THE IMPACT OF HETEROGENEITY ON THRESHOLD-LIMITED SOCIAL CONTAGION, AND ON CROWD DECISION-MAKING

By

Panagiotis Dimitrios Karampourniotis

A Dissertation Submitted to the Graduate

Faculty of Rensselaer Polytechnic Institute

in Partial Fulfillment of the

Requirements for the Degree of

DOCTOR OF PHILOSOPHY

Major Subject: PHYSICS

Examining Committee:

György Korniss, Dissertation Adviser

Boleslaw K. Szymanski, Dissertation Adviser

Vincent Meunier, Member

Toh-Ming Lu, Member

Chjan Lim, Member

Rensselaer Polytechnic Institute
Troy, New York

June 2017
(For Graduation August 2017)

© Copyright 2017
by
Panagiotis Dimitrios Karampourniotis
All Rights Reserved

CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGMENT

A successful completion of Ph.D. Dissertation requires a number of conditions to be satisfied the most important being the work of the Ph.D. Candidate, and the guidance and support of his/hers advisors. Prof. György Korniss and Prof. Boleslaw K. Szymanski are the ideal Ph.D. advisors and educators anyone could ask for. Kind, patient, straight to the point, and flexible with new ideas. I thank them deeply for their guidance, help, and support. Furthermore, I would like to thank my committee members Prof. Chjan Lim (who was also my academic advisor for my MS in Applied Mathematics), Prof. Vincent Meunier, and Prof. Toh-Ming Lu, as well as our collaborator Prof. Brian Uzzi from Northwestern University. Furthermore, I would like to thank the postdoctoral researchers Dr. Sameet Sreenivasan, Dr. Noemi Derzsy, and Dr. Pramesh Singh for their support. I also express my gratitude to my family members, friends, and colleagues for their constant encouragement. And above all, I would like to thank my mother for being so supportive in my absence.

This work was supported in part by the Army Research Office grant No. W911NF-12-1-0546, by the Army Research Laboratory under Cooperative Agreement No. W911NF-09-2-0053 (the ARL Network Science CTA), by the Office of Naval Research Grant Nos. N00014-09-1-0607 and N00014-2640, and by the Northwestern University Institute on Complex Systems (NICO) grant No. NU SP0033419. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies either expressed or implied by the Army Research Office, Army Research Laboratory, Office of Naval Research or the U.S. Government.

ABSTRACT

Recent global events and their poor predictability are often attributed to the complexity of the world event dynamics. A key factor generating the turbulence is human diversity. Here, we study the impact of *heterogeneity* of individuals on opinion formation and emergence of global biases. In the case of opinion formation, we focus on the heterogeneity of individuals' susceptibility to new ideas. In the case of global biases, we focus on the aggregated heterogeneity of individuals in a country.

First, to capture the complex nature of social influencing we use a simple but classic model of contagion spreading in complex social systems, namely the threshold model. We investigate numerically and analytically the transition in the behavior of threshold-limited cascades in the presence of multiple initiators as the distribution of thresholds is varied between the two extreme cases of identical thresholds and a uniform distribution. We show that individuals' heterogeneity of susceptibility governs the dynamics, resulting in different sizes of initiators needed for consensus.

Furthermore, given the impact of heterogeneity on the cascade dynamics, we investigate selection strategies for accelerating consensus. To this end, we introduce two new selection strategies for *Influence Maximization*. One of them focuses on finding the balance between targeting nodes which have high resistance to adoptions versus nodes positioned in central spots in networks. The second strategy focuses on the combination of nodes for reaching consensus, by targeting nodes which increase the group's influence. Our strategies outperform other existing strategies regardless of the susceptibility diversity and network degree assortativity.

Finally, we study the aggregated biases of humans in a global setting. The emergence of technology and globalization gives raise to the debate on whether the world moves towards becoming flat, a world where preferential attachment does not govern economic growth. By studying the data from a global lending platform we discover that geographical proximity and cultural affinity are highly negatively correlated with levels of *flatness* of the world. Furthermore, we investigate the

robustness of the flatness of the world against sudden catastrophic national events such as political disruptions, by removing countries (nodes) or connections (edges) between them.

