KEEPING UP WITH THE JONESES: AN EMPIRICAL INVESTIGATION OF CONTAGION ON CRBT ADOPTION

Bin Zhang
iLab, Heinz College
Carnegie Mellon University
Pittsburgh

Second Research Paper

Committee:
Dr. Christos Faloutsos
Dr. David Krackhardt (Chair)
Dr. Ramayya Krishnan
Dr. Daniel Neill

May 3, 2010

Abstract

Cohesion and structural equivalence are two competing social network models to explain diffusion of innovation. While considerable work has been done on these models, the question of which network model explains diffusion has not been resolved. This paper examines diffusion of Caller Ring Back Tones in a cellular telephone network. Since these societal scale networks are very large (e.g., our call detail record data set has more than one million customers and one billion calls over a three months period from a large cellular service provider in India), the study of diffusion in these settings require the development of methods to extract connected subpopulations from the network. We develop a novel technique to detect densely connected and self-contained components of the network and demonstrate through random restarts that the technique can enumerate distinct connected components in the network. Using a standard network auto-correlation model, we study the competing effects of cohesion and role equivalence on each of the distinct connected components detected using our sampling technique. The comparison of the results from the two models shows cohesion is more statistically significant than role equivalence. The results have consistent pattern across different sizes of subpopulation. We also conducted meta-analysis to summarize the size for both network effects. We find a significant summarized effect and its size changes as E/I index changes.

Contents

1 INTRODUCTION

The core question in this research is whether cohesion (direct contact) or role equivalence (structural equivalence) plays a more influential role in caller ring-back tone (CRBT) diffusion. When forming opinions or making decisions, people usually use someone they know or someone in their social network as their frame of reference, taking their opinions into account. This progress, in which an actor adapts his behaviors to those of alters in his social network, is known as contagion or social influence (Leenders, 1997). There are two social network models, cohesion and role equivalence, to analyze contagion. Cohesion is made through communication, which is direct contact between actor and alter; while role equivalence is created through comparison, which occurs when an actor competes with other alters who he considers in a similar position to him in the social network. Both models have been used to explain the progress of contagion. Coleman et al. (1966) studied diffusion of medical innovation and found medical doctors adopted new technology at the early stage because of cohesion. Burt (1987) reanalyzed Coleman et al's data and concluded that contagion does not happen through cohesion but rather through role equivalence. Burt used a network autocorrelation model to conduct the analysis but could only include one network effect at a time because of the limitations of the method at that time. Since then, quantitative methods for social networks have been developed. For example, Doreian (1989) developed a two regimes of network effect autocorrelation model, which can accommodate two networks structures in one model. But such a method has never been applied to this classic research question.

In this paper, we would like to investigate the effect of cellular phone communication social networks on diffusion of CRBT adoption. CRBT is becoming one of the most attractive mobile content with a projected worldwide revenue of \$4.7 Billion by 2012 [1]. CRBT replaces plain ring-back tones with music a caller will hear as he/she waits for the receiver to answer. The analysis of social network based on cellular phone communication is an exciting area of research. Cellular phones are a ubiquitous communication medium. They have become a strategic component of modern life and economies. Cellular phones are expected to become the major medium of communications, including voice calls, internet access, email and file transfer etc. Phone call networks can be a good representation of one's true social network, because unlike other examples of large social networks, which are often extracted from online networking sites, the interaction between two individuals entails a stronger notion of intent-to-communicate. Furthermore, phone call networks can provide detailed information on the spatio-temporal behavior of users, especially on the social networks they build and maintain, as reflected by their phone calls. Several recent studies have used cellular phone call data to characterize the social interactions of cell phone users, with a focus on understanding the structural properties of the graph, its evolution, and the evolution of social groups. It can be important to understand how the dynamics of adoption are likely to unfold within the underlying phone call social network: the extent to which people are likely to be affected by decisions of their friends and colleagues. However, how new products spread within such networks has not been well studied.

Our data set is from one of India's largest cellular phone services. Each record includes phone numbers

---

[1] fiercewireless.com

(hash values) from the sender and receiver, the date and time of phone call. There are over one million subscribers and over one billion phone call records in our data after preprocessing. Given the size of our data, it is impossible to analyze the whole data size using the current computing power available. Hence in order to investigate CRBT diffusion, we need to solve another substantial problem – finding subpopulation to make our analysis computationally tractable. The extracted subpopulations should be isolated subnetworks so they can show pure network effect on CRBT diffusion. A community has the features we want because, according to Newman (2006), community is a densely connected group of vertices in a network, with only weak connections between groups. We can use community for this matter because it can represent real social groupings in a social network (Girvan and Newman, 2002). Our new method – Two-stage Clustering and Pruning (TSCP) algorithm is based on a commonly used sampling method – chain-referral sampling, also known as snowball sampling. The algorithm is designed to identify subpopulations having the same features as communities. Our algorithm identifies self-contained subpopulations on the order of 1000 fast, without input of the parameters of the global network.

The rest of the paper is organized as follows. We discuss the literature on diffusion, cohesion and role equivalence models in Section 2; the data set, our subpopulation extraction method that can identify local communities fast and with low computing cost, and two network effects autocorrelation model, and meta-analysis for our research are presented in Section 3. In Section 4 we present our empirical results. Discussion, conclusion and limitation complete the paper in Section 5.

2 THEORY

2.1 Diffusion

The theoretical foundation for our research is social influence on the diffusion process. Diffusion is the "process by which an innovation is communicated through certain channels over time among the members of a social system ... a special type of communication concerned with the spread of messages that are perceived as new ideas" (Rogers, 1962). The network model has been widely used to study diffusion since the Bass (1969) model. A history of network models used to study diffusion of innovations is reviewed by Valente (2005). He categorized the evolution of network models as three stages: macro models, spatial autocorrelation and network (effect) models. In the early era, the probability of adoption is only related to the time that an actor gets exposed to the object of diffusion. In 1969, Bass proposed a famous model to include both rate of innovation and imitation. This model can estimate both the influence from the social network and innovativeness, and is shown in the equation below. Let Y_{t-1} be the proportion that has adopted at time period $t-1$; y_t be the proportion of new adoption at t; b_0 be the coefficient of innovation, which is the probability of initial adoption; and b_1 be the coefficient of innovation.

$$\frac{y_t}{1-Y_{t-1}} = b_0 + b_1 Y_{t-1}$$

$$y_t = b_0 + (b_1 - b_0)Y_{t-1} - b_1 Y_{t-1}^2$$

Bass' model is still at the population level. Its assumption is that everyone in the social network has the same probability of interacting. Such an assumption is not realistic because given a large social network, the probability of any random two nodes connecting to each other is not the same. It seems fair to assume people with closer physical distance communicate more and exert greater influence on each other thus spatial autocorrelation was brought in to the models used in the literature. It is still a measure of diffusion at the population level, and does not account for whether one actor is more or less likely to adopt based on his network position. It does not show how network structure influence diffusion either. So researchers turned to network models to more accurately reflect these influences.

The diffusion network model explains that the initial adoption is based on actor's innovativeness and exposure to sources of influence, and that this influence originates from alters who have already adopted and are able to persuade nonusers to adopt. Before moving on, we need to take a look at event history analysis, which offers some quite useful tools for network analysis. Sometimes the factors that affect adoption also affect the formation of network. Thus it is necessary to collect data at different time periods (panel data), and bring in event history analysis. The purpose of event history analysis is to explain why certain individuals are at a higher risk of experiencing the event of interest than others. The most commonly used analysis methods include failure-time models, survival models, and hazard models etc. (An event is a transition from one status to another, e.g. from non-adoption to adoption.)

Network influences are captured by contagion model. Social contagion is the interpersonal connection over which innovation is transmitted (Burt, 1987). The probability of each actor's adoption increases when the number or proportion of the adopters in his network increases. The network exposure is defined as below, where E_i is the proportion of actor i's neighbors who have adopted; y is the variable of adoption; and w is social network structure matrix.

$$E_i = \frac{\sum\limits_{j=1}^{N} w_{ij} y_j}{\sum\limits_{i=1}^{N} w_i}$$

2.2 Cohesion Model

In the diffusion network model, network influence drives diffusion. Network influence can be measured by exposure, which can be modeled as three processes, direct ties, structural equivalence (role equivalence), and centrality (Valente, 2005). Direct ties and structural equivalence both belong to physical proximity. Direct ties is also called the cohesion model and structural equivalence is also called the competition model or role equivalence model.

