(i;\bar{g}) = \{j : \bar{g}_{ij} = 1\}$ as a set of agents with whom i has been connected. The payoff to agent i is

$$\pi_i(x_i, g_i) = f(x_i + \sum_{j \in N(i;\bar{g})} x_j) - cx_i - |N(i;g)|k \qquad (1)$$

where $c > 0$ reflects the unit cost of purchasing the non-rival goods, x_i refers to the number of unit player *i* purchased, $k > 0$ is the cost of sending one link and $|N(i;g)|$ refers to the cardinality of the set $N(i;g)$.

Different specifications for *f* define different types of games. In this paper, we follow Galeotti and Goyal (2010) and assume *f* is a step function

$$\begin{cases} f(y_i) = 1 & \text{if } y_i \geq 1 \\ f(y_i) = 0 & \text{if } y_i < 1 \end{cases} \qquad (2)$$

where $y_i = x_i + \sum_{j \in N(i;\bar{g})} x_j$. The above return function $f(y_i)$ resembles the payoff structure of best shot game in the widely studied public good games literature. The advantage to using a step function is that it provides sharp equilibrium predictions that can be more easily tested in the laboratory[17].

It can be shown that every equilibrium of the network best shot game is a star network when $k < c$ (Galeotti and Goyal, 2010)[18]. The intuition is as follows: if in equilibrium the sole investor deviates and does not invest, then the group obtains no information, implying a lower payoff for everyone including the investor. Similarly, if a person who has linked to the investor deviates by not linking, choosing to link to another

[17] Instead of star network, the general prediction is a so-called core-peripheral network, where a few interconnected agents invest in information and the rest of agents connect to them. A star network is a special core-peripheral network with a single agent core.

[18] When $k > c$, the unique equilibrium is an empty network.

(who in equilibrium cannot have the information), or becoming an investor oneself, in all cases such deviations clearly lead to lower payoffs. Therefore, the star network is a Nash equilibrium. Note also that all star network equilibria in the best shot game are efficient (in the sense that equilibria are not Pareto ranked). This feature of the network best shot game, as well as its clean equilibrium predictions, leaves it ideal for laboratory testing. In the following section we detail our design, which follows the network best shot game closely.

IV. Experiment design and hypothesis

Our experiment is designed to examine how naturally-occurring institutions affect star network formation with ex ante homogenous agents. Institutional characteristics such as sequential decisions, investment limits and the RFR often coexist with star networks. We conjecture that these institutional characteristics may be important conditions for the formation of star networks in naturally occurring environments. Our laboratory study brings these institutional features into a controlled laboratory setting and examines the effect of each on star network formation.

IV.1 Experiment Design

IV.1.1 General Environment

Our design is based on the best shot game introduced in the appendix of Galeotti and Goyal (2010). This modification leads to the sharp prediction that a star network is the unique Nash equilibrium configuration, and is also efficient. To the best of our knowledge, our study is the first to examine the network formation process where agents make simultaneous linking and information investing decisions.

Each experimental session includes 16 subjects randomly divided into four groups. All subjects participate in three stage games. Each stage game consists of a random number of rounds[19]. Groups are fixed during each stage game, and each group member holds a unique ID: J,K,L or M. We avoid using "A" as an ID because it may be focal[20].

In each round, decision-makers decide whom to link to among the other three group members and also whether to purchase information. If a participant purchases information, she pays a cost of E$0.9 and earns the value of information, E$3, with certainty. On the other hand, if a player decides to send a link to another player, she pays a cost of E$0.5 per link. When a subject links to another subject who has purchased information, the subject who chooses to link also earns E$3. Subjects who link to other subjects that have not purchased the information pay a cost of E$0.5, but earn nothing. Costs and payoffs remain fixed throughout all three stage games and all treatments.[21].

Subjects submit their decisions using the decision screen (see Appendix A, Fig. 1). Then, a display screen informs all players of the current network outcome and each group member's payoff (see Appendix A, Fig.2).

Within each of the three stage games, the payoff is determined by the accumulated earnings over all rounds. Players are informed about their own stage payoff

[19] There are always at least 4 rounds per stage. After round 4, the game has a random stopping probability of 0.04 at any given round. To keep control over the length of the real experiment, we use the predetermined length 16, 44 and 24 for experimental stages I, II and III respectively. Those numbers are generated using a randomization device. Each practice stage includes 8 rounds.

[20] Some have suggested that the first mover J might also be in a focal position. It turns out that, for all treatments, J is not statistically more likely to be the center of the star than any of the other positions ($p<0.195$ in all bivariate comparisons across all treatments, two-sided Mann-Whitney tests)

[21] These parameter values ensure that the equilibrium and efficient networks are star-shaped.

at the end of each stage. They are also reminded that they will be re-matched with players with whom they have not played previously, and that their stage payoff will not be carried over to the new stage. Each subject's earnings for the experiment are determined by one randomly-determined stage game.

IV.1.2 Treatment Design

Within the general experimental environment described above, we study the effects of three institutional characteristics of network formation. We examine sequential decisions and investment limits on network formation, both individually and jointly, using a two-by-two treatment design. A fifth treatment then studies the effect of the "right of first refusal."

In a two-by-two design, we vary the sequence of decisions in one dimension to be either sequential or simultaneous. In simultaneous treatments, subjects from the same group make their decisions at the same time, not knowing what other subjects would choose. In sequential treatments, only one subject makes a decision per round. Players make decisions according to the alphabetical order of their ID (first J, then K, then L and finally M) with full knowledge of the choices made by earlier decision makers. Further, players earn money even on rounds for which they do not make a decision, with their payoff determined by their most recent previous choice in combination with the choices of others[22].

The second dimension of our design varies the existence of investment limits. Absent investment limits, players can invest in information and links at will,

[22] Note that in relation to the simultaneous game, participants make fewer decisions in the sequential game. This is done to ensure that payoff incentives remain identical between the simultaneous and sequential games.

independently of other players' decisions. With investment limits, the following three conditions hold: (i) in each round, each player can either send a link or invest in information, but cannot do both; (ii) each player can send at most one link; and (iii) at most one player can invest in information at any given time. We refer to the treatment without investment limits as the "baseline", and denote treatments with investment limits as "limits."

Notice that Seq_L and Seq_B differ in two ways: while investment is limited in Seq_L, it also implies the "right of first refusal", by which we mean that a person who currently invested in information has the right to continue his/her investment. The reason is that in Seq_L, a subject who has invested in information will continue to hold it until their next decision, and nobody else will be able to invest in additional information. Consequently, the only way they can lose the information is if they give up the information. It follows that comparing Seq_L to Seq_B measures the total effect of the investment limits combined with the RFR.

While these two effects cannot be separated in our sequential environment, it is possible to achieve separation in a simultaneous setting. To do this, we construct a fifth treatment that builds on Sim_L but eliminates the RFR. In any given round, agents who choose to invest in information will have an equal chance to obtain the information, regardless of whether he/she invested the information in the previous round[23]. This treatment is denoted as "simultaneous-limits with no RFR" (Sim_L_NoRFR).

[23] In Sim_L, if a previous investor chooses to invest, he/she will be able to continue the investment with 100% certainty. The first period investor is randomly determined if multiple players choose to invest.

In summary, we investigate network formation in five treatments that differ in terms of the sequence of moves, whether investment is limited, and the existence or nonexistence of the RFR. We list the properties of these five treatments in Table 1.

Table 1 Properties of Treatments

Treatment	Decision sequence	Investment limits?	RFR?
Seq_B	Sequential	N	N
Seq_L	Sequential	Y	Y
Sim_B	Simultaneous	N	N
Sim_L	Simultaneous	Y	Y
Sim_L_NoRFR	Simultaneous	Y	N

IV.2 Equilibria and Hypotheses

IV.2.1 Equilibria prediction[24]

For all three simultaneous treatments (Sim_B, Sim_L and Sim_L_NoRFR), the stage-game equilibrium is identical to the one described in Section III and in Galeotti and Goyal (2010). It is easy to see that adding investment limits to this environment does not change the equilibrium predictions, because these limits only rule out certain non-equilibrium actions. Similarly, the existence of the right of first refusal does not affect the stage game equilibrium predictions since it only makes it (weakly) more likely for people to hold beliefs consistent with equilibrium outcomes.

For two sequential treatments (Seq_B and Seq_L), it is necessary to modify the stage game into its extensive form. It is easy to show that, under the parameter values as

[24] Our experiment includes repeated games with a random stopping rule, but we focus only on the analysis of the stage game equilibria as it is easy to show that a sequence of stage-game Nash equilibrium strategies is also a subgame-perfect equilibrium in the repeated game. NE strategies other than those found in the stage game might exist in repeated game environments (as demonstrated by Folk Theorems). It is beyond the scope of this paper to provide a characterization of these additional equilibria.

specified above, the unique subgame perfect Nash equilibrium for the extensive form game occurs when the first mover invests, and each subsequent player links to that investor. Further, because investment limits only rule out certain non-equilibrium play, it is straightforward to verify that this remains the unique SPNE in this case as well. Similarly, RFR would not alter the equilibrium prediction.

All in all, all five treatments in our study share a common equilibrium: the star networks.

IV.2.2 Hypothesis

We expect to see a positive effect on star network formation in environments that include sequential decisions, investment limits, and the "right of first refusal"[25]. To measure the effect, we first construct a measure of equilibrium frequency. We count a network graph as a star if and only if there is one member who chooses to invest in information and the other three agents send exactly one link to the sole investor. For each stage of the experiment, equilibrium frequency is calculated by dividing the total number of star networks by the total number of rounds in that stage game. Then, for each treatment, the mean frequency of star networks is calculated by averaging the measure over all the partner-matched repeatedly-played stages.

In light of the above discussion, we expect the mean frequency of star to follow the order below (where $Freq_X$ denotes the frequency of star networks in treatment X):

[25] The effects can be demonstrated in multiple ways. We discuss a particularly intuitive measure of equilibrium frequency in the text. Further discussion of network stability and individual rationality can be obtained from the authors on requests.

Hypothesis 1.1 Sequential decisions increase the frequency of star networks:

$$Freq_{Seq\_B} > Freq_{Sim\_B}; Freq_{Seq\_L} > Freq_{Sim\_L} \quad (3)$$

Hypothesis 1.2 investment limits and the RFR combined increase the frequency of star networks:

$$Freq_{Sim\_L} > Freq_{Sim\_B}; Freq_{Seq\_L} > Freq_{Seq\_B} \quad (4)$$

Hypothesis 1.3 In simultaneous environments, RFR alone increases the frequency of star networks:

$$Freq_{Sim\_L} > Freq_{Sim\_L\_NoRFR} \quad (5)$$

Hypothesis 1.4 In simultaneous environments, investment limits alone increases the frequency of star networks:

$$Freq_{Sim\_L\_NoRFR} > Freq_{Sim\_B} \quad (6)$$

We select these three institutions because we believe they will stabilize the decision environment, therefore we also hypothesize that when sequential moves, investment limits and the "right of first refusal" are present, subjects change their decisions less often. Since each player makes four decisions per period (one purchasing and three linking decisions), we determine the percent of choices an individual changes between rounds. The percentages of changed choices are then averaged over each stage, and this average is our measure of stability. We hypothesize this mean percentage of changed choices will be ordered as follows (where $Change_X$ denotes the percentage of choice change in treatment X):

Hypothesis 2.1 In sequential treatments, subjects make less choice changes in environments with investment limits and the RFR than in baseline:

$$Change_{Seq\_L} < Change_{Seq\_B} \quad (7)$$

Hypothesis 2.2 In simultaneous environments, subjects make the least choice change in environments with investment limits and the RFR. Subjects make the most choice changes in baseline:

$$Change_{Sim\_L} < Change_{Sim\_L\_NoRFR} < Change_{Sim\_B} \quad (8)$$

IV.3 Experimental Procedure

The experiment sessions were conducted between December 2010 and March 2011 in the ICES laboratory at George Mason University. Subjects were recruited via email from registered students at George Mason University. Each subject participated in only one session and none had previously participated in a similar experiment.

In total, 160 subjects participated in the computerized experiment programmed with z-Tree (Fischbacher, 2007). Each experimental session lasted between 120 and 150 minutes. Subjects' total earnings were determined by the Experimental Dollars (E$) earned at the end of the experiment, which were then converted at a rate of E$3 per US dollar. The average earnings were $25.28, ranging from a maximum of $53 to a minimum of $8 across all sessions.

In all treatments, before a session started, subjects were seated in separate cubicles to ensure anonymity. They were informed of the rules of conduct and provided with detailed instructions. The instructions were read aloud. In order to guard against confusion, after subjects finished reading the instructions, they were asked to complete a quiz. An experimenter checked their answers. Then the experiment worked through the quiz questions on a white board in front of the laboratory. The experiment began after all subjects confirmed they had no further questions.

We ran 2 sessions for each treatment condition. Thus, in the end, we obtained 672 network graphs for each treatment (excluding the practice stage). Most of our analysis assumes 24 observations (eight groups each of which plays three stage-games with perfect strangers) for each treatment.

V. Results

We present results in the order of hypotheses listed in Section IV.2.2. First, we discuss results concerning the frequency of star networks. Then we investigate the stability of choices: how often individuals change their linking and investing decisions.

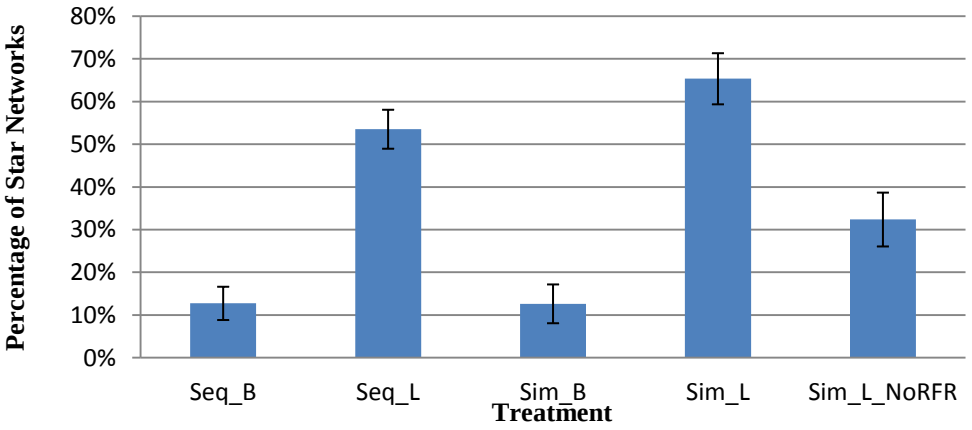


Figure 1 Mean frequency of star networks (in %) by treatment
***Note: standard error shown in marks**

The mean frequency of star networks in each of our treatments is shown in Figure 1. It is clear from this figure that star networks emerge at different rates, with baseline treatments displaying the lowest frequency of star networks. More formally, our findings are as follows:

Result 1. (Test of hypothesis 1.1) Sequential decisions do not increase the frequency of star networks.

We found star networks to emerge with frequency 12.6% and 12.7% in Sim_B and Seq_B, respectively ($p=0.667$[26]). On the other hand, when investment limits and the RFR are both present, 53.5% of networks formed in Seq_L are star shaped in comparison to 65.3% in Sim_L; this is also insignificant at standard levels ($p=0.054$).

Result 2. (Test of hypothesis 1.2) More star networks emerge when investment limits combined with the RFR are present.

Sim_L generated 65.3% of star networks, while only 12.6% of networks in Sim_B are star shaped. This difference is significant ($p< 0.001$). Agents in Seq_L form star networks 53.5% of time. When compared with the 12.6% in Seq_B, the difference is again significant ($p< 0.001$). Thus, our data provide clear evidence supporting the positive impact of investment limits and the RFR on star network formation.

Result 3. (Test of hypothesis 1.3) The RFR promotes star network formation in simultaneous decision environments.

The right most two bars in Figure 1 correspond to Sim_L and Sim_L_NoRFR. The only difference between these two treatments is that the RFR is present in the former but absent in the latter. In Sim_L star networks emerge at a rate of 65.3%[27]. The frequency in Sim_L_NoRFR treatment (32.4%) is significantly lower than this (p

[26] Unless otherwise indicated, all p-value refer to two-tailed Mann-Whitney tests.
[27] The frequency of star networks found here, 65.3%, is at least as high as any of the star-network frequencies reported by Goeree et al (2009) under ex-ante agent heterogeneity.

=0.0016). This is evidence that the RFR alone promotes star networks in simultaneous environment.

Result 4. (Test of hypothesis 1.4) In simultaneous environments star networks emerge more frequently with than without investment limits.

Sim_L_NoRFR generates star networks at a rate of 32.4%, while the frequency in Sim_B is 12.6% (p=0.0034). Thus, investment limits alone promote star networks in simultaneous environment.

The following section investigates the stability of individual choices. The mean percentage of changed decisions between period *t-1* and *t* is plotted in Figure 2 by treatment condition.

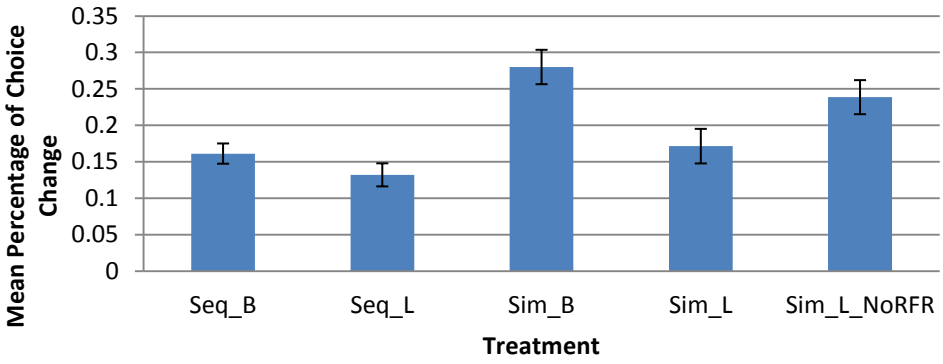


Figure 2 Mean percentage of choice change by treatment

Result 5. (Test of hypothesis 2.1)Among the two sequential treatments, institutions that generate more equilibrium networks also exhibit greater choice stability.

The average percentage of choice change in Seq_B and Seq_L is 16.11% and 13.19% respectively. The difference between them are statistically significant (p=0.001)

Result 6. (Test of hypothesis 2.2)Among the three simultaneous treatments, institutions that generate more equilibrium networks also have exhibit greater choice stability.

Players in Sim_B change on average 27.99% of their choices each period, significantly higher than the frequency of change, 17.13%, in Sim_L (p=0.0035). Players in Sim_L_NoRFR make 23.85% of their choices per period. This is higher than Sim_L (p=0.0254) and lower, although not statistically significantly, than Sim_B (p=0.1546).

VI. Behavioral Rules

The purpose of this section is to draw inferences about the behavioral rules of individuals in our various treatments. Our maintained assumption is that behavioral rules in all treatments are formed using elements from a menu of information that are finite and identical, but that different treatments lead to rules that differ at the level of usage on the information. Without ex ante knowledge of what kind of rules may exist, we use cluster analysis to detect them[28]. Compared to regressions, cluster analysis can better explore patterns within a complex environment where the classification structure may not be well defined. It allows us to explore behaviors among individuals without the need to pre-define the nature or number of possible rules (see also Houser et al, 2004). For the purpose of this study, we implement k-means cluster algorithm.

[28] Cluster analysis, as a numerical method for classification, serves the function of organizing a large and complicated data set into a smaller number of groups of objects. Cluster analysis is widely used in fields such as astronomy, biology and marketing, and increasingly in economics (Fisher, 1963; Hirschberg et al, 1991; Houser, et al, 2004).

Our analysis proceeds in two steps. First, we estimate for each individual the parameters that characterize the way they make decisions given information. Then, we use cluster analysis to group similar individuals into behavioral rules. In particular, we run a linear regression for each individual with the decision to invest (or not) as a binary dependent variable, on a constant, a dummy for whether investing is rational and an index characterizing the subjects investing behavior in the previous two rounds (see also Kurzban and Houser, 2005). Then, we use the k-means algorithm to cluster these estimates into groups of behavioral rules. We repeat the above analysis for the linking decision.

