EMPIRICALLY VALIDATING THE THRESHOLD DISTRIBUTION ASSUMPTION IN COMPLEX CONTAGION

By

Qi Hao

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Communication—Doctor of Philosophy

2020

ABSTRACT

EMPIRICALLY VALIDATING THE THRESHOLD DISTRIBUTION ASSUMPTION IN COMPLEX CONTAGION

By

Qi Hao

Diffusion research based on the threshold model made assumptions about the threshold distribution. Yet the assumptions were rarely tested. This study estimated the threshold distribution of people in networks using four empirical datasets. The results showed that the threshold distributions in these datasets were not as usually assumed. The research suggests that many of the conclusions of simulation work done with the threshold model could be based on untrue assumptions.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ALGORITHMS

KEY TO SYMBOLS

[0,1]	A range with boundaries included
t_i	Timestamp for an activation
T_i	Timestamp for a communication between two people
$\boldsymbol{t} = (t_1, t_2 \ldots)$	A list of all the timestamps of individual activations
$\boldsymbol{T} = (T_1, T_2 \ldots)$	A list of all the timestamps of communications
x_{k,t_i}	The activity of the individual k, at timepoint t_i
$c(k, l, T_i)$	A communication from individual k to individual l at a timepoint T_i
CA(x_{k,t_i}, x_{k,t_j})	Two activations of the same individual, x_{k,t_i} and x_{k,t_j} are *communicatively adjacent*
PS(x_{k,t_i}, x_{k,t_j})	A *purchasing section* of individual, k
$\boldsymbol{t_i}$	The i-th *frontier*'s timestamp
$y_{k,\boldsymbol{t_i}}$	Individual k's *cumulated social influence* by frontier $\boldsymbol{t_i}$
k_k	Degree centrality of individual, k
δ (a,b)	Kronecker's delta operator, equals 1 if a = b, 0 if a≠b
∅	The empty set
|{…}|	The number of elements in a set
¬	not
∧	and
A ==> B	From A one can deduce B
A <== B	A is deduced from B

A <==> B	A and B are equivalent , from either one we can deduce the other
τ	Threshold
σ	Value of a random error
$P()$	A process
$p()$	Model representation of a process
S, S'	System's states
s, s'	Model representation of a system's states

CHAPTER 1. Introduction

Granovetter's (1978) paper on the threshold model was influential. It was cited by scholars from a wide range of fields. The original model was developed and expanded by later scholars and applied to many fields including social psychology, economics, communication, computer science, epidemiology, chemistry, robotics, biomedicine, engineering, and applied mathematics. As of January 2020, the original paper had 5835 citations according to Google Scholar. The author of the present research reviewed a sample of the recent (after 1999) relevant (involving the distribution of thresholds) journal papers (not including books) in this citation list[1]. The following is a very brief summary of how later scholars treated Granovetter's (1978) original assumption of the threshold distribution.

The literature review covers papers from multiple fields written by people with very different backgrounds. The diverse body of literature comes with different ways of describing models based on very different world views and research paradigms.

Threshold Model

The threshold model (**TM**) was developed by Granovetter (1978). The original version claimed that:

1) people choose to adopt a new behavior because of the influence of other people
2) a person needs to see a certain number of adopters before deciding to adopt
3) this number is a fraction of the total number of people this person could observe, and this fraction is called a threshold

[1] To be more specific, the literature review focused on the notion of threshold as in "individual threshold for activation." Some scholars who cited this paper used the word "threshold" to mean "tipping point" or "critical mass" in the population before large scale diffusion and adoption happens. Some other scholars who cited the paper used the word "threshold" to mean "a critical value of a model parameter or a critical point in a process". Neither of these uses of the word was about individual agents' behavior activation thresholds and these kinds of research were excluded from the literature review. Models that were in the field of biomedical research and robotics were excluded as well due to the author's lack of domain knowledge to evaluate their relevance.

4) when people see more adopters than their threshold, the person adopts; otherwise not
5) the threshold could differ for different people

Granovetter's model weighed each influencer equally. The model could be easily developed to put different weights on each different friend of the focal person. For example, if a close friend adopts, that could count as two casual friends' influences. Either way, the threshold can be defined as

1) counting the number of activated friends and standardizing by the total number of friends
2) counting the weighted sum of activated friends and standardizing by the total weighted sum of all friends.

The model aggregates adopters' influence by computing a linear combination of influence. This class of model is referred to as the Linear Threshold Model (**LTM**) (Acemoglu, Ozdaglar & Yildiz, 2011; Kempe, Kleinberg & Tardos, 2003) The model was generalized to a General Threshold Model (**GTM**) by removing the linearity constraint and allowing the total influence created by a set of activated friends to be any arbitrary monotone function. For example, when the second influencer adopts, the influence of the first two influencers could be the product of the two influencers' weights. (Kempe, Kleinberg & Tardos, 2003). Another variation of the model was the Competitive Linear Threshold Model (**CLT**) (Chen, Lakshmanan & Castillo, 2013). This model allowed each individual to have two independent thresholds and modeled the diffusion of two behaviors on the same social network. Additionally, Acemoglu Ozdaglar and Yildiz (2011) proposed a stochastic linear threshold model (**SLT**). In this version of the model, the diffusion process is not deterministic. When an individual's threshold is overcome, the individual does not have to activate. Instead, there is a chance that the individual could reject this time's activation by considering whether to activate or not.

Threshold Distribution

The threshold distribution was a critical element of the threshold model. Without knowledge of the actual threshold values of the members in a network, it would be impossible to apply the threshold model. However, such an important distribution was almost always assumed and never empirically measured. Scholars at the beginning of the last decade shared this view. For example, in a critical review (Peres, Muller and Mahajan, 2010) the authors concluded: "The empirical literature on network externalities, surprisingly, lacks evidence on individuals' adoption threshold levels… The shape of the distribution of the thresholds within a population is of utmost importance to the speed of diffusion" and "Given that social threshold modeling is already well-grounded in the sociological literature on collective action, one would imagine that the issue of the distribution of thresholds is by now well established. Unfortunately, this is not the case". Goldenberg, Libai and Muller (2010) said "An important input for this modeling approach relates to the distribution of thresholds in the population… Unfortunately, there is scant empirical evidence regarding threshold distributions, since few attempts have been made to empirically measure thresholds. (p.7)". Libai et al. (2010) said "More empirical evidence is needed as a base for building robust agent-based models … while there is a rich literature in various disciplines using diffusion thresholds to model customer to customer processes (e.g., Granovetter 1978), it is mostly theoretical and lacks empirical support in the individual level. (p. 277)".

Valente (1996) empirically estimated individual thresholds based on the Granovetter (1978) model. The author categorized the threshold values into high, medium and low categories and described how many of the early, middle and late adopters had high, medium and low

thresholds. However, the findings in the paper was more of a description of the diffusion result and did not directly inform modeling.

Social psychological literature studying the predictors (or usually called determinants) of individual thresholds could shed light on the formation of individual thresholds. Although these studies may not be able to draw causal conclusions, the distribution of the predictors of thresholds could partially and indirectly inform and validate the distribution of the thresholds. Braun (1995), from an economic perspective, claimed that the thresholds of people are driven by benefit-cost distributions as well as the perceived network structure. Thresholds should be higher in larger, less-connected social systems where the benefit from a collective good is lower. A theoretical paper (Matsueda, 2006) claimed that the most important determinants for behavioral thresholds would be a conception of self or identity, followed by the availability of alternatives, and sensitivity to opinions of others. McGloin and Rowan (2015) found that gender, ethnicity, normative belief, and impulsivity significantly predicted the self-reported thresholds for taking part in group crimes. Thomas and Marie (2013) used the dual-process theory to explain individual activation thresholds. The situational factors that give rise to a behavior require immediate decisions based on simple information processing, whereas the lasting norms require deliberate decisions considering long-term consequences. So, both situational factors and perceived social norms are determinants of the thresholds to specific actions. Studies of choice shifts suggested that a person's threshold to participate in risky behaviors is not preset and constant but may depend on whether the person is alone or at the presence of peers (e.g. Dodoiu, Leenders, & van Dijk, 2016; Gardner & Steinberg, 2005). The same effect of presence of others not only exists with risky behaviors, but also with helping behaviors (e.g. Latané & Nida,1981). These findings are consistent with the methodological holism mentioned by Matsueda (2006),

that society is prior to individuals and the individuals are to be defined and interpreted as parts of society. The social nature of thresholds requires that the concept be defined and discussed in social contexts, with special attention paid to people's behaviors in the presence of and in relation to others.

Literature on social influence could shed light on the threshold distribution in a similar manner. The factors that affect strength of social influence could be integrated to model effects of social influences and, in turn, estimate thresholds and threshold distributions. Evidence showed that physical distance, immediacy, number of sources, group size and other factors could affect social influence strength (e.g. Latane, 1996; Nowak, Szamrej & Latané, 1990; Latané & Nida, 1981)[2].

Until the last decade, scholars (e.g. Peres, Muller and Mahajan, 2010, p. 101) could barely identify any attempts to verify the empirical threshold distribution: "We know of only a few (partial) empirical verifications of Granovetter's (1978) original claim that the distribution is truncated normal" (Ludemann, 1999; Goldenberg et al., 2010). The Ludemann (1999) paper used self-report to measure people's thresholds with a single item survey question administered to 247 participants by asking: "How many members of your community, in percentage, would have to put their waste glass in a public recycling bin before you would do so?" The finding indicated that the threshold distribution was truncated normal. The Goldenberg et al. (2010) paper conducted four survey studies (asking for self-reported thresholds) with a total of 180 students with some network effect designed into the study (allowing videoconferencing of participants) and "found in all four studies that the externalities distributions (threshold distribution) were

[2] Note that the objectives of social influence in these studies were mixed. Some were about influence on behaviors, some about influence on attitudes and some both. The current research only focuses on behaviors' influence on behaviors.

truncated bell-shaped. In one study, this distribution was symmetrical, and thus, a truncated normal distribution is a reliable working assumption. In the other three cases, the distribution was somehow skewed. In two studies, negative skewness was evident, and in one study, moderate positive skewness was observed. Overall, these results support the threshold distribution used in this study. However, we believe that given the importance of threshold distribution presented herein, future empirical research is needed to gain insights into how to assess the distribution of thresholds and the shapes of the distribution under various market scenarios (p. 7)."

These attempts to measure empirically the threshold distribution were worth following up. However, two things need to be noted. First, these studies were relying on self-report, but thresholds may not be accurately accessible by people. Estimates based on external behavior records may be a better approach. Second, Goldenberg et al. (2010) mentioned that it was Granovetter's (1978) original claim that the distribution was truncated normal. This statement may mislead people to think that Granovetter had an original statement about what shape the threshold distribution should take. That is not true. Actually, Granovetter (1978) only used truncated normal distribution as an example. He also used the random uniform distribution as another example in the same paper. Granovetter claimed that the threshold distribution should be issue-specific so he was not trying to claim that the distribution of thresholds should follow a certain distribution for some theoretical or empirical reasons.

Scholars' Treatment of the Threshold Distribution

As pointed out by Goldenberg et al. (2010), "While threshold modeling has served as a major tool in the collective action literature, nearly all studies have been based on either analytical assessment or simulations, with rare examples attempting to infer thresholds from

indirect behavioral data. (p.7)" Scholars sometimes intentionally stayed vague about the threshold distribution. Gruhl, Guha, Liben-Nowell and Tomkins' (2004) paper described the threshold model as: "Each node u in the network chooses a threshold tau ∈ [0, 1], typically drawn from some probability distribution… (p. 492)".

For analytical research (e.g. Dodds & Payne, 2009), if done with enough mathematical generality, the variation of threshold distribution should not be a problem because the conclusions do not necessarily depend on the assumption of a specific threshold distribution. For example, Watts (2002) obtained a general solution for the global cascade condition that could apply to any threshold distribution. Watts (2002) demonstrated the result with a random uniform threshold distribution whereas others (e.g. Gleeson, 2008; Payne, Dodds & Eppstein, 2009) replicated and generalized Watts' (2002) conclusions with a constant threshold distribution.

Mossel and Roch (2010) showed that as long as the individual level activation function is monotonic increasing and submodular, the influence function would be submodular and monotonic increasing as well regardless of threshold distribution. So, the preservation of submodularity in diffusion processes is not influenced by the threshold distribution[3].

For another example, Acemoglu, Ozdaglar and Yildiz (2011) analytically showed that a small degree of clustering (a small number of closely-knit sub-communities in the network) might help diffusion because the upper bound of the expected number of nodes ultimately reachable by the diffusion is inversely proportional to the number of clusters. The proof did not

[3] Submodularity means the marginal gain in the output of a function decreases for the same amount of extra input when the total input increases. For example, the total number of indirect friends a new friend can bring is submodular, namely, the more friends one already has the fewer extra indirect friends a new friend can bring as compared to knowing this new friend without already having a lot of friends. Monotone increasing means as long as the number of inputs increases, the output will increase. In the case of the individual activation function, monotonic increasing means as long as the number of active friends of a focal person increases, the total social influence the person receives increases. The individual activation function is how the set of active friends (input) leads to the individual's activation (output). The influence function here is how a set of a few initial active individuals (input) leads to the final set of active individuals in the whole network (output).

rely on a specific threshold distribution. Although later in the simulation, the authors chose a random uniform distribution of thresholds for demonstration, the conclusion from the analytical part of this paper would hold if the random uniform distribution assumption is violated.

As for the simulations, threshold distribution is a key factor that affects the simulation results. According to Valente (1995), before 1995 much of the threshold modeling literature has implicitly or explicitly assumed that thresholds are normally distributed in the population. Later, since 2000, much of the simulation research reviewed by the author of the present paper used a random uniform threshold distribution (e.g. Acemoglu, Ozdaglar & Yildiz, 2011; Kempe, Kleinberg & Tardos, 2015; Leskovec, Adamic & Huberman, 2007; Mossel & Roch, 2010; Watts, 2002). Some simulations used constant thresholds for all nodes in the network (e.g. Centola, Eguiluz & Macy, 2007; Centola & Macy, 2007; Nematzadeh, Ferrara, Flammini & Ahn, 2014; Valente & Davis, 1999).

The scholars made their choices of the threshold distributions for various reasons. Sometimes, scholars used the random uniform distribution of thresholds in their simulations just so that the simulation would be manageable. For example,

> "Unfortunately, we show in Section 3.2 that hard-wired thresholds make the optimization problem hard to approximate to within a multiplicative factor of n1−ε for any ε > 0; the same hardness results naturally apply to the case in which the thresholds are part of the input. For this reason, …we instead assume that the thresholds θv are chosen independently and uniformly at random from the interval [0, 1] …" (Kempe, Kleinberg & Tardos, 2015, p. 111)

Scholars who used the random uniform threshold distribution argued that because we have no knowledge of the threshold distribution, we should use the random uniform. For example,

> "… to model our lack of knowledge of the thresholds, we instead assume that the thresholds θv are chosen independently and uniformly at random from the interval [0, 1]" (Kempe, Kleinberg & Tardos, 2015, p. 111)

> "The threshold values are random. This is to account for our lack of knowledge of the exact threshold values. KKT assume that the thresholds are uniformly random." (Mossel & Roch, 2010, p. 2177)

The trend of using a random distribution due to lack of knowledge could be changed as scholars accumulate more knowledge of the threshold distribution and the factors influencing it. Most of the simulations that used the random uniform threshold distribution also had analytical proofs that did not depend on the threshold distribution. Admittedly, the purpose of the simulations in these studies was only to demonstrate the results and provide validation for the analytical results.

Centola, Eguiluz and Macy (2007, p. 450) used a constant threshold distribution in their simulations, and argued that the usage of a constant threshold for all nodes could isolate out the effect of threshold distribution so they could focus on the effect of network topology:

> "In order to isolate the effects of network topology from the effects of the threshold distribution, we assign every node an identical threshold T…"

This argument is wrong on many levels. First, to eliminate the effect of the threshold distribution, one needs to keep constant the effect of the threshold distribution rather than keep constant the threshold value for every person. The conclusions of diffusion are on the network/population level, not on the individual level. Making all individuals have the same threshold does not equal controlling the effects of the threshold distribution on diffusion. Second, using a constant threshold for every person still retains a threshold distribution, the constant distribution. The authors claimed that doing so controlled the effect of the threshold distribution. On the contrary, the authors' findings were exactly due to the threshold distribution selected. If any other threshold distribution were used, none of the results reported in this paper would be possible as they were presented. Third, saying "to isolate/control the effect of something," implies that the effect of that thing isolated out could be added in later. But Centola et al.'s (2007) study *relied* on the constant distribution of thresholds and the effects of non-constant threshold distributions *cannot* be added to his conclusions. To consider the effects of any other threshold distribution, one would need to redesign the simulations to a point that it would be a totally different experiment.

Ways of Realizing Random Uniform Distribution

Among the scholars who used the random uniform threshold distributions, there are mainly two ways of employing it. One way is to draw the threshold for each node randomly from the same uniform distribution on [0,1], and the other way is to draw each threshold value from an independent random arbitrary distribution. The first method automatically means the threshold distribution used in the simulation is a random uniform distribution. Watts' (2002) paper generated thresholds in the second way, with each threshold value drawn from an independent arbitrary distribution standardized on [0,1]. This meant that there were as many distributions as

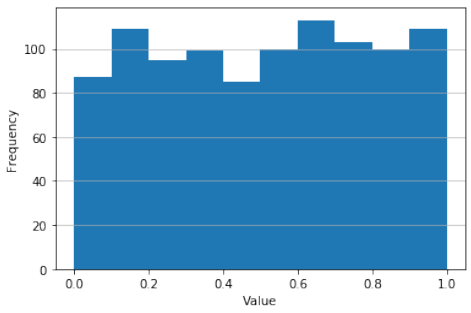
thresholds. Each distribution could be any distribution (e.g. normal, random, Poisson etc.). This notion of distribution is the distribution of possibilities rather than the distribution of probabilities.

This paper refers to the distribution of probabilities in a population of individuals rather than the possibility sense used in Watts' (2002) paper. The second way of describing the threshold generation process used in research like Watts (2002), which describes the distribution of each node, belies their assumption of the threshold distribution (in the first sense). So, what is their assumption of threshold distribution in the first sense? This question can be translated into the following question. Suppose:

1) there are many values, represented by X and indexed by i, so each value is X_i
2) each X_i can be drawn from a distribution D_i (in the second sense),
3) all the values of X_i can form a distribution called distribution D (in the first sense),
4) the $D_i's$ are arbitrary distributions and independent for different i's, then:
5) what distribution is distribution D?

And the answer is: D would be random uniform. Following is a computational experiment to demonstrate the conclusion. First, 1000 independent random arbitrary distributions are specified. Then from each distribution a value is drawn. Finally, the histogram of the 1000 values is plotted subsequently, which approximates closely a random uniform distribution.

Figure 1. Shape of distribution D, realized by drawing from independent distributions



Note. The intuition behind the result in Figure 1 is each specific distribution, except for the random uniform distribution favors certain areas of the value axis in terms of the high frequency of values in that area. For example, the normal distribution favors values in the area close to the mean. So, if a value is drawn from a normal distribution it is more likely to be a value close to the mean. The many random arbitrary distributions are expected to favor equally each area of the value axis, and as a result, the distribution (in the first sense) of all the values is expected to approximate the random uniform distribution.