The cohesion and role equivalence models are two competing models in diffusion theory. In the cohesion model, a focal person's adoption is influenced by his neighbor who he directly connects to. Those connections are communications between actors. A focal person could be informed by, persuaded by, or receive suggestion from the people in his network. The most famous piece of work about cohesion model might be that of Coleman et al. (1966). In their seminal work, they found that medical doctors prescribe a new drug because of directed ties with other doctors. The greater a doctor's connection to his colleagues, the earlier he prescribes that drug for the first time. Coleman et al. gave a very reasonable explanation: when there is the need to make a decision in an ambiguous situation, the doctor would ask for suggestions and advice from those who they usually discuss questions with or get advice from. "The more frequent and empathic the communication is between actor and alter," the more likely alter's adoption will affect actor's. Such adoption could be a result of discussions about benefits and costs between actor and alter (Burt, 1987).

Rogers and Kincaid (1981) also investigated cohesion's effect on innovation diffusion. Different from Coleman et al., Rogers and Kincaid used personal network density as the measure of cohesion. Their result is similar to that of Coleman et al. though, they found that personal network density is positively related to the adoption.

The conclusions from the literature discussed above suggest that people notice and understand new product through discussion and observation with those who are in their social network (Harkola and Greve, 1995). Thus a focal person's direct ties to adopters influences his decision about adoption. So in the hypothesis below we want to test whether the average number of CRBT subscriptions by people whom a focal person calls influences the number of subscription by that person.

Hypothesis 1: Cohesion and CRBT Diffusion
The number of CRBT subscription a focal person purchases is positively related to the average number of CRBT subscription the focal person's neighbors purchase

2.3 Role Equivalence Model

Role equivalence is also known as structural equivalence model. It is a positional model (Burkhardt, 1994). An actor is structurally equivalent to an alter if they connect to the same others. Structural equivalence model describes the competition between actor and alter that have same position in the social network. "Structural equivalence model were developed ... explicitly as a vehicle for describing the structure of role relations defining statuses across multiple networks." (Burt, 1987) For example, a medical doctor wants to maintain an image of innovativeness. After another doctor who he/she shares common friends or advisee with adopts a new technology, the doctor believes the adoption of a new technology will enhance his/her innovative reputation and effective power in the social network, so he wants to adopt before others who are in the same position as him.

In our research, we do not emphasize the competition relationship between actor and alter that are structurally equivalent. Instead, we want to emphasize that they have same or similar pattern of relations with other nodes in the network. In the role equivalence model, "the trigger to actor's adoption is adoption by the people with whom he jointly occupies a position in the social structure." (Burt, 1987) Similarity in decisions may happen when actor and alters connect to the same people. In addition to occurring when people adopt the behaviors, attitudes, and beliefs of those with whom they interact, similarity may occur when people interact with the same others (Burt, 1982, 1987; Coleman et al., 1966; Krackhardt, 1989). The degree to which a focal individual and another person interact with the same others reflects the extent to which the focal and the other are structurally equivalent (Burt, 1982). In the relational model, individuals have to interact directly in order to be similar to one another; but according to the structural equivalence model, they do not. Rather, structural equivalence is a measure of the extent to which individuals communicate with the same other people, not necessarily with one another. Thus, two individuals may be structurally equivalent even if they never communicate with one another. Actor would infer judgement of alters that have the same position in the influential flow of the network, and in order not to lose its influential power, actor would make same judgement and eventually adopt as well.

Many researchers have reanalyzed Coleman et al's seminal work. Burt (1987) reanalyzed their Medical Innovation data and drew the conclusion that contagion was not the only or the dominant factor driving diffusion. Personal preferences about adoption was also a significant factor. More importantly, Burt observed that structural equivalence alters of adopters are more likely to adopt, concluding that the effect of contagion was through structural equivalence instead of cohesion. Strang and Tuma (1993) found strong influence from doctors that are role equivalent and little influence from cohesion. Burkhardt (1994) also compared two effects, cohesion and role equivalence, with regard to users' attitude, self-efficacy belief and frequency of use of computers. He found "when people evaluate their own personal skills or self images, they rely on those close to them; when they determine job-related attitudes, they are more likely to rely on role equivalents. " Van den Bulte and Lilien (2001) reanalyzed Burt's analysis. They also compared effects from cohesion and structural equivalence. Different from Burt, who used Euclidean distance to measure structural equivalence, Van den Bulte and Lilien used proportion of exact alters' matches as their measure. Their results show, without considering marketing effort, both network effects are significant, with structural equivalence being more significant.

We conclude from the literature that people in the same position in a social network will "use each other as a frame of reference for subjective judgments and so make similar judgments even if they have no direct communication with each other" (Harkola and Greve, 1995). Our hypothesis about the role equivalence effect on the diffusion is described as below:

> **Hypothesis 2: Role Equivalence and CRBTs Diffusion**
> *The number of CRBT purchased by a focal person is positively related to the extent to which the focal person's role equivalent alters have purchased many CRBT*

In summary, we have seen that determining figuring out whether the cohesion or role equivalence model is the real driving factor of diffusion has been a research subject that many scholars have devoted in it for a long time. Different scholars have found either effect significant in their research. This subject continues to be a phenomenon that keep attracting scholars' attention and effort. In our research, we hope to see which network effect is responsible for deciding the diffusion of CRBT among cellular phone users. So far all the empirical researches only accommodate one network effect in one model, and compare the coefficients from two models using Q-test (Leenders, 2002). Although Doreian (1989) has proposed two regimes of network effect autocorrelation model, a method appropriate for this context. Such method has not been applied to this research question yet.

3 METHOD

3.1 Data

Caller ring back tones (CRBT) have become more and more popular among cellular phone users globally. CRBT gives a called party the ability to decide what a calling party will hear as he/she waits for the called party to answer. For example, a CRBT subscriber selects a popular song as his ring-back tone. When someone calls, the caller will not hear the standard plain ring-back tone but instead will hear the song until the called party (the subscriber) answers the phone or the mailbox takes over. CRBT replaces standard ring-back tones with any tune the subscriber chooses, such as a song or a joke. With the ability to set up personalized ring-back tones, subscribers can instantly express their own individuality. They also make a fashion statement by allowing other callers to hear their own personalized CRBT. Their self-satisfaction can also be fulfilled if purchasers believe others will enjoy their CRBT.

Our data were obtained from a large Indian telecommunications company (source and raw data confidential). We have cellular phone call records and CRBT purchase records over a three-month period, and phone account holders' demographic information such as age and gender. The entire data set contains approximately 26 million unique users. Since we have both the hash values of phone number for the calling party and the called party, the network we have is a directed one, where calling party is the initial node and called party is the terminal node, and the phone call is the edge between the two parties. There were about 1 billion phone calls initially in our data set. It includes all the phone calls received by the company's customers in three months. Those phone calls were initiated by customers both inside and outside of the company. Our analysis is only constrained to phone calls that occur between customers using the same provider. As a result, we have some of their demographic information such as age and gender.

Identifying reciprocated calls is important in our data set. Since the phone call social network is directed, asymmetry can exist between callers. Any asymmetric connection between two callers may indicate a weak connection, and may be less likely to indicate a relationship that provides social influence of either party on the other. An asymmetric connection indicates an unstable relationship while symmetric connections imply equal and stable connections (Hanneman and Riddle, 2005). We define reciprocity for dyads (A, B) as the condition in which A calls B and B calls A in same calendar month. We interpret reciprocity as an increase probability that the two parties are acquaintances. Thus we further constrain our analysis to include only the data that involve reciprocal dyads. Constrained by these requirements, the size of our phone call record becomes to about 197 million, calls from 1.4 million customers. Although the network yielded by these constraints is much smaller than the original one, the coefficient of structure matrix of the network still can not be estimated by maximum-likelihood, given the current computing power available. Therefore a subpopulation extraction looks necessary.

A detailed description of the preprocessed data is listed in Table 1. The dependent variable was measured as an integer variable, indicating number of CRBT a caller downloaded given a three-month period. There were 1.3 million CRBT purchases during this period. We only lost 494 observations if we limit our data to those with demographic data. There are about 7,000 distinct ring tones downloaded.

Table 1: Data Description

Variable	**Description**
Y	Dependent variable, number of CRBT purchased by caller
Gender	Gender of cellular phone account holder
Age	Reported age of cellular phone account holder
Degree	Outdegree of phone account,
W_c	cohesion effect matrix, normalized
W_{re}	role equivalence effect matrix

All these CRBT purchases in three months belong to about 580,000 distinct customers.

The independent variables included in our models are gender, which is the gender of the cellular phone account holder; age, which is the age of the account holder. Since we do not know whether the account holder is the real user, the age and gender information might not be the real information about the caller. We also include the outdegree of the caller, which is the number of unique users an account calling to, to observe the exogenous effect of number of connections. Our independent variables also includes network measures. Cohesion is defined as callers who make phone calls to each other. Cohesion assumes callers who make phone calls to each other will hear the called party's CRBT thus more likely to buy that ring-back tone or get interested in CRBT and eventually adopt the technology. Since the number of people a caller calls are drastically different, we normalize the cohesion matrix by dividing each row by the total number of adopters, to make the matrix element to be the percentage of adoption. Structural equivalence is defined as the Euclidean distance between two callers. It measures how many common friends two callers share. The more common friends two callers share, the smaller the Euclidean distance between them. However, in order to make the parameter of role equivalence have a positive relationship with high role equivalence, we use the inverse of role equivalence plus one.