VI.1 Behavioral rule parameters

The independent variables we include in our regressions are meant to capture a person's: (i) base rate willingness to invest or link to others (captured by the regression's constant); (ii) consistency with individual rationality (captured by the a dummy variable that takes value one if it is optimal to invest (or link)); and (iii) propensity to form a "habit" of choice in the sense that they do what they did before (captured by the variable indicating the lagged decisions for the past 2 rounds). Equations 3 and 4 specify our regression equations for investing and linking respectively:

$$invest_{i,t} = \beta_1 * rational^{p}_{i,t} + \beta_2 * \sum_{s=1}^{2} invest_{i,t-s} + \beta_3 + \varepsilon_{i,t} \quad (9)$$

$$linksending_{i,t} = \gamma_1 * rational^{l}_{i,t} + \gamma_2 * \sum_{s=1}^{2} linksending_{i,t-s} + \gamma_3 + \delta_{i,t} \quad (10)$$

where

$$rational^{p}_{i,t} = \begin{cases} 1, \text{ if subject i should have purchased information at round t} \\ \quad \text{according to individual rationality} \\ 0, \text{ otherwise} \end{cases}$$

$$rational^{l}_{i,t} = \begin{cases} 1, \text{ if subject i should have sent link at round t} \\ \quad \text{according to individual rationality} \\ 0, \text{ otherwise} \end{cases}$$

$$invest_{i,t-s} = \begin{cases} 1, \text{ if subject i invested in information in round t-s} \\ 0, \text{ otherwise} \end{cases}$$

The above regressions are repeated for each individual. We end up with 142 and 152 subjects in our sample for the investing and linking regressions, respectively[29]. Each individual's estimates can be represented by a point in 3-space (See Appendix B).

VI.2 K-means clustering

We implement our k-means cluster analysis, as well as cluster number selection, using R. Based on the C-H index, we find three clusters in both investing and linking decisions[30] (See Appendix C).

[29] We drop 18 subjects for the investing decisions analysis, as there is zero variation in their dependent variables. For the same reason, we drop 8 subjects for the linking decisions analysis.

[30] Decision rules for investing and linking differ both in interpretation and range of measurement. Hence, we discuss them separately.

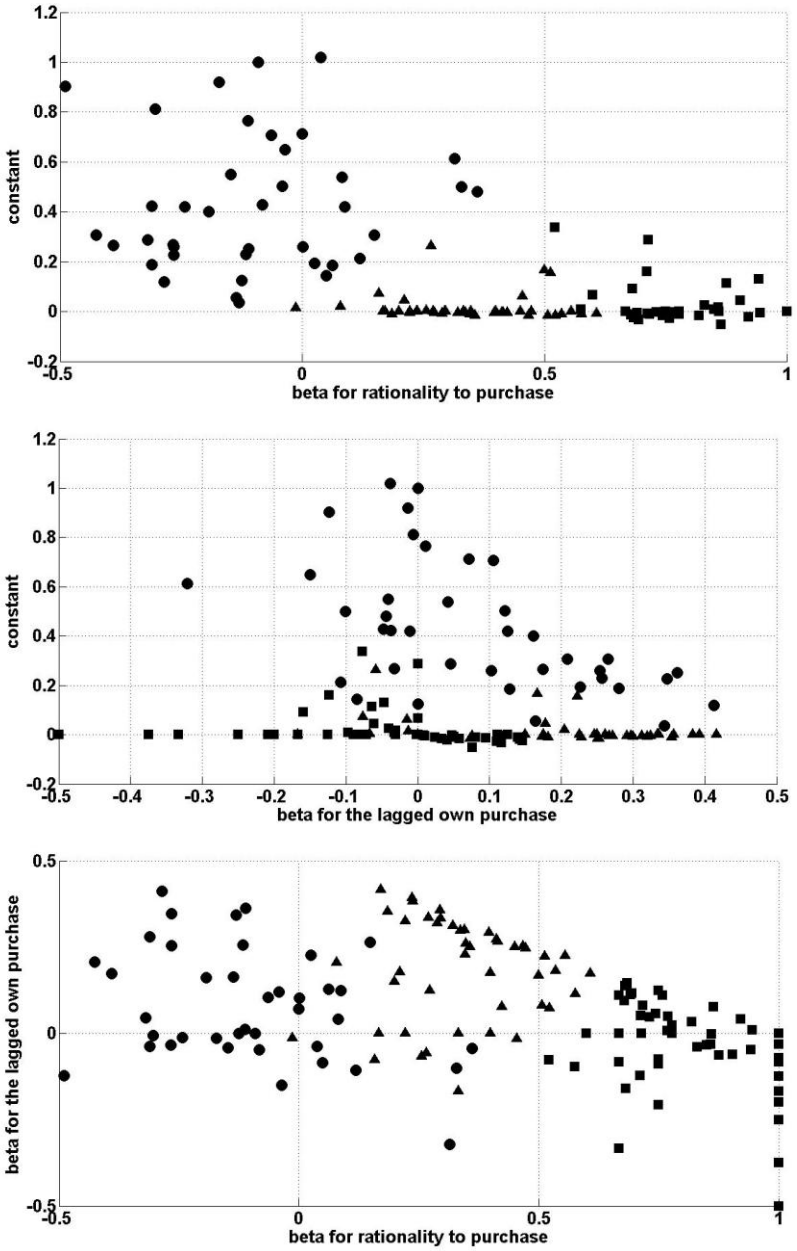


Figure 3 Projections of Estimates from Investing Decision

a) Investing decisions

The three panels of Figure 3 are the three 2-space projections of the estimates $\{\beta_1, \beta_2, \beta_3\}$ from regression on investing decisions (Equation 9) into corresponding 2-space. Each point represents an individual's estimates from his/her investing decisions regression. Points with the same marker belong to the same cluster.

It is clear from visual inspection that our clusters are well-separated. To provide statistical evidence on the strength of this separation, we analyze the separation along each independent variable's axis. Mann-Whitney tests find significant differences between all pairs of clusters in each axis ($p<0.001$), with the exception of the constants in the triangle and round clusters.

Not only are the clusters clearly separated, the location of the clusters also carries meaningful interpretation in our sample. Table 2 provides the mean estimate for each independent variable and for each cluster, and also reports whether that mean is significantly different from zero.

Table 2 The Mean of Estimates from Regression on Investing Decision

	Square Cluster	Triangle Cluster	Round Cluster
Rational to invest	0.8190 (0.0000)	0.3411 (0.0000)	-0.0978 (0.0054)
Lagged choice	-0.0408 (0.1480)	0.1745 (0.0000)	0.0782 (0.0120)
Base rate(constant)	0.0175 (0.2589)	0.0137 (0.7066)	0.4279 (0.0000)
Number of subjects	57	46	39

Note: p-value from Wilcoxon signed-rank test in parentheses

Based on the results from Table 2, we summarize the characteristics of the three clusters that define the three behavioral rules used by our subjects.

We define the cluster indicated with round markers as the "Rational" type. People that belong to this cluster are guided by the rationality of the current opportunity to invest. They focus less on their past choices, and their base rate of investing is near zero.

We define the cluster indicated by triangle markers as the "Habit" type. Subjects in this cluster are guided by rationality, but relatively less than the Rational type. Instead, their current decisions follow closely their past decisions.

We define the cluster indicated by square markers as the "Dogmatic" type. We find that these subjects have the highest base rate of investing among all three types.

We now investigate how the institutional characteristics in our various treatments affect the type of behavioral rules subjects use. Table 3 reports the frequency of types by treatment.

Table 3 Number of Individuals in each Treatment and each Type According to Investing Decisions

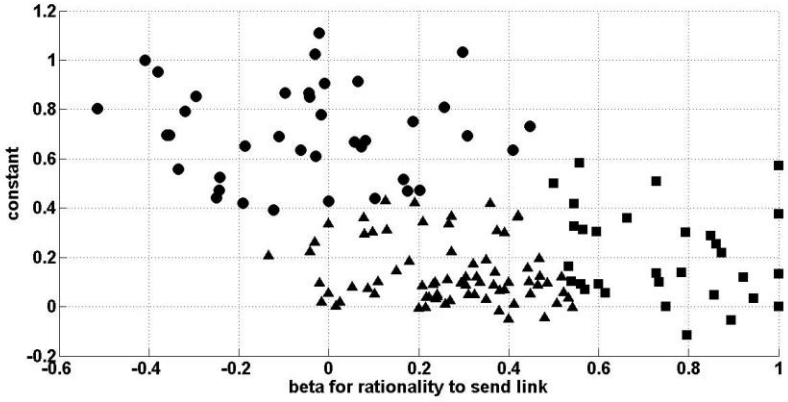
	Seq_B 12.6% of star	Seq_L 53.5% of star	Sim_B 12.7% of star	Sim_L 65.3% of star	Sim_L_NoRFR 32.4% of star
"Rational" Round	9 (31.03)	23 (92.00)	0 (0.00)	3 (11.11)	22 (70.97)
"Habit" Triangle	8 (27.59)	2 (8.00)	3 (10.00)	24 (88.89)	9 (29.03)
"Dogmatic" Square	12 (41.38)	0 (0.00)	27 (90.00)	0 (0.00)	0 (0.00)
Total	29 (100.00)	25 (100.00)	30 (100.00)	27 (100.00)	31 (100.00)

Note: percentage in parenthesis

As noted above, star networks emerge in fewer than 13% of our two baseline treatments (Seq_B and Sim_B). This low level of star network formation coincides with a concentration of Dogmatic type subjects (41.38% and 90% respectively). That is to say, having a concentration of players using the Dogmatic investing rule is not conducive to star network formation.

On the contrary, for the Seq_L treatment, which generates a relatively high percentage of star networks, the large majority of subjects (92%) choose to behave rationally. The other highly effective treatment, Sim_L, generates 65.3% of star networks. Its success at generating star network coincides with a high level of Habit typed subjects (88.89%), a few Rational subjects (11.11%) and no Dogmatic subjects.

The Sim_L_NoRFR treatment generates a medium level of star networks (32.4%). No subject in this treatment belongs to the Dogmatic type. In particular, most of them (70.97%) follow rational behavioral rules.



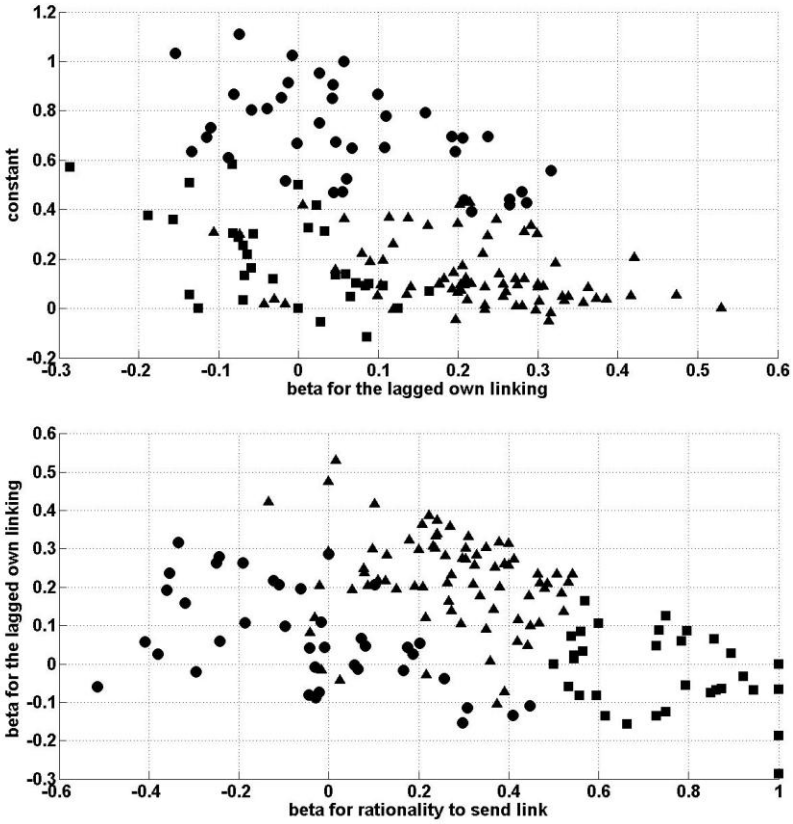


Figure 4 Projections of Estimates from Linking Decision

b) Linking decisions

Similar to the above analysis, the three panels of Figure 4 project each 3-vector estimate $\{\gamma_1, \gamma_2, \gamma_3\}$ from regression on linking decisions (Equation 10) into corresponding 2-space.

Again, we find clear visual separation between our clusters from estimates on linking behavior, and Mann-Whitney tests support significant differences between all pairs of clusters on all three axes ($p<0.001$), again with the exception of the estimates of the constants between the round and triangle clusters ($p=0.5279$).

Interestingly, the location of the clusters closely resembles those found for investing decisions. Consequently, we assign the same labels, Rational, Habit and Dogmatic, for each of these clusters as well. Table 4 reports the mean of each estimate for each cluster and the Wilcoxon signed-rank p-value for the test of whether the cluster's mean is significantly different from zero.

Table 4 The Mean of Estimates from Regression Analysis of Linking Behavior

	Round cluster	Triangle cluster	Square cluster
Rationality	0.7746 (0.0000)	0.2693 (0.0000)	-0.0461 (0.2112)
Lagged choice	-0.0187 (0.4091)	0.2179 (0.0000)	0.0745 (0.0041)
Base rate(constant)	0.1735 (0.0000)	0.1331 (0.0000)	0.6980 (0.0000)
Number of subjects	37	75	40

Note: p-value from Mann-Whitney test in parenthesis

Based on the characteristics of the three clusters described in Table 4, we define three behavioral rule types as follows:

We define the round cluster to be a "Rational" type. People who belong to this cluster make decisions that are guided largely by the rationality of their current choice.

We define the triangle cluster as a "Habit" type. People in this cluster make choices that resemble their previous choices.

We define the square cluster as a "Dogmatic" type. Subjects in this group send links to others at a high base rate (69.8%, statistically significantly higher than either of the other types ($p<0.001$)).

To see how institutions interact with types, we report types by treatment in Table 5. We were surprised that the clusters found in the linking analysis resemble so closely the clusters found in our analysis of investing behaviors. In both cases, the two baseline treatments with the lowest frequency of star networks also have the highest percentage of subjects belonging to Dogmatic type (40.63% and 60% for Seq_B and Sim_B respectively).

Result from Seq_L shows that the majority of subjects are the Rational type. And while Sim_L has the most frequent star network formation, it also has a high percentage of Habit type subjects.

Table 5 Number of Individuals in Different Treatments and Types According to Linking Decisions

	Seq_B 12.6% of star	Seq_L 53.5% of star	Sim_B 12.7% of star	Sim_L 65.3% of star	Sim_L_NoRFR 32.4% of star
"Rational" Round	13 (40.63)	20 (71.43)	0 (0.00)	2 (6.67)	2 (6.25)
"Habit" Triangle	6 (18.75)	4 (14.29)	12 (40.00)	24 (80.00)	29 (90.63)
"Dogmatic" Square	13 (40.63)	4 (14.29)	18 (60.00)	4 (13.33)	1 (3.13)
Total	32 (100.00)	28 (100.00)	30 (100.00)	30 (100.00)	32 (100.00)

Note: percentage in parenthesis

To investigate the relationship between the behavioral rules players use when making linking or investing decisions, Figure 5 plots the percentage of players belonging to each type in linking decisions conditional on each type in investing decisions[31]. 78% of

[31] Note that by construction, players in the investment-limits treatments who make rational investing decisions also necessarily make rational linking decisions. This is true only for the players who behave

subjects that are the Habit type in investing decisions are also Habit type in linking decisions. Similarly, there is substantial overlap among participants classified as Rational and Dogmatic between linking and investing decisions. Indeed, a Pearson Chi-square test rejects that type classifications are independent between investing and linking decisions ($p<0.001$).

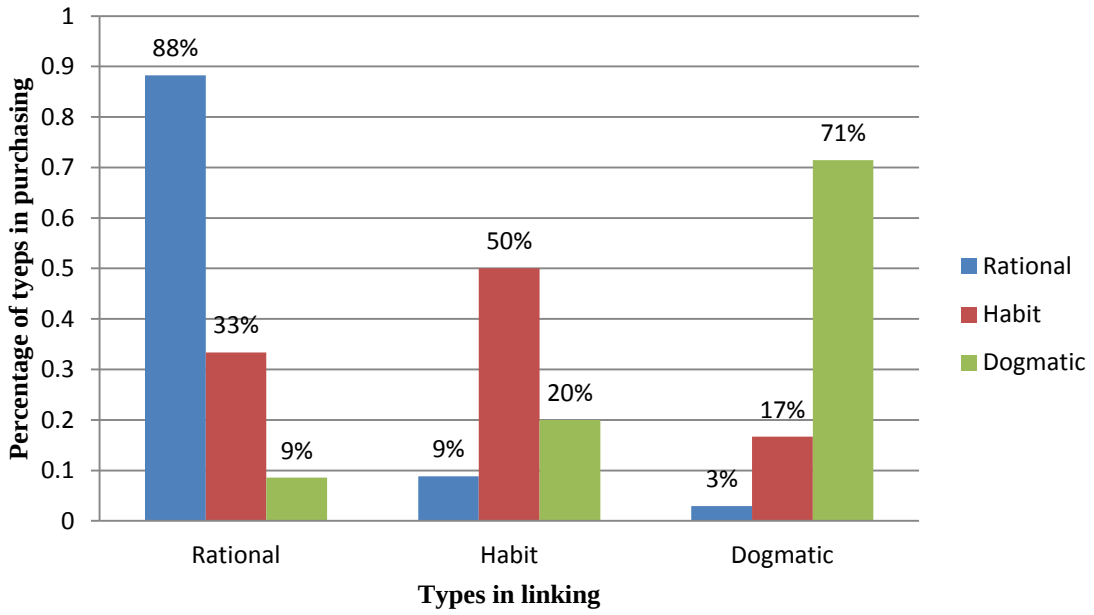


Figure 5 Frequency of Each Type in Investing Decision Conditional on Types in Linking Decision

VII. Conclusion

Star networks emerge naturally in many social environments, and theory indicates star network equilibria are efficient. Based on a model suggested by Galeotti and Goyal (2010), we study star network formation in the laboratory. Previous studies indicate that

---

perfectly rationally in the investment-limits treatments. Our design does not otherwise imply any correlation between the behavioral rules followed by players when making linking or investing decisions.

persistent star networks emerge in the lab, but only under ex ante agent heterogeneity (Goeree et al, 2009). This contrasts with natural environments, where star networks frequently emerge even when agents are ex ante homogeneous (Feick and Price, 1987; Conley and Udry, 2010). We conjectured that sequential decisions, investment limits, and the "right of first refusal," may stabilize the decision environment and promote the emergence of star networks, even in the presence of ex-ante homogeneous agents.

Our main finding is that investment limits and the "right of first refusal" promote star network formation. In comparison to baseline treatments, we find that environments with those features realize increased star-network frequency and decision stability.

In order to shed light on the impact of institutions at the individual level, we use a cluster analysis to draw inferences about behavioral rules used by participants in different environments. We find players clearly separate into clusters using "Rational", "Habit" and "Dogmatic" rules. Moreover, "Rational" and "Habit" are used more often in the presence of institutions that promote star networks. Further, we were comforted in finding that type-classifications for investing and linking decisions were tightly correlated, arguing for the validity of this behavioral characterization.

It is worthwhile to note that Falk and Kosfeld (2003) and Georee et al (2006) discovered the importance of inequality aversion in preventing star networks from forming under standard theoretical conditions. This suggests that finding approaches to subsidizing investors might promote star network formation. In our environment, investors earn less per round than linkers, yet stars form in our environment absent

subsidies[32]. One explanation may be that institutions that enhance decision stability allow participants to rotate their network position and thus maintain a high level of overall network efficiency while concurrently equalize earnings.

It seems clear that focal points can improve coordination and promote the emergence of star networks. Goeree et al (2009) may have in part provided such a focal point by assigning heterogeneous payoff functions. Our study shows that focal points in network environments may emerge endogenously. This finding could be of value especially when policy makers are either unwilling or unable to assign focal points to specific people.

Our focus on ex-ante homogenous agents is both an advantage and a limitation of our research. One important question we are unable to address is whether the successful institutions we discovered might promote the "right" star in the presence of heterogeneity, in the sense that the person best suited to be in the center is most likely to hold that position. Similarly important is to understand how attitudes towards risk and uncertainty impact the particular star formed under different institutional arrangements. Would those with a greater tolerance for risk be more likely to become a star's center? Designing studies to answer these questions would be valuable next steps towards a deeper understanding of the formation and efficiency of social networks in natural environments.