CHAPTER 1
Introduction

Networks allow for the reduction in the representation of natural and artificial complex systems by transforming objects to nodes and the interaction between objects to edges between the nodes. This representation reveals and allows the study of various properties about the structure and interactive processes of a system. In this representation a node can carry any number of attributes and the edges can have all kinds of weights assigned to them, all depending on the system we aim to represent and its dynamic processes [1]. The study of networks, namely Network Science, has been ongoing over the past two hundred years and was mainly driven by mathematicians and sociologists. In the past few decades a significant number scientists from other disciplines, like physics, computer science, and biology have been involved. Nowadays, Network Science is remarkably multidisciplinary [2]. In fact, physics [3–6] lies at the heart of the field of Network Science. The physics approach has been different in three significant ways compared to other disciplines [3], as we will elaborate further down.

1.1 Networks

1.1.1 Artificial Networks

A network represents more than a group of nodes. It connects them together with a specific pattern. That pattern is in most cases not accidental. In the case of artificial networks, the functionality of a network is typically clear, and predefined. Artificial networks typically are designed to serve a specific global functionality, hence nodes are not independent actors choosing their connections as the network grows. For instance a group of people may serve no functionality on their own. Yet, standing next to each other in a line, connected as a chain, creates a bucket brigades. A bucket brigades is not an accidental design, since it allows for a faster transfer of water for firefighters, or remove of debris from a collapsed building for rescue teams. Another example of a artificial network is message passing in case of

an emergency. It used to be that when an unpredicted event occurred, like a sudden change of plans, the first few people who knew about it, would share the message by calling a small number of members of the group. Then, they would also pass the message and so on. In these cases the functionality of the networks was clear and simple. Yet there are other artificial networks that have much higher complexity of functionalities and structure. An example of a higher level complexity network is transportation networks. They are designed to transport any type of vehicles optimally, given some constraints. Those constraints are typically geographical conditions, cost of construction and safety. Other examples of artificial networks are power grid networks, computer networks, telecommunication networks etc.

1.1.2 Natural Networks

Natural networks are also easily recognizable and designed to serve some functionalities, yet those functionalities are not necessarily serving the interest of the system as a whole, but the interest of each node in particular. For instance, the specific pattern of networks formed from research paper citations is formed through the choices of citations for each research paper. In these cases, the growth function is defining the pattern. Biological networks [7] are at the core of natural ones. For example, cell metabolic networks cover all the metabolic processes that define the biochemical and physiological properties. Another example is ecological networks. They describe the biotic interactions in an ecosystem. Here, nodes represent the various species and edges between them represent interactions such as trophic or symbiotic. An interesting case of natural networks is our brain. It consists of nearly one hundred billion nerve cells (neurons), billions of nerve fibers (dendrites and axons); with neurons being connected by hundreds of trillions of synapses. It is estimated that a three year old child has in average around one quadrillion synapses. The size of the (human) brain's neural network is the largest network known to humans, and perhaps the most complex.

In this Dissertation, we will focus more on social networks, and mainly Online Social Networks (OSN) [9], which are the facilitators of a number of traceable social dynamics [8], topic that this work is mostly focused on. OSN are considered natural,

regardless of the fact that current technology is required for their access. The network structure that they have is not beforehand designed to serve a great goal, but arises from each person's preferences of online connections. The goal of the members (nodes), is to reach out to other connections, such as friends, professionals or random people, and extend their online fingerprint. That fingerprint can be used to keep their friendships warm with updates, by sharing moments sharing of the lives, or to extend their professional network and being up to date with the current research/professional/skills trends, and possible opportunities. In addition to the dynamics that take place between members of social networks (which we will elaborate further down), the popularity of OSN has given rise to a large number of external influences pursuing to benefit from the exposure that social dynamics can offer. Some of the external influences are viral marketing and advertising, multiple types of online services, politics, etc. And thus, the interest of studying them, goes beyond the purposes of understanding the human condition. It expands on studying their dynamics in order to target members of OSN for advertisements of products [10], services such as Healthcare [11–14], political influence [15,16], ideology spread [17, 18], as well as militaristic operations [19, 20] among others.