3.2 Extracting Subpopulations

Before we can address the question of which network effect, cohesion or role equivalence, has the greater influence over the diffusion of CRBT, we face another challenging problem: making our analysis tractable with respect to our data. Our data size is huge. We have three months of phone call records and three months of CRBT purchase records. The number of phone call is about 200 million, and distinct phone call is more than 11 million, from about 1.4 million unique callers. That is to say, the number of nodes in our network is 1.4 million and the number of edges is close to 11 million. It is not feasible to inverse an adjacency network of this size. So we have to extract a subpopulation from the full data. Two obvious methods come to our minds: snowball sampling and spectral modularity clustering.

3.2.1 Snowball Sampling

Since our data size is large, we have to construct a subpopulation. However, random sampling with respect to these phone calls will not preserve the social network's structure at the local level. A more legitimate and efficient method to build a subpopulation in this situation is snowball sampling. This method was introduced by Coleman (1958) and Goodman (1961). Individuals in the sample are collected through a chain-referral process. The sampling procedure starts with predefining the number of steps of sampling s, then randomly drawing some nodes from the population. For each node i, we get k nodes that it connects to, so k is number of nodes that each node in one step will link to. Sampling stops after s steps. Since s and k are all defined before the sampling, it is also called s stage k name snowball sampling. It is possible that nodes linked from one node will be linked from other nodes again. This method has been extended by Salganik and Heckathorn (2004). In their method, an individual sample is formed by randomly selecting a user from the network and returning the connected component containing this user, repeating this on the remaining users until some maximum number of users is attained. This method is desirable for social network data sampling since it allows researchers to have larger sample size than other methods given available resources (Semaan et al., 2002). Snowball sampling is also referred as chain-referral sampling, link-tracing sampling, and random-walk sampling. Samples can be used to make estimates about the network connecting the population. Using information about networks constructed from snowball samples, we can "derive the population proportion in different groups" (Salganik and Heckathorn, 2004).

We acknowledge that snowball sampling has shortcomings. For example any bias in the seed selection would lead to biased sample. The estimates drawn from snowball samples are biased and cannot be use to infer from the whole population. However, it may be still suitable in our subpopulation extraction task because we are more interested in the impact of network structure on actor's adoption. It was pointed out by Snijders (1992) that snowball sampling is more appropriate for inference about the structure of the network. Since our research concentrates on the network effect of each node, snowball sampling's breath-first search principle can be part of our subpopulation extraction algorithm.

3.2.2 Spectral Clustering (Modularity)

We also can find subpopulation extraction methods in the cluster detection literature. This is a field attracting researchers in physics, mathematics, and computer science. It has been found that many networks are inhomogeneous, consisting of distinct groups, "with dense connections within groups and only sparser connections between them" (Newman, 2004). Such groups are called "communities". Researches show that communities at the local level can be quite different from each other, and even different from the global network (Newman, 2006). So analyses on local communities can give us more properties of, and information about, the network. In our paper we would like to concentrate on modularity clustering, one of the most commonly used community detection methods, and one that has been proven to be effective. The detail of the method is shown below: modularity, Q, is difference between the actual number of edges in communities and the expected number of edges. If it is large, meaning the edges are more dense then expected, the communities is cohesive. Note that i and j must belong to the same group. If the difference between actual and expected edge between two nodes that belong to same group is large then Q is large. The definition of Q is described as below. Let m be the total edges of the graph, A_{ij} be the actual number of edges between nodes i and j, P_{ij} be the expected number of edges between i and j; g_i be the group's number that node i belongs to; g_j be j's group number; and $\delta(g_i, g_j)$ indicates whether i and j belong to the same group, where 1 stands for the same group and 0 otherwise.

$$Q = \frac{1}{2m} \sum_{i}^{n} \sum_{j=1}^{n} [A_{ij} - P_{ij}] \delta(g_i, g_j)$$

The time complexity of modularity is not fast if we have huge network. For example, the complexity of spectral clustering method to bisect a graph is $O(n^3)$, where n is the number of nodes in the network. A faster method using the min-cut, Kernighan-Lin algorithm, can reach complexity of $\Theta(n^2 log(n))$. The fastest algorithm still have $O(nm)$ time complexity, where m is the number of edges in the network. Modularity methods need to know the number of communities in advance, and most of the methods can only divide the network into two communities. A notable work in community detection that achieves better complexity is by Clauset (2005). His method improves the local modularity maximization by using the greedy algorithm. This method's complexity is $O(n^2 d)$, where n is the number of nodes traversed and d is the mean degree of node. However, if the number of nodes is at the level of, or beyond one million, such a method is still not tractable. Summarizing the current status of community detection method complexity, Newman writes:

> "I don't think the spectral algorithm will work for such a large network [N=1 million]. Remember that the algorithm is $O(mn)$, where m is edges and n is vertices, so you're talking about 10^{15} operations to perform the eigenvector calculation, which is not feasible with current computers.
>
> Basically, if you're working with networks that large, you are limited to $O(n)$ or $O(nlogn)$ algorithms, of which there are only a few, none of which work very well. There was a new multi-scale method that came out in the last year that might be worth looking at, and there's an improved version of the greedy algorithm of Clauset et al. that can do a good job in some circumstances. But overall the situation is not very promising for extremely large networks at present."

Such concerns highlights the challenge of this topic; there is no method that can really solve this problem with a large-scale network. Some well-accepted methods have to make a trade-off between speed and stability. As Newman argues about the performance of his spectral clustering method.

> "I should point out that this code should not be used for gauging the speed of the algorithm. It uses dense-matrix methods to do the eigenvector calculation, which are very slow. A much faster implementation is possible using sparse matrix methods. I have such an implementation, but it's not very stable. For merely testing the efficacy of the algorithm on small networks this implementation is better."

Given these problems, neither snowball sampling nor modularity can extract subpopulations with the ideal features. So we developed our own method – Two-Stage Clustering and Pruning (TSCP) Algorithm.

3.2.3 Two-Stage Clustering and Pruning (TSCP) Algorithm

In this section, we present our subpopulation extraction method that can identify a dense and isolated local community without processing the whole network. We attempt to find subpopulations that have the following characteristics: first, they have high density within the community. Internal density shows strong connections among actors in the community, so contagion is likely to happen. Second, these network are reasonably dense but not fully connected. We need to find dense subpopulations to feature network influence, but not so dense that we lose variation of degrees of each actor. Third, relatively few ties from within the community to the outside, so the subpopulation is relatively isolated and integral itself.

Our method is designed to find subpopulations having the preferred characteristics as above and consists of two steps, clustering and pruning.

Clustering Our method starts from a random node in the global network, and gets a snowball sample of a predefined size n. Our sampling will also include the leaf nodes (pedants) at the same level as the n-th node, if they were not included in the n nodes yet. The size of such set is N_1, and $N_1 \geq n$. The reason for getting N_1 nodes instead of just getting n is that we can get a more completed subset, and record more completed connection among nodes to be included in our subpopulation. After getting a snowball sample, define the following variables: N_2 is number of nodes connecting to those added in the the last step of snowball sampling, defined as boundary nodes, I is the total number of connections in the sample, E is the number of connections (ties) between nodes in last step of snowball and N_2. For illustrations of each variable, see Figure 1.

We then construct the adjacency matrix of nodes we get from snowball sampling. In adjacency matrix $\mathbf{A}$, element A_{ij} shows whether node i connects to node j or not, which means whether i has made phone call to j or not in the context of this research.

$$A_{ij} = \begin{cases} 1 & \text{if node } i \text{ connects to node } j \\ 0 & \text{otherwise} \end{cases}$$

$$\text{where } i \in [1, N_1], j \in [1, N_1 + N_2]$$

This adjacency matrix records information of reciprocated phone calls. A value of 1 in columns 1 to N_1 indicating phone calls between both parties in the snowball sample, and 0 meaning no phone calls existing between those two parties; a 1 in columns $N_1 + 1$ to $N_1 + N_2$ indicating phone calls between caller inside of the snowball sample and outside of it, and 0 meaning no phone calls.