[32] The standard deviations of payoff in Seq_L, Sim_L and Sim_L_NoRFR are 7.05, 7.12 and 7.02, respectively. They are significantly lower at 1% level pairwisely compare to the ones in Seq_B and Sim_B (10.48 and 9.05 respectively).

Appendix A. Z-tree experimental interface

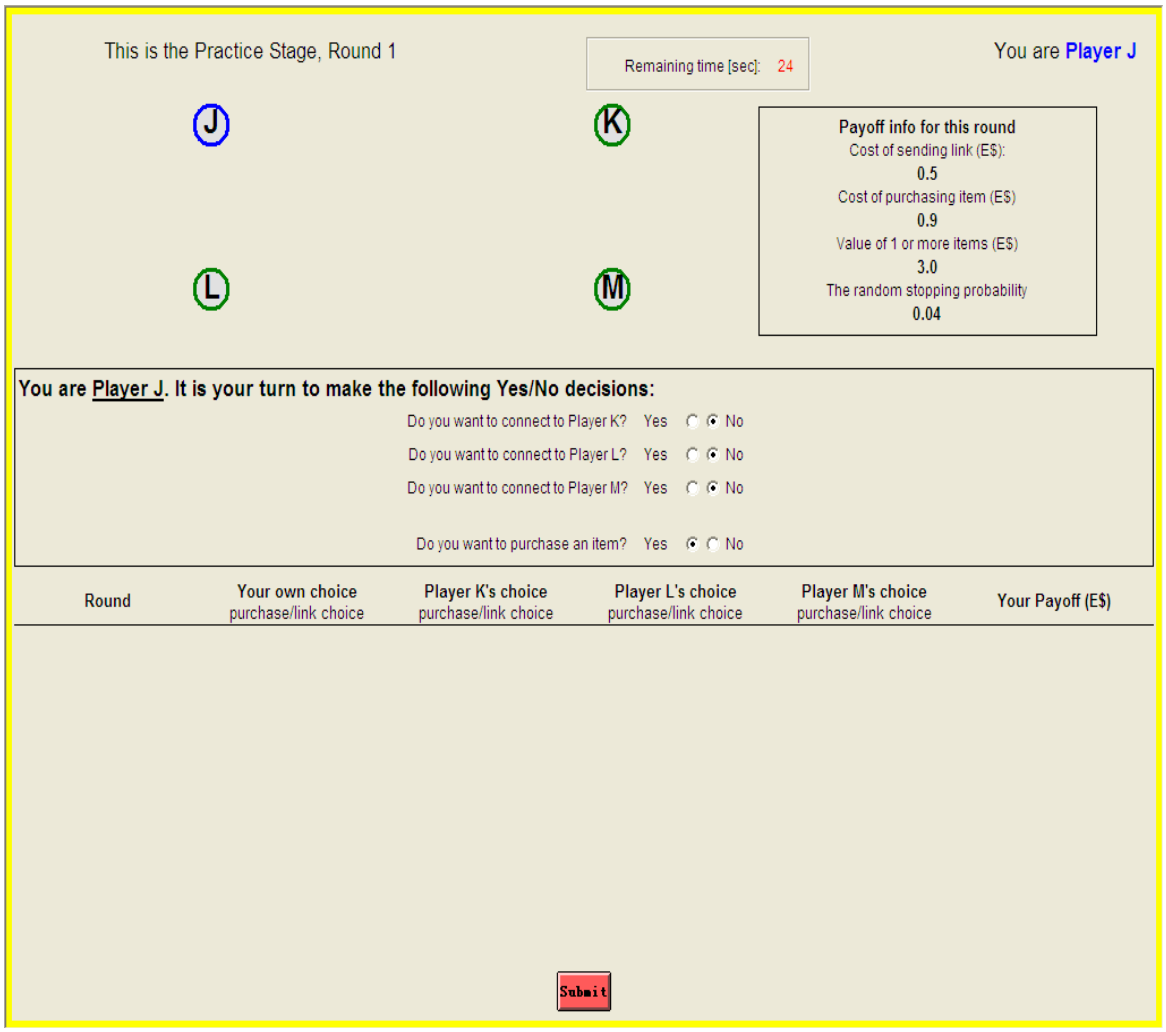


Figure 6 An example of the decision screen

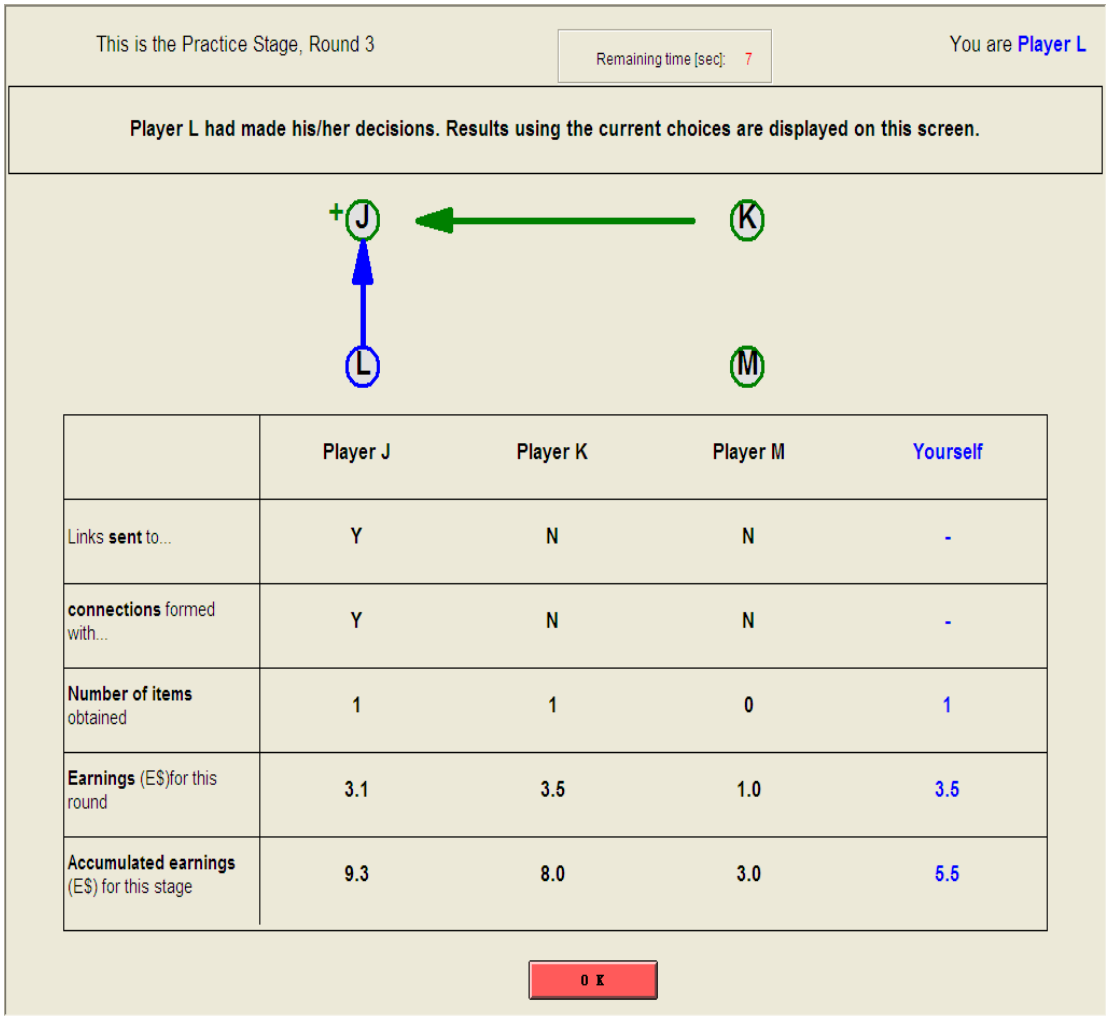


Figure 7 An example of the display screen

Appendix B. 3-space Plot for Individuals' Estimates by Treatment

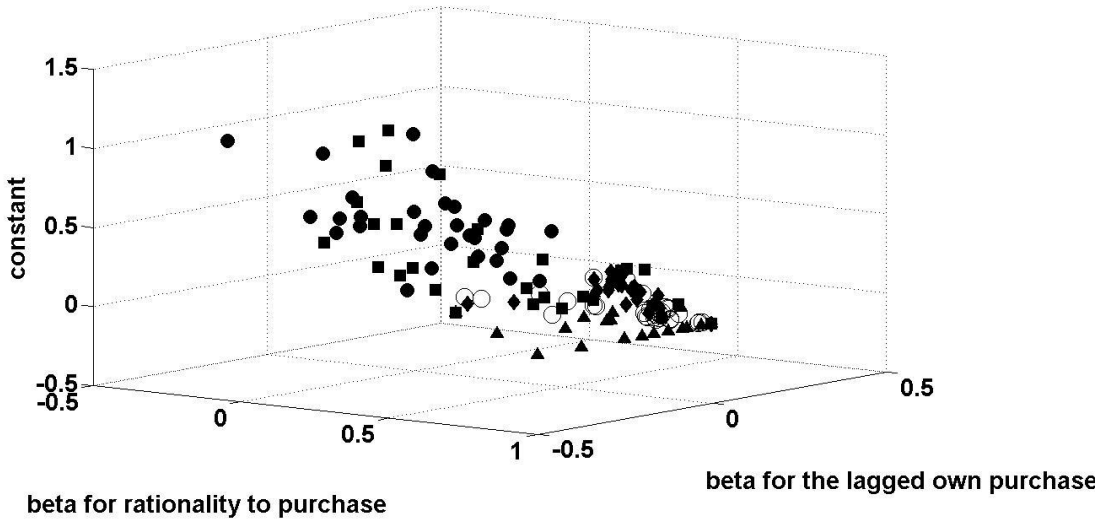


Figure 8 Investing decisions

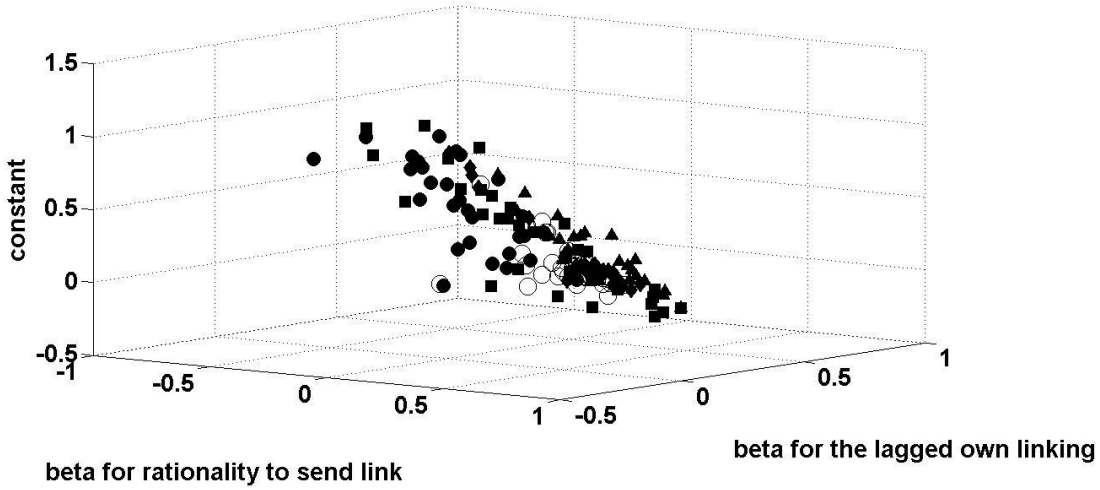


Figure 9 Linking decisions

Note: different markers represent different treatments
■ -- Seq_B ; ▲ -- Seq_L; ● -- Sim_B; ♦ -- Sim_L; ○ -- Sim_L_NoRFR

Appendix C. 3-space Plot for Individuals' Estimates by Cluster

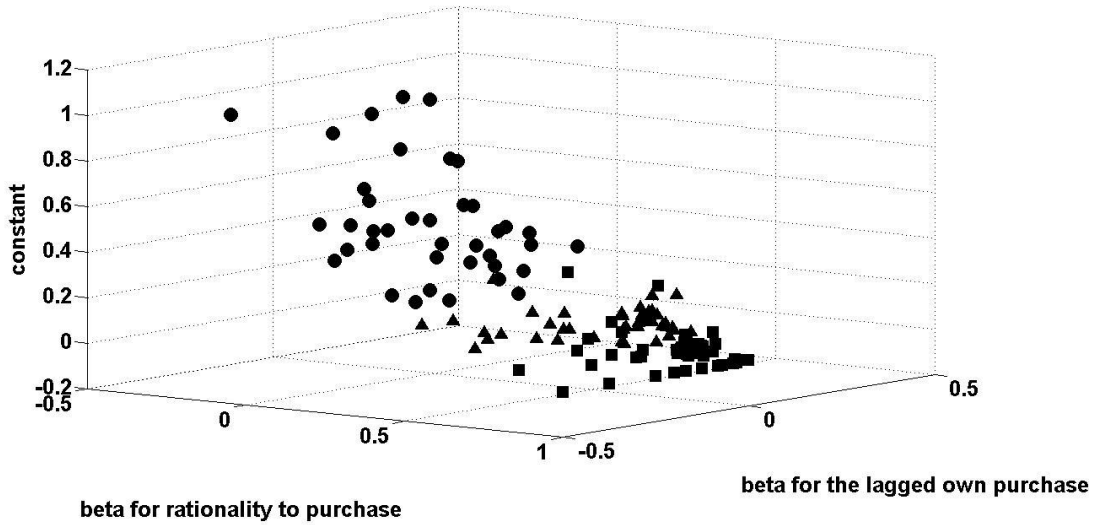


Figure 10 Investing decisions

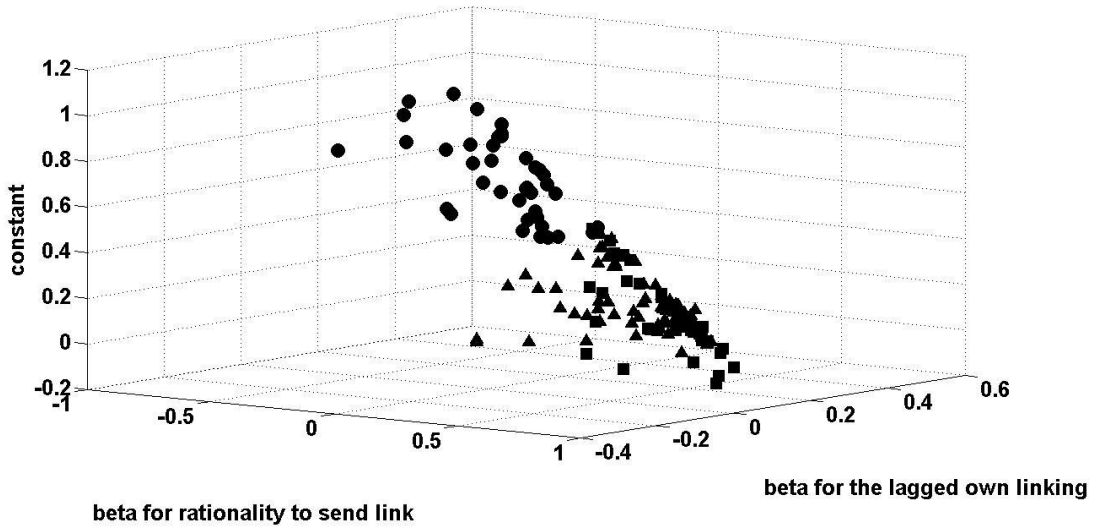


Figure 11 Linking decisions

CHAPTER TWO: MONEY OR FRIENDS: SOCIAL IDENTITY AND DECEPTION IN NETWORKS

I. Introduction

Groups with different financial incentives often deceive each other. Within a firm, for example, people from different departments often manipulate the information they send to each other so that the executive decisions could be in favor of themselves. This phenomenon has been discussed often in popular writings (Cloke and Goldsmith, 2000; Cowan, 2003; Tobak, 2008) and studied widely in the field of industrial and organizational psychology as well as management (Colb et al, 1992; Rahim, 2000; Dreu and Galfand, 2007; Conrad and Poole, 2011; Miller, 2011). It exists at many important industries including high-tech research and development, mass media, health care, etc (Gupta et al, 1985; Eckmen and Lindlof, 2003; Pirnejad et al, 2008). In one recent case, Apple's CEO, Tim Cook, has to issue his official apology to the users of the Map application as Apple's marketing team has made promises that could not be fulfilled by its engineering team. The results of misinformation can be detrimental to an organization.

The above cases can be described by sender-receiver games, which are environments characterized by two groups of people with misaligned monetary incentives. The seminal work by Crawford and Sobel (1982) describes a one sender and one receiver case (it is also called strategic information transmission game, or cheap talk

game)[33]. Many other studies have looked into the variation of this model, but not until Galeotti, Ghiglino and Squantani (2011), players either make decision as a sender or as a receiver, but never both[34]. Galeotti et al (2011) look at N player communication in a networked setting where one can send cheap-talk messages to others and also received messages from others. Their model generates sharp predictions when the players are divided into two groups. In the "two group model", players share the same payoff function within a group and differ in the payoff between groups. The model predicts that the truthfulness of the messages will react to the membership of the monetary group. In particular, truth-telling within a monetary group is more frequent than between groups.

Our study aims to test these predictions of the "two group model". We choose to follow the model since it resembles real world examples that we are interested in. It includes the intra-organizational conflict case we mentioned earlier, the diffusion of political opinion in a large population as well as the case of advice giving on financial and medical decisions. We design our baseline experiment to test the two-group cheap talk model in the lab.

Deceptive communication may respond not only to monetary but also to social incentives. Many past studies provide evidence that social identity impacts economic behaviors including charity and envy (Chen and Li, 2009), punishment (Bernard et al, 2006), cooperation (Goette et al, 2006; Charness et al, 2006; Brewer, 1999), self-esteem (Shih et al, 1999) and contributions to public goods (Eckel and Grossman, 2005). Would

[33] Experimental studies support this prediction includes Dickhaut et al (1995), Blume et al (1998), Blume et al (2001), Cai and Wang (2006) and Wang et al (2010).
[34] An exception is Hagenbach and Koessler (2010).

players' social identity motive affect one's decision to lie? This remains an open question in identity literature[35].

This paper designs a laboratory experiment to investigate the joint impact of social identity and monetary incentives on deception in group environments. First, we investigate whether one will lie to achieve higher monetary gain in the network sender-receiver environment. We further study how the choice of deceiving others is impacted by introducing (non-monetary) social identity.

A laboratory analysis is ideal for our study. The reason is that in natural environments it can be difficult to identify separately the effects of monetary incentives and social identity since (1) shared social identity may form around similar monetary incentives; (2) people have many social identities (e.g., gender, ethnicity, age) and it can be difficult to know which identity is salient during a decision process. Our laboratory study enables us to overcome these problems because identity is induced (Tajfel et al, 1971; Chen and Li, 2009). Then, by randomly assigning players with different identities to different incentive groups, we observe choices made under all relevant incentive-identity scenarios[36]. This design, therefore, enables us to identify the separate effects of "money" and "friend" on deception, and to compare the relative sizes of these two effects.

Our main findings are as follows.

[35] Many studies found gender different in deception or how people perceive deception. Using fMRI data, Marchewka et al (2012) suggests that gender different in deception may be independent from the identity aspect.

[36] The four scenarios are: same-incentive-same-identity, same-incentive-different-identity, different-incentive-different-identity, and different-incentive-same-identity.

1. Absent identity, consistent with theory, truth-telling nearly always occurs among those with identical monetary incentives. Truth-telling also occurs to a great extent when monetary incentives are mis-aligned. In particular, while theoretically people should tell the truth exactly half the time in these cases, we find truth-telling to occur at rate 74.5%.

2. Introducing social identity may not promote truth-telling. We find that sharing an identity does not increase the frequency of truth-telling. One is more willing to lie, however, to those holding a different identity, and this is true regardless of the nature of the monetary incentive.

3. Players become more trusting in the presence of social identity. This seems to suggest people are unable to recognize that introducing identity leads to more lies.

To our knowledge, we are first to provide empirical evidence on behavior in sender-receiver games with multiple senders and multiple receivers[37]. Despite the many insights gleaned from one-sender-one-receiver cases, extending the strategic information transmission to a group context is also important. The reason is that much communication occurs in groups of multiple people who may hold divergent preferences. Our study informs such environments and may aid in the design of institutions and organizations to foster more truthful transmission of information and reduce conflict within organizations.