Summary

A brief review of the literature showed that the threshold distribution was an important factor affecting the results of research based on the threshold model. But, such an important distribution has not been adequately measured and validated. Simulation studies made assumptions about this distribution (e.g. random uniform, truncated normal, constant) for various reasons, and based their conclusions on unverified assumptions. To ground the threshold model, it is meaningful to generate an empirical estimate of the threshold distribution in datasets collected from actual networks.In the next few chapters, the author will identify some datasets, estimate each individual's cumulated social influences (chapter 2), estimate individual thresholds based on their cumulated social influences before activations (chapter 3), plot out the distribution of individual thresholds for different datasets (chapter 3) and discuss the empirical distributions' meaning in the threshold model literature (chapter4).

CHAPTER 2. Computing Cumulated Social Influence

A Description of the Empirical Data Structure Used in the Study

To apply the threshold model requires a very specific kind of data structure. First, the dataset needs to provide information on the activation of people. Without activation, a threshold makes no sense because a threshold is by definition (Granovetter, 1978) the threshold of individual activation. Second, the dataset needs to contain information about the interaction of people on social networks. Without network and interaction information, it is hard to infer the social influence happening between people (Ugander, Backstrom, Marlow & Kleinberg, 2012). Third, the activation behaviors and interaction behaviors need to be timestamped so that it is possible to find out which action happened earlier, and which happened later. This is necessary because if interaction happened after activation, it makes no sense to reason that the interaction influenced the activation (cf. Figure 2).

Four datasets will be used to estimate the empirical threshold distribution. The first three datasets are about the in-game purchasing behaviors of the players in a multiplayer online game. The fourth dataset is the music liking behaviors of the last.fm music app. The timestamped behavior in the game datasets is in-game purchasing and the interaction is in-game chatting. The timestamped behavior of the music datasets is clicking the “like” button for songs and the interaction is the app’s function of showing users their friends’ loved songs (so it is assumed that the “like” behavior of each user is observed by their friends).

The algorithm in this chapter will be described using the online game data as an example. The algorithm will easily generalize to the music data. Before describing the algorithm, some assumptions and definitions are made.

Assumptions and Definitions in the Study

Assumption 1: **[social influence assumption]:** Activation of behavior is due to social influence and social influence alone. (The words "activation" and "adoption" are used interchangeably.)

Definition 1: **Social influence** is adopters' communication with future adopters.

The present study equates the effects of "communication with previous adopters" with the effects of "social influence of previous adopters." The data used in this study are all from online settings, where other forms of influence such as direct observation are not possible. In this study, when an adopter communicates with another person, it is sometimes worded as "the adopter 'sends' social influence to another person" and "the other person 'receives' social influence."

Assumption 2: **[cumulative influence assumption]**: Social influences can be cumulated. Receivers keep records and keep updating the social influence they receive over time. The threshold model implicitly assumes memory (Dodds &Watts, 2005).

Definition 2: **Cumulated social influence** is the number of previous adopters with whom a focal person communicates.

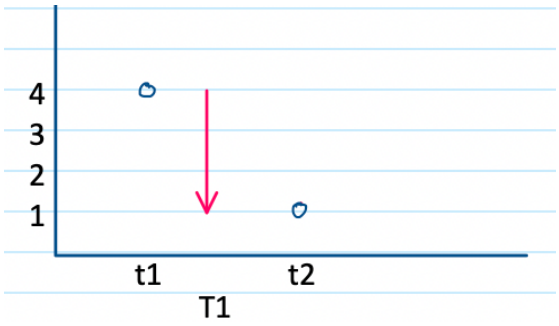
Corollary 1**:** [from assumption 1 and definition 1] Activations are preceded by communications with previous adopters. The sequence of events of "influence and then activation" is called the influence-activation process.

Assumption 3**: [repeated influence assumption]**: When the influence-activation process is repeated, the social influences one accumulates are reset after activation. The sequence of "influence, activation, reset influence, new influence, new activation…" is called the repeated influence-activation process.

Corollary 2: [from assumption 1,3 and definition 1] In the repeated influence-activation process, activations are preceded with communications with previous adopters.

The following is a visualization of the fundamental unit of the influence-activation process of an individual. The whole paper is based on identifying and analyzing this fundamental pattern in datasets.

Figure 2. Representation of the fundamental unit of the influence-activation process



The x-axis in Figure 2 is time. Lower case letters (t1 and t2) represent the timestamps of individual activation. Capital letters (T1) represent the timestamps of communication. The y-axis is individual indexes. This figure describes the activation of individual 4 at t1, activation of individual 1 at t2 and the communication of 4 to 1 at T1. The arrow (communication) in this figure constitutes social influence because it comes from an (earlier) adopter and it goes to someone who has not adopted by the time of communication. If, say, activation of 1 happened before 4's communication (i.e. t2<T1), then the communication does not count as social influence according to definition 1.

An Algorithm to Compute Cumulated Social Influence

To model the diffusion of any product, behavior, or innovation, there are two series of timestamps that need to be analyzed, taking the in-game product diffusion as an example,

1) the series of timestamps recording people's **purchasing behaviors** and
2) the series of timestamps recording people's **interactions (i.e. communication)**.

It is assumed that if two people have not communicated with each other within the game through text messages during a time period, they do not send or receive social influence from each other

in that time (players may have texted or called each other via cellphone off line, but this paper assumes that is not happening and the game is the only platform on which players communicate). A behavior is considered to be an activation due to social influence when and only when:

1) this behavior happened after communication with others and

2) the "others" that the focal individual communicated with had the behavior prior to their communication.

An intuitive approach is to find all the activations (i.e. purchases) and count the cumulated number of influences before activation for each person. But it is possible that some activations are not preceded with influences, thus violating corollary 1 or corollary 2. The way out of this mismatch between the data and the definitions is to treat multiple activations without social influence between them as one activation (called a *purchasing section*). And then count the cumulated social influence before each purchasing section. The following goes through the algorithm step by step:

1) Individuals in the network will be sorted arbitrarily and assigned an index number from 1 through *J*; where *J* is the total number of individuals.
2) For any product, list all the timestamps of individual purchases and sort them from early to late in a vector $\boldsymbol{t} = (t_1, t_2 \ldots)$;
3) List all the timestamps for communications between individuals in a vector $\boldsymbol{T} = (T_1, T_2 \ldots)$;
4) For any timepoint t_i, any individual is either active (made the purchase at this time) or inactive (not making a purchase at this timepoint). Let x_{k,t_i} represent the activity of the individual k, at timepoint t_i, and x_{k,t_i} takes values of either 1 or 0 for any k.
5) Let any communication from individual k to individual l at a timepoint T_i be defined as a tuple $c = (k, l, T_i)$, or denoted $c(k, l, T_i)$.

6) Define two activations of the same individual, x_{k,t_i} and x_{k,t_j} to be *communicatively adjacent*, CA(x_{k,t_i}, x_{k,t_j}), if there is no communication to the individual, k, from any other individual who activated between t_i and t_j. Namely,

CA(x_{k,t_i}, x_{k,t_j}) <== $\{c = (l,k,T_j) | t_i < t_m < T_j < t_j,\ x_{l,t_m} = 1\} = \emptyset$

Define $\neg CA(t_i, t_{i+1})$ to be that the two timepoints are not communicatively adjacent.

7) Define a *purchasing section* (PS) of an individual to be a largest time period in which the first and last activations of the individual are *communicatively adjacent*. Namely,

PS(x_{k,t_i}, x_{k,t_j}) <== $\text{CA}\left(x_{k,t_i}, x_{k,t_j}\right) \wedge \neg CA\left(x_{k,t_m<t_i}, x_{k,t_j}\right) \wedge \neg CA(x_{k,t_i}, x_{k,t_n>t_j})$

8) Define the left edge of a purchasing section t_i, the starting time point of a purchasing section, to be called the *frontier* of a purchasing section. Because of 6) and 7), it follows that

9) The cumulated social influence for an individual at any time point within any one of its purchasing sections is the same as the amount of cumulated social influence of that individual at the frontier of that purchasing section.

Proof: If

the individual is at t_i, the frontier,

then

the conclusion in 9) is obvious.

If

the individual is at time point t_k such that $t_i < t_k \leq t_j$,

then

(By definition of CA in 6)

none of the purchasers since time point t_i have sent any new communication to the individual,

none of those who sent communication to the individual purchased during this time

so (by definition 1), the individual does not receive social influence during a Purchasing Section

so, 9) follows.

10) The two series of timestamps, the communication timestamps, and the purchasing time stamps, can be then put together onto one timeline, although indexed separately (communication timestamps by T_i and purchasing timestamps by t_i). Because of 9), the diffusion of purchasing due to social influence can be thought of as purchasing sections separated by communications between them. Let the bold $\boldsymbol{t_i}$ represent the *i*-th frontier's timestamp, to distinguish it from normal purchasing timestamps. For example, $\boldsymbol{t_3}$ represents the timestamp for the 3rd frontier, whereas t_3 represents the 3rd timestamp of individual purchases**.**

(See Appendix 4 for an example visualizing step 1 through step 10)

11) Individual *k*'s ***cumulated social influence*** by each frontier ($y_{k,\boldsymbol{t_i}}$) is the number of adopters the individual has communicated with ($c_{n,k,\boldsymbol{t_i}}$) since the last frontier, $\boldsymbol{t_{i-1}}$, standardized by the focal individual's degree centrality (total number of connected individuals). Mathematically,

$y_{k,\boldsymbol{t_i}} = (\sum_{n=1}^{J-1} c_{n,k,\boldsymbol{t_i}})/k_k$, where, k_k is the degree centrality of individual k.

Each n who communicated with k in the time period is weighted equally, so for any n,

$$c_{n,,k,\boldsymbol{t_i}} = \begin{cases} 1, if\{c(n,k,T_i)\big|\boldsymbol{t_{i-1}} \leq t_i < T_i \leq \boldsymbol{t_i}, x_{n,t_i} = 1\} \neq \emptyset \\ 0, otherwise \end{cases}$$

Write the exposure index $c_{n,,k,\boldsymbol{t_i}}$more compactly using the Kronecker's delta as

$$c_{n,,k,\boldsymbol{t_i}} = 1 - \delta(\left|\{c(n,k,T_i)\middle|\boldsymbol{t_{i-1}} \leq t_i < T_i \leq \boldsymbol{t_i}, x_{n,t_i} = 1\}\right|, 0)$$

So, each individual k's cumulated social influence at each generation $\boldsymbol{t_i}$ can be estimated as:

$$y_{k,\boldsymbol{t_i}} = \frac{\sum_{n=1}^{J-1} c_{n,k,\boldsymbol{t_i}}}{k_k} = \frac{\sum_{n=1}^{J-1}(1-\delta(\left|\{c(n,k,T_i)\middle|\boldsymbol{t_{i-1}} \leq t_i < T_i \leq \boldsymbol{t_i}, x_{n,t_i} = 1\}\right|,0))}{k_k}$$

(See appendix 5 for an example of step 11)

Summary

By running from step 1 through 11 on every single person in the dataset, the algorithm computes the cumulated social influence before each activation (i.e. purchasing section) of each person. For example, suppose individual 1 had three activations and individual 2 had two activations, then the output would look like the following:

Table 1. Example Output Format for Results After Step 11

Individual	Cumulated social influence before 1st activation	Cumulated social influence before 2nd activation	Cumulated social influence before 3rd activation
1	1/3	2/3	1/3
2	3/10	1/2	

(See Appendix 6 for evidence validating the definition of purchasing sections and Appendix 7 for the pseudo-code of the algorithm)

CHAPTER 3. Estimating Threshold Values and Threshold Distributions

Datasets

Four datasets were used to test previous scholars' assumptions about the threshold distribution. The first three datasets were from an online game, *A Chinese Ghost Story*, hosted by NetEase. The data included the players' timestamped communication behaviors with their co-players and the users' timestamped purchasing behaviors within the game. For more details of the game and the data collection process, see Xu et al.'s (2017). The original owners of the dataset shared three subsets of the data. The duration of each dataset was one month. The descriptive data for the datasets are presented in Table 2.

Table 2. NetEase Data Description

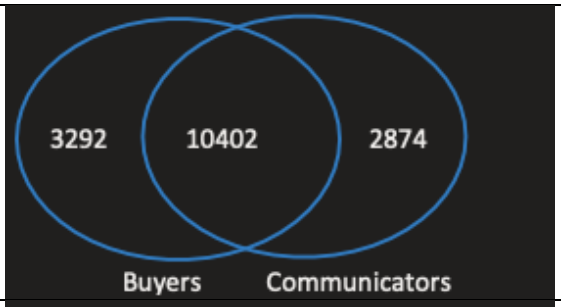
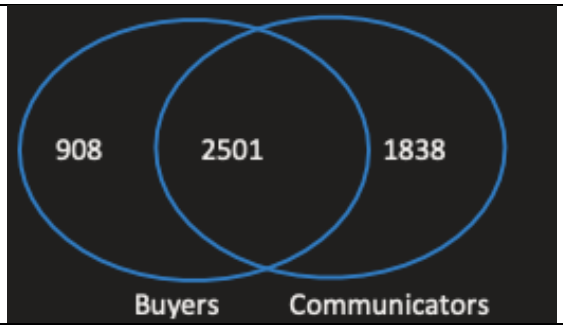
Year and Month	**Days of Month**	**Number of Buyers and Communicators**	**Number of Purchases, Communications, Products.**
2013.04	**4.20 - 4.30**		**# purchases:** 403,467 **# communications:** 2,035,672 **# products:** 743
2013.09	**9.01 - 9.30**		**# purchases** 272,348 **# communications** 2,745,734 **# products** 1210

Table 2. (cont'd)

2014.04	**4.01 - 4.30**		**# purchases** 121,520 **# communications** 1,201,608 **# products** 1078

Consider the 3rd dataset as an example. The data were collected on users from April. 1 through April 30, 2014. The data contained 121,520 purchases of 5,247 players of 1,078 products. Out of the 5,247 players, 908 made purchases but did not communicate with any other player during this time, 2,501 communicated and made purchases, and 1,838 communicated with others but did not make any purchases. In total, the players communicated 1,201,608 times.

The fourth dataset, used by Sharma and Cosley (2016), was about users of the last.fm app from April through June 2014. The dataset contained the timestamped listening, liking, and banning behaviors of each user. Additionally, the dataset had information about the friend set of each user. Those users without friends were excluded from the current analysis due to the paper's focus on diffusion. The app had the function of showing users their friends' liked songs. The descriptive data for this dataset are as follow:

- Number of users who had friends (connected users): 141,346
- number of connected users who had liking behaviors: 141,346
- number of connected users who had listening data: 0
- number of connected users who had banning data: 0
- number of unique songs liked by connected users: 5,791,799
- number of liking behaviors of users: 43,753,706

As shown from the preceding descriptive data, the dataset may differ from what was described in the original paper. According to the current author's examination of the dataset, none of the users who had friends had listening or banning data, and none of the users who had listening or banning data had friends. This data structure is puzzling because the original paper's finding, that users' behavioral similarity to friends is approximately the same as their similarity to non-friends, so that social influence does not occur, would require a subset of users both to have user behavior data and friendship data. Yet such a subset does not exist for listening and banning behaviors. Another deviation from the original paper's description was that the friending relationship was not mutual. For example, in the dataset, user "30093" was a friend of user "1" but not vice versa.

Despite these deviations from the original author's description, the liking behaviors of the users linked by friendship still make suitable data for the estimation of the threshold distribution. In the dataset, a song can be liked by more than one user, and on average a song is liked by 7.55 users. The percentage of a focal user's friends who liked a song before the focal user liked the song could be considered as the cumulated social influence for that person before the liking behavior. Then each user would have a list of cumulated social influence scores for all the songs he or she liked. The cumulated social influence scores before activation can then be used to estimate individual thresholds.

Estimation of Individual Thresholds

Each individual may have multiple adoptions. Namely, in the NetEase datasets each player could have multiple purchases and in the last.fm dataset each user could like multiple songs. Before each adoption, each user should have cumulated social influences by seeing their friends adopting before them. After running the algorithm described in chapter 2, each individual

will have as many values of cumulated social influence as their number of adoptions. Then comes a question: which of these values better approximates the individual's threshold? This paper will use three different ways to estimate the threshold value of an individual based on the individual's values of cumulated social influence.

Minimum Value Approximation. Assume that every time a person activates this person's threshold must have been exceeded (Valente, 2012). Under this assumption, even the minimum cumulated social influence before the activation of an individual must exceed the individual's threshold. So, the minimum value of the person's cumulated social influence scores before activation is used as an approximation for the individual's threshold.

Average Value Approximation. Assume every time a person activates, the cumulated social influence does not necessarily have to overcome the person's true threshold, τ, but only has to overcome the person's true threshold plus a random error, σ, which has a mean of zero and a small standard deviation. This assumption implies that a person's acting threshold (i.e. true threshold plus random error, $\tau + \sigma$) is not constant. This fact, in turn, means sometimes it would take a bit more social influence to activate the same person and sometimes, by chance, social influence a bit less than the true threshold is also able to activate the person. Let us further assume this random error, σ, is normally distributed and for each activation its value is independently drawn from its distribution. Under such assumptions, the job of estimating an individual's true threshold, τ, becomes the job of estimating the mean of the distribution of $\tau + \sigma$ given the observed cumulated social influence scores. This task can be treated as a maximum likelihood estimation problem: given multiple scores drawn from the distribution of $\tau + \sigma$, what is the most likely mean for $\tau + \sigma$? And the answer is: the most likely mean for $\tau + \sigma$ is the average of the observed scores. Derivations of this conclusion can be found from the following

webpage: https://towardsdatascience.com/maximum-likelihood-estimation-explained-normal-distribution-6207b322e47f. So, the average value of the person's cumulated social influence before activation is used as an approximation for the individual's threshold.

Maximum value approximation. This method uses each individual's maximum cumulated social influence before its adoption as an approximation for this individual's threshold. This value was selected as a matter of conceptual symmetry, given that I am examining minimum values.

Estimations of the Threshold Distribution

The estimated thresholds of individuals in the datasets (based on minimum, average, maximum scores of cumulated social influences before activation) were plotted in histograms to visualize the distribution of thresholds. The plots used all the data in each of the first three datasets. A sampling process was applied for the fourth dataset. Because some members liked more than 131,000 songs and some members had more than 131,000 friends (131,000 was approximately the default reading length limit of a list in python in Jupiter notebook), to find the number of friends who liked a song before a user liked it for each song for each user would require a prohibitively long runtime. Thus, a subset of all users was sampled, and for each user, if that user had more than 100 friends, a random sample of 100 friends was drawn as the user's friend set. If a user liked more than 20 songs, a random sample of 20 songs was drawn to represent the user's liked songs. Then the threshold distribution was estimated and plotted based on the sample. The sample size of users was increased incrementally until the distribution stabilized and a further increase in sample size did not change the pattern of the distribution. Eventually the researcher ended up using 200 sample users.

Results

Based on the cumulated social influence values for each individual in each of the four datasets, three histograms were plotted for each of the four datasets, generating 12 figures in Appendix A.