We denote the matrix with N_1 nodes in snowball sample as $\mathbf{I}$, since it has all internal connections of a subpopulation, and denote the matrix with N_2 nodes $\mathbf{E}$, since the latter has external connections to the network outside of the subpopulation. (See Figure 2) The elements of $\mathbf{E}$ have the form $E_{i,j} = 1$ if i connects to j, and 0 otherwise. The product $\mathbf{C} = [\ \mathbf{I}\ \mathbf{E}\ ] \cdot [\ \mathbf{I}\ \mathbf{E}\ ]^T$ gives us the total common connection between nodes in $\mathbf{I}$ (Figure 3). The product matrix $\mathbf{C}$ is at the size of $N_1 \times N_1$. The diagonal

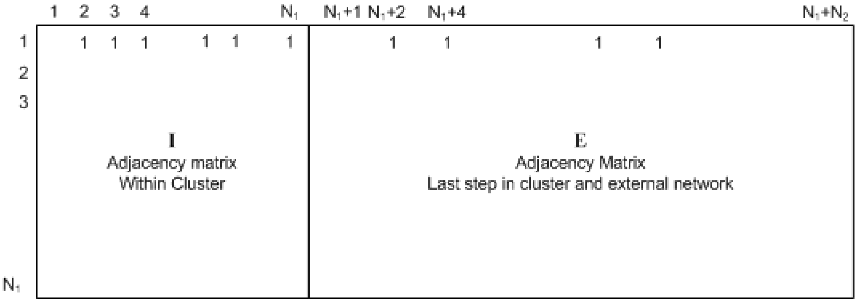


Figure 1: Adjacency matrix of a cluster **I** and boundary nodes **E**

element of **C**, C_{ii}, is the degree of each node (caller) i, while the off The diagonal of Hadamard product (element-wise product) $\mathbf{N} = \mathbf{I} \circ \mathbf{C}$ gives us the number of common called parties between nodes in **I**, and the two nodes must also be connected. We then find a subset of nodes that have the highest degree in $\mathbf{I} \circ \mathbf{C}$ at the size of N_p, We call this step pruning and call the set of nodes $\mathbf{N}_{PT}$.

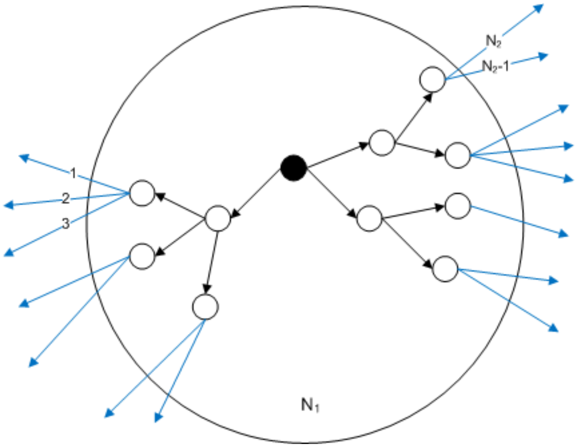


Figure 2: Nodes and edges included in matrices **I** and **E**

Figure 4(a) below is an example of a snowball sample at the size of 200, before the pruning step.

Using this method we can get a smaller more cohesive cluster, denote its size as N_3. Next we use the nodes in $\mathbf{N}_{PT}$ to start our next round snowball sampling. Likewise, we set a predefined number n and include all the pedants at the same level of n-th node again. We get the matrices **I** and **E**, and start the second round of pruning and get **C** for this round. Then we start another round of clustering. After several rounds of clustering, we will fall into a relatively more densely connected community compared to the region in the network with which started sampling. In the last round of clustering, after getting $\mathbf{N}_{PT}$, we snowball sample a larger set than n_1, which is intended to include more nodes in the global network, we then move on to pruning stage.

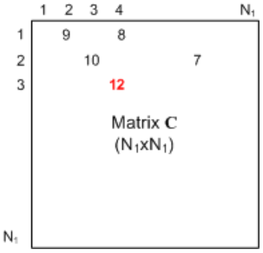
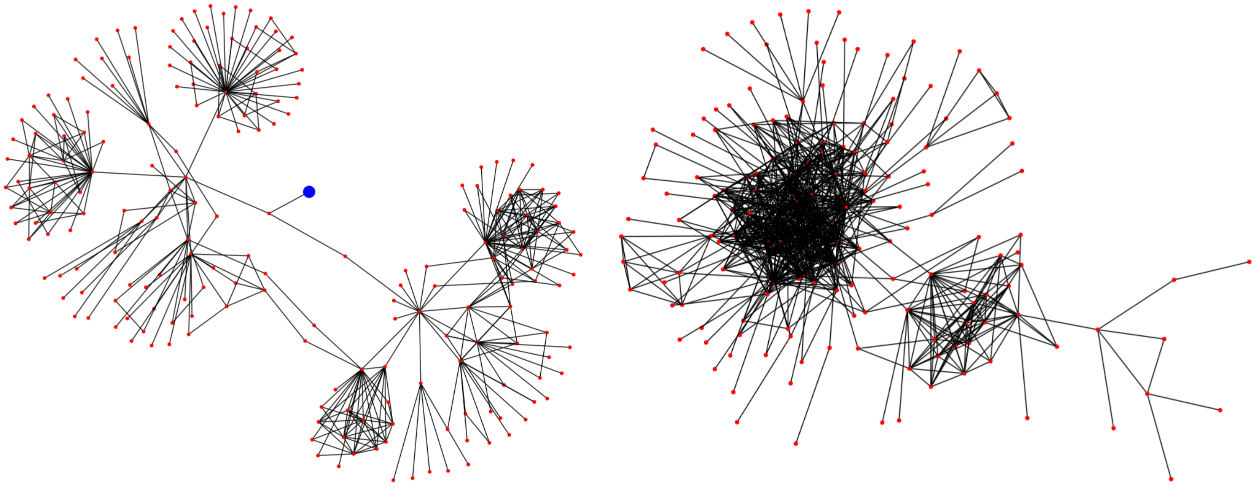


Figure 3: Product of $[\,\mathbf{I}\ \mathbf{E}\,] \cdot [\,\mathbf{I}\ \mathbf{E}\,]^T$

Pruning From the last round of clustering, we get a final pair of matrices **I** and **E**. Then we can use $\frac{E}{I}$ (Krackhardt and Stern, 1988) to evaluate the cohesiveness of the final cluster. We prune the nodes that have the largest $e_j - i_j$, where e_j is the number of nodes in **E** that j connects to, and i_j is the number of nodes in **I** that j connects to. If $\frac{E}{I} < t$ ($t = 0.1$ for now) then we stop snowballing. A subpopulation obtained through three rounds of snowballing and pruning is shown in Figure 4(b). We can tell that the subpopulation is much more dense than at the very beginning. The pseudocode of our algorithm is provided in the Appendix.



(a) A subpopulation at size of 200, before pruning

(b) A cluster at the size of 200, after pruning

Figure 4: Initial subpopulation and final subpopulation after pruning

The complexity of our TSCP algorithm is $O(N_1 N_2)$, where N_1 is the size of initial snowball sample, and N_2 is the size of the boundary nodes. If we want to identify a subpopulation at the size of 1000, we set N_1 to be 200 based on our empirical experience, if the mean degree of each node is 10, N_2 is about 1000, so the complexity of TSCP algorithm is $O(10^5)$. On the other hand, Clauset's greedy maximization of local modularity algorithm, which is considered as one of the fastest community detection algorithms, has a complexity of $O(10^7)$. Our algorithm is asymptotically faster than most of the current methods. Empirically we have compared the running speed of our method and Newman's spectral clustering. It took spectral clustering two hours to detect a subpopulation of size 1000, while only took TSCP 15 minutes to detect a subpopulation at the same size. A rigorous benchmark comparison will be conducted in our future works.

3.3 Two Regimes of Network Effects Autocorrelation Model

We use a network autocorrelation model in order to investigate network effects on CRBT diffusion. The network autocorrelation model takes both interdependence of actors and their local effect such as demographics into consideration. Such interdependence are described by a weight matrix. Examples of such weight matrices are the adjacency matrix (Coleman et al., 1966) and the Euclidean distance between two actors (Burt, 1987). Most of the models can only accommodate one network effect, for example Burt's model, and Leenders' model. The goal of our research requires investigation of which effect, cohesion or role equivalence, plays a more significant role in CRBT diffusion, thus we adopted Doreian (1989) two regimes of network effects autocorrelation model, which incorporates two network weight matrices. Doreian's model can capture both actor's intrinsic opinion and influence from alters in his social network. The model is described as below:

$$y = X\beta + \rho_1 W_c y + \rho_2 W_{re} y + \varepsilon \tag{1}$$

where y is the dependent variable, an integer variable representing the number of CRBT a caller purchased; X is a vector of explanatory variables including age, gender and outdegree (number of different people that a customer places phone calls to); W represents the social structure underlying each autoregressive regime. W_c is the matrix of weight for cohesion; W_{re} is the matrix of weight for role equivalence; ρ_1 and ρ_2 are the parameters of two network effect respectively; ε is normally distributed disturbance term. With this model, we can investigate how networks affect people's decisions about the number of CRBTs purchased.