[37] A few studies look at environment where there are one sender and two receivers (Battaglini and Makarov ,2011) or where there are two senders and one receiver (Minozzi and Woon, 2011; Lai, Lim, and Wang, 2011). Those studies differ from ours as players in those experiments make decisions as either a sender or a receiver, but never both. We focus on a game that better describes the environment of intra-organizational communication, which is characterized by having each player as both sender and receiver.

As we discuss further below, our results may also suggest ways to design strategies for maximizing the transmission of truthful political information.

The remainder of the paper is organized as follows: The next section briefly reviews the theoretical and experimental literature. Section 3 lays out the theoretical background of the study. Section 4 presents the experimental design and procedure. Section 5 sets up the hypothesis and reports experimental results. Section 6 concludes.

II. Literature on Deception and Social Identity

There have been a number of economic theories and experimental tests on sender-receiver games. First, we review these theories. Then, we discuss the experimental evidence, in particular, the recent literature on deception. Finally, we review economic experiments on social identity.

II.1 Theory of Cheap Talk Game

Information is often delivered in a strategic way. When the information holders do not have the same incentive as the uninformed decision maker, they tend to hold back some but not all of the information so that they gain advantage in the transaction. This important economic phenomenon is first described in the seminal model by Crawford and Sobel (1982). In their paper, a sender has the full knowledge of the state of the world and can send messages to influence a receiver's belief so that he or she may make a choice that benefits the sender. The receiver, or course, reacts to the possibility of manipulation in senders' messages and chooses an action that maximizes his or her own earnings.

The model predicts a partitioned equilibrium, where the larger are the payoff differences between the two players, the coarser is the partition, meaning senders hold

back more truthful information. In the extreme case where the payoff difference is too large, senders are predicted to send random messages. That is to say, senders engage in cheap talk.

The seminal work by Crawford and Sobel (1982) has been extended in many directions. For example, Milgrom and Roberts (1986), Gilligan and Krehbiel (1989), Austen-Smith (1993), Krishna and Morgan (2001a, b) investigate the case where there are more than one sender for each receiver. Battaglini (2002) and Ambrus and Takahashi (2008) further extended the analysis to environment where the senders are giving advice on multidimensional issues. Morgan and Stocken (2008) study the case of polling where each sender has a different information and ideology. Also, Farrell and Gibbons (1989) discussed the case where there are two receivers and two states of the world. In all of these cases, however, each agent plays either as receiver or as sender. Note that this contrasts with our study, where each person is both a sender as well as a receiver.

We are aware of only two models of strategic information transmission in networks, where each person can act as both sender and receiver (Hagenbach and Koessler, 2010; Galeotti, Ghiglino and Squantani, 2011). Many real world environments would seem to require this framework. For example, workers from different departments at the same company often talk and listen to each other, and people of different political opinions may also mutually exchange information.

Hagenbach and Koessler (2010) investigated a case where each individual receives some information and the aggregation of all private signals equals to the truth. In their environment, a player earns more when (1) choosing a number that is closer to the

true state of the world plus an individual bias and (2) choosing a number that is closer to others' choices. The first part incentivizes the individuals to make the best guess of the truth, while the second half requires coordination of choices between players. Like in other cheap talk games, players send messages free-of-cost before choosing the numbers and all messages are non-verifiable.

Galeotti et al (2011) also models group communication but the earnings in their model are defined in a different way. In particular, their model assumes that a player earns the highest payoff if everyone in the game, including him/her self, chooses a number that matches the truth plus his/her own bias. Since different players have different biases, one may try to affect others' beliefs using cheap talk messages. The predictions of this and Hagenbach and Koessler (2010) are quite similar. Since we build mostly from Galeotti et al (2011), we make clear the details of their model in Section III. We choose to use Galeotti et al (2011) because it resembles an intra-organizational communication and decision making environment where everyone's choice directly affects each player's payoff, the situation in many important environments discussed above and of interest to us.

II.2 Deception Experiments

The earlier experimental literature on cheap talk game studies how well the empirical data matches the prediction of Crawford and Sobel (1982). In those games, sender can choose vague messages by sending a range of possible states (e.g. sending (1-3) when signal is 2.). Dickhaut, McCabe and Mukherji (1995) confirms the comparative statics of the model by showing that the senders' messages become vaguer and the

receivers' actions deviate more from the true state as preferences between sender and receiver diverge. Cai and Wang (2006) replicated the above finding and further show that the average payoffs of senders and receivers are very close to the predicted level for the most informative equilibrium. Their data also suggest that senders over-communicate and receivers over-trust the message. Wang, Spezio and Camerer (2010) study the source of over-communication using eye-tracking data.

Some recent experimental studies use a simplified sender-receiver game to study deception behavior in the lab. Gneezy (2005) analyzed an experiment where there are only two states of the world. They found that people are sensitive to both their own gain and others' losses when deciding to lie. Lundquist et al (2009) modified the game further into a labor contract context, where the senders have information on their ability level and face an incentive to lie so the receiver will agree to hire. With this design, they can observe not only whether a player has lied but also the size of the lie. They found that lie aversion increases with the size of the lie and also the strength of the promise. The data also show evidence that free form messages leads to fewer lies and more efficient outcomes. Typically in this literature[38], messages are considered deceptive if a sender's message contains other than true state of the world. We also use this method to analyze deceptive messages[39].

[38] Sutter (2009) and Xiao (2012) take into account "sophisticated" deception if a deceptive sender chooses the true message with the expectation that the receiver will not follow his/her message.

[39] To identify a players exact strategy (truth-telling or cheap talk) requires repeated observations. The result may be ambiguous if individuals switch between different strategies during the game. We do not try to identify players' strategy but simply study the frequency of deceptive messages.

II.3 Parochial Altruism & Social Identity

Whether non-monetary gains can lead to less deception has not yet been studied. However, many experiments on the effect of social identity suggest that might be the case. Chen and Li (2009) used artist preference to divide people into identity groups and found that people are more altruistic towards those of the same group. In particular, they show artificial identities make people reward more and punish less towards in-group members. People of same identity also choose more social-welfare maximizing actions, which results in higher expected earnings. "Parochial altruism" is also found among indigenous people in Papua New Guinea (Bernhard, Fischbacher and Fehr, 2006). Their subjects tend to favor people of the same tribe by giving a higher transfer amount in dictator games and punishing more when the unfair dictator is from another tribe. Identity also affects cooperation. Eckel and Grossman (2005) found that with strong team identity priming, players of similar identity could achieve higher levels of contributions in public goods experiments.

Charness, Rigotti and Rustichini (2007) found making social identity salient leads to aggressiveness. In particular, the presence of inactive players with the same identity results in more coordination in the Battle of Sex game and less cooperation in the Prisoners' dilemma game. However, part of their study used shared monetary payoffs to strengthen social identity. Since we are trying to distinguish the effect of monetary and social identities, we adopt the artist preference method used in Chen and Li (2009) so that the assignment of social identity is unrelated to monetary payoffs.

It is plausible that in-group altruism, cooperation and aggressiveness may lead to fewer lies within an identity group and more lies between groups in a game of strategic

information transmission. However, to our best knowledge, there has not been any study that directly looks at the effect of social identity on deception in network environments. Our study fills the gap.

III. Theoretical Background

Our experiment design follows the two-group communication model in Galeotti, Ghiglino and Squantani (2011). We review the detail of the model in this section.

The sets of players is denoted by $N=\{1,2,\ldots,n\}$ partitioned into two groups, N_1 and N_2, with size n_1 and n_2, respectively, where n1+n2=n. Without loss of generality, assume $n_1 > n_2 \geq 1$. Player i's individual bias is b_i. In two-group communication model, each members of group 1 has a bias normalized to 0; members of group 2 have a bias $b_i=b>0$. The state of the world θ is uniformly distributed on [0, 1]. Every player *i* receives a private signal $s_i \in \{0,1\}$ where $s_i=1$ with probability θ.

Communication among players is exogenously restricted by a communication network $g \in \{0,1\}^{n\times n}$ where player *i* can send message to *j* if $g_{ij}=1$ with $g_{ii}=0$ for all $i \in N$. The communication neighborhood of *i* is the set of player to whom *i* can send his signals and it is denoted by $N_i(g)=\{ j \in N: g_{ij}=1\}$. In this study we focus on the case where *g* is a complete network, meaning players can send a message to every other player.

Communication mode describes to what extend the technology of communication allows to target messages. In a private message setting, player *i* chooses what message to send to each other player *j*. A communication strategy profile for each signal $s_i \in \{0,1\}$ is defined as $m=\{m_1,m_2,\ldots,m_n\}$ in which $m_i(s_i)=\{m_{ij}\}_{j\in N, j\neq i}$.

After communication occurs, each player chooses an action. Agent *i*'s action strategy, based on his/her own signal and messages received from others, is y_i:$\{0,1\}^{n-1}\times\{0,1\}\rightarrow R$; $y=\{y_1,y_2,\ldots y_n\}$ denotes an action strategy profile. Given the state of the world θ and a profile of actions $\hat{y} = \{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n\}$, the payoff of *i* is:

$$u_i(\hat{y}|\theta) = -\sum_{j\in N}(\hat{y}_j - \theta - b_i)^2 \qquad (11)$$

That is, agent *i*'s payoffs depend on how close his own action y_i and the actions taken by other players are to her ideal action $b_i + \theta$.

A communication network g together with a strategy profile (m, y) induces a subgraph of g in which each link involves truthful communication. They refer to this network as the equilibrium truth-telling network denoted by c(m,y|g), a directed graph where $c_{ij}(m, y|g) = 1$ if and only if j belongs to i's communication network and $m_{ij}(s) = s$ for every s={0,1}. Given c(m,y|g) and that the agents are divided into two groups, the in-degree of an arbitrary player in group i, k_i , is defined as the number of agents who send a truthful message to him/her. Among all the truthful messages, the amount sent by members of the same group is denoted by k_{ii}, while the amount sent by members of opposite group is k_{ij}.

Their analysis focuses on pure strategy Bayesian Nash equilibrium. They provide a full characterization of the utility-maximizing equilibrium networks[40] with a focus on the natural subclass of those networks where there is complete intra-group communication. In our experiment setting, the bias we choose will yield the same

[40] It is a tradition in strategic information transmission models to characterize the utility maximizing equilibrium as babbling is always an equilibrium solution but not meaningful in most contexts.

prediction whether we decide to use the full characterization or the subclass. The following equation describes the in-degree of an arbitrary player in group *i* in the utility-maximizing equilibrium truth-telling network:

$$k_{ii} = n_1 - 1 \qquad (12)$$

$$k_{ij} = max\left\{min\left\{\left\lfloor\frac{1}{2b} - n_i - 2\right\rfloor, n_j\right\}, 0\right\}, i, j = 1,2, i \neq j \qquad (13)$$

That is, if $b < \frac{1}{2(n+2)}$, both intra-group and inter-group communication is complete; and if $b > \frac{1}{2(n_2+2)}$, there is complete intra-group communication and no inter-group communication. When b takes the intermediate value, inter-group communication also takes intermediate value[41].

Given this type of communication, in equilibrium all players trust all intra-group communication. They treat inter-group messages as true signals whenever $< \frac{1}{2(n+2)}$, and as no information whenever $b > \frac{1}{2(n_2+2)}$. That completes the equilibrium prediction of the model.

IV. Design and procedure

IV.1 Baseline treatment

The design of our baseline treatment is based on a game introduced by Galeotti et al (2011). There are multiple games in their paper, and we adopt the one that is characterized by private communication between two groups with different biases in their payoff function. We choose to study this game as it is highly relevant to the real world

[41] Specifics related to intermediate biases can be found in Galeotti et al (2011).

network communication problem that interests us. To our best knowledge, our paper is the first to examine how people choose to transmit information in this environment.

Each experimental session includes 15 subjects. Each five are randomly assigned to play the game. All subjects participate in three stage games. Each stage game consists of a random number of rounds[42]. Players know that the other four players are fixed during each stage game, and each of them holds a unique ID: J, K, L, M or N. Player J, K and L belong to Group 1. Player M and N belong to Group 2. Group 1 and Group 2 players differ in their payoff function by only one parameter: the bias. And the biases are common knowledge for all players.

Each round of the experiment is a guessing game. Before a round starts, the computer generates a random integer ***r*** between 0 and 5 (including 0 and 5). The number is unknown to all players. At the beginning of each round, each player receives a private signal that is either 0 or 1. Players do not see others' signals. However, they are told that the sum of the five signals received by all five players equals to the random integer[43]. Before players guess the number, they are given the opportunity to exchange "cheap-talk messages" between each other. The messages are constrained to be either 0 or 1 to match the space of the signal. Moreover, messages are group specific, so each player decides on what message to send to Group 1 and Group 2 players rather to the message for each

[42] There are always at least 4 rounds in a stage. After round 4, the game has a random stopping probability of 0.04 at any given round. To keep control over the length of the real experiment, we randomly generated predetermined round lengths of 19, 28 and 32 for experimental stages I, II and III respectively. The practice stage lasts 3 rounds.

[43] This part of design follows Hagenbach and Koessler (2010). We deviate from Galeotti et al (2011) for two reasons: (1) the former involves less uncertainty therefore is an easier task for our subjects and (2) the main predictions that we test in this paper remain the same between the two models.

player[44]. After all players submit their messages, they observe the messages that are sent to them and are asked to guess the value ***r*** randomly chosen at the beginning of the round. They also choose a number ***x*** based on their guess of ***r*** to determine everyone's payoff for that round. The payoff function for Group 1 and Group 2 players are as follows[45]:

$$Payoff_{J,K,L} = 20 - \sum_{i=1}^{5} (x_i - r - b_1)^2 \qquad (14)$$

$$Payoff_{M,N} = 20 - \sum_{i=1}^{5} (x_i - r - b_2)^2 \qquad (15)$$

Player J, K and L share the same payoff function as shown in equation &&&. The payoff is maximized when **all five players**, including him/herself, choose the number ***x*** that equals the true value of the random number ***r*** plus a group-specific bias b_1 . Player M and N share the same payoff function as shown in equation %%%. The difference between their payoff functions and the ones for Group 1 players is the group-specific bias b_2. As indicated in the theory, this payoff structure incentivizes every player to (1) choose a number ***x*** that is as close as possible to their best guess of ***r*** plus their own group's bias ***b*** and (2) make other players, both in the same group and in the different group, to choose the same ***x***. The presence of cheap talk messaging makes it possible for players in one group to manipulate the choice of ***x*** made by players in the other group. In our experiment setting b_1 and b_2 can only takes four different values, that is (0, 0), (0, 1), (1, 0) or (1, 1). Note that (1, 1) appears always and only in the practice stage, and therefore is not included in our data analysis. The other three combinations appear in random order

[44] The message is group specific in order to simplify the decision problem for the subjects.

[45] The payoff differs from the theory section since we give 20 experimental dollars as an endowment per period. This ensures subjects do not earn negative amounts during the experiment. This change does not alter the theoretical predictions.

for the three experimental stages. The structure of the game and all payoff-related information, including the value of b_1 and b_2, are common knowledge. Players also know that the value of b_1 and b_2 remain fixed within a stage game, but change between stages.

The following three screens implement this design. First, subjects send messages using the "messaging screen" (see Appendix A, Fig. 1). Then, subjects make guesses on the random integer r and choose the payoff relevant value of x on the "guessing screen" (See Appendix A, Fig 2). While they are making these two choices, the same screen also shows them the messages they received from others graphically. Finally, the "result screen" (see Appendix A, Fig 3) reveals the true value of the random integer and displays all the actions taken by the other four player and their current payoff.

Payoffs accumulate within, but not between, each of the three stage games. Players are informed about their accumulated payoff at the end of each stage. They are also reminded that they will be re-matched with a new set of players, and that their stage payoff will not be carried over to the new stage. Each subject's earnings for the experiment are determined by one randomly-determined stage game according to a die roll at the end of the experiment.

IV.2 Identity Treatment

The identity treatment differs from the baseline treatment described above in that it includes an identity priming stage at the beginning of the experiment. We are using the artist preference as the method of priming, as first introduced by Tajfel and Turner (1979) and then reintroduced by Chen and Li (2009). The procedure and the paintings we are using follow the latter (see Appendix 2 for the paired paintings). Subjects are presented

with five paired paintings sequentially. Within each pair, one painting is a Kandinsky and the other a Klee. Subject can indicate their preference for each pair and are told that they will be assigned to an "artist team" if they choices show that they prefer one artist more frequently (above 3 out of 5 choices). After their team assignment, they are given another two new paintings by Kandinsky and Klee and are asked to guess the correct ownership of the artwork within 5 minutes. Each correct answer earns an additional E$40. People who are assigned to the same artist team can exchange free form text through a chat window[46]. This chat design is used in Chen and Li (2009) to strengthen the identity.

Once the identity priming is completed, instructions for the baseline game are distributed. The only difference in the instructions is that subjects are told that their artist team identity will become public information and will be displayed during the entire "guessing number game".

Note that players in different "groups" differ in their monetary incentives. Players that are assigned to different artist "teams" only differ in their preference about painters and are randomly assigned to two incentive groups. Therefore, within an incentive group, there may be people on the same or different teams. In this study, we use the word "group" and "team" to distinguish monetary (the former) from non-monetary (the latter) affiliation.

IV.3 Procedures

The experiment sessions were conducted between May 2012 and September 2012 in the ICES laboratory at George Mason University. Subjects were recruited via email

[46] Subjects are told not to reveal information regarding their name, race, age or anything that can reveal their identity.

from registered students at George Mason University. Each subject participated in only one session and none had previously participated in a similar experiment.

In total, 75 subjects participated in the computerized experiment programmed with z-Tree (Fischbacher, 2007). Each experimental session lasted between 120 and 150 minutes. Subjects' total earnings were determined by the Experimental Dollars (E$) earned at the end of the experiment, which were then converted at a rate of E$20 per US dollar. Average earnings before adding the $5 show up fee were $18.40, ranging from a maximum of $29.3 to a minimum of $4.8 across all sessions.

In all treatments, before a session starts, subjects are seated in separate cubicles to ensure anonymity. They were informed of the rules of conduct and provided with detailed instructions. The instructions were read aloud. In order to ensure there is no confusion, after subjects finished reading the instructions they were asked to complete a quiz. An experimenter checked their answers and corrected any mistakes one by one. Then the experimenter worked through the quiz questions on a white board in front of all subjects. The experiment began after all subjects confirmed they had no further questions.

We ran 3 sessions for baseline condition and 2 sessions for treated condition. Within each session, we obtained 97 message sending decisions for each subject (excluding the practice stage). Our analysis conservatively assumes 45 independent observations (27 in the baseline condition and 18 in the treated condition) unless otherwise specified.

V. Hypothesis and Results

V.1 Hypothesis

Our hypotheses stem from Galeotti et al (2011) as well as the literature on social identity. We begin by indicating hypotheses based on Galeotti et al (2011), which will be tested using data from the baseline treatment. Second, we list hypothesis on social identity effects which will be tested using data from both baseline and identity treatments.

Based on the theory discussed in section III, we can make following hypothesis:

Hypothesis T1: Players always tell truth to those in the same monetary group.

Hypothesis T2: Players tell truth to others in a different monetary group if the bias is (0, 0), and always send random message (babble) if the bias is (0,1) or (1,0).

Hypothesis T3: Players fully trust the message sent by their group members and also fully trust the message sent by other group members if bias is (0, 0). They disregard those between group messages if the bias is (0,1) or (1,0).

Since social identity may have a positive effect on truth-telling, we hypothesize as follows:

Hypothesis S1: After priming identity, players in the same monetary group will continue to tell 100% truth. Emphasizing the difference in identity, however, may reduce within-group truth-telling. Therefore, introducing social identity priming can have only a negative effect on within-group truth-telling in comparison to the baseline condition.

Hypothesis S2: The same team identity may increase truth-telling between monetary groups but different team identities may reduce truth-telling between groups. The overall effect may depend on the relative size of these two effects as well as the identity composition of each group.

Hypothesis S3: Hypothesis S.1 states that both within and between group messages are no less trustworthy if the message senders and receivers share the same identity. S.2 states that trustworthiness decreases if senders and receivers share different identities. If receivers fully anticipate this behavior then, in relation to an environment without identity, their guesses will deviate more from the messages they receive.

V.2 Results

We lay out the results in the order of the hypotheses above. First, we discuss the baseline observation in comparison to theoretical prediction (from T1-T3). Then, we compare the results from the social identity treatment with the baseline results (from S1-S3).

Result T1: Without identity, most within-group messages are truthful.