Figure 3 is the histogram of individuals' thresholds using maximum value approximation based on the first dataset. There are three spikes (values with frequencies higher than the values immediately left of it and right of it) at threshold values of 0, 0.5 and 1. If we ignore the spikes at thresholds of 0 and 1, the rest of the plot looks like a normal distribution. There are two dips in the distribution (area where values' frequencies drop low): threshold values between 0 and 0.2, or between 0.8 and 1. It is very improbable for thresholds to take values in these tworegions.

Figure 4 is the histogram of individuals' thresholds using average value approximation based on the first dataset. There are three spikes of threshold values: 0, 0.5 and 1. Except for these three spikes, the area at a threshold below 0.5 approximates a normal distribution with a mean of 0.2 (standard deviation of about 0.1). There are threedips: The threshold rarely takes values between 0 and 0.1, between 0.4 and 0.5, or between 0.5 and 1.

The pattern in these two plots can be summarized as: "low, medium and high spikes plus bell-shaped distributions with different means between the spikes." Figure 5 is the histogram of individuals' thresholds using minimum value approximation based on the first dataset. There are still three spikes at 0, 0.5 and 1, but the first spike at 0 is much higher than the other two, indicating that most people activated (i.e. made a purchase) at least once without any of their friends' influence. The rest of the distribution had lower probabilities, but with higher probabilities between 0 and 0.5 than between 0.5 and 1. The distribution in this picture is closer to a power-law distribution (or even a constant distribution) than a random uniform distribution.

The distribution with minimum values approximation can be described as: “spikes plus long-tail distributions”.

The three plots for the second dataset and the three plots for the third dataset are similar in pattern to the first dataset. Similar conclusions can be drawn after observing the histogram distributions. The third plot for the second and third dataset (Figure 8 and Figure 11) showed patterns closer to a power-law distribution than the third plot (Figure 5) of the first dataset.

Figure 12 is the histogram of individuals’ thresholds using maximum value approximation based on the fourth dataset. The histogram generally showed a long-tailed distribution, which is better described by a power-law distribution than a random uniform distribution. Most people’s threshold values lied between 0 and 0.5, with very few values between 0.5 and 1. There are still three spikes at 0, 0.5, and 1, except the spike at threshold equaling 0 is much higher than the other two.

Figure 13 is the histogram of individuals’ thresholds using average value approximation based on the fourth dataset. The histogram generally showed a long-tailed distribution, which is better described by a power-law distribution than a random uniform distribution. Most people’s threshold values lied between 0 and 0.01, with very few values above 0.01.

Figure 14 is the histogram of individuals’ thresholds using minimum value approximation based on the fourth dataset. The distribution was a constant distribution, with a single value at threshold equaling 0. This indicated that most people activated (i.e. liked a song) at least once without any of their friends liking the song before them (without influence).

It is safe to eyeball the histograms, without fitting a distribution line to each of the 12 histograms, and say that the thresholds in these datasets did not follow a random uniform distribution (or truncated normal) no matter how the threshold was approximated (by each

person's maximum cumulated social influence, average cumulated social influence or minimum cumulated social influence before activations). In addition, Kolmogorov-Smirnov (KS) tests were performed for each histogram to see if their distributions were significantly different from data drawn from a uniform distribution. The results (see Appendix A) for all datasets showed a p-value of less than 0.001. Thus, the KS-tests indicated the observed distributions in the histograms did not follow uniform distributions. The histograms showed high spikes and deep valleys. These patterns indicate that there are certain values of the threshold that are popular (highly probable) in the population and there are values that are rarely taken by individuals. If we have to estimate someone's threshold, according to the distribution pattern, it would be better to guess it is of value 1, 0, 0.5 or some value between 0 and 0.5. If we ignore the spikes, the rest of the histograms look similar to either the normal distribution or the power-law distribution. The observed threshold distribution could be a result of aggregating several random variables that follow certain classic distributions. Future research should explore what these random variables might be and how they could aggregate to produce the observed distribution of thresholds. The current study can claim little on this matter, but one thing is visually suggested by the plots: the threshold distribution is not likely to be random uniform in many empirical datasets and there are clear non-random patterns that could be potentially explained. Claiming that we have no knowledge of the threshold distribution so we should assume it is random seems to be treating the issue too perfunctorily.

Additional Analyses

Some additional analyses have been done to further explore the distribution of thresholds. The analyses so far used percentages (of active contacts) to characterize the threshold and treated repeated purchases of the same product as new activations. Instead of percentages, raw numbers

of active contacts could be used to characterize how much influence one received before activation. Instead of allowing repeated purchases to be new activations, one can count only the purchase of new products as new activations. The additional analyses will try to explore the distribution of threshold in two new ways:

Exploration 1: Using raw numbers (of active contacts before activation) to characterize and estimate the threshold while allowing repeated activations (counting repeated purchases of the same product as new activations); and

Exploration 2: Counting only purchases of new products as new activations while using both percentages and raw numbers to characterize thresholds.

The result of Exploration 1 can be found in APPENDIX C. For each of the four datasets, three histograms were plotted using same methods except that the raw number of active friends were plotted instead of the percentage of friends. The thresholds form the x-axis of the histograms, measuring the raw number of friends that influenced activation (based on maximum number, average number and minimum number over multiple times of activation). The y-axis is the number of people who had each threshold.

The threshold distribution in terms of raw number of influencers before activation seems to follow a long-tail distribution. Most people were activated after interaction with about 5 to 10 influencers. Because most people do not have more than 10 friends (see Figure 15 in APPENDIX C for degree distribution), the activations usually happened after less than 10 influencers. The limited number of friends (less than 10) and the small number of influencers (usually less than 5) for most people showed that the threshold distribution is highly influenced by and dependent on the degree distribution in that: 1) The degree distribution puts a cap on the maximum number of influencers most people could have; and 2) When most people have similar

small numbers of influencers, the variation in the degree (distribution) will highly affect the variation in the threshold (distribution) due to the way threshold is conceptualized (i.e. number of influencers divided by total number of friends, degree).

Exploration 2 is supposed to explore what happens if only first-time purchases are counted as a new purchase and not repeated purchases of the same product. After deleting the repeated purchases of the same products and re-analyzing the purchasing sections, it is found that the results showed exactly the same purchasing sections for all datasets. Because the purchasing sections are the same, all following analyses of the activations without repeated behaviors (purchases) are all the same as the ones with repeated behaviors. This finding results because in every new purchasing section, the players bought at least one new product. As a result, although repeated purchases were deleted, no purchasing sections were. This finding, in addition to APPENDIX F, once again in a different way, validated the paper's definition and usage of *Purchasing Sections*. Each purchasing section is a new period of activation, not only because it is a new temporally separate period, but also because it involves purchases of new products never bought in previous purchasing sections.

After some additional explorations, the threshold distribution was still not shown to be uniform. Evidences showed that the normal distribution may be a better approximation. Future research could try to use normally distributed determinants of threshold to explain and predict the distribution of thresholds.

By imposing the threshold model on data, the author defined *purchasing sections* (or it can be called *activation section* to apply to more general datasets). The concept of a purchasing section was completely based on assumptions and definitions, but the usage of this concept, by cutting time series data into sections and treating each section as one activation, turned out to be

highly informative about and consistent with empirical data patterns. The definition and validation of *purchasing section* showed that: 1) Behavior does not equal activation (if behavior is repeated). Multiple behaviors may be counted as one activation and each new activation may involve multiple behaviors, including new ones and re-occurring ones. 2) Due to its high level of matching to empirical data patterns, the concept of an *activation section* (e.g. purchasing section) can be used as a corner stone in model building in order to achieve isomorphism.

CHAPTER 4. Discussion

Finally, it is worthwhile to discuss how robust the simulation results are when the assumption of threshold distribution is violated. For research that had analytical proofs using simulation only as a demonstration of the proof, as long as the proof was not relying on a specific kind of threshold distribution, the change of threshold distribution should not affect the conclusions of the research. For example, Kempe, Kleinberg and Tardos (2015) discussed how the optimal solution of the influence maximization problem could be approximated. The influence maximization problem tries to find the best set of initial seeds before diffusion that can generate the largest number of adopters after diffusion. The authors utilized the submodularity of the influence function and built an algorithm that could identify the initial seeds whose ultimate influence approximates the optimal seed set. This algorithm could work on networks with any threshold diffusion theoretically speaking. Although the authors ran their simulation on a network with a random uniform threshold distribution for shorter run time and ease of approximation, their algorithm still holds when other kinds of threshold distributions are used. The only difference is that when the threshold distribution differs, the optimal set of initial seeds identified by the algorithm would differ.

For simulation research whose conclusions relied on its choice of threshold distribution, the violation of the threshold distribution assumption could be problematic. For example, Centola, Eguiluz and Macy (2007) assumed a constant threshold and conducted their simulation experiments by varying the constant threshold value simultaneously for every person. If the threshold is not constant, with each person taking a different threshold value, it is not clear how this experiment could be done, because threshold would not be a parameter anymore but rather a distribution.

The current research has limitations as well. A single study, even with four data sets, is not enough to draw a firm conclusion about the threshold distribution. Further investigations of the empirical distribution of thresholds need to be carried out. The present research bracketed all inter-individual influences with the word "social influence" and assumed it is only social influence that can overcome thresholds and activate behaviors. However, there are more refined ways to characterize different inter-individual influences. For example, there could be exchange of information (e.g. word of mouth, Goldenberg, Libai and Muller, 2001), externality effects (things become more desirable when more people have it), competition and cooperation (e.g. adoption due to social needs rather than self-decision), etc. Each different mechanism of inter-individual influence would suggest a different way of modeling under specific sets of assumptions. The current research bracketed all their effects into one model. Also, activation could be self-driven or need-based and have nothing to do with the influence of others. This fact is ignored completely in the research too. The goal of this research is to impose the threshold model onto datasets and estimate the threshold distribution. So, the correct interpretation of the purpose of this paper is: "suppose mechanisms described in the threshold model is and is the only relevant mechanism, what would the threshold distribution be?"

The threshold distribution only captures which and how many people are susceptible and susceptible to what extent. Other individual attributes such as individuals' popularity may also play an important role in the diffusion process. The dual distribution describing people with which level of popularity (centrality) has which level of susceptibility (threshold) will also need to be discussed in future research.

APPENDICES

APPENDIX A: Threshold Distribution Result

NetEase Data

```
==========================================================================
# Based on Dataset from 2013 04
==========================================================================
```

Table 3. Threshold distribution with maximum values

```
(array([1.361e+03, 2.000e+00, 1.000e+01, 2.600e+01, 3.000e+01, 8.600e+01,
        1.880e+02, 3.200e+02, 2.920e+02, 2.290e+02, 7.550e+02, 2.770e+02,
        6.260e+02, 2.870e+02, 8.900e+01, 1.262e+03, 1.820e+02, 1.790e+02,
        4.250e+02, 4.300e+01, 6.500e+02, 1.240e+02, 3.070e+02, 3.700e+01,
        1.420e+02, 1.190e+02, 3.400e+01, 1.800e+01, 0.000e+00, 2.072e+03]),
 array([0.        , 0.03333333, 0.06666667, 0.1       , 0.13333333,
        0.16666667, 0.2       , 0.23333333, 0.26666667, 0.3       ,
        0.33333333, 0.36666667, 0.4       , 0.43333333, 0.46666667,
        0.5       , 0.53333333, 0.56666667, 0.6       , 0.63333333,
        0.66666667, 0.7       , 0.73333333, 0.76666667, 0.8       ,
        0.83333333, 0.86666667, 0.9       , 0.93333333, 0.96666667,
        1.        ]),
 <a list of 30 Patch objects>)
```

KS-test Result(statistic=0.20, pvalue<0.001)

Figure 3. Threshold distribution with maximum values

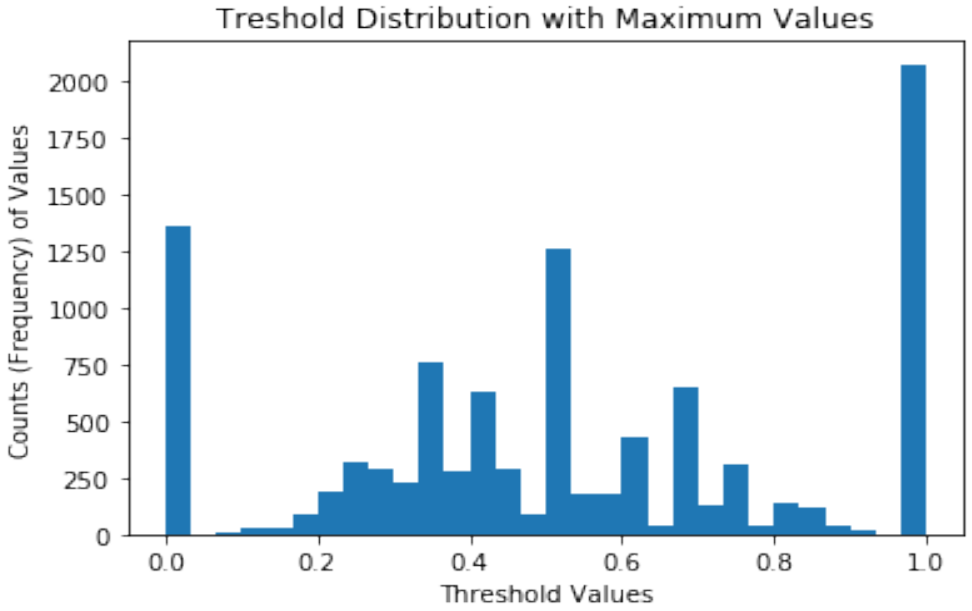


Table 4. Threshold distribution with average values

```
(array([1361.,   23.,  222.,  498.,  553.,  729.,  695.,  808.,  491.,
         329.,  812.,  306.,  319.,  174.,   34., 1292.,   75.,   37.,
         101.,    8.,  220.,   20.,  117.,    4.,   37.,   26.,    9.,
           4.,    0.,  868.]),
 array([0.        , 0.03333333, 0.06666667, 0.1       , 0.13333333,
        0.16666667, 0.2       , 0.23333333, 0.26666667, 0.3       ,
        0.33333333, 0.36666667, 0.4       , 0.43333333, 0.46666667,
        0.5       , 0.53333333, 0.56666667, 0.6       , 0.63333333,
        0.66666667, 0.7       , 0.73333333, 0.76666667, 0.8       ,
        0.83333333, 0.86666667, 0.9       , 0.93333333, 0.96666667,
        1.        ]),
 <a list of 30 Patch objects>)
```

KS-test Result (statistic=0.35, pvalue<0.001)

Figure 4. Threshold distribution with average values

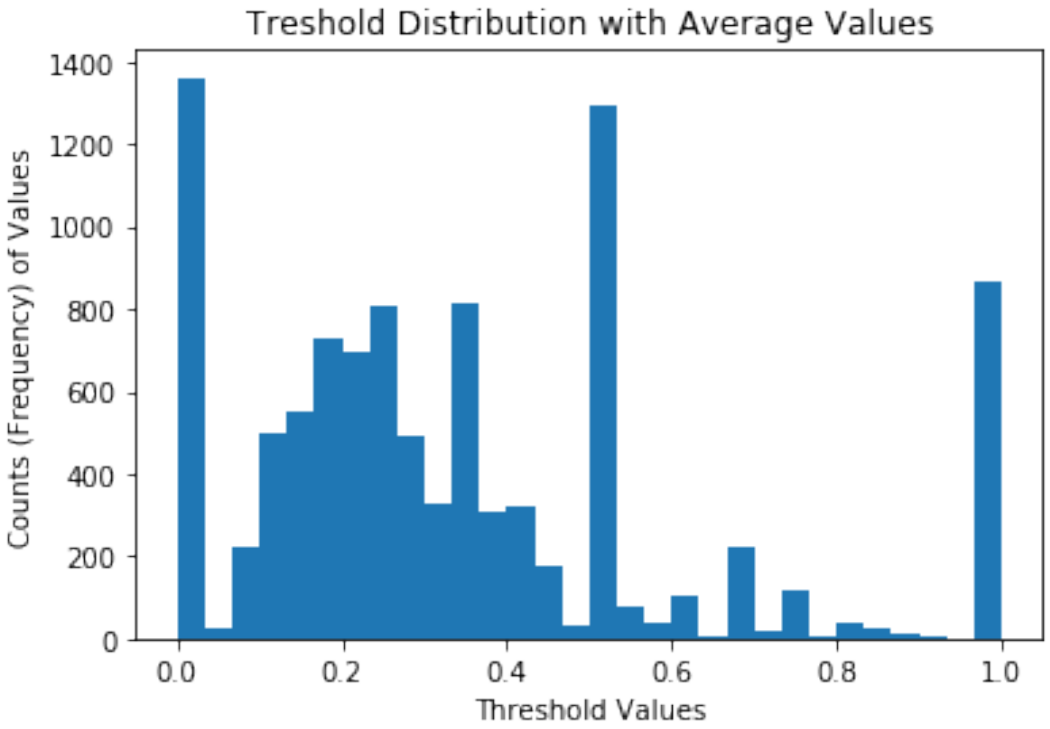
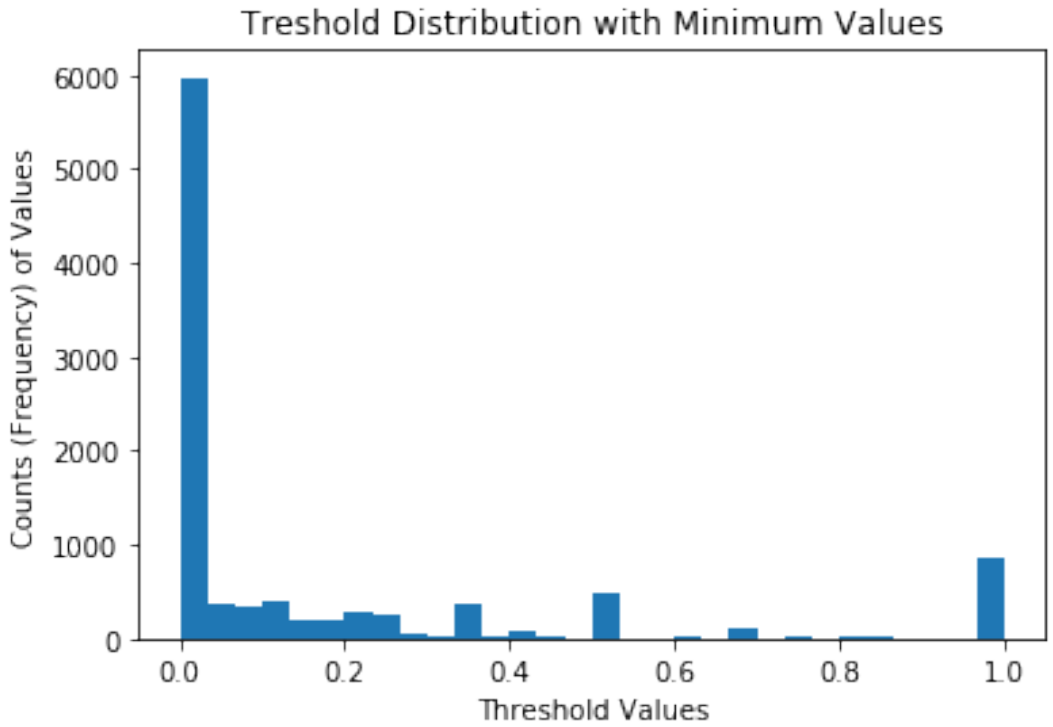