1.1.3 Emerging Properties of Networks

A number of mathematical models have been developed to generate synthetic networks, which are used to understand the structure of numerous artificial and natural networks. The earliest model used is the Erdős-Rényi(ER) graphs [21]. On ER graphs, newly introduced nodes have the same probability of connecting with any other node. The resulting degree distribution is binomial, that is

$$P(k) = \binom{N-1}{k} p^k (1-p)^{N-1-k} \tag{1.1}$$

for a finite system size N, and

$$P(k) = \frac{(Np)^k e^{-Np}}{k!} \tag{1.2}$$

for assymptotically large system size N, where the degree, system size and link probability are given by k, N, and p respectively. In this model, $\langle k \rangle = Np$ is held constant, where $\langle k \rangle$ the average degree. ER graphs do not capture the complexity of most networks. However, because of their simplicity, they are very useful for studying dynamical processes that run through them, since complex networks could unpredictably impact the resulting dynamics. Furthermore, ER graphs (among a few other models) are nicely expressed analytically, hence being used thoroughly in analytical models on dynamics allowing for the study tipping points [22–24], continuous transitions, on non-monotonic behaviors on networks.

Creating analytical models that can capture the properties of networks, and especially the complexity of natural networks is a challenge, while ER graphs are far from representing complex networks. To tackle this, physicists, instead of focusing on random graph models, got inspired by the structure of empirical networks [25,26]. In addition, they focused on the statistical properties of empirical networks. A very useful measure, which was neglected before, is the degree distribution of a network [27–31]. The degree of a node is the number of connections it has. For instance, for OSN it is the number of friends a person has. It has been shown [32] that the degree distribution of the Internet follows a power-law trend, $P(k) \propto k^{-a}$, with degree exponent, $a = 2.1$. Several other natural networks follow a power law degree distribution [27, 33, 34]. This does not seem to be an accidental property, since there are specific growth functions that give rise to their structure [29]. Physicists supported that the power law degree distribution can be explained by two properties. The first one is a scale-free state of organization of natural networks, meaning that the power law distribution does not depend on the system size. The second property, preferential attachment, describes a specific behavior of the network for newly introduced nodes. It states that new nodes will connect with higher probability of connecting with nodes which are already more connected (higher degree nodes). Based on those two properties, they could now construct analytically solvable synthetic networks, named Scale-Free networks (SF).

Social networks carry another property that was revealed by the seminal experiment conducted by Milgram [35–37] back in the 1960', which revealed that the

average number of acquaintances separating any individual on the planet is small, regardless of the geographical distance, race, and other cultural differences. The phenomenon is known as 'small-world'. Interestingly, it also appears in other natural and technological networks, such as the WWW [38], economic networks [39], brain networks [40] and transportation networks [41, 42]. Essentially, small-world networks combine local clustering with random long range ties, giving them their small-world property. A measure of the efficiency of information exchange in networks shows how small-world networks are efficient both locally and globally [43]. Watts and Strogatz [27] introduced a method for synthesizing small-world networks by starting from a regular network and randomly rewiring a small number of edges. By controlling the probability p of rewiring between any two edges, a regular graph with $p = 0$ could be transformed to a random graph with $p = 1$. Then, the average distance L between any two nodes becomes a function of p. The average distance between any nodes for small world network scales as $L \propto \log N$ [44], where N is the system size.

The final property of networks that we will address here is their resilience against random failures and targeted attacks. In case of an error or attack on the node, that node is disconnected from the network. As a network loses its nodes, it becomes more and more disconnected. Thus, to quantify the impact of node disconnections in a network, physicists measure the size of the remaining giant component, that is the largest connected (functional) component of the network [45]. It is observed that many interconnected natural networks are much more resilient to failure than artificial networks. Such a case is brain networks, which are connected in a topology that maximizes stability, making them extremely robust [46, 47]. This finding is important, since it evidences the importance of resilience that living organisms have developed through millions of years of evolution not just under exogenous influences but also under endogenous failures. For artificial networks, examples of random failures are natural disasters on power grid networks, server failures and malfunctions in computer networks etc., while targeted attacks could possibly simulate of terrorist or military actions. Physicists [45] showed that it is scale-free networks that are most resilient to random failures and errors. However, targeted