Our data support hypothesis T1. As shown in Figure 12, we found that 95.3% of the within-group messages are truthful in the baseline treatment. Although the overall level of truth-telling seems high, it is significantly lower than the predicted level of 100% ($p=0.001$). Consistent with the theory, the bias of the opposing group does not affect within-group messages in any statistically significant way (pairwise comparisons, all ps greater than 0.856). Moreover, group size also does not impact the truthfulness of within-group messages (96.2% for Group 1 and 94.4% for Group 2, no statistic different, $p=0.412$).

Result T2: Without identity, players tell less truth to members of different group than to their own group members.

Our data support hypothesis T2. In the baseline treatment, 79.0% of messages sent between two groups are truthful. This level of truth-telling is much lower in comparison to within-group messages ($p<0.001$). This effect is larger if the bias is (0,1) or (1,0). However the effect persists even if the bias is (0,0).

In case where bias is (0, 0) the two groups share the same payoff function, so that truth-telling is predicted to be 100%. However, 87.9% of these between-group messages are truthful, significantly lower than predicted. It is also lower than the truthfulness for within-group messages (compare to 95.3% , $p<0.001$), suggesting that simply dividing subjects into two groups has an impact on their truthfulness regardless of monetary incentive[47].

According to theory, under bias (0,1) and (1,0), there only exists a babbling equilibrium with 50% truthful messages. We observe 74.5% truthful messages between groups[48], which is significantly higher than predicted levels ($p<0.001$). Under either bias, truth-telling is significantly lower than the case where the bias is (0,0) (pairwise comparisons, $p=0.018$ and 0.047 respectively).

[47] Eckel and Grossman (2005), however, suggest that minimal group identity does not affect subjects' behaviors in their experimental setting. Our data suggests that the effectiveness may be sensitive to the environment.

[48] We combine the two cases together as the unequal bias case, as there is no significant difference between them ($p=0.652$)

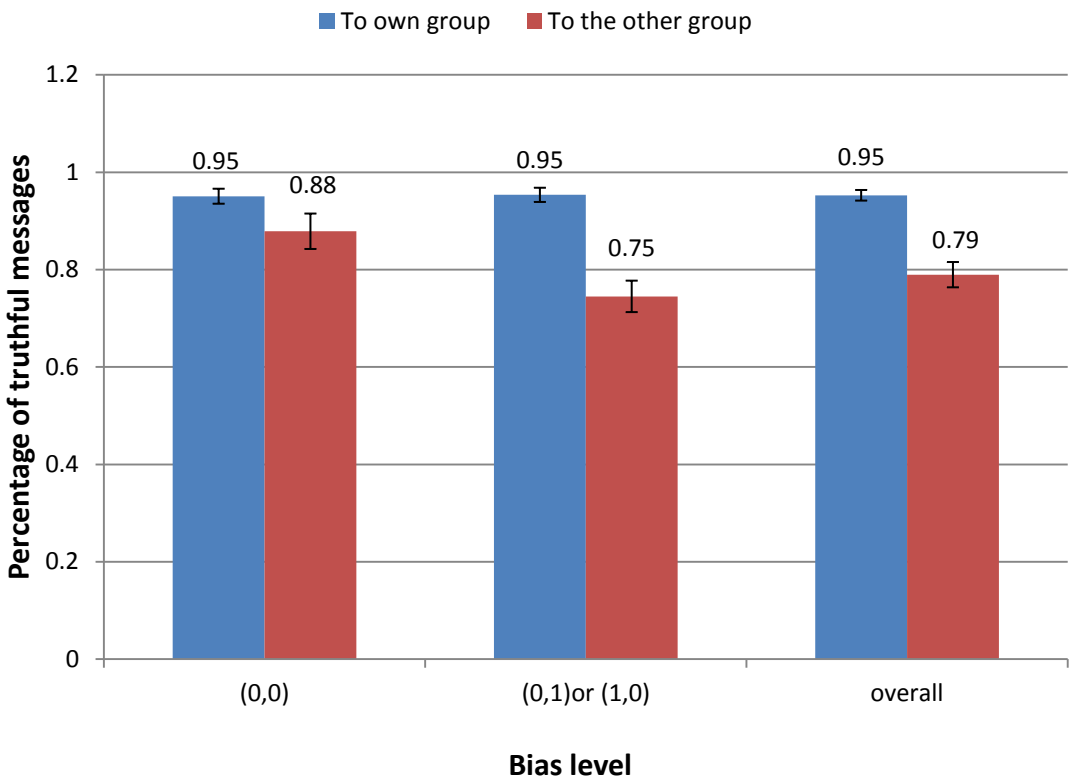


Figure 12 The Percentage of Truthful Messages in Baseline

Result T3: Without identity, players overly trust messages they receive.

To measure whether a player believes the messages s/he receives, we measure the difference between one's guess and the sum of "1" messages received. When the difference is zero, we define the "trust" measure to equal one and set it to zero otherwise. 74.9% of all guesses submitted in the baseline exactly equaled the sum of "1" messages received. When the bias is (0,0), 82.1% of guesses are consistent with the messages received, which is significantly lower than the predicted 100% level of trust.

In equilibrium, when the bias is either (0,1) or (1,0), random choice will lead Group 1 players to appear completely trusting of between group messages 37.5% of the

time. This can be seen as follows. In equilibrium, each player in Group 1 faces four possible message combinations sent at random by two Group 2 players: (0,0), (0,1), (1,0) and (1,1). Each of these four outcomes is equally likely to appear. Group 1 players form beliefs about the true signal that Group 2 players hold independent of these messages: (0,0), (0,1), (1,0) and (1,1). Each outcome is also equally likely to happen. Therefore, out of 16 message-belief pairs with each combination having the same probability, six of the sums can coincide at random (6/16=37.5%). Similarly, Group 2 players may appear to be trusting in 31.25% of time even if they are choosing at random. Overall then, random choice will lead 35% of choices to appear fully trusting. Our data show that 72.1% of guesses are fully trusting, significantly higher than 35%. Moreover, the trust levels between bias (0,0) and bias (0,1) and (1,0) are significantly different ($p<0.001$).

Identity effects are inferred using two metrics. To understand these metrics, note first that any player in the game is related to each other player in the game according to both group affiliation and team identity. Intuitively, the first metric is the frequency with which a player interacts within each type of relationship. It turns out there can be six such relationships: (1) in the same group and shares the same team identity (denoted as SGSI); (2) in the same group but with a different team identity (SGDI); (3) in a different group but shares the same team identity (DGSI); (4) in a different group and with different team identity (DGDI) ; (5) in the same group and have no team identity (SGNI) and (6) in a different group and have no team identity (DGNI). Note that all subjects in baseline will fall into SGNI or DGNI while all identity treatment subjects will belong to the first four categories. We do this categorization for each of the five individuals in the game and then

average across players and finally sum the results over the number of rounds in each stage game.

For example, consider Player J. The relationship between J and K is determined as SGDI if K and J prefer different artists in the identity priming stage. J and L interact as SGSI if L shares the same artist preference with J. However, K and L are always SG in relation to J, while M and N are always DG relative to J. Our procedure determines, for each player, how many times each of the six relationships occurs for each other player in each stage game.

The second metric builds upon the first and calculates, under each of the six scenarios, how many times a particular player sent a true message. If so, we code the message to be one, otherwise zero. For each stage game, we average across all five individuals and all rounds to determine how many times truthful messages are delivered under each of the six scenarios.

Finally, we divide the second metric by the first in order to reveal the percentage of truthful messages sent under each of the six possible relationships. I denote the percentage truth for the same group and same identity as "P_{SGSI}", and similarly for the other five cases. The goal is to compare those percentages across relationships. Our main findings are as follows:

Result S1: Compared to the baseline of no identity, sharing the same identity does not increase within-group truth-telling, but having different identities reduces truthfulness among group members. Introducing identity thus has a detrimental impact on within-group honesty.

This can be shown using the following equation:

$P_{SGDI}<P_{SGNI}=P_{SGSI}$ (16)

$P_{SGWI}<P_{SGNI}$ (17)

Our data support hypothesis S1. As shown in Figure 13, the only significant difference is between the two bars on the right and the four bars on the left, showing that holding different identities reduces truth-telling within a group. Indeed, we observe a significant drop of 31.2% ($p<0.001$) between cases of different identity and no identity.

Further, when breaking down the data by cases of equal or unequal bias, the negative effect of different identity is due to the unequal bias cases. Note that theoretically, within-group messages would not depend on the other group's bias. Our data, however, suggest otherwise: players do consider the other group's incentive when deciding what message to send to their own group. On the other hand, by comparing situations of no identity and of the same identity, we found no significant improvement resulting from introducing the same identity ($p=0.630$). On average, 95.3% of within-group messages are truthful for the former (in baseline data) and the level is 94.2% for the latter. Overall then, the effect of introducing identity is to decrease the truthfulness of within-group messages by 16.2%, and this is significant ($p=0.001$).

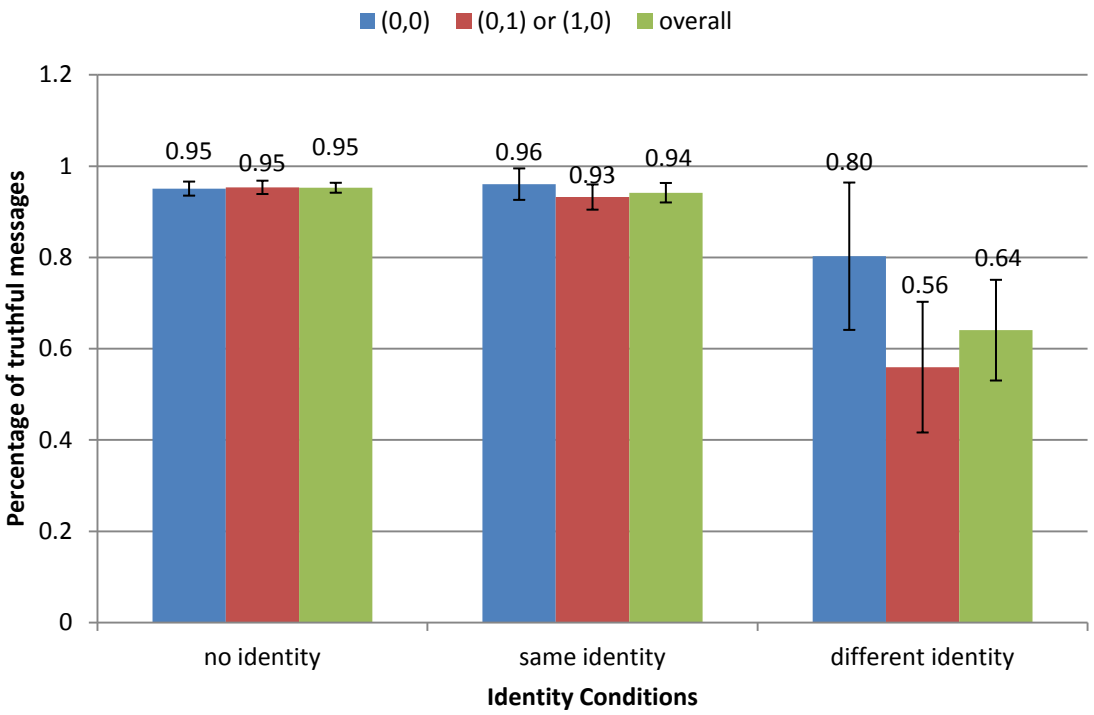


Figure 13 Percentage of Truthful Messages Within Group

Result S2: Introducing identity affects the truthfulness of between group messages in opposite directions. Different identity decreases truthfulness significantly. The same identity increases truthfulness slightly, but the effect is insignificant. Overall, there is no significant change in honesty after introducing identities.

This can be demonstrated as follows:

$$P_{DGDI}<P_{DGNI}=P_{DGSI} \quad (18)$$

$$P_{DGWI}=P_{DGNI} \quad (19)$$

Hypothesis S2 is supported by our data. Introducing different identities decreases the percentage of truth-telling significantly from 79.0% to 56.2% (p=0.013). Holding the

same identity seems to move the measure in the predicted direction: truthfulness increases to 84.3% , but the comparison is insignificant (p=0.281). This result are mainly driven by the cases where bias is unequal, indicating that the decisions based on identity are not independent of the other group's monetary incentives.

This result is consistent "moral wiggle room" (Dana et al 2007). In particular, negative identity may be used as a psychological excuse to lie more to members of another group. Without such an excuse, one may feel guilt as a result of lying for monetary gain. When monetary incentives are aligned, the identity effect is mitigated as players are not able to lie for monetary gain and there is no need to lie or to look for excuses to lie. Figure 14 illustrates this, showing the exact truth-telling percentages for all cases of bias and all relationships.

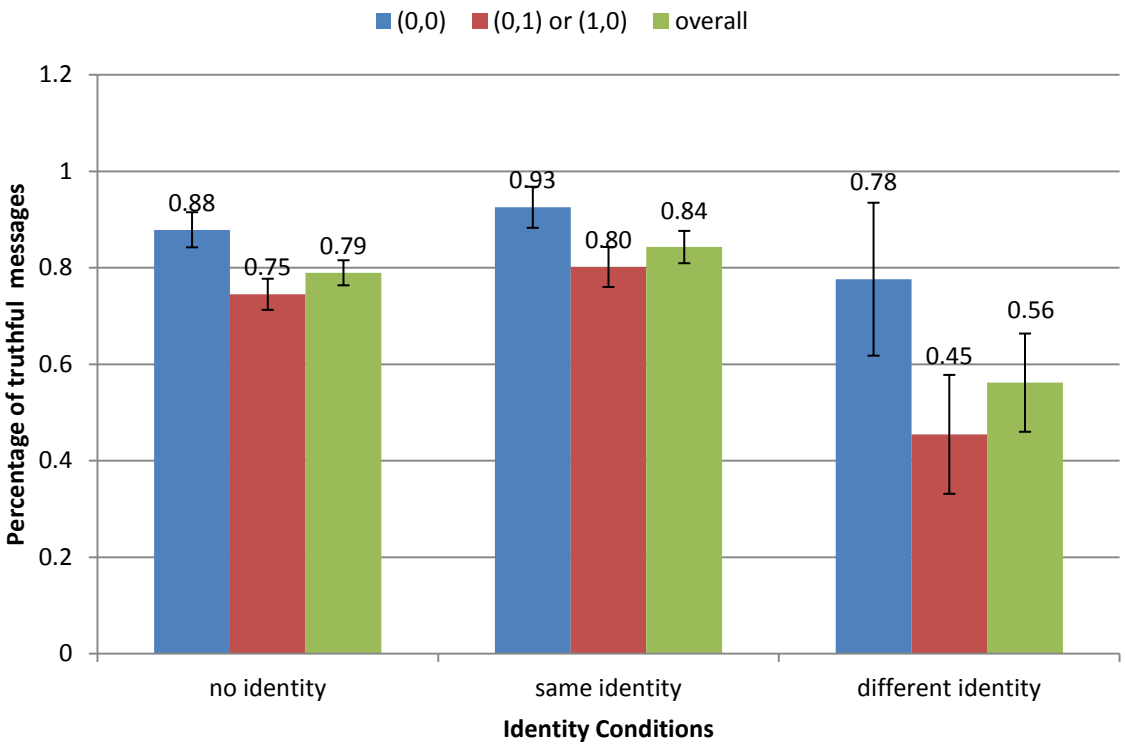


Figure 14 Percentage of Truthful Messages Between Groups

Result S3: Introducing identity leaves players more trusting, in contrast with hypothesis S3.

We use the same measure for "trust" constructed for the analysis of result T3. When the bias is (0,0), we found that 93.5% of guesses are consistent with fully believing in the messages received, significantly higher than the level of trust in baseline treatment ($p<0.001$). When bias is either (0,1) or (1,0), 81.3% of guesses are fully trusting, again significantly higher than the corresponding trust level in baseline ($p<0.001$). Moreover, trust levels under these two conditions are significantly different ($p<0.001$). Combined

with result S2, introducing identity doesn't seem to alter player's overall honesty, but has led to a higher level of trust in others' messages.

VI. Conclusion

Information is transmitted between group members in a strategic way. Both monetary and social incentives may affect the truthfulness of people's messages. Based on a model suggested by Galeotti et al (2011), we conducted a laboratory study of deceptive behavior in an environment of strategic information transmission. We found that absent social identity the message sending behavior of our subjects mostly conformed with theory. In particular, between-group messages were less truthful than within-group ones. However, we found behavior to depart from predictions in that people often overly trusted messages they received, regardless of the presence of social identity. On the other hand, identity has a negative effect on the frequency of truthful information transmission. In particular, we find that the negative effect of holding different identities outweighs the positive effect of sharing the same identity. Interestingly, despite this reduction in truthfulness, subjects trust more in the presence of social identity.