Table 5. Threshold distribution with minimum values

```
(array([5.97e+03, 3.67e+02, 3.51e+02, 3.88e+02, 2.04e+02, 1.96e+02,
        2.87e+02, 2.61e+02, 5.80e+01, 1.50e+01, 3.59e+02, 2.00e+01,
        8.90e+01, 9.00e+00, 0.00e+00, 4.92e+02, 8.00e+00, 8.00e+00,
        3.30e+01, 2.00e+00, 1.08e+02, 7.00e+00, 3.00e+01, 3.00e+00,
        1.50e+01, 1.80e+01, 4.00e+00, 2.00e+00, 0.00e+00, 8.68e+02]),
 array([0.        , 0.03333333, 0.06666667, 0.1       , 0.13333333,
        0.16666667, 0.2       , 0.23333333, 0.26666667, 0.3       ,
        0.33333333, 0.36666667, 0.4       , 0.43333333, 0.46666667,
        0.5       , 0.53333333, 0.56666667, 0.6       , 0.63333333,
        0.66666667, 0.7       , 0.73333333, 0.76666667, 0.8       ,
        0.83333333, 0.86666667, 0.9       , 0.93333333, 0.96666667,
        1.        ]),
 <a list of 30 Patch objects>)
```

KS-test Result (statistic=0.57, pvalue<0.001)

Figure 5. Threshold distribution with minimum values



```
================================================================================
# Based on Dataset from 2013 09

================================================================================
```

Table 6. Threshold distribution with maximum values

```
(array([503.,   2.,   4.,   8.,  15.,  32.,  78., 102., 140., 142., 328.,
        146., 306., 198., 108., 605., 154., 121., 236.,  47., 286., 114.,
        160.,  35., 106.,  90.,  33.,  16.,   5., 757.]),
 array([0.        , 0.03333333, 0.06666667, 0.1       , 0.13333333,
        0.16666667, 0.2       , 0.23333333, 0.26666667, 0.3       ,
        0.33333333, 0.36666667, 0.4       , 0.43333333, 0.46666667,
        0.5       , 0.53333333, 0.56666667, 0.6       , 0.63333333,
        0.66666667, 0.7       , 0.73333333, 0.76666667, 0.8       ,
        0.83333333, 0.86666667, 0.9       , 0.93333333, 0.96666667,
        1.        ]),
 <a list of 30 Patch objects>)
```

KS-test Result (statistic=0.16, pvalue<0.001)

Figure 6. Threshold distribution with maximum values

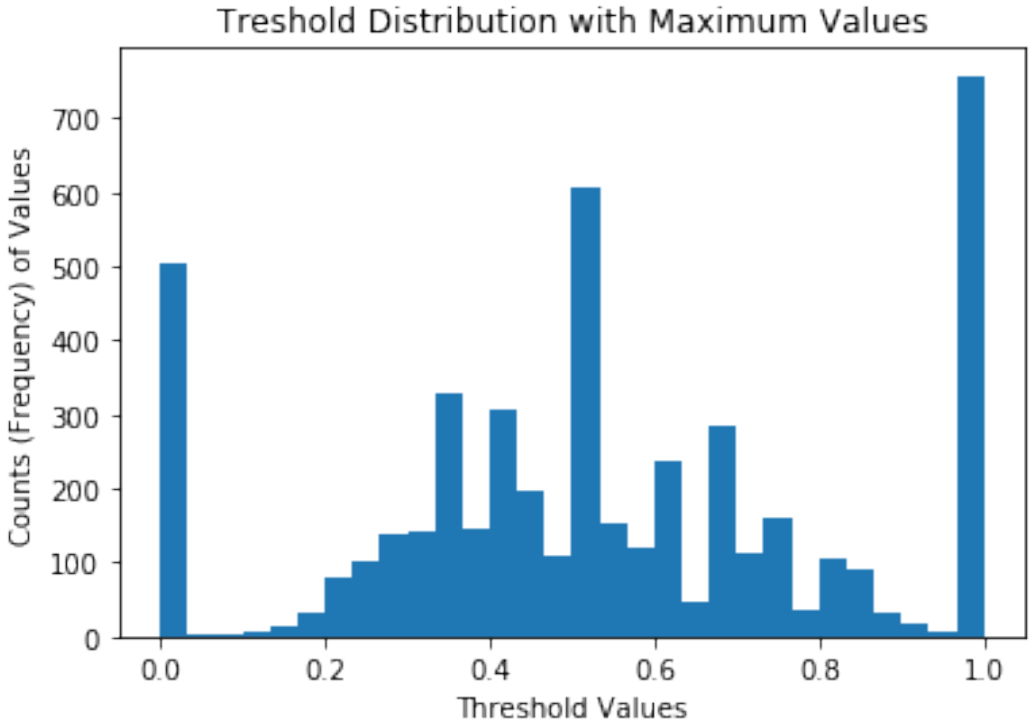


Table 7. Threshold distribution with average values

```
(array([503.,  13.,  92., 216., 262., 351., 341., 366., 250., 194., 368.,
        165., 189., 102.,  32., 527.,  55.,  46.,  67.,  13., 112.,  36.,
         64.,   4.,  36.,  27.,  18.,   6.,   5., 417.]),
 array([0.        , 0.03333333, 0.06666667, 0.1       , 0.13333333,
        0.16666667, 0.2       , 0.23333333, 0.26666667, 0.3       ,
        0.33333333, 0.36666667, 0.4       , 0.43333333, 0.46666667,
        0.5       , 0.53333333, 0.56666667, 0.6       , 0.63333333,
        0.66666667, 0.7       , 0.73333333, 0.76666667, 0.8       ,
        0.83333333, 0.86666667, 0.9       , 0.93333333, 0.96666667,
        1.        ]),
 <a list of 30 Patch objects>)
```

KS-test Result (statistic=0.31, pvalue<0.001)

Figure 7. Threshold distribution with average values

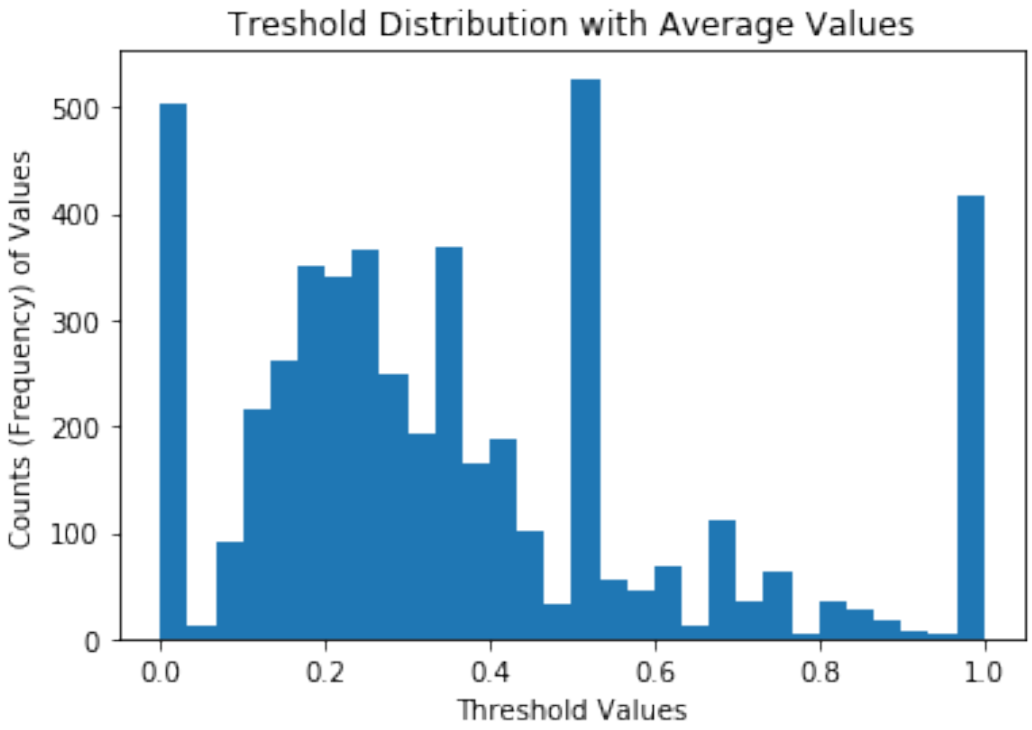
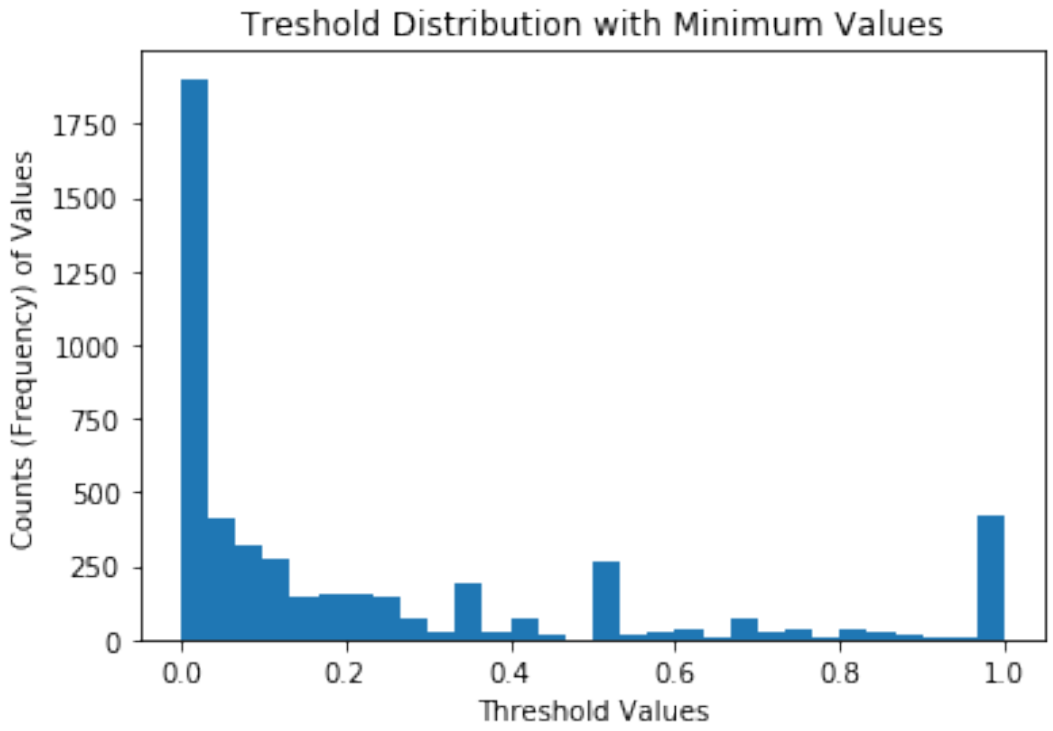


Table 8. Threshold distribution with minimum values

```
(array([1903.,  408.,  322.,  270.,  142.,  155.,  155.,  145.,   70.,
         22.,  194.,   22.,   73.,   15.,    2.,  265.,   12.,   23.,
         34.,    5.,   76.,   26.,   36.,    3.,   32.,   24.,   15.,
          6.,    5.,  417.]),
 array([0.        , 0.03333333, 0.06666667, 0.1       , 0.13333333,
        0.16666667, 0.2       , 0.23333333, 0.26666667, 0.3       ,
        0.33333333, 0.36666667, 0.4       , 0.43333333, 0.46666667,
        0.5       , 0.53333333, 0.56666667, 0.6       , 0.63333333,
        0.66666667, 0.7       , 0.73333333, 0.76666667, 0.8       ,
        0.83333333, 0.86666667, 0.9       , 0.93333333, 0.96666667,
        1.        ]),
 <a list of 30 Patch objects>)
KS-test Result (statistic=0.48, pvalue<0.001)
```

Figure 8. Threshold distribution with minimum values



```
==========================================================================
# Based on Dataset from 2014 04

==========================================================================
```

Table 9. Threshold distribution with maximum values

```
(array([384.,   1.,   4.,   8.,  12.,  32.,  42.,  81.,  67.,  68., 212.,
         76., 141., 100.,  37., 282.,  74.,  67.,  89.,  16., 119.,  38.,
         55.,  15.,  32.,  22.,  12.,   3.,   0., 345.]),
 array([0.        , 0.03333333, 0.06666667, 0.1       , 0.13333333,
        0.16666667, 0.2       , 0.23333333, 0.26666667, 0.3       ,
        0.33333333, 0.36666667, 0.4       , 0.43333333, 0.46666667,
        0.5       , 0.53333333, 0.56666667, 0.6       , 0.63333333,
        0.66666667, 0.7       , 0.73333333, 0.76666667, 0.8       ,
        0.83333333, 0.86666667, 0.9       , 0.93333333, 0.96666667,
        1.        ]),
 <a list of 30 Patch objects>)
```

KS-test Result (statistic=0.16, pvalue<0.001)

Figure 9. Threshold distribution with maximum values

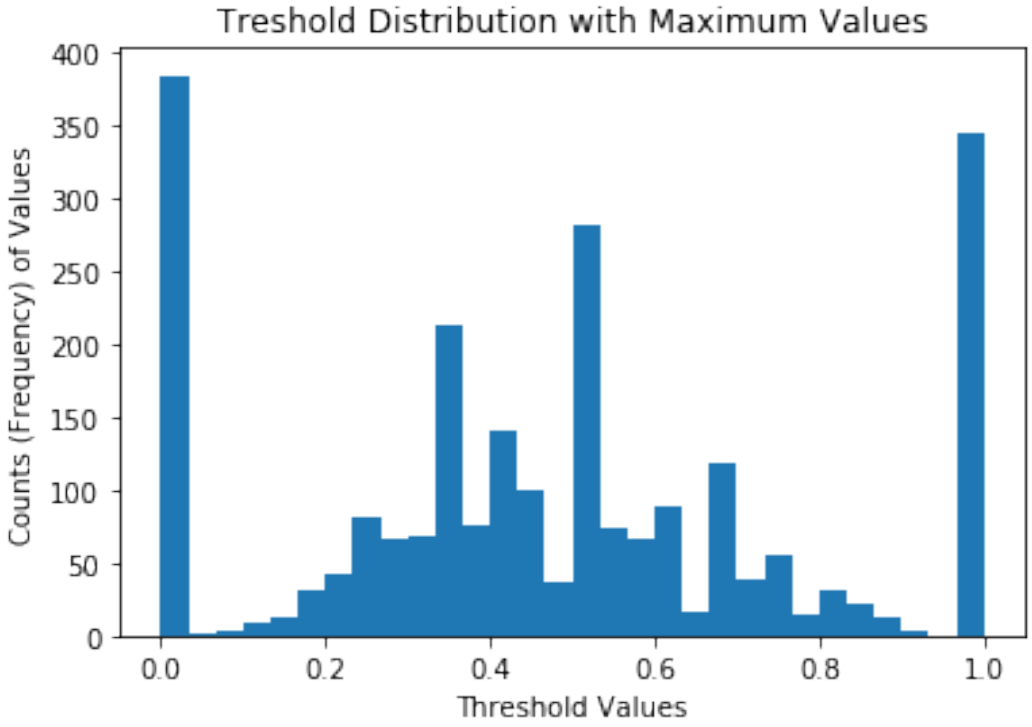


Table 10. Threshold distribution with average values

```
(array([384.,  13.,  48., 107., 135., 178., 176., 189.,  92., 104., 168.,
         71.,  70.,  37.,  10., 260.,  18.,  15.,  29.,   6.,  51.,   7.,
         26.,   5.,  17.,  12.,   5.,   2.,   0., 199.]),
 array([0.        , 0.03333333, 0.06666667, 0.1       , 0.13333333,
        0.16666667, 0.2       , 0.23333333, 0.26666667, 0.3       ,
        0.33333333, 0.36666667, 0.4       , 0.43333333, 0.46666667,
        0.5       , 0.53333333, 0.56666667, 0.6       , 0.63333333,
        0.66666667, 0.7       , 0.73333333, 0.76666667, 0.8       ,
        0.83333333, 0.86666667, 0.9       , 0.93333333, 0.96666667,
        1.        ]),
 <a list of 30 Patch objects>)
```

KS-test Result (statistic=0.34, pvalue<0.001)

Figure 10. Threshold distribution with average values

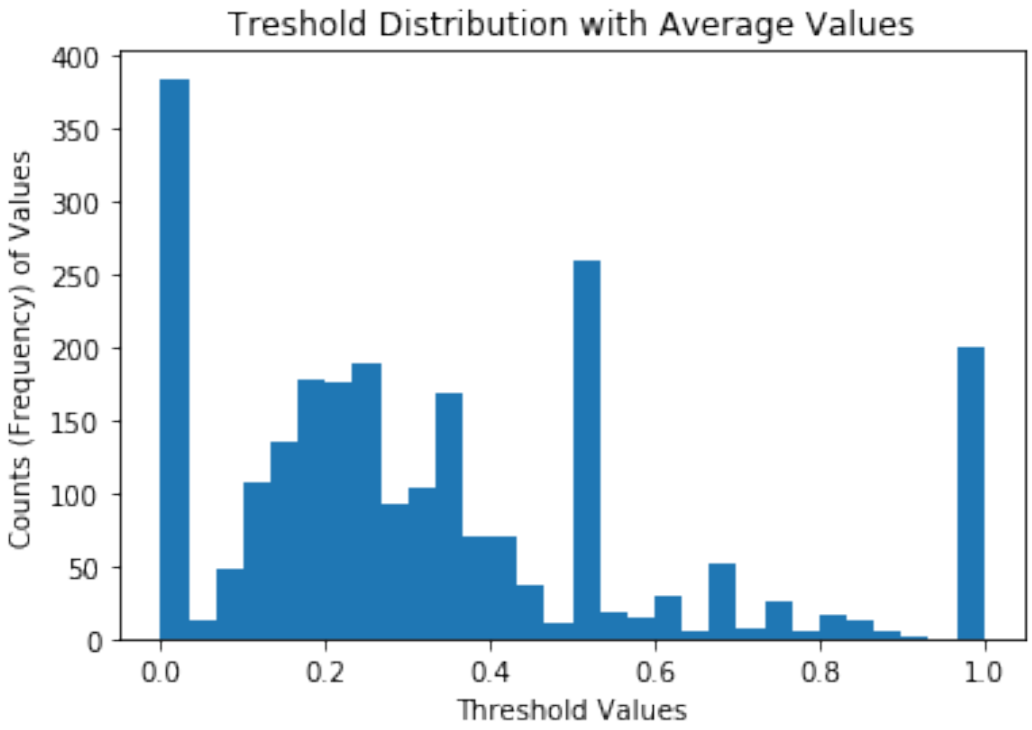
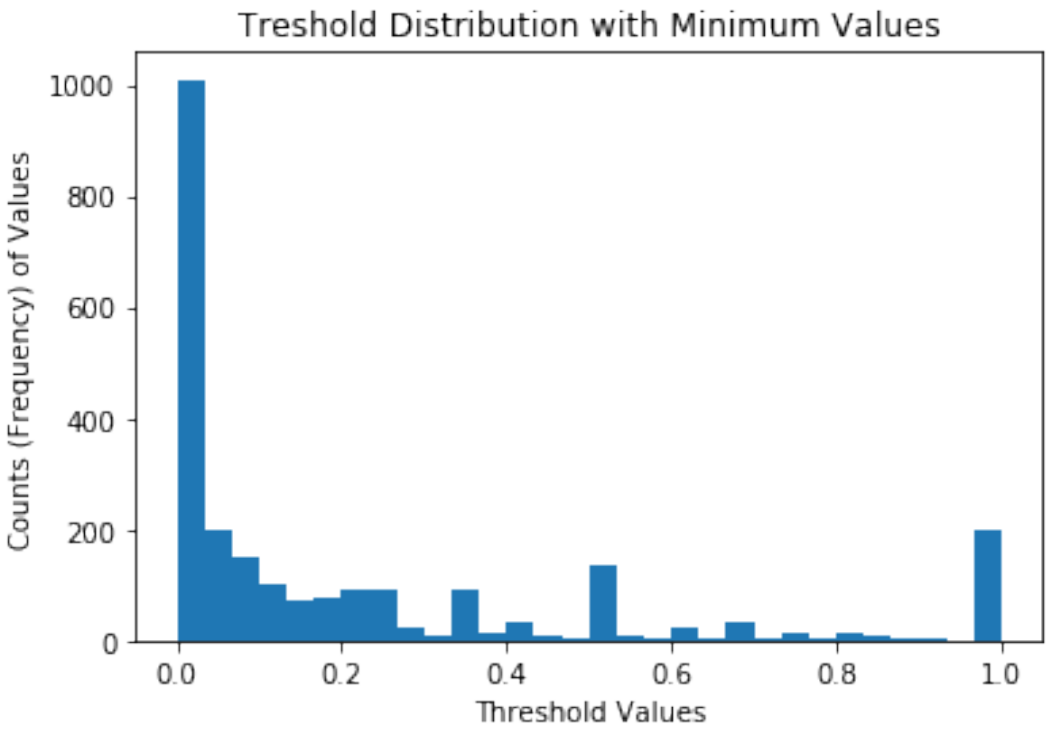