We use Euclidean distance to measure structural equivalence. In a directional network with non-weighted edges the Euclidean distance between two actors i and j is the sum of squared difference between the nodes that i and j connect to respectively, and from all nodes to i and j respectively. The distance is shown in the equation (2).

$$d_{ij} = \sqrt{\sum_{k=1, k\neq i,j}^{N} (A_{ik} - A_{jk})^2} \tag{2}$$

where

$$A_{ik} = \begin{cases} 1 & \text{if node } i \text{ and } k \text{ are neighbors} \\ 0 & \text{otherwise} \end{cases}$$

The larger d between node i and j, the less structurally equivalent they are. We get the inverse of d_{ij} plus one in order to construct a measure with a positive relationship with role equivalence:

$$s_{ij} = \frac{1}{d_{ij} + 1}$$

3.4 Meta-analysis

After collecting of subpopulations and estimating the network effects, it is critical to know whether these effects are consistent across different studies. Thus we turn to meta-analysis for assessing the consistency of these effects. Meta-analysis not only directly takes effect sizes, instead of p-value, into consideration, it also accommodates observed dispersion due to random sampling variation. Meta-analyses are based on two statistical models, the fixed-effect model and random-effects model. The fixed-effect model assumes that effect size is the same across all studies. However, given there are many factors impacting the effects, such as the size of subpopulations, the connection strength within a community versus the strength between a community and outside of it, it is not reasonable to assume all the network effect sizes are the same in all the subpopulations extracted, thus the assumption of a same effect size does not likely to hold in our study. On the other hand, random-effects models address variations across studies. Random-effects models assume that true effects are normally distributed (Borenstein et al., 2009). So we would like to use random-effects model to conduct meta-analysis.

Meta-analysis uses observed effects from different studies to estimate the population effect. In our context, we could use network effects from different subpopulations to estimate the effect for the whole network. The random-effects model calculates a weighted mean as a precise estimate of the overall mean. The method is described as follow:

First, define

$$Q = \sum_{i=1}^{k} w_i \beta_i^2 - \frac{\left(\sum_{i=1}^{k} w_i \beta_i\right)^2}{\sum_{i=1}^{k} w_i} \tag{3}$$

$$w_i = \frac{1}{\sigma_i^2}$$

$$df = k - 1$$

$$k = \text{number of studies (subpopulations)} \tag{4}$$

and

$$C = \sum_{i=1}^{k} w_i - \frac{\sum_{i=1}^{k} w_i^2}{\sum_{i=1}^{k} w_i} \tag{5}$$

we can get the sample estimate of the between-studies variance:

$$\hat{\tau}^2 = \frac{Q - df}{C}$$

After between-variance is obtained, we can get the total variance defined as below:

$$\sigma_i^{*2} = \sigma_i^2 + \hat{\tau}^2$$

We can then get the weight of each study:

$$w_i^* = \frac{1}{\sigma_i^{*2}}$$

Finally we can compute the mean effect:

$$\mu^* = \frac{\sum_{i=1}^{k} w_i^* \beta_i}{\sum_{i=1}^{k} w_i^*}$$

μ^* is the measure we use to estimate the global network effect in our meta-analysis.

4 RESULTS

4.1 Extracted subpopulation and Descriptive Analysis

69% customers have single accounts, 29% have two accounts on same number, 2% have three accounts or more. 7% of the account holders are females. 91% are males, and 2% missing. The average age of customer is 39 years old.

There were 389,964 CRBT purchases in the first month; 518,192 purchases in the second month; and 424,616 purchases in the third month. From the distribution we found the number of CRBT purchases in the second month is higher than the other two months. (See Table 2.) There were 7,176 different CRBTs purchased during this period. There were 4,586 CRBTs purchased in the first month; 5,134 CRBTs purchased in the second month; and 5,122 CRBTS purchased in the third month. There were 576,358 unique customer purchased CRBT in the three months period. 35% of them (259,170) bought only one CRBT, the rest of them bought more than one. 294,210 customers purchased CRBT in the first month; 340,311 customers purchased CRBT in the second month; and 313,227 customers purchased CRBT in the third month. The top three ring tones downloaded are MERE SAPNON KI RANI (50,503 times), CHUKAR MERE MAN KO (43,055 times), and O SAATHI RE (42,548 times). We also observe stable adoption behavior. Among the 447,193 callers who did not buy any CRBT in the first month, there are 169,069, and 164,006 of them bought CRBT in the second month and the third month, respectively. We can conclude that once callers adopt CRBT, they will continue buying CRBT.

We extracted 20 subpopulations in total, which can be categorized into three groups according to their sizes. We have seven subpopulations of size about 200, from 150 to 263; eight subpopulations of size about 500, from 431 to 677. five subpopulations of size about 1000, from 774 to 977. Those subpopulations can also categorized as two groups based on E/I index, where E is the number of

Table 2: Descriptive statistics of CRBT

Month	**Num. of CRBT Downloads**	**Pct.**	**Distinct Tones**	**Pct.**
First month	389,964	29.3%	4,586	30.9%
Second month	518,192	38.9%	5,134	34.6%
Third month	424,616	31.9%	5,122	34.5%
Total	1,332,772	100%	14,842	100%

external caller connection, and I is the number of internal caller connection in the subpopulation. One group has the E/I index at 0.4, the other has the index at 0.1. The former group has stronger connection to other communities, while the latter group are relatively isolated communities. We want to confirm whether there is difference with respective to network effect between these two groups.

Table 3: Extracted subpopulations

Subpopulation	N	$N_{[y>0]}$	I	E	E/I	I/N
1	150	62	730	291	0.4	4.9
2	159	82	1528	610	0.4	9.6
3	171	59	1712	161	0.1	10.0
4	202	116	1452	580	0.4	7.2
5	213	82	652	227	0.4	3.1
6	238	93	780	78	0.1	3.3
7	263	142	902	87	0.1	3.4
8	431	206	2910	291	0.1	6.8
9	447	136	6508	646	0.1	14.6
10	465	312	5779	2287	0.4	12.4
11	485	249	3504	349	0.1	7.2
12	553	285	3962	394	0.1	7.2
13	563	312	2402	960	0.4	4.3
14	597	470	13281	1319	0.1	22.2
15	677	431	6134	612	0.1	9.1
16	774	509	11006	1099	0.1	14.2
17	789	374	3124	312	0.1	4.0
18	894	521	5413	534	0.1	6.1
19	962	489	5920	592	0.1	6.2
20	977	497	5960	594	0.1	6.1
Mean	501	271	4183	601		8.1
S.D	277	171	3422	513		4.8

The descriptive statistics of independent variables for each subpopulation is listed in Table 4.

Table 4: Descriptive statistics of independent variables

	Subpopulations							
Variables	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	
Gender	0.0067	0.16	0.19	0.12	0.033	0.11	0.065	
0=male	(0.082)	(0.51)	(0.50)	(0.45)	(0.18)	(0.40)	(0.29)	
Age	36	37	45	42	38	42	42	
	(11)	(12)	(11)	(12)	(11)	(11)	(12)	
Degree	5.7	15.1	13.6	10.0	3.6	4.0	4.5	
	(7.6)	(14.5)	(14.9)	(10.6)	(5.1)	(5.2)	(5.6)	
N	150	159	171	202	213	238	263	
E/I	0.4	0.4	0.1	0.4	0.4	0.1	0.1	
	(8)	**(9)**	**(10)**	**(11)**	**(12)**	**(13)**	**(14)**	**(15)**
Gender	0.23	0.37	0.57	0.12	0.27	0.12	0.050	0.13
	(0.60)	(0.78)	(0.89)	(0.41)	(0.62)	(0.42)	(0.25)	(0.44)
Age	46	55	48	45	40	42	36	39
	(13)	(4)	(12)	(13)	(13)	(13)	(12)	(13)
Degree	9.9	21.7	18.1	9.3	9.8	5.3	3.4	12.9
	(12.9)	(11.9)	(18.6)	(13.6)	(15.5)	(6.7)	(1.8)	(10.9)
N	431	447	465	485	553	563	597	677
E/I	0.1	0.1	0.4	0.1	0.1	0.4	0.1	0.1
	(16)	**(17)**	**(18)**	**(19)**	**(20)**			
Gender	0.27	0.15	0.038	0.062	0.062			
	(0.63)	(0.44)	(0.20)	(0.29)	(0.29)			
Age	40	41	39	41	41			
	(14)	(13)	(12)	(12)	(12)			
Degree	21.0	4.8	8.5	8.0	8.0			
	(14.7)	(6.8)	(12.9)	(12.6)	(12.6)			
N	774	789	894	962	977			
E/I	0.1	0.1	0.1	0.1	0.1			

4.2 Analysis of Network Autocorrelation Model

We first show the results of network autocorrelation model for small subpopulations, where the dependent variable is the number of CRBT purchases. Table 5 present the results for subpopulations at the size of 200 level, ranging from 150 to 263. We find that a community at the size of about 200, role equivalence effect is consistently insignificant, while cohesion effect shows inconsistent pattern. Four out of seven subpopulations show significant cohesion effect at level of $p < 0.05$. However, if we separate these seven subpopulations by E/I ratio, we found two insignificant effects belong to high E/I ratio subpopulations, and only one insignificant cohesion effect belong to low E/I ratio group. Thus if a small community (subpopulation) at the size of about 200 and high $E/I = 0.4$, the network effects are relatively not significant. If a subpopulation has strong connection with the global network, the network effects of are not significant. When we have isolated subpopulation, with $E/I = 0.1$, most of the cohesion effects are positive and statistically significant. Such result confirms our hypothesis. It show that callers receive strong influence through direct connections to alters in the same community who have already adopted. An example illustrates: if a caller calls more CRBT subscribers, he gets exposure to more ring-back tones, and is more likely to hear ring tones interest him, thus he is more likely to try different ring tones. In subpopulation 1, 4 and 7, the effects of outdegree are positive and significant at 0.0001 level, which suggest if a caller calls to more people, the probability of purchasing a CRBT increases.