Appendix A. Z-tree Interface

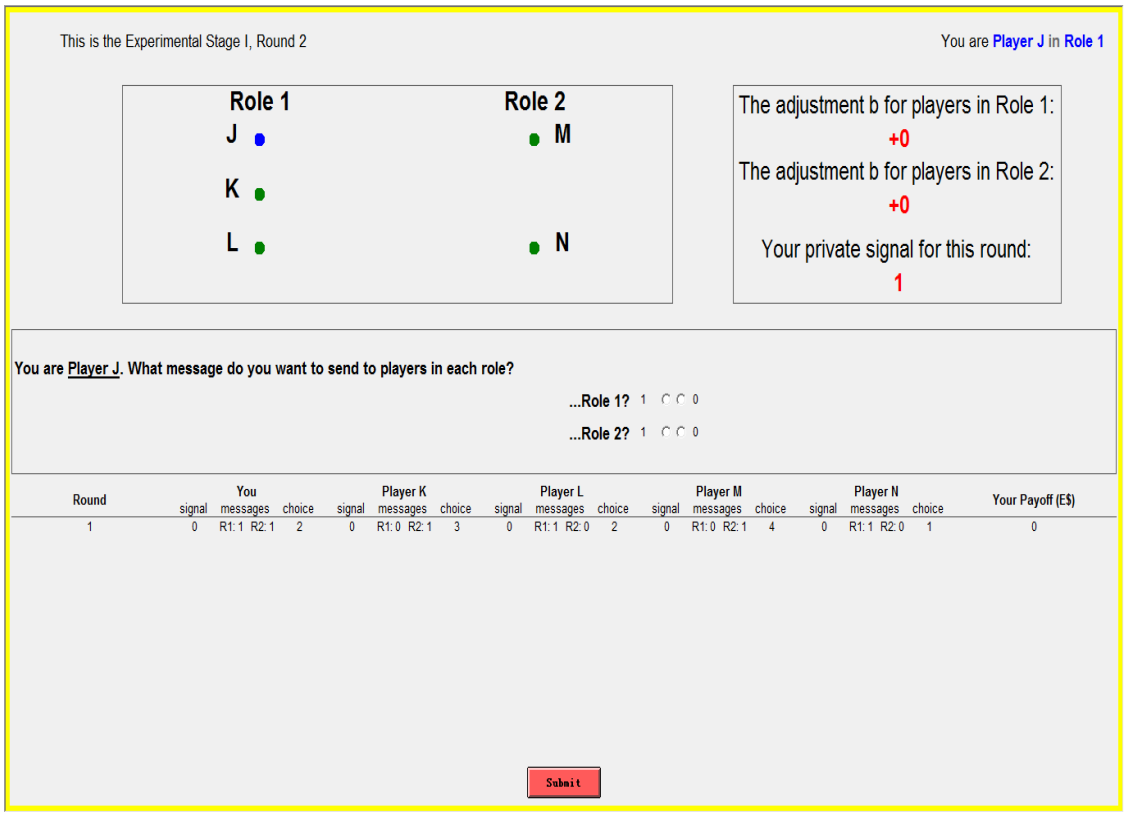


Figure 15 Messaging Screen

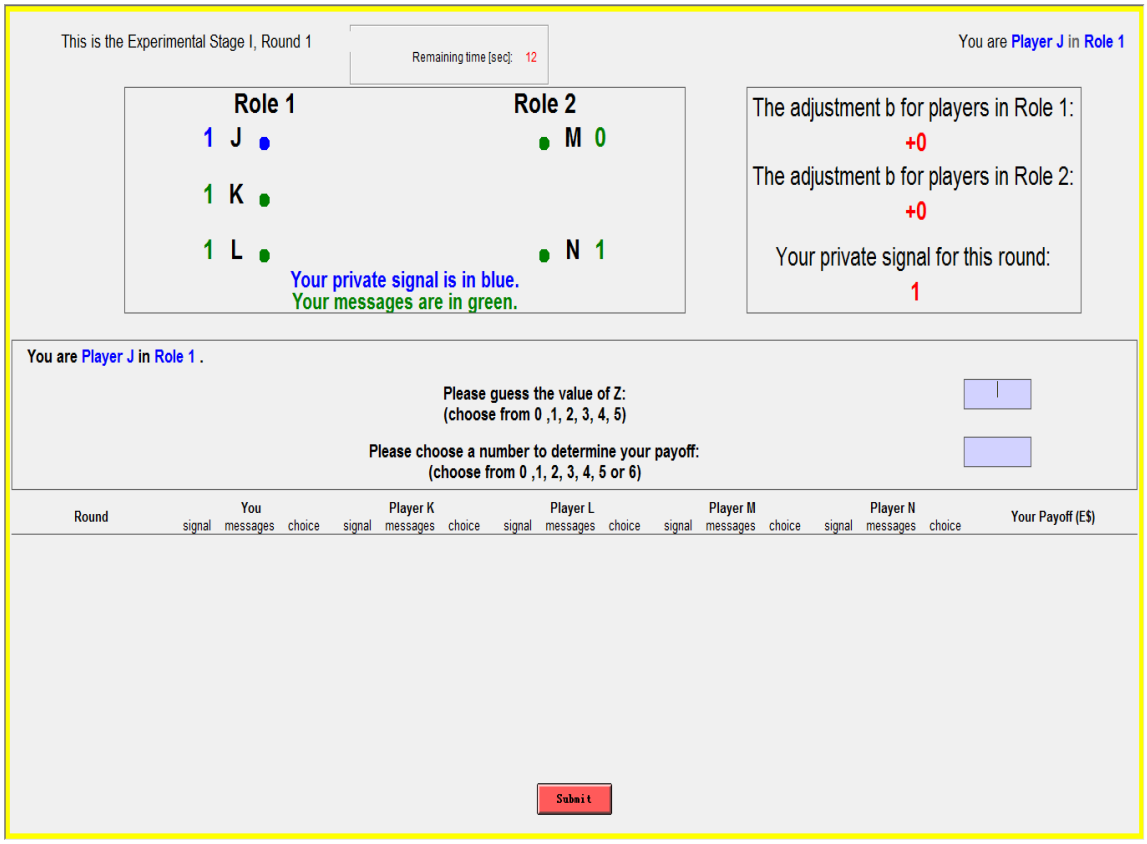


Figure 16 Guessing Screen

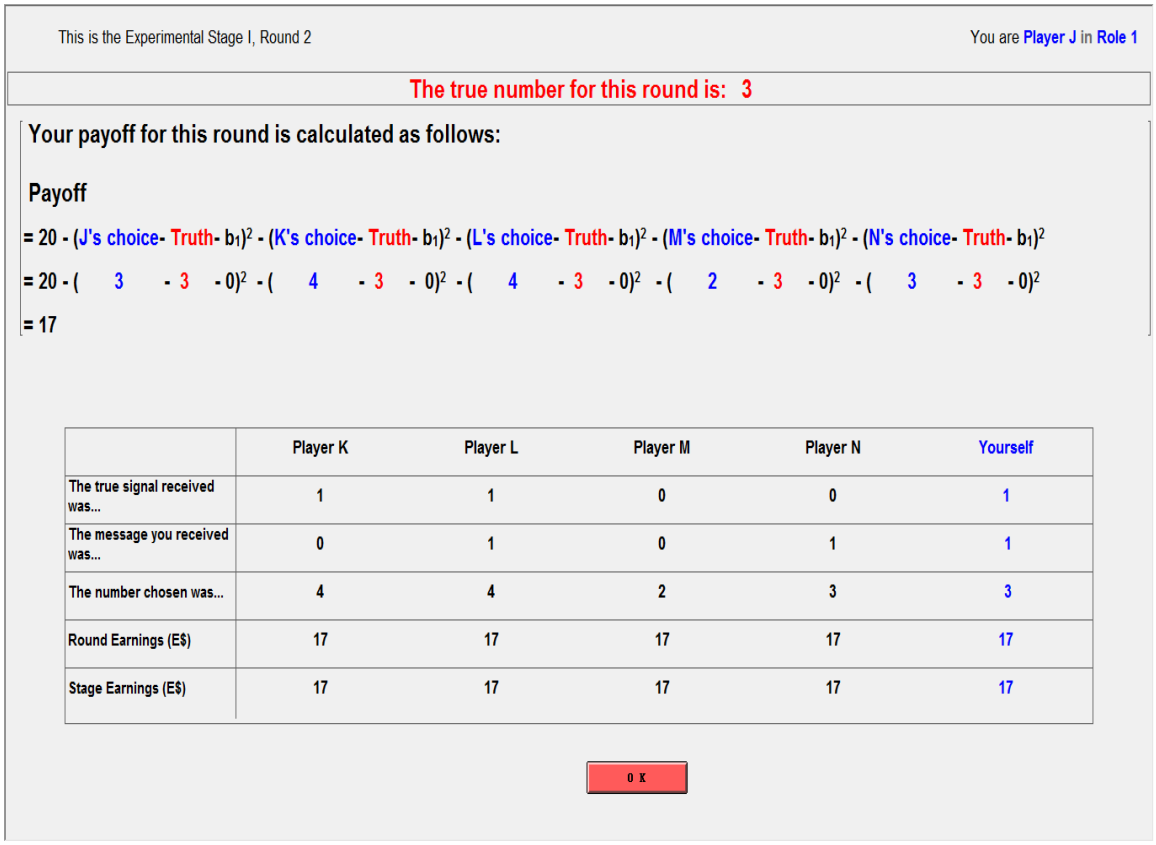


Figure 17 Result Screen

CHAPTER THREE: EXPLORING BEHAVIORAL DATA USING CLUSTER ANALYSIS

I. Introduction

A general question facing researchers in social science is how to classify observations into meaningful groups so that we can better understand the structure of data. When natural features, such as gender, age or income, are obviously driving the change of the variable of interest, we can hypothesize on the direction of change between groups and use statistical methods such as ANOVA or regression analysis to validate or reject such hypothesis. However, such a priori interpretations of data are not always available.

Cluster analysis, as a numerical method for classification, allocates large and complicated datasets into a small number of groups with no need to make arbitrary ex ante assumptions. As early as the 1920s, psychologists were interested in the composition of ability. Some claimed all ability could be explained using two factors (Spearman, 1904), others argued that there were more divisions, such as verbal, arithmetic, memory and spatial. Left unanswered were the number of low-level abilities and the way they relate to each other. This question inspired Robert Tryon to develop the first cluster analysis algorithm, then leading to the development of the first cluster analysis software BC TRY in the 1960s (Tryon, 1932; Tryon, 1935; Tryon and Beiley, 1966).

Since then, numerous mathematical algorithms have been proposed to improve the performance of clustering (Everitt et al 2011). Due to its simplicity and wide applicability, cluster analysis has been commonly used for data analysis in fields ranging from astronomy (Rosenburg, 1910; Babu and Feigelson, 1996 for a review), biology (Kerr and Chirchill, 2001;Witten and Tibshirani, 2010), psychology (Johnson, 1967; Farmer et al, 1983; Borgen and Barnett, 1987; Hay et al,1996) and anthropology (Clarke, 1968; Sutton and Reinhard, 1995), marketing (see Punj and Stewart, 1983 for a review), to increasingly in economics (Fisher, 1969; Hirschberg et al, 1991; El-Gamal and Grether, 1995; Slater and Zwirlein, 1996; Houser, et al, 2004; Yamamori et al, 2008; Adomavicius et al, 2012).

Walter Fisher was the first economist to systematically study the problem of classification. In his 1969 book *Clustering and Aggregation in Economics*, he foretold the increasing complexity of quantification in social variables and stressed "the need for systematic and scientific simplification" of social science data through clustering[49]. The discussion regarding the methods of clustering disappeared in economics for a long time after Fisher's book was published. In 1960s and 1970s, the fields that saw new developments and applications using clustering methods were largely psychology and anthropology.

El-Gamal and Grether (1995) revived economists' interest in uncovering behavioral strategies from complex data. They developed a pseudo-baysisan approach to

[49] The methods reviewed in Fisher (1969) is somewhat different from the cluster analysis defined by its current literature. The author did relate these clustering and aggregation methods to the general literature of cluster analysis.

classify behavioral strategies used by individuals in games. The method is loosely related to finite mixture density clustering. Houser et al (2004) developed a related method in which the nature and the number of decision rules are determined simultaneously.

Substantial time elapsed from Fisher's original work to the time empirical economists began to apply cluster analysis to real-world datasets. Among the few studies that implement cluster analysis, a variety of topics are included. Hieschberg et al (1991) identify clusters for welfare measures across countries using multiple hierarchical agglomerative clustering methods. Slater and Zwirlein (1996) adopt a slightly different hierarchical method using Ward's minimum variance as clustering criteria[50]. They allocated 303 S&P 400 companies into 8 distinct groups in which some were classified as "stable maintainers" and others "leveraged strategists".

Recently, a few experimental economists started to use cluster analysis to identify behavioral patterns among subjects. De Rubeis et al (2007) investigates the difference on the transmission pattern of sexually transmitted disease. The authors clustered individuals based on their demographic and clinical characteristics and separated the social network analysis for each cluster. Yamamori et al (2008) found three types of dictators in a modified dictator game with communication using Ward's minimum variance hierarchical clustering. Adomavicius et al (2012) found that bidders in their auction experiment could be categorized into three behavioral groups using k-means clustering.

The goal of this paper is to review cluster analysis methods that are straightforward and easily implementable. Two key questions must be answered before

[50] The difference and relations between cluster method and cluster criteria will be detailed in Section 2.

implementing any clustering procedure[51]: the method to be used for clustering and the method to find the "correct" number of clusters. As these two decisions are made independently, we review them in separate section of the paper.

We begin with a discussion of various distance measures, separated into measures for categorical data and continuous data. With a particular distance measure, different dissimilarity indices and clustering criteria are developed to formulate the goal of optimization. Since finding the optimal solution can be extremely computationally burdensome, semi-optimal clustering algorithms, such as k-means and k-median algorithms are discussed. Section 2 reviews procedures for cluster analysis and discusses different methods used in each procedure. In addition to the choice of clustering methods, one also needs to choose how to determine the "correct" number of clusters. Section 3 reviews two major approaches to doing this, the Silhouette width and the Calinski-Harabatz index. The final section concludes.

II. Methods of clustering

With optimization cluster analysis one develops indices and criteria to know in a mathematically precise way how "close" or far apart objects are to each other. There are many schools of thought regarding clustering.

One method adopts a bottom-up approach where the closest two objects are grouped first and then a third objects that are closest to the two[52] are added, so on and so forth. This method gradually forms a tree-like cluster result which gives its name

[51] An exception arises when one uses finite mixed density approaches for cluster analysis. In this case both questions are answered at the same time.

[52] Depending on the sub-school of thought, the similarity of an object to a group of objects could be evaluated by the distance of the object from the mean, the centroid, or the farthest or the closest object of the group.

"Hierarchical clustering". The hierarchical cluster analysis has a natural implication in taxonomy where objects bear similarity at different levels and join groups that are not necessarily horizontally comparable. An example is the classification of plants where genus, family and variety are groups formed at different levels of similarity. However, when studying clusters in social science data, researchers are often interested in parallel group structures that contain the entire dataset. This specific goal is achieved with another clustering method, optimization clustering.

The goal of optimization clustering is to allocate optimally all objects into a few groups[53] so that the aggregate distance within a group is small and the distance between groups is large. As this method provides a way to place individuals into flexible decision rule categories, and is straightforward and easily applicable to almost all behavioral datasets, we believe that the method bears relevance to the current discussion.

We introduce optimization clustering by describing each step of the clustering procedure. It starts with distance measures which calculate how close and far apart an object (or a group) is from another object (or another group). Built on the distance measures, we then discuss a variety of (dis-)similarity indices developed to aggregate these distance measures for any particular group. Different similarity indices are then combined to become the goal of the maximization (or minimization) problem. We introduce these goals (also known as optimization criteria) one by one. Finally, we demonstrate how clustering algorithms, like k-means and k-median, provide quasi-optimal solutions for the computationally impossible clustering problems.

[53] The number of groups is a choice variable for the researchers. Methods to choose the number of groups are discussed in Section 3.

II.1 Distance Measures

The starting point of many clustering investigations is an $n \times p$ multivariate matrix X with n observations each of which are described with p distinct characteristics. For behavioral datasets, this can be interpreted as a matrix of n individuals with each individual having p descriptive variables, such as gender, age, choices, etc.

A variety of distance measures have been proposed to measure quantitatively the distance between objects from a set of categorical or continuous observations (see, e.g., Jajuga et al, 2003). Categorical data are usually measured in terms of similarity, while continuous data are commonly measured in dissimilarity (or distance). These two types of measures are mostly interchangeable as they carry the same amount of information regarding distance.

When individual measures are binary, one may use the Matching Coefficient or Jaccard Coefficient as a distance measure. For each pair of individuals, the following table counts the matches and mismatches in the p variables.

Table 6 Counts of matches and mismatches for two individual *i* and *j*

		Individual *j*		
		1	0	Total
Individual *i*	1	*a*	*b*	*a+b*
	0	*c*	*d*	*c+d*
	Total	*a+c*	*b+d*	*p=a+b+c+d*

The Matching Coefficient approach simply calculates the ratio of one-one and zero-zero matches over the total number of characteristics p.

$s_{ij}=(a+d)/(a+b+c+d)$ (20)

Alternatively, the Jaccard Coefficient ignores the zero-zero matches when calculating the similarity. Therefore, the Jaccard Coefficient is:

$$s_{ij} = a/(a+b+c) \qquad (21)$$

This is particularly useful when the absence of a large number of attributes may not necessarily lead to a high degree of similarity. For example, in biology, lacking similar attributes when comparing certain plants with certain insects does not lead to a high degree of similarity between them. Therefore, the principle to choose between the above two coefficients depends on the characteristics of the variables. When co-absence is considered informative, one may use the Matching Coefficient, otherwise the Jaccard Coefficient should be used[54].

When each variable has more than two categories, the similarity measure s_{ijk} is constructed for each variable: when two individual i and j are the same on the kth variable, s_{ijk} equals one, and is zero otherwise. The measure is then averaged over all p variables. The over-all similarity measure between individual i and j is calculated as:

$$s_{ij} = \frac{1}{p}\sum_{k=1}^{p} s_{ijk} \qquad (22)$$

Alternatively, one can also divide multiple categories into two subsets, then convert the original data into binary datasets and finally apply the Matching Coefficient or Jaccard Coefficient approach as in equation 2 and 3. However, whether it is proper to divide categories into two subsets may depend on the specific dataset and the research question one wishes to address.

[54] Similar coefficients have been proposed by Rogers and Tanimoto (1960), Sneath and Sokal (1973) and Gower and Legendre (1986). Their proposed coefficients vary the weight on the mismatches.

When each individual has their characteristics measured as a continuous variable, distance between two individuals i and j are typically quantified by a dissimilarity index d_{ij}. A variety of dissimilarity measures are proposed, among which Euclidean distance is the most commonly used one:

$$d_{ij} = \left[\sum_{k=1}^{p} (x_{ik} - x_{jk})^2 \right]^{1/2} \qquad (23)$$

where x_{ik} and x_{jk} are, respectively, the *k*th variable value of the *p*-dimensional observations for individual i and j. This distance measure has the appealing property that the d_{ij} can be interpreted as physical distances between two *p*-dimensional points $x_i = (x_{i1}, x_{i2} ... x_{ip})$ and $x_j = (x_{j1}, x_{j2} ... x_{jp})$ in Euclidean space. Alternatively, city block distance measures the dissimilarity of individuals on a a rectilinear configuration[55].

$$d_{ij} = \sum_{k=1}^{p} \left| x_{ik} - x_{jk} \right| \qquad (24)$$

Where x_{ik} and x_{jk} are defined in the same manner as it is in Euclidean distance. Both of the above two measures are special cases of the general Minkowski distance with r=2 and r=1 respectively:

$$d_{ij} = \left(\sum_{k=1}^{p} \left| x_{ik} - x_{jk} \right|^r \right)^{1/r} \quad (r \geq 1) \qquad (25)$$

In some cases, the data may contain both categorical and continuous variables. It is possible to construct a single measure by combining distance measures either with or without certain weighting function.

[55] It is also known as the Manhattan distance or taxicab distance as it is measures the travelling distance between two points on the street when city blocks are organized chess-board style.

Notice that even though the distance measures mentioned above for categorical data are measuring distance in similarity while those for continuous data is in dissimilarity, in most cases, these two measure are interchangeable using the following formula[56]:

$$d_{ij} = \sqrt{1 - s_{ij}} \qquad (26)$$

In the following discussion, we assume the distance is measured in, or has been converted to, dissimilarity.

II.2 dissimilarity index

Whichever distance measure one may choose, one can form the dissimilarity matrix D by stacking the distance between all pairs of objects. In behavioral datasets, therefore, each row or column of a dissimilarity matrix corresponds to an individual. Each entry reflects a quantitative measure of dissimilarity between a particular pair of objects.

An informative clustering should include groups such that the distance between objects in the same group is small, while the distance between groups is large. Based on this simple principle, a variety of so-called "dissimilarity indices" (formed by taking combinations of distance measures) have been suggested.

With d_{lv}^{qk} defined as the dissimilarity between the *l*th object in the *q*th group and the *v*th object in the *k*th group, the following equations gives a simple example of an index that measures heterogeneity within group m:

[56] Gower (1966) showed that if a similarity matrix S, with element s_{ij}, is nonnegative definite, then the matrix D, with elements d_{ij} defined by equation 5 is Euclidean.

$$h_1(m) = \sum_{l=1}^{n_m} \sum_{v=1, v\neq l}^{n_m} (d_{lv}^{mm})^2 \qquad (27)$$

Intuitively, this index is the sum of squared dissimilarities between two objects that belong to the same group m.

Another commonly used similar index measures the sum of squared dissimilarities between an object in a cluster group m and the mean of objects in group m. It is also known as the trace of within-group dispersion matrix[57]. This index comprises the foundation for the k-means clustering algorithm which we will discuss later.

$$h_2(m) = \frac{1}{2n_m} \sum_{l=1}^{n_m} \sum_{v=1}^{n_m} (d_{lv}^{mm})^2 \qquad (28)$$

The final index we note here uses the smallest sum of distances to quantify dissimilarity of a group:

$$h_3(m) = \min_{v=1,\ldots n_m} \left[\sum_{l=1}^{n_m} d_{lv}^{\mathrm{mm}} \right] \qquad (29)$$

where a reference object *v* is connected with all other objects in the group *m* to form a star, which then determines the sum of distance of the group. Since the smallest sum of distance is achieved when the reference object *v* is at the center of the group, the index is often referred to as the "star index". $h_3(m)$ index is used in the k-median algorithm.

All three indices mentioned above measure the dissimilarity within the group m and ignore the information about the distance between group m and other groups. Separation indices are designed to capture this information. One commonly used

[57] The dispersion matrix is derived from multivariate matrix X directly without constructing the dissimilarity matrix D. These two methods are mathematically equivalent, hence we omit the discussion of the other method.

separation index takes form $h_1(m)$ but now instead of summing over within group distance, the distance $d_{ml,kv}$ captures the dissimilarity between the object l from group m and the object v from a different group k.

$$h_4(m) = \sum_{l=1}^{n_m} \sum_{k \neq m} \sum_{v=1}^{n_k} (d_{lv}^{mk})^2 \qquad (30)$$

As separation indices are mostly capturing the same information as in dissimilarity indices[58] and that the current computer algorithms tend to use the latter, we will refer readers who are interested in other separation indices to Everitt et al (2010).

II.3 clustering criteria

Having chosen an index to represent a group's dissimilarity, clustering criteria can be defined by aggregating these group measures over all groups. The aggregation can be defined as the sum of dissimilarity over all groups as in $c_1(n,g)$, or as the maximum or minimum dissimilarity among groups as in $c_2(n,g)$ or $c_3(n,g)$ below:

$$c_1(n,g) = \sum_{m=1}^{g} h(m) \qquad (31)$$

$$c_2(n,g) = \max_{m=1,\ldots g} [\mathrm{h(m)}] \qquad (32)$$

$$c_3(n,g) = \min_{m=1,\ldots g} [\mathrm{h(m)}] \qquad (33)$$

One of the most commonly used clustering criteria combines $c_1(n,g)$ with dissimilarity index $h_2(m)$ to represent the total sum of within group dissimilarity. The criterion can also be shown equivalent to the within-group sum-of-squares criteria derived directly from the $n \times p$ multivariate matrix X.