Table 11. Threshold distribution with minimum values

```
(array([1009.,  197.,  148.,   99.,   71.,   79.,   93.,   91.,   23.,
          11.,   92.,   14.,   31.,    7.,    3.,  135.,    7.,    6.,
          21.,    5.,   34.,    6.,   15.,    5.,   16.,   11.,    4.,
           2.,    0.,  199.]),
 array([0.        , 0.03333333, 0.06666667, 0.1       , 0.13333333,
        0.16666667, 0.2       , 0.23333333, 0.26666667, 0.3       ,
        0.33333333, 0.36666667, 0.4       , 0.43333333, 0.46666667,
        0.5       , 0.53333333, 0.56666667, 0.6       , 0.63333333,
        0.66666667, 0.7       , 0.73333333, 0.76666667, 0.8       ,
        0.83333333, 0.86666667, 0.9       , 0.93333333, 0.96666667,
        1.        ]),
 <a list of 30 Patch objects>)
```

KS-test Result (statistic=0.48, pvalue<0.001)

Figure 11. Threshold distribution with minimum values



LastFM Data - threshold distribution visualization

```
==========================================================================
# Based on last.fm Dataset liking

==========================================================================
```

Table 12. Threshold distribution with maximum values

```
(array([62., 30., 29., 21.,  8., 15.,  8.,  2.,  7.,  0.,  3.,  3.,  1.,
         0.,  0.,  6.,  1.,  0.,  1.,  0.,  1.,  0.,  0.,  0.,  0.,  0.,
         0.,  0.,  0.,  2.]),
 array([0.        , 0.03333333, 0.06666667, 0.1       , 0.13333333,
        0.16666667, 0.2       , 0.23333333, 0.26666667, 0.3       ,
        0.33333333, 0.36666667, 0.4       , 0.43333333, 0.46666667,
        0.5       , 0.53333333, 0.56666667, 0.6       , 0.63333333,
        0.66666667, 0.7       , 0.73333333, 0.76666667, 0.8       ,
        0.83333333, 0.86666667, 0.9       , 0.93333333, 0.96666667,
        1.        ]),
 <a list of 30 Patch objects>)
```

KS-test Result (statistic=0.66, pvalue=4.06e-85)

Figure 12. Threshold distribution with maximum values

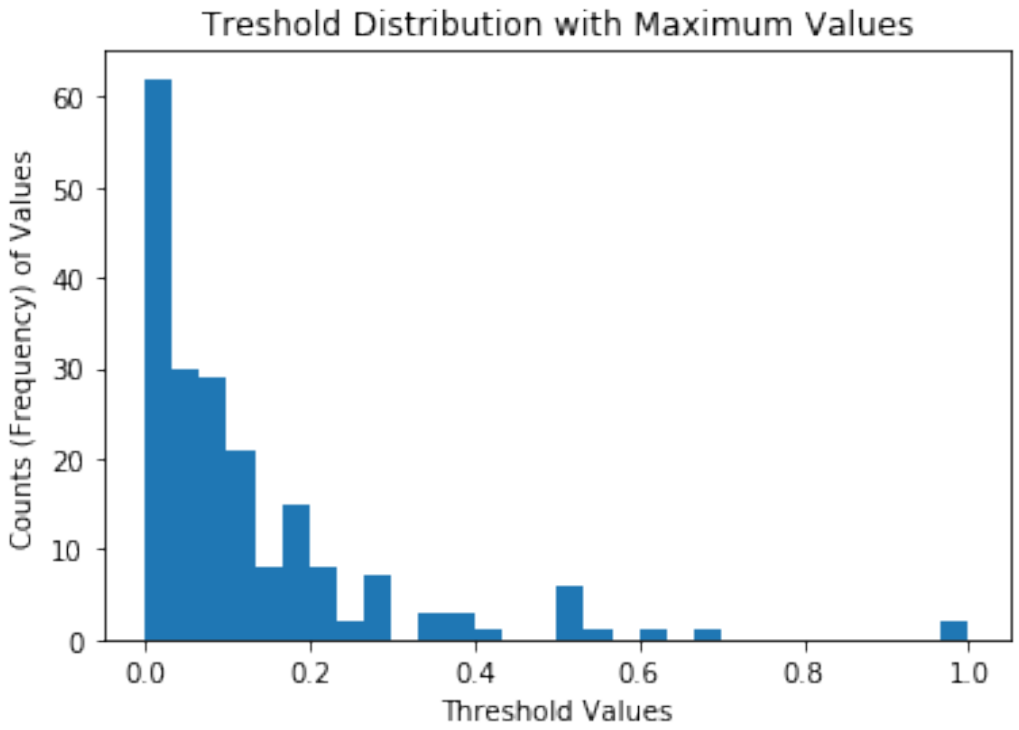


Table 13. Threshold distribution with average values

```
(array([83., 30., 29., 13., 10.,  6.,  9.,  2.,  4.,  2.,  0.,  1.,  2.,
         1.,  1.,  0.,  2.,  0.,  1.,  1.,  0.,  0.,  1.,  0.,  0.,  0.,
         0.,  0.,  1.,  1.]),
 array([0.        , 0.00809524, 0.01619048, 0.02428571, 0.03238095,
        0.04047619, 0.04857143, 0.05666667, 0.0647619 , 0.07285714,
        0.08095238, 0.08904762, 0.09714286, 0.1052381 , 0.11333333,
        0.12142857, 0.12952381, 0.13761905, 0.14571429, 0.15380952,
        0.16190476, 0.17      , 0.17809524, 0.18619048, 0.19428571,
        0.20238095, 0.21047619, 0.21857143, 0.22666667, 0.2347619 ,
        0.24285714]),
 <a list of 30 Patch objects>)
```

KS-test Result (statistic=0.88, pvalue=4.51e-185)

Figure 13. Threshold distribution with average values

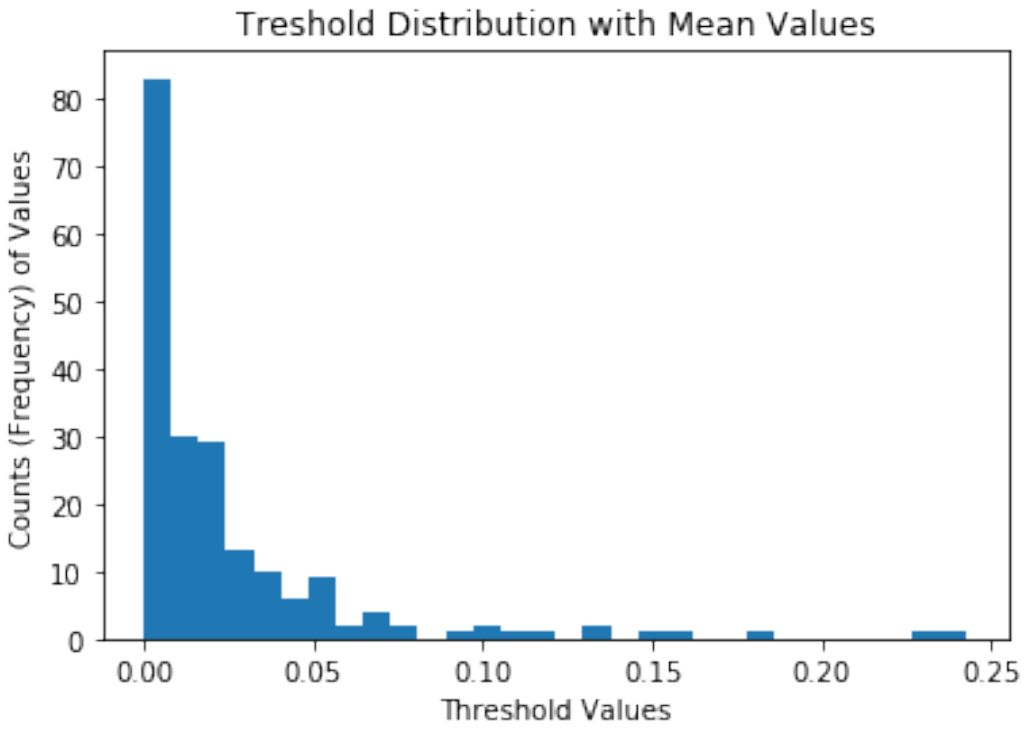
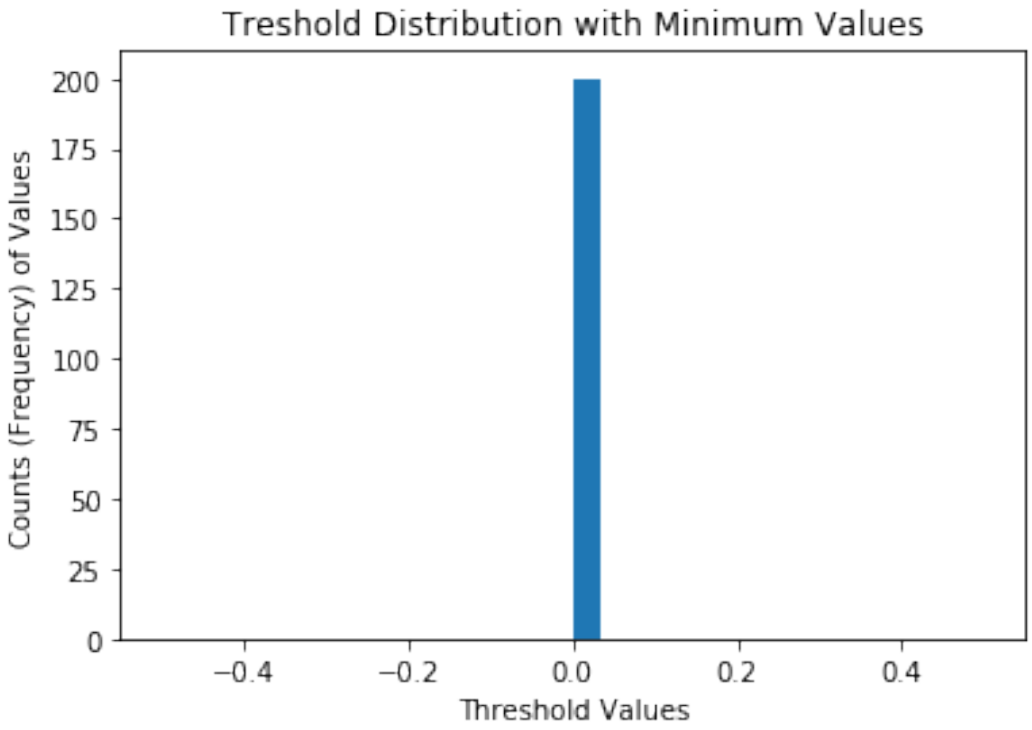


Table 14. Threshold distribution with minimum values

```
(array([ 0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,
         0.,  0.,  0.,  0., 200.,  0.,  0.,  0.,  0.,  0.,  0.,
         0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.]),
 array([-0.5       , -0.46666667, -0.43333333, -0.4       , -0.36666667,
        -0.33333333, -0.3       , -0.26666667, -0.23333333, -0.2       ,
        -0.16666667, -0.13333333, -0.1       , -0.06666667, -0.03333333,
         0.        ,  0.03333333,  0.06666667,  0.1       ,  0.13333333,
         0.16666667,  0.2       ,  0.23333333,  0.26666667,  0.3       ,
         0.33333333,  0.36666667,  0.4       ,  0.43333333,  0.46666667,
         0.5       ]),
 <a list of 30 Patch objects>)
```

KstestResult(statistic=0.99, pvalue<0.001)

Figure 14. Threshold distribution with minimum values



APPENDIX B: Model, Theory and Simulation

Models and Modeling

This section is going to discuss what we mean when we say "model"; what models do; what modeling (effort) is; whether a "path model" or a "structural equation model" is a model.

We refer to people in ideal shape with proportional and symmetric body parts who demonstrate the dresses made by designers as models. Medical students study skeletons of humans and call them models. We refer to miniature airplanes made of wood as models. When we go to the realtor's office and our agent points a laser pointer to an apartment in a small vivid model of the neighborhood on a huge table, we also experience the usage of models and would require that this model is a good honest one. The basic meaning of "model" is **representation**, a simplified idealized representative presentation of a class of complex variations of things that are hard to fully present. A model is supposed to be a typical instance of the class of things it represents by capturing the most defining features of that class of entities. If a model does not achieve good representation, it is not a (good) model.

A model can not only represent entities but also **processes**. Models of the revolutions of the solar system, models of photosynthesis, models of evolution, models of climate change, models of spiral of silence, etc. are all examples of models of processes. Two more concepts are brought into consideration when it comes to the models of processes: **system** and **time**. The set of involved parts in a process together form a system. The sun plus the eight planets revolving around it form a system. The genes of animals, the animals competing with each other and the environment they exist in altogether form a system. In physical sciences, a system can be open, closed or isolated. An open system does not contain all the materials that contribute to the change of the system's state and has material exchange with the outside of the system. An unsealed fish tank is an open system because air is pumped into the water and food is dropped in from the outside. A community of interacting people with their opinions and behaviors, say a university, is analogous to an open system because people come and go from the community. A closed system contains all the materials that are related to the system but not all the energy, so a closed system can have energy exchange with the outside but not material exchange. For example, the earth is still largely a closed system (if we ignore the spaceships that we send out and meteorocks that come in). A small group of people having a secret meeting behind closed doors can be considered as a closed system. An isolated system is a system that does not have material exchange or energy exchange with the outside. The universe can be seen as an isolated system. Some government secret chambers that can be well sealed and blocks all radio transmissions from the outside can be an isolated system when it is sealed up. For any model of systems, it is critical that the model makes clear what system is being modeled.

In addition, processes can only be made sense of over time because processes are about changes and changes happen over time by definition. The language of social psychology research methods is misleading in this context. Scholars analyzing cross-sectional data uses the word "increase" and "decrease" to refer to the difference of the mean scores between groups. For example, people might say "compared to the placebo group, the treatment group's mean increased by …". Here, by "increase", they don't mean "change". They mean "vary". The same applies to when people interpret regression coefficients. By saying "increase" in "for every unit increase in the predictor, the dependent variable increases by …" people mean "vary", not "change". The word "predict" in regression interpretation misleads people even more. People think of "predict" as in "predict the future state" while the word "predict" in simple linear regression only means "explain the variance of" or "covary with". The usage of the word "increase" or "predict" does not necessarily mean a model is about a process. How exactly should processes by modeled and what identifying feature marks that a study is about processes? A paradigm is introduced below.

Processes are about changes of the state of systems over time. This puts another requirement on models (in addition to good representation) if a model is a model of process: **isomorphism**. A model of process needs to achieve (a high level of) isomorphism (Miller & Page, 2009).

The idea of isomorphism is formalized by mathematicians. Its mathematical definition and verbal definition are put below:

Mathematics: Identify and define a system. Then let *S* be the actual state of the system at time point one; let *S*' be the actual state of the same system at time point two which is later than time point one. Let P() be the process that describe how the system changes from one state to another over time, such that P(*S*) = *S*'. Let *s* be the model's simplified and idealized representation of S, let s' be the model's simplified idealized representation of S' and let p() be the model's simplified idealized representation of P(). Then a model (with a consistent way of such simplified idealized representations) achieves isomorphism with the actual system is one in which

$$P(S) = S' \Leftrightarrow p(s) = s'$$

holds for any pair of (S, S') and their corresponding counterparts (s, s').

English: A model of a system is isomorphic with the actual system it models if

1) The representation of whatever change in the actual system state is exactly the change in the representation of the actual system state, and
2) The actual interpretation of whatever change in the model's representation of system state is exactly the change in the interpretation of the model's representation of system state.

The action of simplifying and idealizing reality into model counterparts is called representation and the action of translating model's language into its reality counterparts is called interpretation. Isomorphism basically says the "representation of change must be the change in the representation" and the "interpretation of change must equal the change in the interpretations". Or put simpler, the model and the reality must change hand in hand. Here is a trivial example of an isomorphic model of reality. Suppose initially there is a bowl and three things outside of the bowl: a bottle of water, a piece of burning coal and a piece of dry wood. We are going to randomly pick two or three things outside of the bowl and put them into the bowl. We want to know whether after the process one of the initial things outside of the bowl gets destroyed. In this case, a modeler can represent water and wood with number positive one, represent fire with negative one, and represent the process of putting things in the bowl as multiplying the corresponding numbers together. Positive result means nothing destroyed and negative result means something's destroyed in the process. This model is isomorphic to the reality because,

1) If we put, say, water and fire into the bowl, the process in reality will have water destroy fire and the representation of that result is -1, which is exactly the result of multiplying 1 (representing water) and -1 (representing fire), and
2) If we multiply, say, 1 and 1 in the model, then we get 1, indicating this combination won't destroy the things we put in the bowl, which is exactly what happens if we put in water and wood in bowl.

We can check 1) and 2) for all possible outcomes of the system with all combinations of the three elements allowed and see that both will hold for all combinations.