attacks, such as aiming for the nodes with highest degree, on those systems can greatly disrupt their connectivity, and essentially destroy them, due to the very large number of connections a few nodes (hubs) have in SF networks. The understanding of the resilience of natural, i.e. mass extinctions in ecological networks, and artificial systems, i.e. power grid cascading failures, to errors and attacks is of huge importance. Recently, to capture the unpredictability of resilience (and its implications) of complex systems, a new analytical framework was proposed [48] which was taking into account the systems dynamics and topology to reduce the multi-dimensional parameters the to one parameter. The authors' framework "unveils the network characteristics that can enhance or diminish resilience, offering ways to prevent the collapse of ecological, biological or economic systems, and guiding the design of technological systems resilient to both internal failures and environmental changes." In Chapter 4 we will address the impact on random failures and perturbations, and targeted attacks on the nodes and edges of an empirical global network.

1.2 Cascades in Social Systems

Networks exist to facilitate dynamical interactions between nodes. Those interactions typically lead to a change of the different states of those nodes. Cascades are caused from the propagation of those changes in a network. On their purest form they are self-amplifying processes that depend solely on the type of interaction between the nodes and the network structure, just like domino effects do. Here, we will focus on cascades, and in particular cascades in social systems, a defining characteristic of state change from one node to another in networks, yet not strictly defined. A small perturbation, e.g. the change of the state of a node, can lead to unpredictably large cascades, with catastrophic or beneficial impact depending on the processes and its implications. For instance, a perturbation as little as a malfunction of a power line in power grid networks, can build up to a vast size power shut down, such as the 2003 blackout in Italy [49]. Similarly, a build up of traffic congestions in road systems can start even by small speed variations of the going vehicles; or in the case of financial networks, a financial institution may

cause the failure of another, as we observed in 2008. In other cases, networks are designed to facilitate cascades, such as neural brain networks, designed to allow massive triggering of neurons through voltage collection, or online social networks and communities, where opinions and marketing become viral. Moreover, the cascading process of a state of an attribute may trigger changes of states of other attributes of nodes participating in interconnected systems, which make it even harder to predict their outcome [50–52]. There are three important properties defining cascades: non-locality, non-additivity, and disproportional impact [53]. Non-locality refers to a node changing its state by a non-neighboring node, without the neighboring nodes being required to change state. Most opinion diffusion models do not carry that property, since the nodes adopt a new opinion through interaction with their direct (first) neighbors. Epidemic models of disease spread naturally have a locality component, while cascading failures hold the non-locality feature. That property typically is responsible for the high unpredictability of nodes that will be affected by the change in the state of another node in the network. Non-additivity refers to the property of a spread mechanism where the change of the state of a node is not enough to cause a change of the state of a neighboring node. Threshold Models of spread of influence and cascading failures in power grid networks typically follow this property. However, epidemic disease cascades do not, since the probability of a node getting infected when a neighboring node does, is always non-zero (unless vaccinated). Finally, disproportional impact refers to the effect the change of state of a node will have in the system, which is based not on the network structure properties of the spreader node, but on the intrinsic characteristics of that node. For example, typically in epidemiology models the importance of all nodes is considered the same in the system, and the models are based on the assumption that in average each person has the same spreading probability. Contrary to that, the catastrophic impact of failure of power grid nodes depends on their capacity.

1.3 Modeling Cascades

Even when the mechanism of the cascading spread is well understood and quantified, due to the aforedescribed properties, modeling the cascades can be chal-

lenging. Yet, the benefits on understanding them, predicting their location and size, and finally controlling them are tremendous. Let us mention some of those types of cascades.

1.3.1 Cascading Failures

Cascading Failures [54] describe the overload of nodes and the propagation of that overload to other nodes. They are typically appearing in systems with flow distribution, such as current redistribution leading to blackout transmission in power grid networks, as systemic risk in networks of financial institutions [55, 56], as traffic overload in computer networks, road traffic networks [57] etc. The failures can occur at the node [50] or edge level and they are non-local, non-additionary, and with disproportional impact [53]. Cascading Failures can have tremendous impact on our lives [49, 58–61]. Several researchers focus on developing methods to control such failures and to mitigate against them [62, 63].