[58] Roughly speaking, the sum of squared distance of the sample comprises two parts: the within group sum of squares and the between group sum of squares. Since the total sum of squared distance is constant, minimizing within group sum of squares, the dissimilarity index mentioned earlier, is equivalent to maximizing the between group sum of squares, the separation index.

$$c_1^*(n,g) = \sum_{m=1}^{g} h_2(m) = \sum_{m=1}^{g}\sum_{l=1}^{n_m} (d_l^{m\bar{m}})^2 = \sum_{m=1}^{g}\sum_{l=1}^{n_m} (x_l^m - \bar{x}^m)'(x_l^m - \bar{x}^m) \qquad (34)$$

Intuitively, when the above $c_1^*(n,g)$ clustering criterion is minimized, agents put into the same cluster share descriptive variables most similar to each other as compared to when they are allocated based on any other alternative clustering outcome.

There are a few features of the above clustering criterion of which any user should be aware. First, the method is scale dependent. For data that contains variables measured on different scales, one may reach different solutions from the same raw data standardized in different manners. Second, this clustering criterion imposes a "spherical" structure on the clusters and is unlikely to find clusters of other shapes, for example, agents that are separated into a few layers. Other clustering criteria exist to circumvent these two features[59]. However, any clustering approach has its advantages and disadvantages, and one must evaluate approaches within the context of particular applications.

II.4 iterative algorithms—k-means and k-median clustering

Ideally, one would consider all combinations of objects and choose the one that yields the lowest dissimilarity index within each group[60]. However, when the number of objects is large, it becomes infeasible to do this. Indeed, Liu (1968) provides the exact number of possible partitions one must consider in order to cluster n objects into g groups:

[59] Attempts to create clustering criteria less restrictive regarding the cluster's shape include Scott and Symons(1971), Symons(1981), Murtagh and Raftery(1984), Banfield and Raftery(1993) and Celeux and Govaert(1995)

[60] Indices that measure the separation between groups are also used in many other methods. We refer interested readers to Everrit et al (2011)

$$N(n,g)=\frac{1}{g!}\sum_{m=1}^{g}(-1)^{g-m}\binom{g}{m}m^n \qquad (35)$$

That is, in order to partition 100 network agents into 5 groups, the number of possible combinations to examine is about 6.6×10^{67}. The task becomes impossible even with modern computational power when the population under analysis comprises hundreds, if not thousands, of agents. This excessive computational burden has led scholars to develop numerical search algorithms to approximate clustering solutions. Here we review the two most commonly used numerical algorithms, k-means and k-median, both of which involve iterative updating processes for partitions and group centroids.

- **K-means algorithm:**

As stated in its name, the k-means algorithms emphasize the mean of the clusters. Generally speaking, all k-means algorithms involve iterative updates of clusters by simultaneously relocating objects into the cluster whose ***mean*** is closest and then recalculating cluster means. Particularly, all k-means algorithms contain the following four steps:

(1) g initial seeds are defined for each cluster by a p-dimensional vector, $\tilde{x}^m=(\tilde{x}_1^m,\tilde{x}_2^m,...,\tilde{x}_p^m)$ where $\tilde{x}_k^m$ stands for the kth characteristic of the initial seed of cluster m. The squared Euclidean distance between the ith object and the initial seed of cluster m is simply calculated as:

$$d_{i\tilde{x}^m}^2=\sum_{k=1}^{p}(x_{ik}-\tilde{x}_k^m)^2 \qquad (36)$$

By comparing the result of equation (X) for an object with each initial seed (there are g of them), we allocate the object to the cluster where the result is minimized.

(2) After all objects have been allocated to one cluster or another, the mean of the cluster is obtained by taking average over all objects that falls into each cluster. This is done for each dimension of the p characteristics:

$$\bar{x}^m = (\bar{x}_1^m, \bar{x}_2^m, ..., \bar{x}_p^m) \qquad (37)$$

The above mean of clusters $\bar{x}^m$ can then replace the initial seeds $\tilde{x}^m$ and be used to calculate the squared distance between each object and each cluster centroid as in equation (X). Objects are again moved to the cluster which yields the lowest squared distance measure.

(3) The step (2) is repeated. For each repetition, the old cluster mean is replaced by the one calculated from the latest membership. The process repeats until no objects change membership.

Although all k-means algorithms attempt to minimize within-group sum of squared deviations from (group) mean, they may differ from each other in details. Depending on the specific dataset used, these differences may have substantial impact on the clustering results[61]. Here we trace a few important differences of these most popular algorithms.

First, the methods of initialization affect the final clustering results. The simplest suggestion, currently used in SPSS, chooses g random data points as initial cluster seeds

[61] We have found substantial differences in K-means clustering results produced by the standard packages in Stata, R and Matlab. We traced it to differences in the specific numerical algorithms used by each package.

(MacQueen, 1967). A slightly different method randomly partition all data points into g mutually exclusive groups and use the group mean as initial seeds (Steinley 2003). These two methods both rely on the random process, therefore may yield a different clustering result each time the algorithm is performed.

Various deterministic methods also exist. Astrahan (1970) suggest a two parameter method as follows: before initialization, two distance d1 and d2 are specified. Then for each data point, a density index is calculated as the number of objects that are less or equal to d1 distance away from the object. The object that yields the highest density is selected as the first seed. Objects that are within the distance of d2 to the first seed are removed from the consideration. A second seed is selected if it has the highest density among the remaining objects. The objects that are within distance d2 to the second seeds are removed. The process continues until all g seeds are determined. A similar process was suggested by Ball and Hall (1965) and implemented in the PROC FASTCLUS procedure in SAS. Although other types of random or deterministic processes exist (see Milligan, 1980 and Bradley and Fayyad 1998 for examples), Steinley (2003) suggest that the most robust method that outperform most of the arbitrary initialization rules is to use multiple random restarts (in order of thousands) and pick the one result that gives the smallest clustering criteria value. *Kmeans* package in R allow the user to specify the number of restart.

Second, to further minimize the squared distance as in equation (X), some algorithm suggests to introduce an additional stage of single-object reallocation process after the group reallocation has been settled (Spath, 1980; Hartigan and Wong, 1979).

Specifically, after performing the standard iterative process (1)-(3) mentioned above, if there is an object in cluster m such that

$$\frac{n_m}{n_m - 1}(d_i^{m\bar{m}})^2 > \frac{n_{m'}}{n_{m'} - 1}(d_i^{m\bar{m}'})^2 \qquad (38)$$

The object *i* should be moved from cluster *m* to cluster *m'* and the squared distance (as in equation (X)) is reduced. The objects will be checked and moved if necessary one after another until no further improvement can be achieved by this process[62].

- **K-median algorithm:**

In more recent years, the k-median algorithm has received increasing attention (Kaufman and Rousseeuw, 1990; spath, 1985; Hansen and Jaumard, 1997; Kohn et al, 2010). This algorithm relocates an object to a group whose ***median*** is the closest to it according to certain distance measure. Numerically, the specific clustering procedure proceeds like k-means except that the clustering criteria in equation (6) is replaced by

$$c_2^*(n, g) = \sum_{m=1}^{g} \sum_{l=1}^{n_m} \left| (x_l^m - \breve{x}^m) \right| \qquad (39)$$

Where $\breve{x}^m$ refers to the median vector of the *mth* cluster. The original idea of using median instead of mean is to reduce the influence of outliers. However, Garcia-Escudero and Gordaliza (1999) pointed out that k-median method can also be as affected by outliers as k-means since the "joint" selection of two medians are unlikely to be as robust in terms of centralization as when only one random variable is involved.

[62] The *kmeans* package in Matlab and R adopt this two-phase iterative algorithm.

Variations of k-median algorithm also exist in terms of how initial seeds are selected and how objects are swapped between clusters. PAM (Partitioning Around Medoids), developed by Kaufman and Rousseeuw (1990) and implemented in the *pam* package of R language, is one of the most popular one. The algorithm sets the objective function as the sum of distance between each object and its nearest medoid. The initial seeds in PAM are chosen by a greedy built phase[63] where the seed is added one after another and only the one that brings the largest improvement on the objective function will be selected.

Once the built phase completes, a multi-iteration swapping stage begins. For each iteration, a medoid object i and a non-medoid object j will be selected that brings the largest improvement on the objective function if i and j are switched. The iterations continue until no improvement is possible. Since in both built phase and swapping phase, there are many pairs of objects to go through to find the largest improvement, the original PAM algorithm is very time consuming with large dataset and increasing number of clusters[64].

III. Methods for choosing the number of clusters

Independent of the choice of clustering criteria and algorithms introduced above, one also needs to choose the method to determine the number of clusters. The past literature has recommended many methods that are algorithmic, graphical or formulaic.

[63] In programming, greedy algorithms refer to the ones that are based on heuristics who find locally optimal choice.

[64] The same authors also developed a similar but less deterministic method CLARA (Clustering LARge Applications), implemented in R language. This method could reduce the computing time significantly when a dataset is large. Meanwhile, STATA implements its *cluster kmedians* command in a similar way as in the basic k median algorithm as described at the beginning of this subsection.

All of these methods are based on some logical heuristics. To judge which method is better at recovering the number of clusters, Milligan and Cooper (1985) conducted a Monte Carlo analysis to compare 30 of the most popular ones and concluded that the top performer is the one suggested by Calinski and Harabasz (1974) (which we denote by C-H)[65]. Another popular method readily available in many commercial packages is Silhouette Width. The output of this method includes a visualization giving direct clue on the performance of clustering under different numbers of clusters. We review Silhouette Width in this paper as well.

III.2 C-H index

C-H (1974) suggested that the optimal number of clusters, g*, should maximize the following value C(g):

$$C(g) = \frac{trace(B)}{g-1} \Big/ \frac{trace(W)}{n-g} \qquad (40)$$

where

$$B = \sum_{m=1}^{g} n_m (\bar{x}^m - \bar{x})(\bar{x}^m - \bar{x})' \qquad (41)$$

representing the between-group dispersion matrix, and

$$W = \sum_{m=1}^{g} \sum_{l=1}^{n_m} (x_l^m - \bar{x}^m)(x_l^m - \bar{x}^m)' \qquad (42)$$

representing the within-group dispersion matrix, both of which derive from the original multivariate matrix X.

[65] Another successful technique developed by Duda and Hart (1973) works with hierarchical cluster methods. The network data do not fit these types of cluster analysis.

III.3 Silhouette Width

The Silhouette Width index is first mentioned in Rousseeuw(1987). His paper argues that due to the absence of visualization for the quality of cluster, it is hard to tell whether an object is well-classified or misclassified. He then proposed the index and the plot of Silhouette Width to visualize the quality of cluster. Interestingly, the Silhouette Width Index has become increasingly popular as a way to choose the number of clusters and has been adopted by most commercial packages along with the Calinki-Harabatz Index we introduced above.

For a given clustering result, the Silhouette width indices, denoted by s(i), are calculated for each object i=1,2,…,n, which are then combined into a Silhouette plot. Individual silhouette width s(i) is defined as:

$$s(i) = \frac{\min_{C \neq M(i)} \frac{1}{n_C} \sum_{\substack{k \notin M(i) \\ k \in C}} d(i,k) - \frac{1}{n_{M(i)}} \sum_{\substack{j \in M(i) \\ j \neq i}} d(i,j)}{\max[\frac{1}{n_{M(i)}} \sum_{\substack{j \in M(i) \\ j \neq i}} d(i,j), \min_{C \neq M(i)} \frac{1}{n_C} \sum_{\substack{k \notin M(i) \\ k \in C}} d(i,k)]} \quad (43)$$

where M(i) refers to the cluster that contains object i, $n_{M(i)}$ refers to the number of objects in cluster M(i) and C refers to any cluster other than M(i).

The first term in the numerator refers to the minimum average distance of an object to all members of another cluster. It calculates the average distance from i to all members of an arbitrary cluster C. After the average distance is calculated for all arbitrary clusters, the closest cluster (in terms of distance to object i) is used.

The second term in the numerator refers to the within cluster average distance for object i. The term simply calculates the distance between object i and each other object in

the same cluster and then takes an average. The denominator is the maximum of the two terms that appear in the numerator.

From the above formula, it is easy to see that s(i) would increase as object i is closer to other objects in the same group and farther away from objects in other groups. However, more characteristics of the index are revealed by evaluating s(i) under three different conditions.

First, note that if $\frac{1}{n_m}\sum_{\substack{j\in m(i)\\ j\neq i}} d(i,j) < \min_{c\neq m(i)} \frac{1}{n_c}\sum_{\substack{k\notin m(i)\\ k\in m(c)}} d(i,k)$, then s(i) can be simplified as $1-\frac{\frac{1}{n_m}\sum_{\substack{j\in m(i)\\ j\neq i}} d(i,j)}{\min_{c\neq m(i)} \frac{1}{n_c}\sum_{\substack{k\notin m(i)\\ k\in m(c)}} d(i,k)}$.That is s(i) is always positive and approaches 1 as the measure of within dissimilarity (the numerator) is much smaller than the measure of the smallest between dissimilarity (the denomenator).

Similarly, consider the opposite case where $\frac{1}{n_m}\sum_{\substack{j\in m(i)\\ j\neq i}} d(i,j) > \min_{c\neq m(i)} \frac{1}{n_c}\sum_{\substack{k\notin m(i)\\ k\in m(c)}} d(i,k)$.

Under this condition, s(i) can be simplified as $\frac{\min_{c\neq m(i)} \frac{1}{n_c}\sum_{\substack{k\notin m(i)\\ k\in m(c)}} d(i,k)}{\frac{1}{n_m}\sum_{\substack{j\in m(i)\\ j\neq i}} d(i,j)}$-1, which is always a negative number and approaches -1 if within dissimilarity is large and the between dissimilarity if small. That is to say that the silhouette width index defined as in

Rousseeuw(1987) is an index between -1 and 1 with a higher positive number indicating a better clustering quality.

In practice, one should choose the number of clusters that maximizes the average Silhouette Width across all objects.

VII. Summary

Cluster analysis is an intuitive method to analyze complicated data sets. Without making assumptions on the properties of the data, the method divides observations into small number of groups based on patterns of similarity. We reviewed the key procedure of cluster analysis in this paper. First, we reviewed several distance measures that fit for different types of measures (binary, categorical or continuous). We then illustrated how distance measure can be combined into (dis-)similarity matrix and how these matrices are further used in forming clustering criteria. We also discussed the detail of two popular algorithms: k-means and k-median. Finally, we reviewed two indices, Calinski-Harabatz Index and Average Silhouette Width, used to discover the number of clusters prior to the implementation of cluster analysis. We argue that the decision data from laboratory experiments are often generated by complex behavioral rules that can be difficult to specify a priori. Therefore, these data may particularly benefit from clustering methods.

REFERENCES

For Chapter 1

Bala, V. and S. Goyal (2000): "A noncooperative model of network formation", Econometrica, 68(5) 1181-1229

Barabasi, A. and R. Albert (1999): Emergence of Scaling in Random Networks, Science, 286 (5439), 509-512.

Berninghaus, Siegfried K., K. Ehrhart and M. Ott (2006): A network experiment in continuous time: The influence of link costs, Experimental Economics, 9:237–251

Berninghaus, Siegfried K., K. Ehrhart, M. Ott and B. Vogt (2007): Evolution of networks—an experimental analysis, Journal of Evolutionary Economics, 17: 317–347

Berninghaus, S.K.; Ehrhart, K.-M.; Ott, M. (2011): Forward-Looking Behavior in Hawk-Dove Games in Endogenous Networks: Experimental Evidence, Games and Economic Behavior, 75: 35-52

Bramoulle, Y. & R. Kranton (2007): Public goods in networks, Journal of Economic Theory,135(1) 478-494

Bramoulle, Y., D. Lopez-Pintado, S. Goyal and F. Vega-Redondo (2004): Network Formation and Anti-coordination Games, International Journal of Game Theory, 33 1-19

Callander, S., and C. Plott (2005): Principles of Network Development and Evolution: An Experimental Study, Study of Public Economics 89:14691495

Calinski, T. and Harabasz, J. (1974): A dendrite method for cluster analysis, Communications in Statistics 3, 1–27

Conley, T. and C. Udry (2010): Learning about a New Technology: Pineapple in Ghana, American Economic Review,100(1) 35-69

Cooper, R., Dejong, D., Forsythe, R. and Ross,T. (1993): Forward Induction in the Battle-of-the_Sexes Games, American Economic Review, 83,1303-1316

Corbae, D. & J. Duffy (2008): Experiments with network formation, Games and Economic Behavior, Elsevier, 64(1) 81-120

Deck, C. and C. Johnson (2004): Link bidding in a laboratory network.. Review of Economic Design, 8 (4): 359-372.

Dodd, P., J. Watts and F. Sabel (2003): "Information exchange and the robustness of organizational networks", Proceedings of National Academy of Science USA, 100 12516-12521

DiMasi, J., R. Hansen and H. Grabowski (2003): The price of innovation: new estimates of drug development costs, Journal of Health Economics, 22 151–185

DiMasi, J and H. Grabowski (2007): The Cost of Biopharmaceutical R&D: Is Biotech Different? Managerial and Decision Economics, 28: 469–479

Droste, E., R. Gilles and K. Johnson (2000): Endogenous interaction and the evolution of conventions, working paper

Duda, R and P. Hart (1973): Pattern classification and scene analysis, Wiley, New York

Everitt, B, S. Landau, M. Leese and D. Stahl (2011): Cluster Analysis, Wiley Press, 5th edition

Falk, A., and M. Kosfeld (2003): It's All About Connections: Evidence on Network Formation, Institute for the Study of Labor (IZA) Discussion Paper 777, Zurich IEER

Feick, L and L. Price (1987): The Market Maven: A Diffuser of Marketplace Information, Journal of Marketing, 51(1) 83-97

Fisher, W (1698): Clustering and Aggregation in Economics, The Johns Hopkins University Press

Fischbacher, U. (2007): z-Tree: Zurich Toolbox for Ready-made Economic Experiments, Experimental Economics, 10(2) 171-178

Foster, A. and M. Rosenzweig (1995): Learning by Doing and Learning from Others: Human Capital and Technical Change in Agriculture, Journal of Political Economy, 103(6) 1176-1209,

Gladwell, M (2000): The Tipping Point: How Little Things Can Make a Big Difference, Little Brown

Galeotti A., S. Goyal and J. Kamphorst (2006): Network Formation with Heterogenous Players, Games and Economic Behavior, 54(2), 335-372

Galeotti, A. and S. Goyal (2010): The law of the few, American Economic Review, 100(4) 1468–92.

Geissler G. and S. Edison (2005): Market Mavens' Attitudes Towards General Technology: Implications for Marketing Communications, Journal of Marketing Communications, 11, 2, 73-94.

Godes, D. and M. Dina (2009): Firm-Created Word-of-Mouth Communication: Evidence from a Field Study, Marketing Science, 28 (4), 721-739.

Goeree, J., A. Riedl and A. Ule (2009): "In search of stars: Network formation among heterogeneous agents", Games and Economic Behavior, 67(2) 445-466

Goyal, S. and F. Vega-Redondo (2005): Network Formation and Social Coordination, Games and Economic Behavior, 50 178-207

Guth, W., Huck, S. and Rapoport, A.(1998): The Limits of the Positional Order Effect: Can it Support Silent Threats and Non-Equilibrium Behavior, Journal of Economic Behavior and Organization, 34(2) 313-325

Hansen, P. and B. Jaumard (1987): Minimum sum of diameters clustering, Journal of Classfication, 4 215-226.

Hartigan, J. A. and Wong, M. A. (1979): A K-means clustering algorithm, Applied Statistics, 28, 100-108.

Hirschberg, J., E. Maasoumi and D. Slottje (1991): Cluster analysis for measuring welfare and quality of life across countries, Journal of econometrics, 50(1-2) 131-150

Houser, D and M. Keane and K. McCabe (2004): Behavior in a dynamic decision problem: An analysis of experimental evidence using a Bayesian type classification algorithm, Econometrica,72:3, 781-822.

Iyengar, R., Van Den Bulte, C., & Valente, T. W. (2010): Opinion Leadership and Social Contagion in New Product Diffusion, Marketing Science, 30(2), 195-212.