Theoretical physics is possible because the field achieved almost perfect isomorphism between physical systems and mathematical systems. The existence of many particles was proved mathematically long before they were discovered in reality. The field of Chemistry uses equations of chemical symbols to isomorphicly describe the process of chemical reactions and they were able to design the synthesis process of a product on paper before actually synthesizing it. The advantage of finding an isomorphic model of the system of interest is that we can represent a system and its processes more simply so we can describe them, comprehend them and do reasoning about them more efficiently than direct experiments. More importantly, this action of isomorphic modeling transforms many problems about the real world into problems of symbols, mathematics, logic and computation. It allows the scholars in a field to channel in the knowledge of scholars in many other fields. For example, representation theory says every group can be represented with a permutation group. We may not have much knowledge about a specific group of entities, but as long as we find a way to define it as a group, then we can isomorphicly represent it with permutation groups, which mathematicians already know a lot about. The same applies to networks. Many real-world problems are better solved when translated into a modeling problem of networks.

Moreover, isomorphic model representations with symbol systems support long reasoning chains far better than direct description or non-isomorphic models of reality.

Long chain reasoning is the kind of reasoning that makes many statements, each of which is either based on previous statements or directly based on the model's assumptions, evidences and inputs. The number of statements that base on each other is, loosely speaking, the depth (levels or length) of the reasoning chain. For example, one can use the model to reason about the system's state based on the current state and make a statement to describe the result of that reasoning, then based on this statement (expressed in the model's way of representation), one can make another statement, or a meta statement about the first statement, and call that statement a second statement, etc. This process cannot go long without an isomorphic model. Imagine people are doing such reasoning with natural language, the vagueness of words and the arbitrary choice of words will soon mislead the direction of the reasoning. One may argue that the reasoner can stick to reality by trying to use natural language to represent the process as closely and clearly as possible. But note that many relations and processes don't have counterparts in natural language that are close in their meaning to what the processes do.

Take the "water-fire-wood" mind experiment as an example again. When we allow multiple times of input one after one, say, we want to input "water, wood, wood, water, fire, wood, fire" in the order. Would there be something destroyed in the process? The model, once with its isomorphism proved, will just multiply many 1's and -1's without introducing any new assumptions. But to explicitly explain the result verbally, one would have to go through the chain, talk about the current state of the system one by one until something is destroyed. One may be attempted to shorten the explanation by saying "as long as there are inconsistent elements in the chain there will be things destroyed". But this statement is not given and needs to be proved. And the reasoner used the word "inconsistent" without defining it. Soon the reasoner will find that defining "inconsistent" incurs other new words and statements and proving an intuitive statement is not always easy. Actually, by saying this, the reasoner was verbally saying something similar to a commutative law of multiplication. The reasoner wants to use it but neither realizes it nor knows how to prove it. The advantage of modeling is that this law is given once the isomorphism between the three-element system and the multiplication of 1's and -1's is established. Actually, all the mathematical knowledge about negative numbers, multiplication, identity elements etc. are now given to the modeler.

Now, imagine people are doing long chain reasoning with a non-isomorphic model. The modeler will be making representations that deviate from the reality. The further statements the modeler makes will be based on deviated representations of states and according to deviated representation of processes. The more steps the reasoning takes, the more possible the reasoner deviates further from reality. So, isomorphism is the key to solving *the problem of long-chain reasoning* in modeling.

Admittedly, isomorphism is hard to achieve. It requires the modeler to be well-acquainted with the system's behaviors and well-versed with symbol systems. It requires the system to be well defined and by definition possible to be accurately measured over time. It sometimes requires the modeler to find an existing symbol system that has the perfect coincidence to be used as isomorphism to the system to be modeled or the modeler will have to create one. Let this problem be called *the problem of need for coincidence* in isomorphic modeling.

It is hard to create models in this sense under the current social psychology tradition paradigm. Individual attitudes, as currently defined and measured, is impossible to be accurately measured over time. Attitudes don't always exist throughout the time period. Sometimes it only exists when it is asked about. Attitudes cannot be measured over and over again too frequently with exactly the same items because people may remember what they answered and quickly become unsensitized. In addition, attitudes cannot be measured over and over again too frequently with exactly the same items (say 4 time in 30 min while one watches a video) because the participants would probably have to start answering the same questionnaire again right after they have just finished answering it. For yet a third reason, attitudes cannot be measured over and over again too frequently with exactly the same items (say 4 time in 30 min while one watches a video) because the measurement of attitude (items to answer) will disturb the

treatment (video watching). Attitude, by definition, and by the way it is measured, is not meant to be used to characterize the system's (individual's) state.

Some research in the social psychology paradigm are about overtime change. For example, studies that pretest-measure the sample mean, give treatment and posttest-measure the sample mean are indeed studying change over time. Are these studies doing modeling in the sense described above? (No.) To answer this question we need to ask, what is the system of such studies and are their models trying to achieve isomorphism? The answers are negative.

First, one attempt is to say that the system of study in social psychology is an "average person". Conclusions of social psychologies studies are usually on the variable level about what may change the population mean of a construct. One may think of these studies as treating the whole society as a single person called an "average person", and this "average person's" measures on constructs are inferred with sample means. The conclusions of social psychology are of the form "if … treatments are done to this average person, … is how this average person will change over time." Effect of any single treatment is considered a constant-valued, deterministic effect on this average person. The system social psychology studies is the psychological system of the "average person". Theories of social psychology are theories about the relations of the "average person's" psychological constructs. However, we quickly realize, social psychologists also talk about the (cor)relation strengths of the average person's different psycho constructs. If their system is one "average person" and this average person has one score for each construct, how is correlation computed? The fact that the correlations are computed means that the level of analysis is not this average person, but the individuals the average person is made up of. Then the question is: which one is the system of study, the individual persons or the average person? Social psychology tried to explain away this hidden inconsistency in its methodology worldview by saying analyses can be done on multiple levels with different unit of analysis. When discussing main effects, the unit of analysis is populations/average persons (e.g. "for different populations receiving treatments of different effect sizes, their changes in their means differs"); when discussing correlations, the unit of analysis is individuals (e.g. "individuals who are high in measure A are also mostly high in measure B"). This argument seemed fine under the perspective of social psychology but now becomes a problem under the system modeling perspective: social psychology cannot answer this question, "what is the system you are trying to model?" Social psychology does not have a consistent system. Some of its conclusions think of the system as the set of populations that are compared, some of its conclusions treat the system as a set of single individuals that are compared, and the two conceptions are usually confused and entangled.

Because the social psychology approach is unclear with its identification of the system it studies, many theories and scholars, as a result, do not distinguish between population (average person) and individual person. Many conclusions and interpretations confuse the two levels completely.

For example, in Kahneman and Tversky's famous (1986) paper about framing effects, they found that with the "gain" framed messages, more people preferred a sure gain and avoided taking chances, while in the "loss" framed condition, more people preferred risk seeking and took a chance to avoid loss. The conclusion was in the form of "people's behaviors differed for different frames". The conclusion applies particularly to populations rather than individuals. Although in both conditions, both risk aversion and risk seeking behaviors existed, the researchers still concluded that something differed. The only thing that differed was the majority opinion (proportion of opinions), which was a concept that could only apply to populations. This way of conceptualizing was consistent with the "average person" world view of social psychology. What differed in the two frames was the "average preference" of risk seeking (in other words, the behavior of the "average person" who participated in both conditions differed). A hidden assumption was that every single person was this average person plus some idiosyncrasies. So, if this average person's behavior is understood, it will inform us about every single person's individual preferences. Following this logic, a further step is taken: if "the average person preferred risk aversion in gain frame condition and risk seeking in loss frame condition", then "each and every single individual should behave in the same way more or less". The population level conclusion is directly applied to the individual level.

Some may argue that social psychology is treating the system of interacting variables (the boxes linked by arrows) as its system of study. The variables in the system can be correlated with each other and their means can be compared to each other. This argument is wrong in a more fundamental way: the system that we model needs to be an empirical system in reality rather than a system of symbols or concepts. It can be the eco system, the economic system, the climate system, etc. A system of interrelated variables is not something that exist objectively for us to model. In some sense, the system of variables *is* the model rather than the system to be modeled, except that it is not a model either in the sense of model described in this research.

Path Models Are Not Models

Do longitudinal studies about the change of mean over time for the "average person" count as models? No, because they do not try to achieve isomorphism. For example, one can think of the pretest measure of a psychology construct as *s*, the representation of the system's (average person's) earlier state, *S*, and the posttest measure as *s*', the representation of the system's later state, *S*'. The problem is there is no representation of the process. One may think the treatment is the representation of the process. But actually, the treatment is the process in reality, P(). What a model needs is a representation of it, p(), to capture how this treatment exerts effects. One may also argue that the equation describing the main effect on the mean serves as p(). It does not. Because isomorphism requires that the equivalence below holds for all pairs of *S* and *S'*. An equation about a main effect obviously does not satisfy this requirement. If *S*'' changed more (or less) from *S than S'*, the main effect equation will have to use a corresponding bigger (or smaller) coefficient for the main effect, then that will be a different equation. No one equation will satisfy the following equivalence for all pairs of *S* and *S*'. In statistical words, the main effect equation only captures how one main effect changes the mean but does not capture how the main effect itself can vary.

$$P(S) = S' \Leftrightarrow p(s) = s'$$

Summary

Social psychology studies single individuals by studying populations. The current dominating paradigm of social psychology is not consistent with the approach of process modeling because it neither models a consistent empirical system nor tries to achieve isomorphism. The current dominating paradigm of social psychology is, in my words, descriptive.

Definition of a "Model"

A model is a simplified idealized isomorphic representation of the process of a system. Modeling or model building is the integrated actions of defining a system, looking for good simplified representations to characterize the states and processes of the system and the building and proving of isomorphism. According to this definition of model, a path model or a structural equation model is neither a model nor a system.

"Theory" vs "Model"

The word "theory" is also commonly used by scholars. The relation between "model" and "theory" depends on disciplines and contexts. In mathematics, the word "theory" is used to refer to a general subfield of research. For example, "number theory", "network theory", "graph theory", "game theory", "set theory", "group theory" etc. are all branches of mathematics. According to the mathematicians' use of the word, a theory is a class of studies about a well-defined topic and these studies share exactly the same set of assumptions (axioms). A theory is rarely one theorem, but a set of interrelated axioms, theorems, corollaries and conclusions. A theory usually contains many models and a model can incorporate many theories. In physics, the word "theory" is used to refer to a fundamental world view of the physical nature and the set of equations that can describe the world's fundamental laws under this world view. For example, Newton's theory saw the physical world (e.g. celestial entities, physical bodies etc.) as "motion of mass caused by forces", and he had three equations to describe his three laws of motion and one equation to describe his law of gravity; Einstein's special relativity theory took Newton's world view with some changes: that the things that motion happens in, space and time, are actually related and in nature different dimensions of one and the same thing (spacetime) and that mass

does not hold constant but instead depend on speed. Einstein added some more equations, with the Lorentz transformation at the core. In physics, "theory" and "model" are almost inter-exchangeable. People also say, "in Einstein's model of the universe…". The only nuance is that model usually has an object, as a model of something, and theory does not. Sometimes, the word "model" is used to refer to a theory applied to a specific case. For example, people sometimes say "use the theory of … to model …", Sometimes, the word theory has a connotation of "uncertain". To launch a rocket, people usually say "according to our model, the rocket is expected to …", instead of saying "according to our theory, the rocket is expected to…" probably due to such connotation of the word "theory". In biology, the word "theory" is used to refer to a fundamental conjecture about the bio-nature that is not yet confirmed. For example, people usually say "evolution is only a theory" with the implication that it might be wrong. To emphasize their uncertainty, general public almost never call evolution a model (although computational biologists have built what they call "models" about evolution).

How do social scientists especially communication scientists use the word "theory"? In the social psychology tradition, the things named in the form of "social … theory" usually use a construct to name a social phenomenon and use interrelated constructs to explain the phenomenon through verbal based reasoning. In this context, the use of the word "theory" definitely differs from physicists or mathematicians. To social scientists, a theory does not need the level of generality and abstractness as in natural science theories. A theory is like a descriptive summary of a specific kind of phenomenon. The following saying is what the author of this paper used to captures the difference.

Physicists want to find a "theory for everything" while social scientists want to "find a theory" for everything.

This difference roots in their different ways to see and do science. Physical scientists believe that their phenomena can be ultimately explained by a small set of fundamental rules and everything is different versions, variations, applications, compounds and deductions of these fundamental rules. As a result, their methodologies try to describe those fundamental rules and look for ways to vary, apply, compound and do deductions with these rules. Social scientists, especially social psychologists, see the social phenomena (constructs) as direct or indirect results of other social phenomena (constructs). As a result of this view, their methodology describes causal links and summarizes evidences (e.g. survey data sets, experiment results etc.). The theorizing process of physical scientists rely on mathematics, formalized languages (e.g. process algebra) or other forms of abstract symbol system (e.g. chemical equations). The theorizing process of social scientists rely on verbal reasoning. There are advantages and disadvantages of each approach. Social science directly studies many things that mathematics cannot and may never be able to describe. But heavy reliance on verbal reasoning without a formalized symbol system has its drawback: it's hard to do long-chain reasoning or isomorphic modeling. In physical science papers, it is very common that the final conclusion is more than five steps of deduction under the initial assumptions and axioms; with verbal reasoning, a causal chain more than three steps start to seem too far-fetched. As a result, natural sciences build upon previous theories; social scientists build next to previous theories. To self-quote again:

Natural scientists build skyscrapers while social scientists build gardens.

In this paper, "theory" and "model" will not be distinguished. They both refer to models in the "model of a system's process" sense. The paper only uses the word "model" and not use the word "theory" unless referring to fixed names like "game theory".

Simulation as Model Building

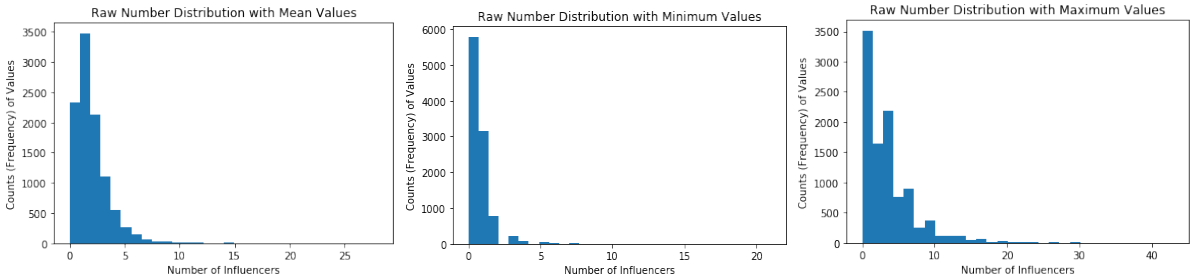
Without going into too much technical details, the author argues that agent based simulations can help solve the long chain reasoning problem for social scientists because the simulator can just code what the reasoning chain argues and play out the results as a demonstration instead of verbally argue for what will happen. For verbal based theorists, simulation is still a good way to validate far-fetched results and

test long reasoning chains. In addition, agent-based modeling can help solve the problem of need for coincidence in isomorphic modeling. With simulation modeling, the simulator does not necessarily need a mathematical representation or design a symbol system to represent the behavior of the system. With non-numerical simulation, the simulator can directly represent many traits, states and behaviors of the system directly.

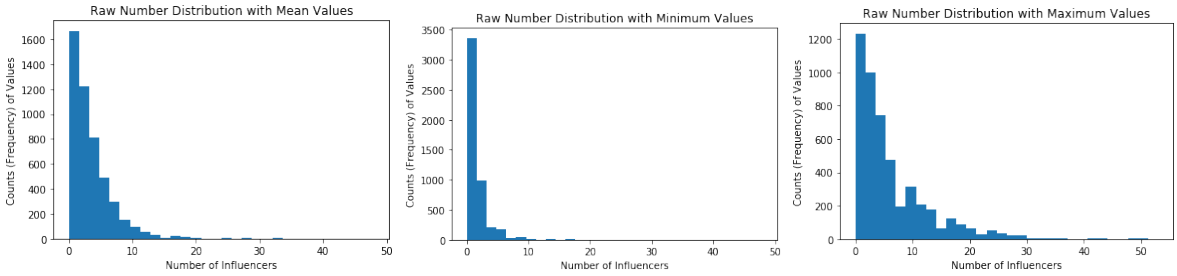
APPENDIX C: Results of Repeated Activation with Raw numbers

Figure 15. Results for additional analyses

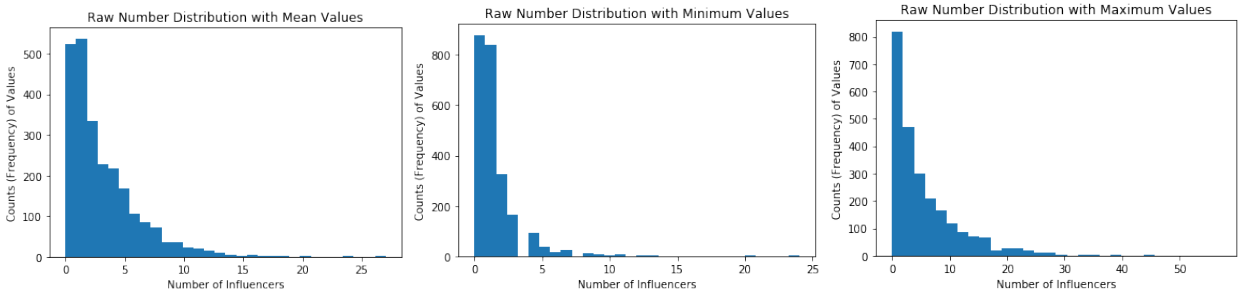
Dataset 201304



Dataset 201309



Dataset 201404



The last.fm dataset

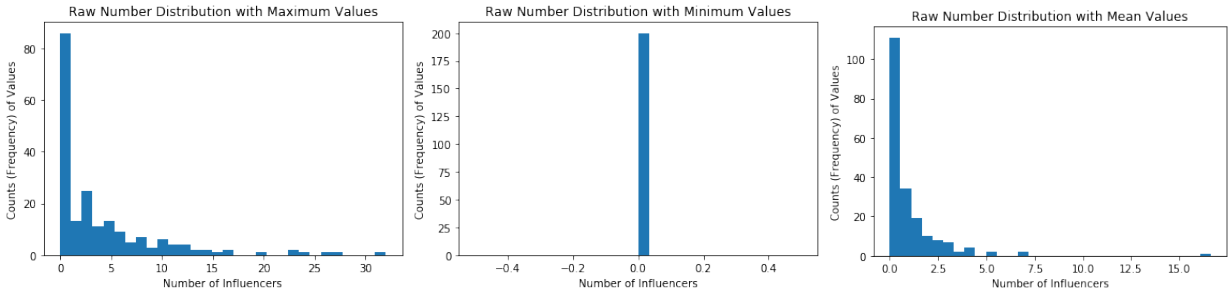
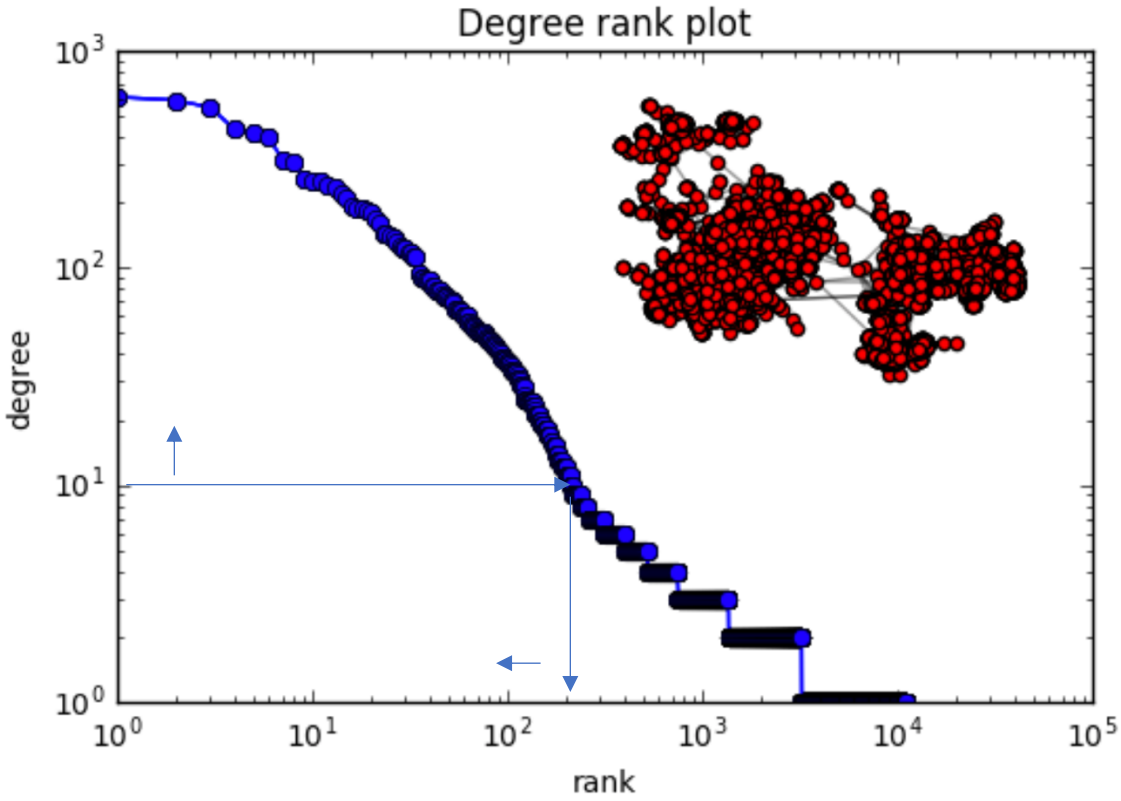