One could argue that the size of subpopulation is likely to have an impact on network effects. So we extract subpopulation at different sizes from the network. If network effects are universal in the global network, when we extract random local communities from the global network, these network effect should be consistent. We then increase the size of subpopulation we want to extract to 500. From the results in Table 6 we find cohesion is consistently statistical significant in all subpopulation at this size level. This again confirms our hypothesis about cohesion, when the network is at the size of about 500, callers' CRBT adoption is influenced by the people who he calls directly. We do not find support for our hypothesis about role equivalence though since the effect from all eight subpopulations are insignificant. Seven out of eight of the subpopulations at 500 level show that cohesion is more significant than role equivalence. For the role equivalence model, caller evaluate alters who are in the same position of a phone call network as him. Same position in a social network means people are in the same kinds of relations, with the same kinds of people. In this case, people who are role equivalent are most immediate competitors in the network. Degree is positive and statistically significant in nearly all subpopulations at the size of 500.

We observed inconsistent effect of cohesion and role equivalence, respectively, across different random subpopulation, at smaller scales. However above a threshold subpopulation size, we did observe consistent effects. Specifically, when the subpopulation is above a threshold in our case 500, we observe a more consistent pattern – cohesion is more significant than role equivalence. (See Table 7). The results show that cohesion effects all are significant at 0.0001 level except one, ranging from 0.072 to 0.56. The role equivalence effect, ranging from 0.00037 to 0.0038, is not significant. The results show that the purchase of ring tone is impacted by solely cohesion. The cohesion effect implies that, of all the alters a caller calls, if the average number of CRBT adopted by alters gets higher, the probability of a caller purchasing CRBT increases as well. The role equivalence effect implies that, if a caller shares more common called parties with an alter who adopted CRBT, the more CRBT those alters purchase, then the more the caller will also purchase. Subpopulation 11 and 15 show significant role equivalence effect at 0.05 level. One explanation is that in a cellular phone call social network, parties who call each other are likely to be friends or belong to same group under some relationship. In this case the enthusiasm of showing others about his adoption of frontier fashion and individuality is higher. The satisfaction of letting friends appreciate his fashion taste or simply an interesting tone is also higher. Motivated by this thought, a perceived competition is created among these subscribers. Actor will know about an alter he does not necessary call to has adopted ring-back tone through common friends they both call. The more ring tones those alters bought, the more CRBT the actor will adopt. Degree is significant at 0.0001 level in all subpopulations at this size, indicating a significant relationship between number of CRBT adoption and number of people call to. Gender and age consistently do not have any relationship with number of CRBT adopted.

Table 5: Network Autocorrelation, , subpopulation 1 to 7

	Subpopulations						
Variables	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**
Gender	-0.20	-0.47	0.38	0.79	-0.78	0.27	2.0^{*}
	(3.3)	(0.43)	(0.51)	(0.63)	(0.77)	(0.51)	(0.75)
Age	-0.026	0.0094	0.023	0.021	0.016	0.024	$0.028^{\dagger}$
	(0.017)	(0.014)	(0.018)	(0.021)	(0.010)	(0.015)	(0.017)
Degree	0.12^{****}	0.0071	-0.0014	0.080^{****}	0.0091	0.0043	0.069^{***}
	(0.016)	(0.0076)	(0.0088)	(0.013)	(0.014)	(0.020)	(0.020)
Cohesion	0.19^{*}	0.22	0.28^{**}	0.18^{*}	0.015	0.056	0.18^{**}
	(0.061)	(0.16)	(0.087)	(0.087)	(0.077)	(0.070)	(0.060)
Role	0.011	$0.034^{\dagger}$	0.0059	-0.0059	0.0062	0.0079	-0.0015
equivalence	(0.020)	(0.019)	(0.022)	(0.015)	(0.012)	(0.011)	(0.0080)
N	150	159	171	202	213	238	263
I	730	1528	1712	1452	652	780	902
E	291	610	161	580	227	78	87
E/I	0.4	0.4	0.1	0.4	0.4	0.1	0.1

†: $p < 0.10$, *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$, ****: $p < 0.0001$

Table 6: Network Autocorrelation, subpopulation 8 to 15

	Subpopulations							
Variables	**(8)**	**(9)**	**(10)**	**(11)**	**(12)**	**(13)**	**(14)**	**(15)**
Gender	−0.034	−0.39	−0.11	0.094	1.4****	0.69†	0.90	0.95*
	(0.33)	(0.26)	(0.31)	(0.46)	(0.36)	(0.39)	(0.87)	(0.47)
Age	0.053***	0.065	0.086****	0.064****	0.037**	0.052****	−0.0080	0.041**
	(0.014)	(0.040)	(0.021)	(0.014)	(0.013)	(0.012)	(0.017)	(0.016)
Degree	0.030**	0.0064	0.040****	0.043****	0.024***	0.058****	0.029****	0.026**
	(0.028)	(0.031)	(0.027)	(0.022)	(0.019)	(0.033)	(0.016)	(0.025)
Cohesion	0.13*	0.68****	0.15*	0.12**	0.13***	0.20****	0.68****	0.30****
	(0.052)	(0.12)	(0.062)	(0.044)	(0.040)	(0.039)	(0.10)	(0.064)
Role	−0.010†	−0.061	−0.0066	−0.0098*	−0.0017	−0.0049†	0.00064	0.0041*
equivalence	(0.0058)	(0.038)	(0.0047)	(0.0041)	(0.0032)	(0.0025)	(0.0052)	(0.0026)
N	431	447	465	485	553	563	597	677
I	2910	6508	5779	3504	3962	2402	13281	6134
E	291	646	2287	349	364	960	1319	612
E/I	0.1	0.1	0.4	0.1	0.1	0.4	0.1	0.1

†: $p < 0.10$, *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$, ****: $p < 0.0001$

Table 7: Network Autocorrelation, , subpopulation 16 to 20

	Subpopulations				
Variables	**(16)**	**(17)**	**(18)**	**(19)**	**(20)**
Gender	0.38	−1.1***	0.97†	−0.13	−0.14
	(0.33)	(0.32)	(0.58)	(0.39)	(0.39)
Age	−0.0014	0.022*	0.032***	0.024**	0.023**
	(0.014)	(0.0091)	(0.0091)	(0.0082)	(0.0081)
Degree	0.029****	0.069****	0.048****	0.034****	0.034****
	(0.0064)	(0.010)	(0.0050)	(0.0050)	(0.0050)
Cohesion	0.56****	0.17****	0.072*	0.15****	0.15****
	(0.066)	(0.030)	(0.033)	(0.033)	(0.032)
Role	0.0038	0.00037	0.00078	0.00051	0.00072
equivalence	(0.0028)	(0.0016)	(0.0015)	(0.0015)	(0.0015)
N	774	789	894	962	977
I	11006	3124	5413	5920	5960
E	1099	312	534	592	594
E/I	0.1	0.1	0.1	0.1	0.1

†: $p < 0.10$, *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$, ****: $p < 0.0001$

4.3 Meta-analysis

Our meta-analysis for network effect shows some interesting results (Table 8). The pooled mean cohesion effect is 0.21, with a 95% confidence interval at (0.16, 0.26). This result shows the mean cohesion result across all subpopulations is significantly different from 0, and thus confirms our hypothesis about the network cohesion effect on purchase decision. It shows the strong tie to people who already adopted product could increase the number of an actor's adoption.