Jackson, M. and B. W. Rogers (2007): Meeting Strangers and Friends of Friends: How Random Are Social Networks?, American Economic Review, 97(3), 890-915

Jackson, M and A. Watts (2002): On the Formation of Interaction Networks in Social Coordination Games, Games and Economic Behavior, 41 265-291

Jackson, M. and A. Wolinsky(1996): A Strategic Model of Social and Economic Networks, journal of economic theory, 71 44-74

Jackson, M. an D. Lopez-Pintado (2011): Diffusion in Networks with Heterogeneous Agents and Homophily, working paper

Jackson, M. (2003): A survey of models of network formation: Stability and efficiency, Game Theory and Information 0303011, EconWPA.

Jackson, M. (2009): An Overview of Social Networks and Economic, Handbook of Social Economics Applications

Jajuga, K., M. Walesiak and A. Bak (2003): On the general distance measure, in Exploratory Data Analysis in Empirical Research, Springer-Verlag, Heidelberg. , 104–109.

Kaufman, L. and P. Rousseeuw (1990): Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, New York

Katz, E. and P. Lazarsfeld (1955): Personal Influence; the Part Played by People in the Flow of Mass Communications, Free Press.

Kelly, J., et al. (1992): Community AIDS/HIV risk reduction: The effects of endorsements by popular people in three cities. American Journal of Public Health, 82, 1483-1489.

Knigge, A. and V. Buskens (2010): Coordination and Cooperation Problems in Network Good Production, Games, 1(4): 357-380.

Köhn HF, D. Steinley and M. Brusco (2010): The p-median model as a tool for clustering psychological data, Psychologic Methods.15(1):87-95.

Kurzban, R and D. Houser (2005): Experiments investigating cooperative types in human groups: A complement to evolutionary theory and simulations, Proceedings of the National Academy of Sciences of the United States of America, 102(5), 1803-1807.

Larrosa, J and F. Tohm (2011): Network Formation with Heterogeneous Agents, working paper

Liu, G. (1968): Introduction to combinatorial mathematics, McGraw Hill

Marsden, P. and K. Campbell (1984): Measuring Tie Strength, Social Forces, 63 (2): 482-501

Merton, Robert K (1968): Social Theory and Social Structure. New York: Free Press.

Milligan, G. and M. Cooper (1985): An examination of procedures for determining the number of clusters in a data set. Psychometrika, 159–159

Rand, W. M. (1971): Objective criteria for the evaluation of clustering methods, Journal of the American Statistical Association, 66 (336): 846–850

Rapoport, A (1997): Order of Play in Strategic Equivalent Games in Extensive Form, Internaltional Journal of Game Theory, 26(1) 113-136

Roch, C. H. (2005): The Dual Roots of Opinion Leadership, Journal of Politics, 67, 110–31.

Rogers, E. (2003): Diffusion of Innovation, Free Press, 5th edition

Schwartz, B., The paradox of choice: Why less is more, New York: Ecco, 2004

Song, L.,S. Appleton and J. Knight (2006): Why do girls in rural China have lower school enrollment? World Development, 34(9), 1639-1653

Steinley, D (1006b): K-means clustering: A half-century synthesis, British Journal of Mathematical and Statistical Psychology, 59(1)

Tran, A. (2009): Can Procurement Auctions Reduce Corruption? Evidence from the Internal Records of a Bribe-Paying Firm, working paper

Ule, Aljaz (2008): Partner Choice and Cooperation in Networks: Theory and Experimental Evidence, Springer Verlag

Valente T. (1995): Network Models of the Diffusion of Innovations, Hampton Press

Vandenbossche, J. and T. Demuynck (2010): Network formation with heterogeneous agents and absolute friction, Working Papers

Weber, R., C. Camerer and M. Knez (2004): Timing and Virtual Observability in Ultimatum Bargaining and "Weak Link" Coordination Games, Experimental Economics, 7:25–48

Weimann, G. (1994): The influentials: The people who influence people. New York: State University of New York Press.

Wiedmann K.P., G. Walsh and V.W. Mitchell (2001): The Mannmaven: an agent for diffusing market information, Journal of Marketing Communications, 7, 185-212.

Williams, T.G. and M.E. Slama (1995): Market mavens' purchase decision evaluative criteria: implications for brand and store promotion efforts, Journal of Consumer Marketing, 12, 3, 4-21.

For Chapter 2

Ambrus, A. and S. Takahashi (2008): Multi-sender Cheap Talk with Restricted State Spaces, Theoretical Economics, 3:1-27

Austen-Smith, D. (1993): Interested Experts and Policy Advice: Multiple Referrals under Open Rule, Games and Economic Behavior, 5(1): 3-43

Battaglini, M. (2002): Multiple Referrals and Multidimensional Cheap Talk, Econometrica, 70(4): 1379–1401

Battaglini, M., and U. Makarov (2011): Cheap Talk with Multiple Audiences: an Experimental Analysis, working paper

Bernhard, H., E. Fehr and U. Fischbacher (2006): Group Affiliation and Altruistic Norm Enforcement, American Economic Review, 96 (2): 217–221

Blume, A., D. DeJong, Y. G. Kim and G. Sprinkle (1998): Experimental Evidence on the Evolution of Meaning of Messages in Sender-Receiver Games, American Economic Review, 88: 1323-1340

Blume, A., D. DeJong, Y. G. Kim and G. Sprinkle (2001): Evolution of Communication with Partial Common Interest, Games and Economic Behavior, 37: 79-120

Brewer, M. B.(1999): The Psychology of Prejudice: Ingroup Love and Outgroup Hate?, Journal of Social Issues, 55 (3): 429-444.

Cai, H and J.T. Wang (2006): Overcommunication in Strategic Information Transmission Games, Games and Economic Behavior, 56 (1):7-36

Charness, G., L, Rigotti, and A, Rustichini(2007): Individual Behavior and Group Membership, American Economic Review, 97: 1340-1352

Chen, Y. and S. X. Li, (2009): Group Identity and Social Preferences, American Economic Review 99(1): 431-457

Cloke, K., J. GoldSmith (2000): Resolving Personal and Organizational Conflict: Stories of Transformation and Forgiveness, Jossey-Bass

Conrad, Charles R., M.S. Poole (2011): Strategic Organizational Communication: In a Global Economy, Wiley & Sons

Cowan, David. (2003): Taking Charge of Organizational Conflict: A Guide to Managing Anger and Confrontation, Personhood Press

Crawford, V.P. and J. Sobel (1982): Strategic Information Transmission, Econometrica, 50(6):1431-1451

Dana, J., R. A. Weber and J. X. Kuang (2007): Exploiting moral wiggle room: experiments demonstrating an illusory preference for fairness, Economic Theory, 33: 67–80

De Dreu, C.K.W., M. J. Gelfand (2007):The Psychology of Conflict and Conflict Management in Organizations, Psychology Press

Dickhaut, J.W, K.A. McCabe and A. Mukherji (1995): An Experimental Study of Strategic Information Transmission: Economic Theory, 6:389-403

Eckel, C. and P. J. Grossman (2005): Managing Diversity by Creating Team Identity, Journal of Economic Behavior & Organization, 58 (3): 371–392

Eckman, A., T. Lindlof (2003): Negotiating the Gray Lines: an ethnographic case study of organizational conflict between advertorials and news, Journalism Studies, 4:65–77

Farrell, J. and R. Gibbons (1989): Cheap Talk with Two Audiences, American Economic Review, 79(5): 1214-23

Galeotti, A, C. Ghiglino and F. Squantani (2011): Strategic Information Transmission in Networks, Working paper

Gilligan, T. and K. Krehbiel (1989): Asymmetric Information and Legislative Rules with a Heterogeneous Committee, American Journal of Political Science, 459-90

Gneezy, U. (2005): Deception: The Role of Consequences, The American Economic Review, 95(1): 384-394.

Goette, L., D. Huffman, and S. Meier (2006): The Impact of Group Membership on Cooperation and Norm Enforcement: Evidence Using Random Assignment to Real Social Groups, American Economic Review, 96 (2): 212–216.

Gupta, A. K., S. P. Raj, and D. L. Wilemon (1985): R&D and Marketing Dialogue in High-Tech Firms. Industrial Marketing Management 14, 289–300

Hagenbach, J. and F. Koessler (2010): Strategic Communication Networks, Review of Economic Studies, 77(3):1072-1099

Kolb, D. M., L.L. Putnam, J. M. Bartunek (1992): Hidden Conflict in Organizations: 1st Edition, SAGE Publications

Krishna, V. and J. Morgan (2001a): A Model of Expertise, Quarterly Journal of Economics, 116(2):747-775

Krishna, V. and J. Morgan (2001b): Asymmetric Information and Legislative Rules: Some Amendments, American Political Science Review, 95(2):435-452

Lai, E. K., W. Lim, and J. T.-Y. Wang (2011): Experimental Implementations and Robustness of Fully Revealing Equilibria in Multidimensional Cheap Talk, working paper

Lundquist, T., T. Ellingsen, E. Gribbe and M. Johannesson (2009): The Aversion to Lying, Journal of. Economic Behavior and Organization, 70(1–2):81–92

Marchewka, A., K. Jednorog, M. Falkiewicz, W. Szeszkowski, A. Grabowska and I. Szatkowska(2012): Sex, Lies and fMRI—Gender Differences in Neural Basis of Deception, Plos One, Aug, 2012

Milgrom, P.R. and J. Roberts (1986): Relying on the Information of Interested Parties, Rand Journal of Economics, 17: 18-32

Miller, Katherine (2011): Organizational Communication: Approaches and Processes, Cengage Learning

Minozzi, W., and J. Woon (2011): Competition, Preference Uncertainty, and Jamming: A Strategic Communication Experiment, working paper

Morgan, J. and P. C. Stocken (2008): Information Aggregation in Polls, American Economic Review, 98(3): 864-96.

Pirnejad, H., Z, Niazkhani, M.Berg and R. Bal (2008): Intra-organizational Communication in Healthcare--Considerations for Standardization and ICT Application, Methods of Information in Medicine, 47(4): 336-45

Rahim, Afzalur (2000): Managing Conflict in Organizations: 3rd Edition, ABC-Clio, LLC

Shih, M., T. L. Pittinsky and N. Ambady (1999: Stereotype Susceptibility: Identity Salience and Shifts in Quantitative Performance, Psychological Science, 10 (1):81–84

Sutter, M.(2009): Deception through Telling the Truth? Experimental Evidence from Individuals and Teams, Journal of Econonomics. 119, 47 ‐ 60.

Tajfel, H. and J. Turner (1979): An Integrative Theory of Intergroup Conflict, in Stephen Worchel and William Austin, eds., The Social Psychology of Intergroup Relations, Monterey, CA: Brooks/Cole

Tobak, S. (2008): Marketing v. Sales: How To Solve Organizational Conflict, CBS, MONEYWATCH

Wang, J.T., M. Spezio and C.F. Camerer (2010): Pinocchio's Pupil: Using Eyetracking and Pupil Dilation to Understand Truth Telling and Deception in Sender-Receiver Games, American Economic Review, 100 (3): 984-1007

Weinrauch, J. D., R. Anderson (1982): Conflicts Between Engineering and Marketing Units, Industrial Marketing Management, 11(4): 291-301

Xiao, E. (forthcoming): "Profit Seeking Punishment Corrupts Norm Obedience, Games and Economic Behavior

For Chapter 3

Adomavicius, G., S. P. Curley, A. Gupta and P. Sanyal (2012): Effect of Information Feedback on Bidder Behavior in Continuous Combinatorial Auctions, Management Science,58:811-830

Anne E. Farmer, a, Peter McGuffinb, Edward L. Spitznagelc (1983): Heterogeneity in Schizophrenia: A Cluster-analytic Approach, Psychiatry Research, 8(1): 1–12

Astrahan, M. M. (1970): Speech Analysis by Clustering, or the Hyperphome Method, Stanford Artificial Intelligence Project Memorandum AIM-124. Stanford, CA: Stanford University.

Babu, G.J., E. D. Feigelson(1997):Statistical Challenges in Modern Astronomy II, Springer

Ball, G. H. and D. Hall, D. J. (1965): ISODATA: a Novel Method For Data Analysis and Pattern Classification Menlo Park, CA: Stanford Research Institute.

Banfield, J. D. and A. E. Raftery (1993): Model-based Gaussian and Non-Gaussian Clustering, Biometrics, 49: 803–821.

Borgen, F. H. and D.C. Barnett (1987): Applying Cluster Analysis in Counseling Research, Journal of Counseling Psychology, 34(4):456-468

Bradley, P. S. and U. M. Fayyad (1998): Refining Initial Points for k-means Clustering, Machine Learning: Proceedings of the fifteenth International Conference edited by J.Shavlik (pp. 91–99). San Francisco: Morgan Kaufmann.

Bushel, P. R., R.D. Wolfinger and G. Gibson (2007): Simultaneous Clustering of Gene Expression Data with Clinical Chemistry and Pathological Evaluations Reveals Phenotypic Prototypes, BMC Systems Biology, 23: 1–15.

Calinski, R. B. and J. Harabasz (1974): A Dendrite Method for Cluster Analysis, Communications in Statistics, 3:1–27.

Celeux, G. and G. Govaert (1995): Gaussian Parsimonious Clustering Models, Pattern Recognition, 28(5):781–793.

Clarke, D. L.(1968): Analytical Archaeology. Methuen

DeRubeis E, J.L. Wylie, D.W. Cameron, R.C. Nair and A.M. Jolly (2007): Combining Social Network Analysis and Cluster Analysis to Identify Sexual Network Types, International Journal of STD & AIDS, 18(11):754-9.

Duda, R. O. and P. E. Hart (1973): Pattern Classification and Scene Analysis, John Wiley & Sons, Inc., New York.

El-Gamal, M.A. and D. M. Grether (1995): Are People Bayesian? Uncovering BehavioralStrategies, Journal of the American Statistical Association, 90: 1137–1145.

Everitt, B.S., S. Landau, M. Leese, D. Stahl (2011): Cluster Analysis, John Wiley & Sons

Fisher, W.D. (1969): Clustering and Aggregation in Economics, The Johns Hopkins University Press

Garcia-Escudero, L. A. and A. Gordaliza (1999): Robustness of Properties of K-means and Trimmed K-means, Journal of the American Statistical Association, 94: 956–969.

Gower, J. C. (1966): Some Distance Properties of Latent Root and Vector Methods Used in Multivariate Analysis, Biometrika, 53: 325–338.

Gower, J. C. (1971): A General Coefficient of Similarity and Some of its Properties, Biometrics, 27: 857–872.

Gower, J. C. and P. Legendre (1986): Metric and Euclidean Properties of Dissimilarity Coefficients, Journal of Classification, 5: 5–48

Hansen, P. and B. Jaumard (1997): Cluster Analysis and Mathematical Programming, Mathematical Programming, 79: 191–215.

Hartigan, J. A. and M. A.Wong (1979): Algorithm AS 136: A k-means Clustering Algorithm, Applied Statistics 28: 100–108.

Hay, P. J., C.G. Fairburn and H.A. Doll (1996): The Classification of Bulimic Eating Disorders: A Community Based Study, Psychological Medicine, 26(4):801-812.

Hirschberg, J. G., E. Maasoumi and D. J. Slottje (1991): Cluster Analysis for Measuring Welfare and Quality of Life Across Countries", Journal of Econometrics, 50: 131-150.

Houser, D., M. Keane and K. McCabe (2004): Behavior in a Dynamic Decision Problem: an Analysis of Experimental Evidence using a Bayesian Type Classification Algorithm, Econometrica, 72(3): 781-822

Ichino, M. and H. Yaguchi(1994): Generalized Minkowski Metrics for Mixed Feature Type Data Analysis, IEEE Transactions on Systems, Man and Cybernetics, 24: 698–708.

Jajuga, K.,M. Walesiak and A. Bak(2003), On the General Distance Measure, Exploratory Data Analysis in Empirical Research (M. Schwaiger and O. Opitz, eds.) Springer-Verlag, Heidelberg.

Johnson, S. (1967): Hierarchical Clustering Schemes, Psychometrika, 32(3): 241-254

Kaufman, L. and P. Rousseeuw (1990): Finding Groups in Data: An Introduction to Cluster Analysis, New York: Wiley.

Kerr, M.K. and G.A. Churchill (2011): Bootstrapping Cluster Analysis: Assessing the Reliability of Conclusions from Microarray Experiments, 98(16):8961-5

Kohn, H. F., D. Steinley and M. J. Brusco (2010): The p-median Model as a Tool for Clustering Psychological Data, Psychological Methods, 15: 87–95.

Legendre, P. and A. Chodorowski (1977): A Generalisation of Jaccard's Association Coefficient for Q-analysis of Multi-state Ecological Data Matrices, Ekologia Polska, 25: 297–308.

Liu, G. L. (1968): Introduction to Combinatorial Mathematics, McGraw Hill, New York.

MacQueen, J. (1967): Some Methods of Classification and Analysis of Multivariate Observations Proceedings of the fifth Berkeley symposium on mathematical statistics and probability (edited by Cam, Le, L. M. NeymanJ. (1) 281–297 Berkeley, CA: University of California Press.

Milligan, G. W. (1980): An Examination of the Effect of Six Types of Error Perturbation on Fifteen Clustering Algorithms, Psychometrika, 45:325–342.

Milligan, G. W. and M. C. Cooper (1985): An Examination of Procedures for Determining the Number of Clusters in a Data set, Psychometrika, 50: 159–179.

Murtagh, F. and A. E. Raftery (1984): Fitting Straight Lines to Point Patterns, Pattern Recognition, 17:479–483.

Punj, G. and D. W. Stewart (1983): Cluster Analysis in Marketing Research: Review and Suggestions for Application, Journal of Marketing Research , 20(2):134-148

Rosenburg, H. (1910): On the Relation Between Brightness and Spectral Type in the Pleiades [title translated in English], Astronomische Nachrichten, 186:71

Rousseeuw, P. J. (1987): Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis, Journal of Computational and Applied Mathematics, 20: 53–65.

Scott, A. J. and M. J.Symons (1971): Clustering Methods Based on Likelihood Ratio Criteria, Biometrics, 27: 387–398.

Slater, S. F. and T.J. Zwirlein(1996): The Structure of Financial Strategy: Patterns in Financial Decision Making, Managerial and Decision Economics, 17(3):253-266

Sneath, P. H. A. and R. R.Sokal(1973): Numerical Taxonomy, W. H. Freeman

Späth, H. (1980): Cluster Analysis Algorithms for Data Reduction and Classification of Objects, New York: Wiley

Späth, H. (1985): Cluster Dissection and Analysis: Theory, FORTRAN Programs, Examples, New York: Wiley.

Spearman, C. (1904): General intelligence, objectively determined and measured, American Journal of Psychology, 15: 201–292.

Steinley, D. (2003): K-means Clustering: What You Don't Know May Hurt You, Psychological Methods, 8: 294–304.

Sutton, M.Q. and K. J. Reinhard (1995): Cluster Analysis of the Coprolites from Antelope House: Implications for Anasazi Diet and Cuisine, Journal of Archaeological Science,22(6):741–750

Symons, M. J.(1981): Clustering Criteria and Multivariate Normal Mixtures, Biometrics, 37: 35–43

Tryon, R.C. (1932): Multiple Factors Vs Two Factors as Determiners of Ability, Psychological Review, 39: 324-51

Tryon, R. C. (1935): A Theory of Psychological Components—An Alternative to "Mathematical Factors," Psychological Review, 42: 425–454.

Tryon, R. C., and D.E. Bailey (1966): The BCTRY Computer System of Cluster and Factor Analysis, Multivariate Behavioral Research, 1:95-111

Witten, D.M. and R. Tibshirani(2010): Supervised Multidimensional Scaling for Visualization, Classification, and Bipartite Ranking, Journal Computational Statistics & Data Analysis archive, 55(1):789-801

Wright, C., T. Burns, P. James, et al. (2003): Assertive Outreach Teams in London: Models of Operation, British Journal of Psychiatry, 183: 132–138.

Yamamori, T, K. Kato, T. Kawagoe and A. Matsui (2008): Voice Matters in a Dictator Game, Experimental Economics, 11: 336–343

CURRICULUM VITAE

Rong Rong graduated from Hefei No.1 High School at her hometown in China, 2003. She received her Bachelor of Science in economics from Shanghai Jiao Tong University in 2007. She received financial support from Interdisciplinary Center for Economic Science (ICES) at George Mason University for her graduate study, during which she received her Master of Arts in Economics from George Mason University in 2009. Prior to her completion of this dissertation, she accepted a tenure-track assistant professor position at Weber State University at Ogden, Utah.