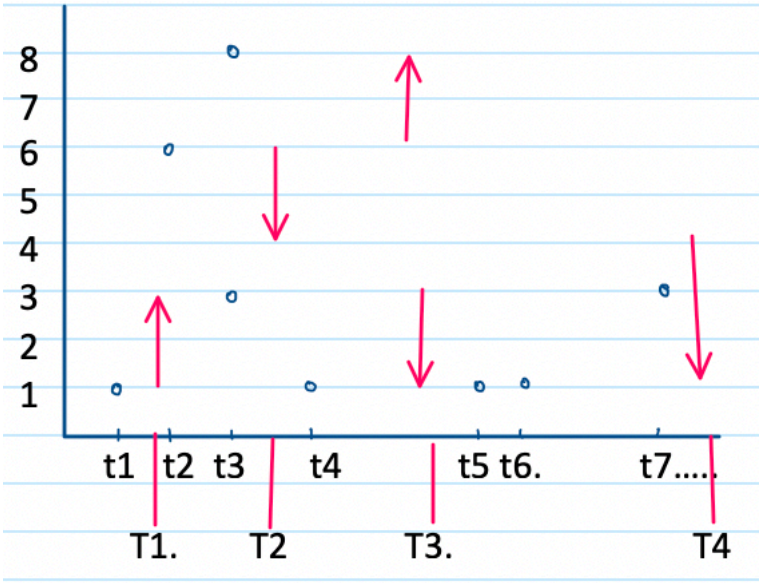
Figure 16. Degree distribution of the whole NetEase dataset



Note: This figure is the degree distribution of the players in the NetEase dataset aggregated across all datasets. In the figure, degree (y axis) is the number of contacts players have and if we put the players in order based on their number of contacts from large to small, every player would have a rank (x axis). In the above picture, players who have 10 friends rank about 200th. The people who have more than 10 contacts (above 10 in the degree axis) will rank higher than 200 (left to 200 on the rank axis). In other words, only about 200 people have more than 10 contacts. The rest of the (about) 10000 players have less than 10 friends.

APPENDIX D: Step by Step Visualization of Chapter 2, 1)-10)

Figure 17. Visualization of activations (circles) and communications (arrows)



Note. The X-axis is the timestamps and the Y-axis are the individual indexes. Individual 1 adopted at timestamp t1, t4, t5, t6…, Individual 3 adopted at timestamp t4, t7…,Individual 6 at t2 and 8 at t3… Individual 1 talked to individual 3 at time T1, 3 to 1 at T3, 6 to 4 at T2, 6 to 8 at T3…

Suppose the picture above is a visualization of the data, according the operationalization defined in 1) – 10), which adoption behaviors belong to the same purchasing section? How many purchasing sections are there? What is the timestamp for the frontier of each purchasing section?

For each individual, (here take individual 1 for example)

First, look at each pair of timestamps of behaviors to see if they are communicatively adjacent. By definition in 6), x_{1,t_1} and $x_{1,t_4}$are communicatively adjacent because no adopter talked to individual 1 between its two adoptions at t1 and t4.

Mathematically,

because

$$\{c = (l, 1, T_j) | t_1 < t_m < T_j < t_4,\ x_{l,t_m} = 1\} = \emptyset$$

By definition in 6), it follows that

$$\text{CA}(x_{1,t_1}, x_{1,t_4})$$

Similarly, $\text{CA}(x_{1,t_5}, x_{1,t_6})$. But $\neg\text{CA}(x_{1,t_1}, x_{1,t_5})$, $x_{1,t_1}$and $x_{1,t_5}$are not communicatively adjacent because individual 3 activated at t3 and communicated to individual 1 at T3 during the time between t1 and t5. Mathematically,

Because

$$t_1 < t_3 < T_3 < t_5 \text{ and } x_{3,t_3} = 1$$

$$\{c = (l, 1, T_j) | t_1 < t_m < T_j < t_4,\ x_{l,t_m} = 1\} = \{c(3,1,T_3)\} \neq \emptyset$$

By definition in 6), it follows that

$$\neg\text{CA}(x_{1,t_1}, x_{1,t_5}),$$

Second, by definition 7), time period t1 through t4 form a Purchasing Section for individual 1. Mathematically,

because

$\mathrm{CA}(x_{1,t_1}, x_{1,t_4})$ and $\neg\mathrm{CA}(x_{1,t_1}, x_{1,t_5})$,

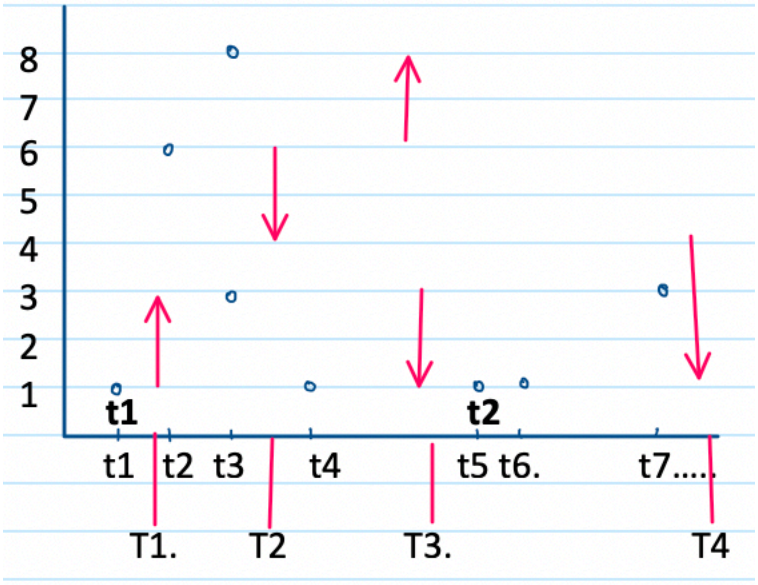
According to definition 7)

$\mathrm{PS}(x_{1,t_1}, x_{1,t_4})$

So, for individual 1, t1 through t4 is a purchasing section and the first two purchasing behaviors belong to the same purchasing section. Timestamps t5 and t6 and the corresponding behaviors belong to the next purchasing section.

According to the definition of frontiers in 10), for individual 1, timestamp t1 is the left edge of the first purchasing section, so it is the first frontier, denoted in bold $\boldsymbol{t_1}$; timestamp t5 is the left edge of the second purchasing section, so it is the second frontier, denoted $\boldsymbol{t_2}$.

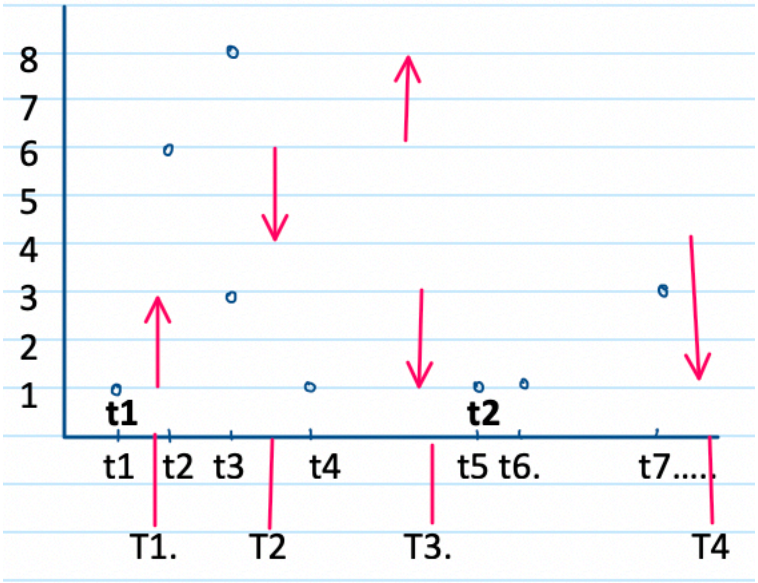
Figure 18. Visualization of activations and communications



Now that we know the timestamp for the frontiers of each individual, by 9), we do not need to update this individual's cumulated social influence until the next timestamp.

APPENDIX E: Step by Step Visualization of Chapter 2, 11)

Figure 18. (cont'd)



Use the above picture as an example visualization of the data. Assume there are only 8 people in the dataset, and individual 1 is connected to 3 people, 3,4 and 8. Then total number of people J = 8, degree centrality of individual 1 is $k_1 = 3$. According to 11), individual 1's cumulated social influence by timestamp $\boldsymbol{t_2}$ is computed as:

$y_{1,\boldsymbol{t_2}} = \frac{\sum_{n=1}^{8} c_{n,1,\boldsymbol{t_2}}}{3}$;

Only 1 person (individual 3) activated and then communicated with individual 1 during the time between $\boldsymbol{t_1}$ and $\boldsymbol{t_2}$, so that only $c_{3,1,\boldsymbol{t_2}} = 1$, and all the other $c_{n,1,\boldsymbol{t_2}}$'s for other values of n equal 0. So, the numerator sums to 1 and the cumulated social influence of individual 1 at $\boldsymbol{t_2}$ is 1/3.

APPENDIX F: Validation of this Paper's Definition of Purchasing Sections

Take the first buyer, with ID '1B31E8C6187' for example.
The paper's algorithm identified the frontiers of purchasing sections for this buyer to be:
[0, 39, 46, 61, 92, 95, 96, 98, 102]
And we can check all the purchases of the buyer sorted by purchasing time from earlier to later:

Table 15. Data for the player

	Buyer ID	Product ID	Purchasing Time
0	'1B31E8C6187'	'2190874'	'2013-09-08 12:32:26'
1	'1B31E8C6187'	'2142018'	'2013-09-09 13:15:19'
2	'1B31E8C6187'	'2112376'	'2013-09-09 13:15:23'
3	'1B31E8C6187'	'2112377'	'2013-09-09 13:15:25'
4	'1B31E8C6187'	'2190982'	'2013-09-09 13:16:21'
5	'1B31E8C6187'	'2190930'	'2013-09-09 13:18:14'
6	'1B31E8C6187'	'2190958'	'2013-09-09 13:18:34'
7	'1B31E8C6187'	'2190940'	'2013-09-09 13:19:01'
8	'1B31E8C6187'	'2190984'	'2013-09-09 13:19:06'
9	'1B31E8C6187'	'2190986'	'2013-09-09 13:19:09'
10	'1B31E8C6187'	'2190982'	'2013-09-09 13:31:43'
11	'1B31E8C6187'	'2190986'	'2013-09-09 13:31:46'
12	'1B31E8C6187'	'2190930'	'2013-09-09 13:32:13'
13	'1B31E8C6187'	'2190984'	'2013-09-09 13:32:15'
14	'1B31E8C6187'	'2190857'	'2013-09-09 14:00:37'
15	'1B31E8C6187'	'2190857'	'2013-09-09 14:01:58'
16	'1B31E8C6187'	'2190858'	'2013-09-09 14:02:28'
17	'1B31E8C6187'	'2190487'	'2013-09-09 14:13:45'
18	'1B31E8C6187'	'2190874'	'2013-09-09 14:16:06'
19	'1B31E8C6187'	'2191808'	'2013-09-09 15:03:53'
20	'1B31E8C6187'	'2191001'	'2013-09-10 03:24:55'
21	'1B31E8C6187'	'2191002'	'2013-09-10 03:25:03'
22	'1B31E8C6187'	'2191003'	'2013-09-10 03:25:12'
23	'1B31E8C6187'	'2191004'	'2013-09-10 03:26:27'
24	'1B31E8C6187'	'2191005'	'2013-09-10 03:26:32'
25	'1B31E8C6187'	'2191002'	'2013-09-10 03:26:58'
26	'1B31E8C6187'	'2191003'	'2013-09-10 03:27:09'
27	'1B31E8C6187'	'2191001'	'2013-09-10 03:27:13'
28	'1B31E8C6187'	'2191006'	'2013-09-10 03:27:54'
29	'1B31E8C6187'	'2190857'	'2013-09-10 04:23:01'
30	'1B31E8C6187'	'2190858'	'2013-09-10 04:23:11'
31	'1B31E8C6187'	'2190858'	'2013-09-10 04:24:10'
32	'1B31E8C6187'	'2190857'	'2013-09-10 04:26:09'
33	'1B31E8C6187'	'2190857'	'2013-09-10 04:26:22'
34	'1B31E8C6187'	'2190477'	'2013-09-10 04:27:50'
35	'1B31E8C6187'	'2190477'	'2013-09-10 04:28:43'
36	'1B31E8C6187'	'2191709'	'2013-09-10 05:01:28'
37	'1B31E8C6187'	'2190476'	'2013-09-10 05:17:15'
38	'1B31E8C6187'	'2190476'	'2013-09-10 05:18:03'
39	'1B31E8C6187'	'2190817'	'2013-09-11 12:41:55'
40	'1B31E8C6187'	'2190816'	'2013-09-11 12:41:59'

Table 15. (cont'd)

	Buyer ID	Product ID	Purchasing Time
41	('1B31E8C6187',	'2190816',	'2013-09-11 12:42:06')
42	('1B31E8C6187',	'2190817',	'2013-09-11 13:13:29')
43	('1B31E8C6187',	'2190815',	'2013-09-11 13:13:36')
44	('1B31E8C6187',	'2190815',	'2013-09-11 13:13:45')
45	('1B31E8C6187',	'2190817',	'2013-09-11 13:13:47')
46	('1B31E8C6187',	'2190617',	'2013-09-12 19:13:01')
47	('1B31E8C6187',	'2190617',	'2013-09-12 19:13:41')
48	('1B31E8C6187',	'2190641',	'2013-09-12 19:13:55')
49	('1B31E8C6187',	'2190641',	'2013-09-12 19:14:55')
50	('1B31E8C6187',	'2114007',	'2013-09-12 19:35:13')
51	('1B31E8C6187',	'2190641',	'2013-09-12 19:43:49')
52	('1B31E8C6187',	'2190641',	'2013-09-12 19:54:41')
53	('1B31E8C6187',	'2190641',	'2013-09-12 19:55:23')
54	('1B31E8C6187',	'2190641',	'2013-09-12 19:56:02')
55	('1B31E8C6187',	'2190641',	'2013-09-12 19:56:04')
56	('1B31E8C6187',	'2190641',	'2013-09-12 19:57:17')
57	('1B31E8C6187',	'2190641',	'2013-09-12 19:57:18')
58	('1B31E8C6187',	'2190858',	'2013-09-12 20:06:04')
59	('1B31E8C6187',	'2190858',	'2013-09-12 20:07:22')
60	('1B31E8C6187',	'2190857',	'2013-09-12 20:08:23')
61	('1B31E8C6187',	'2190482',	'2013-09-14 04:26:09')
62	('1B31E8C6187',	'2190482',	'2013-09-14 04:27:00')
63	('1B31E8C6187',	'2190482',	'2013-09-14 04:27:47')
64	('1B31E8C6187',	'2190482',	'2013-09-14 04:28:38')
65	('1B31E8C6187',	'2190482',	'2013-09-14 04:29:20')
66	('1B31E8C6187',	'2190482',	'2013-09-14 04:30:01')
67	('1B31E8C6187',	'2190482',	'2013-09-14 04:30:39')
68	('1B31E8C6187',	'2190482',	'2013-09-14 04:30:40')
69	('1B31E8C6187',	'2190482',	'2013-09-14 04:32:08')
70	('1B31E8C6187',	'2190482',	'2013-09-14 04:32:10')
71	('1B31E8C6187',	'2190482',	'2013-09-14 04:33:17')
72	('1B31E8C6187',	'2190482',	'2013-09-14 04:33:19')
73	('1B31E8C6187',	'2190482',	'2013-09-14 04:34:40')
74	('1B31E8C6187',	'2190482',	'2013-09-14 04:34:42')
75	('1B31E8C6187',	'2190482',	'2013-09-14 04:35:54')
76	('1B31E8C6187',	'2190482',	'2013-09-14 04:36:34')
77	('1B31E8C6187',	'2190482',	'2013-09-14 04:37:11')
78	('1B31E8C6187',	'2190482',	'2013-09-14 04:37:48')
79	('1B31E8C6187',	'2190482',	'2013-09-14 04:38:32')
80	('1B31E8C6187',	'2190564',	'2013-09-15 10:14:36')
81	('1B31E8C6187',	'2191142',	'2013-09-15 10:48:13')
82	('1B31E8C6187',	'2190982',	'2013-09-15 10:50:57')
83	('1B31E8C6187',	'2190930',	'2013-09-15 10:51:00')
84	('1B31E8C6187',	'2190986',	'2013-09-15 10:51:05')
85	('1B31E8C6187',	'2190984',	'2013-09-15 10:51:08')
86	('1B31E8C6187',	'2190958',	'2013-09-15 10:51:24')
87	('1B31E8C6187',	'2190940',	'2013-09-15 10:52:30')
88	('1B31E8C6187',	'2190930',	'2013-09-15 10:52:34')
89	('1B31E8C6187',	'2190958',	'2013-09-15 10:53:20')

Table 15. (cont'd)

	Buyer ID	Product ID.	Purchasing Time
90	('1B31E8C6187',	'2191233',	'2013-09-15 10:53:38')
91	('1B31E8C6187',	'2190982',	'2013-09-15 10:54:28')
92	('1B31E8C6187',	'2190545',	'2013-09-17 12:26:04')
93	('1B31E8C6187',	'2190083',	'2013-09-17 20:28:01')
94	('1B31E8C6187',	'2190082',	'2013-09-17 20:28:09')
95	('1B31E8C6187',	'2190117',	'2013-09-20 09:26:17')
96	('1B31E8C6187',	'2190360',	'2013-09-21 22:19:12')
97	('1B31E8C6187',	'2190524',	'2013-09-21 22:27:11')
98	('1B31E8C6187',	'2172385',	'2013-09-26 13:20:16')
99	('1B31E8C6187',	'2172385',	'2013-09-26 13:21:41')
100	('1B31E8C6187',	'2172386',	'2013-09-26 13:22:58')
101	('1B31E8C6187',	'2172385',	'2013-09-26 13:23:12')
102	('1B31E8C6187',	'2190781',	'2013-09-30 23:38:22')
103	('1B31E8C6187',	'2190781',	'2013-09-30 23:39:23')
104	('1B31E8C6187',	'2191094',	'2013-09-30 23:42:42')
105	('1B31E8C6187',	'2191094',	'2013-09-30 23:43:18')
106	('1B31E8C6187',	'2191094',	'2013-09-30 23:43:39')
107	('1B31E8C6187',	'2191094',	'2013-09-30 23:44:00')
108	('1B31E8C6187',	'2191094',	'2013-09-30 23:44:15')
109	('1B31E8C6187',	'2191094',	'2013-09-30 23:44:27')

The purchasing section frontier identified by the algorithm (highlighted above) coincide with the starting time of new days' play sessions. This is evidence that the algorithm correctly cut up the sections of purchasing. Note that the algorithm (cf. the pseudo code in Appendix 7) did nothing more than imposing the linear threshold model on timestamped data through mathematical definitions. No assumption about the user's play habits was made. This high congruency between the algorithm-identified purchasing sections and the natural sessions of purchasing shows that the model-derived definition of purchasing sections is a valid one and the dataset is fit for testing the model.