We then break down our subpopulations to two groups. (See Table 9 and 10). One group consists of subpopulations with weaker internal connection within community ($E/I = 0.4$), the other group consists of subpopulations with stronger internal connection within community ($E/I = 0.1$). Through the comparison of the two pooled mean effect, we see quite different ranges for these two groups. The external strongly connected group has a pooled mean at 0.17, with a 95% confidence interval at (0.12, 0.22); while the interval strongly connected group has a pooled mean at 0.23, with a confidence interval at (0.16, 0.30). The results show the cohesion effect is stronger among people in a densely connected community, though the effect is signivicant for both more and less densely connected communities. The result of meta-analysis confirmed our hypothesis that if people within a community has stronger connections, then the diffusion of CRBT will be significantly affected by direct contact, in this context, phone calls.

The pooled effect for role equivalence is −0.000273 (Table 11), but it is not statistically significant; the 95% confidence interval is (−0.0021, 0.0015). So our hypothesis about role equivalence on diffusion is rejected – it has no significant impact on adoption of multiple products. In order to remove the effect that might be caused by heterogenous communities because of different E/I ratio, we separate them into two different groups. As Table 12 shows, the role equivalence effect is statistically insignificant with a size of −0.0036, and a 95% C.I. of $(-0.0092, 0.0020fd)$, when $E/I = 0.4$. The effect is not statistically significant either when $E/I = 0.1$, with a mean at 0.00037, and a 95% C.I. of (−0.0013, 0.0020). This means if the communality is strongly connected then there exists a competition effect among actors (Table 13).

Table 8: Meta-analysis for cohesion effect, all subpopulation pooled

		95% C.I.		
Subpopulation	ρ_c	**Lower**	**Upper**	**Weight**
1	0.19	0.0700	0.31	1.0
2	0.22	−0.094	0.53	0.42
3	0.28	0.11	0.45	0.82
4	0.18	0.0095	0.35	0.82
5	0.15	−0.14	0.17	0.90
6	0.056	−0.081	0.19	0.95
7	0.18	0.062	0.30	1.0
8	0.13	0.028	0.23	1.1
9	0.68	0.44	0.92	0.60
10	0.15	0.028	0.27	1.0
11	0.12	0.034	0.21	1.2
12	0.20	0.12	0.28	1.2
13	0.30	0.18	0.42	0.93
14	0.68	0.48	0.88	0.73
15	0.30	0.17	0.43	1.0
16	0.56	0.43	0.69	0.99
17	0.17	0.11	0.23	1.3
18	0.072	0.0073	0.14	1.2
19	0.15	0.085	0.21	1.2
20	0.15	0.087	0.21	1.3

Summary effect = 0.21, 95% C.I. = (0.16, 0.26)

Table 9: Meta-analysis for cohesion effect, subpopulation with $E/I = 0.4$ pooled

		95% C.I.		
Subpopulation	ρ_c	**Lower**	**Upper**	**Weight**
1	0.19	0.070	0.31	1.1
2	0.22	−0.094	0.53	0.15
4	0.18	0.0095	0.35	0.52
5	0.015	−0.14	0.17	0.66
10	0.15	0.028	0.27	1.0
13	0.20	0.12	0.28	2.6

Summary effect = 0.17, 95% C.I. = (0.12, 0.22)

Table 10: Meta-analysis for cohesion effect, subpopulation with $E/I = 0.1$ pooled

		95% C.I.		
Subpopulation	ρ_c	**Lower**	**Upper**	**Weight**
3	0.28	0.11	0.45	0.82
6	0.056	−0.081	0.19	0.93
7	0.18	0.062	0.30	1.00
8	0.13	0.028	0.23	1.1
9	0.68	0.44	0.92	0.63
11	0.12	0.034	0.21	1.1
12	0.13	0.052	0.21	1.1
14	0.68	0.48	0.88	0.74
15	0.30	0.17	0.43	0.97
16	0.56	0.43	0.69	0.96
17	0.17	0.11	0.23	1.2
18	0.072	0.0073	0.14	1.2
19	0.15	0.085	0.21	1.2
20	0.15	0.087	0.21	1.2

Summary effect = 0.23, 95% C.I. = (0.16, 0.30)

Table 11: Meta-analysis for role equivalence effect, all subpopulation pooled

		95% C.I.		
Subpopulation	ρ_{re}	**Lower**	**Upper**	**Weight**
1	0.011	−0.028	0.050	0.041
2	0.034	−0.0032	0.071	0.045
3	0.0059	−0.037	0.049	0.034
4	−0.0059	−0.035	0.024	0.072
5	0.0062	−0.017	0.030	0.11
6	0.0079	−0.014	0.029	0.13
7	−0.0015	−0.017	0.014	0.24
8	−0.010	−0.021	0.0014	0.44
9	−0.061	−0.14	0.013	0.011
10	−0.0066	−0.016	0.0026	0.64
11	−0.0098	−0.018	−0.0018	0.81
12	−0.0017	−0.0080	0.0046	1.2
13	−0.0049	−0.0098	0.00	1.7
14	0.00064	−0.0096	0.011	0.54
15	0.0041	−0.0010	0.0092	1.6
16	0.0038	−0.0017	0.0093	1.4
17	0.00037	−0.0028	0.0035	2.6
18	0.00078	−0.0022	0.0037	2.8
19	0.00051	−0.0024	0.0034	2.8
20	0.00072	−0.0022	0.0037	2.8

Summary effect = -0.000273, 95% C.I. = $(-0.0021, 0.0015)$

Table 12: Meta-analysis for role equivalence effect, subpopulation with $E/I = 0.4$ pooled

		95% C.I.		
Subpopulation	ρ_{re}	**Lower**	**Upper**	**Weight**
1	0.011	−0.028	0.050	0.12
2	0.034	−0.0032	0.071	0.13
4	−0.0059	−0.035	0.024	0.21
5	0.0062	−0.017	0.030	0.32
10	−0.0066	−0.016	0.0026	1.7
13	−0.0049	−0.0098	0.00	3.6

Summary effect = -0.0036, 95% C.I. = $(-0.0092, 0.0020)$

Table 13: Meta-analysis for role equivalence effect, subpopulation with $E/I = 0.1$ pooled

		95% C.I.		
Subpopulation	ρ_{re}	**Lower**	**Upper**	**Weight**
3	0.0059	−0.037	0.049	0.021
6	0.0079	−0.014	0.029	0.082
7	−0.0015	−0.017	0.014	0.15
8	−0.010	−0.021	0.0014	0.28
9	−0.061	−0.14	0.013	0.0070
11	0.0098	−0.018	−0.0018	0.54
12	0.0017	0.0080	0.0046	0.83
14	0.00064	−0.0096	0.011	0.35
15	0.0041	−0.0010	0.0092	1.2
16	0.0038	−0.0017	0.0093	1.0
17	0.00037	−0.0028	0.0035	2.3
18	0.00078	−0.0022	0.0037	2.4
19	0.00051	−0.0024	0.0034	2.4
20	0.00072	−0.0022	0.0037	2.4

Summary effect = 0.00037, 95% C.I. = $(-0.0013, 0.0020)$

5 CONCLUSION

The debate among researchers about two classes of network models, cohesion and role equivalence, and the impact on diffusion in social networks still persists. Some researchers believe it is direct contact between actor and alter that triggers the adoption of actor. Some believe it is social comparison or competition from the actors who are in the same social positions. Both camps have found empirical evidence of their claims (Coleman 1962, Burt 1987, Leenders 1992). However, other than Coleman's classical *Medical Innovation* data, few new data sets have been used to address this research question. Reconciling these findings is very important because the social network is a key medium of diffusion, and figuring out which network effect drives social influence can help us understand the mechanism of diffusion. In our study, we attempt to readdress this important but unresolved question.

One large challenge from this research is to make our analysis tractable. This challenge comes from the size of our data – both the number of actors and connections in millions. It is impossible to analyze the effects of the whole network given currently available computing power. One way to solve this problem is by analyzing subpopulations extracted from the global network. However, subpopulation or community extraction is not a trivial problem either. It has been long known that there are many open questions in extracting communities from networks. The most significant is performance. The complexity of most of the community detection methods are of $O(nm)$, where n and m are number of actors and connections in the network, respectively. We designed an innovative algorithm to extract local communities (subpopulations) from large scale network. Our method does not require any parameters about the network. It has a complexity of $O(n_1 n_2)$, where n_1 is the size of initial sample, usually with the size of 10^2, n_2 is the size of boundary nodes, with the size of 10^3. So if the size of subpopulation we want to extract is much smaller than the whole population (for example 1000 versus 1 million), our method has a complexity of $O(10^5)$, compared to spectral clustering's $O(10^{13})$, and greedy local modularity maximization's $O(10^7)$. Given the fact that more and more social networks data are large scale, our method provides a solution for extract significant local communities.