One may ask: "then why not just use each day as a purchasing section?". This is not suggested because, first, it is not clear why each day should be used as a purchasing section theoretically speaking. Second, the model-derived definition of purchasing section does not coincide with days perfectly. For example, the 19th purchase and the 20th purchase of the player above happened in two separate days and they were identified to be in the same purchasing section. For some other players, one could have multiple purchasing sections in one day (see example below). So, using date to separate purchasing sections is not a perfect substitute. Third, the model-derived definition of purchasing section is what directly follows from the linear threshold model and fit better the purpose of testing and validating the model.

For player '1A51C1F6173'
The model-derived algorithm identified purchasing section:
[0, 3, 7, 12, 17, 19, 33, 39, 55, 59, 66, 71, 87]
The purchasing data for this player:

Table 16. Data for the player

	Buyer ID	Product ID	Purchasing Time
0	('1A51C1F6173',	'2190185',	'2013-04-20 13:55:14')
1	('1A51C1F6173',	'2191122',	'2013-04-20 14:01:32')
2	('1A51C1F6173',	'2190445',	'2013-04-20 14:58:11')
3	('1A51C1F6173',	'2113013',	'2013-04-20 19:41:33')

Table 16. (cont'd)

	Buyer ID	Product ID	Purchasing Time
4	'1A51C1F6173'	'2191362'	'2013-04-20 19:43:06'
5	'1A51C1F6173'	'2190010'	'2013-04-20 20:41:34'
6	'1A51C1F6173'	'2112010'	'2013-04-20 20:45:13'
7	'1A51C1F6173'	'2108631'	'2013-04-21 10:50:19'
8	'1A51C1F6173'	'2108631'	'2013-04-21 10:50:25'
9	'1A51C1F6173'	'2190185'	'2013-04-21 11:06:47'
10	'1A51C1F6173'	'2190117'	'2013-04-21 13:58:14'
11	'1A51C1F6173'	'2108625'	'2013-04-21 14:40:02'
12	'1A51C1F6173'	'2108622'	'2013-04-21 15:35:43'
13	'1A51C1F6173'	'2108622'	'2013-04-21 15:35:47'
14	'1A51C1F6173'	'2190117'	'2013-04-21 17:17:38'
15	'1A51C1F6173'	'2108631'	'2013-04-21 18:33:42'
16	'1A51C1F6173'	'2113013'	'2013-04-21 18:40:01'
17	'1A51C1F6173'	'2108625'	'2013-04-21 21:37:09'
18	'1A51C1F6173'	'2108622'	'2013-04-21 21:37:15'
19	'1A51C1F6173'	'2190511'	'2013-04-21 22:24:25'
20	'1A51C1F6173'	'2190454'	'2013-04-22 06:32:17'
21	'1A51C1F6173'	'2190445'	'2013-04-22 06:32:20'
22	'1A51C1F6173'	'2108631'	'2013-04-22 07:06:03'
23	'1A51C1F6173'	'2108628'	'2013-04-22 07:15:25'
24	'1A51C1F6173'	'2108622'	'2013-04-22 07:50:53'
25	'1A51C1F6173'	'2190185'	'2013-04-22 17:34:00'
26	'1A51C1F6173'	'2108625'	'2013-04-22 20:49:08'
27	'1A51C1F6173'	'2108622'	'2013-04-22 21:12:02'
28	'1A51C1F6173'	'2190002'	'2013-04-22 21:12:20'
29	'1A51C1F6173'	'2112010'	'2013-04-22 22:28:09'
30	'1A51C1F6173'	'2190445'	'2013-04-23 07:48:54'
31	'1A51C1F6173'	'2190454'	'2013-04-23 07:48:55'
32	'1A51C1F6173'	'2108631'	'2013-04-23 18:10:26'
33	'1A51C1F6173'	'2108628'	'2013-04-23 20:11:31'
34	'1A51C1F6173'	'2108625'	'2013-04-23 22:09:29'
35	'1A51C1F6173'	'2190185'	'2013-04-24 09:04:37'
36	'1A51C1F6173'	'2108631'	'2013-04-24 09:56:13'
37	'1A51C1F6173'	'2190025'	'2013-04-24 10:19:09'
38	'1A51C1F6173'	'2190531'	'2013-04-24 17:35:12'
39	'1A51C1F6173'	'2190445'	'2013-04-25 10:21:07'
40	'1A51C1F6173'	'2190454'	'2013-04-25 10:21:09'
41	'1A51C1F6173'	'2108628'	'2013-04-25 10:26:58'
42	'1A51C1F6173'	'2190147'	'2013-04-25 12:50:57'
43	'1A51C1F6173'	'2190185'	'2013-04-25 13:06:36'
44	'1A51C1F6173'	'2190117'	'2013-04-25 14:25:18'
45	'1A51C1F6173'	'2190454'	'2013-04-25 20:02:00'
46	'1A51C1F6173'	'2190454'	'2013-04-25 20:03:13'
47	'1A51C1F6173'	'2108631'	'2013-04-25 21:03:23'
48	'1A51C1F6173'	'2108625'	'2013-04-26 06:36:53'
49	'1A51C1F6173'	'2190445'	'2013-04-26 07:02:31'
50	'1A51C1F6173'	'2190454'	'2013-04-26 07:02:33'
51	'1A51C1F6173'	'2190454'	'2013-04-26 07:04:36'
52	'1A51C1F6173'	'2190454'	'2013-04-26 07:04:38'

Table 16. (cont'd)

	Buyer ID	Product ID	Purchasing Time
53	('1A51C1F6173',	'2108622',	'2013-04-26 07:27:11')
54	('1A51C1F6173',	'2190185',	'2013-04-26 14:25:03')
55	('1A51C1F6173',	'2190883',	'2013-04-26 18:05:15')
56	('1A51C1F6173',	'2190531',	'2013-04-26 18:35:58')
57	('1A51C1F6173',	'2108625',	'2013-04-26 21:49:55')
58	('1A51C1F6173',	'2108631',	'2013-04-26 21:50:21')
59	('1A51C1F6173',	'2108625',	'2013-04-26 22:35:11')
60	('1A51C1F6173',	'2108622',	'2013-04-27 07:13:36')
61	('1A51C1F6173',	'2108631',	'2013-04-27 11:07:23')
62	('1A51C1F6173',	'2190445',	'2013-04-27 11:10:49')
63	('1A51C1F6173',	'2190454',	'2013-04-27 11:10:51')
64	('1A51C1F6173',	'2190454',	'2013-04-27 13:09:10')
65	('1A51C1F6173',	'2190185',	'2013-04-27 14:39:53')
66	('1A51C1F6173',	'2190881',	'2013-04-27 17:35:52')
67	('1A51C1F6173',	'2190868',	'2013-04-27 17:37:54')
68	('1A51C1F6173',	'2190871',	'2013-04-27 17:41:27')
69	('1A51C1F6173',	'2108625',	'2013-04-27 17:48:51')
70	('1A51C1F6173',	'2108631',	'2013-04-27 18:49:24')
71	('1A51C1F6173',	'2190147',	'2013-04-27 20:27:38')
72	('1A51C1F6173',	'2190011',	'2013-04-27 20:52:16')
73	('1A51C1F6173',	'2190011',	'2013-04-27 20:52:30')
74	('1A51C1F6173',	'2190011',	'2013-04-27 20:52:45')
75	('1A51C1F6173',	'2108622',	'2013-04-27 21:41:27')
76	('1A51C1F6173',	'2190454',	'2013-04-28 11:02:23')
77	('1A51C1F6173',	'2190531',	'2013-04-28 11:25:06')
78	('1A51C1F6173',	'2190454',	'2013-04-28 13:04:25')
79	('1A51C1F6173',	'2190454',	'2013-04-28 13:25:45')
80	('1A51C1F6173',	'2190185',	'2013-04-28 13:58:32')
81	('1A51C1F6173',	'2190117',	'2013-04-28 14:00:52')
82	('1A51C1F6173',	'2190147',	'2013-04-28 17:33:15')
83	('1A51C1F6173',	'2190531',	'2013-04-28 18:15:37')
84	('1A51C1F6173',	'2108625',	'2013-04-28 18:24:43')
85	('1A51C1F6173',	'2190531',	'2013-04-28 18:40:00')
86	('1A51C1F6173',	'2108631',	'2013-04-28 22:19:48')
87	('1A51C1F6173',	'2190531',	'2013-04-29 08:44:42')
88	('1A51C1F6173',	'2190531',	'2013-04-29 09:15:32')
89	('1A51C1F6173',	'2190454',	'2013-04-29 09:24:53')
90	('1A51C1F6173',	'2190531',	'2013-04-29 22:12:34')
91	('1A51C1F6173',	'2190454',	'2013-04-30 07:00:53')
92	('1A51C1F6173',	'2190454',	'2013-04-30 08:52:12')
93	('1A51C1F6173',	'2108625',	'2013-04-30 18:20:32')
94	('1A51C1F6173',	'2190531',	'2013-04-30 18:31:07')
95	('1A51C1F6173',	'2108631',	'2013-04-30 22:38:46')

Algorithm 1: Pseudo Code for Threshold Distribution Estimation

```
# Using information in the raw dataset
Create a list of buyer IDs,
Create a list of player (communicator) IDs
Create a list of product IDs
    Create a list of ValidBuyer IDs, which are the buyers who are also communicators
Create a list of all communications, with each communication in the form (sender, receiver, time),
    Create a list of BuyerCommunications, (sender, receiver, time), in which sender and receiver are
    both buyers
Create a list of all purchases, with each purchase in the form of (buyer, product, time)
# Using the lists created above
# find the frontiers for the purchasing sections of each buyer
For j in buyers:
    While frontier is earlier than the last time of the player's purchase:
        Record frontier and find next frontier by
            Finding the next purchase timestamp that is not communicatively adjacent with
            the current frontier, which is achieved by considering:
                if (No buyers talked to j before timestamp):
                    Then (the timestamp is communicatively adjacent)
                Else if (someone talked to j before timestamp):
                    If (those who talked to j weren't buyers before timestamp)
                        Then (timestamp is Communicatively adjacent)
                    Else (timestamp is not communicatively adjacent)
Creates a dictionary of frontiers of each buyer's purchasing sections
# Find the cumulated social influence for each person before each purchasing section
For j in valid buyers:
    Find the number of influencers, k
    Find all frontiers of the buyer,
    Sort the purchases by time
    If there is only 1 purchasing section:
        Find the timestamp of the frontier of the purchasing section, t
        Find the number of influencers so far till this time, n_t
        Return n_t/k as the cumulated social influence
    If there is more than 1 purchasing section:
        Find the timestamps for the frontier of each purchasing section, t_i
        Find the number of influencers after the (i-1)-th frontier before the i-th purchasing section
        frontier, n_ti
        Return n_ti/k as the cumulated social influence before the i-th purchasing section
# Based on the values of cumulated-social-influence each person has before each purchasing section,
estimate the threshold for this person in three different ways:
    1. Use the maximum value of a person's cumulated social influence as its threshold
    2. Use the average value of a person's cumulated social influence as its threshold
    3. Use the minimum value of a person's cumulated social influence as its threshold
# Based on the threshold values for the sample of people, plot distribution of thresholds
```

REFERENCES

REFERENCES

Acemoglu, D., Ozdaglar, A., & Yildiz, E. (2011, December). Diffusion of innovations in social networks. In *2011 50th IEEE Conference on Decision and Control and European Control Conference* (pp. 2329-2334). IEEE.

Braun, N. 1995. Individual thresholds and social diffusion. *Rationality and Society* 7:167–82.

Dodds, P. S., & Payne, J. L. (2009). Analysis of a threshold model of social contagion on degree-correlated networks. *Physical Review E*, *79*(6), 066115.

Dodds, P. S., & Watts, D. J. (2005). A generalized model of social and biological contagion. *Journal of Theoretical Biology*, *232*(4), 587-604.

Eppes, M. (2019, August 21). *Maximum Likelihood Estimation Explained - Normal Distribution*. Retrieved from https://towardsdatascience.com/maximum-likelihood-estimation-explained-normal-distribution-6207b322e47f

Centola, D., Eguíluz, V. M., & Macy, M. W. (2007). Cascade dynamics of complex propagation. *Physica A: Statistical Mechanics and its Applications*, *374*(1), 449-456.

Centola, D., & Macy, M. (2007). Complex contagions and the weakness of long ties. *American Journal of Sociology*, *113*(3), 702-734.

Chen, W., Lakshmanan, L. V., & Castillo, C. (2013). Information and influence propagation in social networks. *Synthesis Lectures on Data Management*, *5*(4), 1-177.

Dodoiu, G., Leenders, R. T., & van Dijk, H. (2016). A meta-analysis of whether groups make more risky or more cautious decisions than individuals. In *Academy of Management Proceedings* (Vol. 2016, No. 1, p. 16461). Briarcliff Manor, NY 10510: Academy of Management.

Gardner, M, and Steinberg, L. (2005). Peer influence on risk taking, risk preference, and risky decision making in adolescence and adulthood: An experimental study. *Developmental Psychology* 41:625–35.

Gleeson, J. P. (2008). Cascades on correlated and modular random networks. *Physical Review E*, *77*(4), 046117.

Goldenberg, J., Libai, B., & Muller, E. (2001). Talk of the network: A complex systems look at the underlying process of word-of-mouth. *Market. Lett. 3*, 12, 211–223.

Goldenberg, J., Libai, B., & Muller, E. (2010). The chilling effects of network externalities. *International Journal of Research in Marketing*, *27*(1), 4-15.

Granovetter, M. (1978). Threshold models of collective behavior. *American Journal of Sociology*, *83*(6), 1420-1443.

Gruhl, D., Guha, R., Liben-Nowell, D., & Tomkins, A. (2004, May). Information diffusion through blogspace. In *Proceedings of the 13th International Conference on World Wide Web* (pp. 491-501).

Kempe, D., Kleinberg, J., & Tardos, É. (2003, August). Maximizing the spread of influence through a social network. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 137-146).

Latané, B. (1996). Dynamic social impact: The creation of culture by communication. *Journal of Communication*, *46*(4), 13-25.

Latané, B., & Nida, S. (1981). Ten years of research on group size and helping. *Psychological Bulletin, 89*(2), 308.

Leskovec, J., Adamic, L. A., & Huberman, B. A. (2007). The dynamics of viral marketing. *ACM Transactions on the Web (TWEB), 1*(1), 5-es.

Libai, B., Bolton, R., Bügel, M. S., De Ruyter, K., Götz, O., Risselada, H., & Stephen, A. T. (2010). Customer-to-customer interactions: Broadening the scope of word of mouth research. *Journal of Service Research, 13*(3), 267-282.

Matsueda, R, L. (2006). Differential social organization, collective action, and crime. *Crime, Law, and Social Change* 46:3–33.

McGloin, J. M., & Rowan, Z. R. (2015). A threshold model of collective crime. *Criminology, 53*(3), 484-512

Miller, J. H., & Page, S. E. (2009). *Complex adaptive systems: An introduction to computational models of social life*. Princeton university press.

Nematzadeh, A., Ferrara, E., Flammini, A., & Ahn, Y. Y. (2014). Optimal network modularity for information diffusion. *Physical Review Letters, 113*(8), 088701.

Nowak, A., Szamrej, J., & Latané, B. (1990). From private attitude to public opinion: A dynamic theory of social impact. *Psychological Review, 97*(3), 362.

Payne, J. L., Dodds, P. S., & Eppstein, M. J. (2009). Information cascades on degree-correlated random networks. *Physical Review E*, *80*(2), 026125.

Peres, R., Muller, E., & Mahajan, V. (2010). Innovation diffusion and new product growth models: A critical review and research directions. *International Journal of Research in Marketing, 27*(2), 91-106.

Amit, S. & Cosley, D. (2016). Distinguishing between Personal Preferences and Social Influence in Online Activity Feeds." *Proc. CSCW*

Thomas, K. J., & Marie McGloin, J. (2013). A dual-systems approach for understanding differential susceptibility to processes of peer influence. *Criminology, 51*(2), 435-474.

Ugander, J., Backstrom, L., Marlow, C., & Kleinberg, J. (2012). Structural diversity in social contagion. *Proceedings of the National Academy of Sciences*, *109*(16), 5962-5966.

Valente, T. W. (1995). Network models of the diffusion of innovations (No. 303.484 V3).

Valente, T. W. (1996). Social network thresholds in the diffusion of innovations. *Social Networks*, *18*(1), 69-89.

Valente, T. W. (2012). Network interventions. *Science*, *337*(6090), 49-53.

Valente, T. W., & Davis, R. L. (1999). Accelerating the diffusion of innovations using opinion leaders. *The Annals of the American Academy of Political and Social Science*, *566*(1), 55-67.

Watts, D. J. (2002). A simple model of global cascades on random networks. *Proceedings of the National Academy of Sciences*, *99*(9), 5766-5771.

Xu, X., Yang, X., Lu, J., Lan, J., Peng, T., Wu, Y., & Chen, W. (2017). Examining the effects of network externalities, density, and closure on in-game currency price in online games. *Internet Research*, *27*(4), 924-941.