Using the subpopulations extracted, we analyze the effects of cohesion and role equivalence's on number of CRBT purchases by using Doreian's two regimes of network effects autocorrelation model. Our study is one of the very few to investigate multiple network effects on diffusion. Our results show that when the size of community is small (at the levels of 200 and 500) the result is relatively not consistent, when size is large (at the level of 1000), result show consistent pattern. For the number of products purchased, cohesion has a more significant impact. So the strength of communication or connection still drives repetitive purchase behavior. After normalization we see large effect size from role equivalence. The effect is not significant though.

We also use random-effects meta-analysis to summarize the effect from cohesion and role equivalence across subpopulations. Such analysis can tell us whether these network effects are universal characteristics of the whole network, or just local characteristics from subpopulations. The results show that cohesion does have a statistically significant pooled effect across our studies, regardless of whether the subpopulation is isolated. This result leads us to believe there is a universal effect size for cohesion. Strongly internally connected communities see a strong and statistically significant cohesion effect. Also, role equivalence effect is only significant in some internally strongly connected isolated subpopulations. Such result make us believe that competition only exist in a community that members are strongly connected to each other.

There are some limitations to our research that need to be addressed. First, we treat the network effect W as fixed effect, but the acutal effect should be treated as random. Second, we could have a problem of endogeneity. The network effects are possibly affected by other factors. Third, other influences on CRBT adoption are not captured in our model, *e.g.* a marketing campaign for CRBT sales.

For the future work, we plan to explore effect weak ties (asymmetric phone calls), because it might still have effects on diffusion. We want to narrow down the causality between network structure and CRBT adoption, taking time in to account. We will analyze the evolution of the network at different time period. We will also model role equivalence as a disturbance term. We want to consider binary variable, whether a caller will adopt CRBT or not as well. We plan to use two-stage conditional maximum likelihood estimation by Vuong (1984) and Rivers and Vuong (1988). It is a binary probit regression that can handle network autocorrelation effects.

References

Bass, F. M. (1969). A new product growth for model consumer durables. *Management Science*, 15(5):215–227.

Borenstein, M., Hedges, L. V., Higgins, J. P., and Rothstein, H. R. (2009). *Introduction to Meta-Analysis*. John Wiley and Sons.

Burkhardt, M. E. (1994). Social interaction effects following a technological change: A longitudinal investigation. *Academy of Management Journal*, 37(4):869–898.

Burt, R. S. (1987). Social contagion and innovation: Cohesion versus structural equivalence. *American Journal of Sociology*, 92(6):1287.

Clauset, A. (2005). Finding local community structure in networks. *Physical Review E*, 72(2):026132.

Coleman, J. S. (1958). Relational analysis: The study of social organization with survey methods. *Human Organization*, 17(4):28–36.

Coleman, J. S., Katz, E., and Menzel, H. (1966). *Medical innovation: A diffusion study*. Bobbs-Merrill Co.

Doreian, P. (1989). *Two Regimes of Network Effects Autocorrelation*. The Small World. Ablex Publishing.

Girvan, M. and Newman, M. E. J. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826.

Goodman, L. A. (1961). Snowball sampling. *The Annals of Mathematical Statistics*, 32(1):148–170.

Hanneman, R. and Riddle, M. (2005). *Introduction to social network methods*. Online.

Harkola, J. and Greve, A. (1995). Diffusion of technology: Cohesion or structural equivalence? In *Academy of Management Best Papers Proceedings*, pages 422–426, Vancouver, Canada.

Krackhardt, D. and Stern, R. N. (1988). Informal networks and organizational crises: An experimental simulation. *Social Psychology Quarterly*, 51(2):123–140.

Leenders, R. T. (1997). *Longitudinal behavior of network structure and actor atributes: modeling interdependence of contagion and selection.*, chapter Evolution of Social Networks. Gordon and Breach, New York.

Leenders, R. T. (2002). Modeling social influence through network autocorrelation: constructing the weight matrix. *Social Networks*, 24:21–47.

Newman, M. E. J. (2004). Detecting community structure in networks. *The European Physical Journal B – Condensed Matter and Complex Systems*, 38(2):321–330.

Newman, M. E. J. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577–8582.

Rogers, E. M. (1962). *Diffusion of Innovations*. Free Press.

Rogers, E. M. and Kincaid, L. D. (1981). *Communication Networks: Toward a New Paradigm for Research*. Free Press.

Salganik, M. J. and Heckathorn, D. D. (2004). Sampling and Estimation in Hidden Population Using respondent-Driven Sampling. *Socialogical Methodology*, 34:193–239.

Semaan, S., Lauby, J., and Liebman, J. (2002). Street and Network Sampling in Evaluation Studies of HIV Risk-Reduction Interventions. *AIDS Review*, 4:213–223.

Snijders, T. A. B. (1992). Estimation on the basis of snowball samples: How to weight? *Bulletin de Methodologie Sociologique*, 36(1):59–70.

Strang, D. and Tuma, N. B. (1993). Spatial and temporal heterogeneity in diffusion. *The American Journal of Sociology*, 99(3):614–639.

Valente, T. W. (2005). Models and methods for innovation diffusion. In Carrington, P. J., Scott, J., and Wasserman, S., editors, *Models and Methods in Social Network Analysis*. Cambridge University Press.

Van den Bulte, C. and Lilien, G. L. (2001). Medical innovation revisited: Social contagion versus marketing effort. *American Journal of Sociology*, 106(5):1409–1435.

6 APPENDIX

6.1 Pseudocode of Two-stage Clustering and Pruning Algorithm

Step 1: Clustering

Step 1.1

```
Choose a random node $seed$
get the neighbor of node $seed$
$N_1 = \{\}$
$queue = \{\}$
$j = 0$
$n =$ threshold, (e.g. 200)
$m =$ number of $seed$'s neighbor
initialize $nodesOfLevel$ as an empty hash map
$i = 1$
$tempNode =$ first node in the $i$th level
while TRUE do
    if $j > n$
      exit
    fi
    $queue = \{seed\}$
    while $queue$ NOT empty do
        $node =$ pop the top entry from $queue$
        if $node$ in $N_1$
          continue
        fi
        if $node == seed$
          $nodesOfLevel[i] = \{\}$
        fi
        add $node$ to $N_1$
        $j = j + 1$
        if $node != seed$
          if $tempNode == node$
            if $j >= n$
              exit
            fi
            $i = i + 1$
          fi
        fi
        for each neighbor $ii$ of $node$ do
            add $ii$ to $queue$
            if $node == seed$
```

```
                  add $ii$ to $nodesOfLevel[1]$
              else
                  if  node in $nodeOfLevel[i]$
                    if the node is the first node in $nodeOfLevel[i]$
                      $tempNode = ii$
                      $nodeOfLevel[i+1] = \{\}$
                    fi
                    add $node$ to $nodeOfLevel[i+1]$
                  fi
                fi
            od
      od
od
$nodeOfLevel[i+1]$ is the neighbors of all nodes on the same level as the $n-$th node
add $nodeOfLevel[i+1]$ to $N_1$
$ss =$ last step of breath-first sampling
for each node $kk$ added in step $ss$ do
    add neighbor($kk$) to $N_2$
od
```

Step 1.2

```
$matrixI =$ adjacency matrix of nodes in $N_1$
$matrixE =$ adjacency matrix of nodes in $N_2$                    /*source in $N_1$, terminal in $N_2$*/
$I =$ total connections in $matrixI$
$E =$ total connections in $matrixE$
```

Step 1.3

```
$C = [I\ E][I\ E]^T$
comment: /*diagonal element of $C$ is degree, off-diagonal elements are common neighbors */
$N = I \circ C$
find $maxD =$ maximum entry in $N$
$ii =$ row number of $maxD$
$jj =$ column number of $maxD$
add node $ii$ and $jj$ to $N_{PT}$
for $i = 3$ to $n_3$ do
    if neighbor of $ii$ or neighbor $jj$ is not in $N_{PT}$
      $maxii =$ maximum entry in row $ii$
      $maxjj =$ maximum entry in row $jj$
      $maxD =$ maximum$(maxii, maxjj)$
      $ii =$ row number of $maxD$
```

```
        $jj =$ column number of $maxD$
        add node $ii$ and $jj$ to $N_{PT}$
        $i++$
    fi
od
use $N_{PT}$ as new seed, go to Step 1.1
```

Step 2. Pruning

```
set threshold $t = 0.1$
$sumI =$ sum of all elements in $matrixI$
$sumE =$ sum of all elements in $matrixE$
while $sumI/sumE > t$ do
        $arrayI =$ Sum of each row of $matrixI$
        $arrayE =$ Sum of each row of $matrixE$
        for $i = 1$ to length($arrayI$) do
            $arrayDif[i] = arrayI[i] - arrayE[i]$
        od
        $j =$ minimum element of $arrayDif$
        delete row $j$and column $j$ of $matrixI$
        delete row $j$ of $matrixE$
        $sumI =$ sum of all elements in $matrixI$
        $sumE =$ sum of all elements in $matrixE$
od
```