1.3.2 Epidemic Disease Spreading

In epidemiology and network science [64], understanding, modeling and preventing disease spread at a macroscopic level is crucial for the global [65–69] and local [70] public safety. A lot of scientific research has been done to cover various scenarios of disease spread, the basic idea behind them being that a node (unless vaccinated) can get infected with some non-zero probability from each of its neighbors. The spread mechanism of the models is typically probabilistic, local, and each node has the same 'strength' for spreading the disease, thus it does not obey the non-locality, non-additivity and disproportional impact properties. The first mathematical models were excluding the impact of the network structure in the spreading process and were assuming that all individuals (nodes) have the same probability of getting infected in time (full mixing assumption) [71]. For example, a basic disease spread model is the Susceptible-Infected-Recovered (SIR) model, where individuals can be in either of the three mentioned states. Macroscopically, the dynamics of the fraction of nodes on each state (s, i, r) can be described by a system of coupled

differential equations,

$$\frac{\mathrm{d}s}{\mathrm{d}t} = -\beta si \tag{1.3}$$

$$\frac{\mathrm{d}i}{\mathrm{d}t} = \beta si - \gamma i \tag{1.4}$$

$$\frac{\mathrm{d}r}{\mathrm{d}t} = \gamma i \tag{1.5}$$

where, β and γ are the contact-transmission parameter, and the recovery rates from infected to susceptible, respectively. Also, the fractional size of each state is normalized such that $s + i + r = 1$. There isn't an analytically trackable solution to this. For $\gamma = 0$ SIR is simplified to the Susceptible-Infected (SI) model, which has a logistic growth solution. However, the steady state is not hard to calculate, since then the rates are zero, and so

$$r = 1 - s_0 e^{\frac{\beta}{\gamma} r} \tag{1.6}$$

where, s_0 is the initial susceptible fraction ($s_0 \approx 1$ assuming a small disease spread fraction).

Another interesting variation is the Susceptible-Infected-Susceptible (SIS) model, were nodes can become susceptible again after being infected. A mean field theory on networks provides the epidemic threshold λ_c, above which the disease does not die out, as $\lambda_c = \langle k \rangle / \langle k^2 \rangle$ [72]. The case of λ_c for SF networks has been of a great academic interest, due to the particularity of the structure of SF networks; the degree second moment of SF networks diverges for $\gamma \leq 3$. Yet, Castellano and Pastor-Satorras [72] showed that λ_c is not impacted by the "SF nature of the network but stems instead from the largest hub in the system being active for any spreading rate $\lambda > 1/k_{max}$". There are multiple variations of the SIS model, covering all kinds of scenarios from the vaccination/immunization of the nodes, to the impact of the disease duration and more [64]. Finally, identifying the critical nodes that can potentially block the disease spread has been of a huge importance, has been studied in simulated networks [73, 74] and in 'real life' [66, 67]. Interestingly, variations of the above disease spread models can also be used to understand the diffusion of

information in OSN [75].

1.4 Opinion Diffusion Models

1.4.1 Fundamental Social Drivers

Social dynamics is a multidisciplinary field aiming to qualify and quantify the complex human interactions. Research from social psychology identifies multiple social drivers, some of the most fundamental being social influence, structural balance, reciprocity, and homophily. Those qualitative drivers are being studied using controlled experiments, data analysis and quantitative models, as we will describe below. Interestingly enough, there is a number of data analysis research papers clearly distinguishing the different social drivers in social networks [76, 77].

The concept of Structural Balance (or Social Balance) was first introduced by Heider [78], a social psychologist, and later it was introduced in graphs by Cartwright and F. Harary [79] in the 1950s. It focuses on negative, positive, and perhaps neutral relationships [80]; those states are assigned on the edges of the nodes in a network. Structural Balance aims to describe the dynamics arising from four cases 'the enemy (friend) of my enemy (friend) is my friend' and 'the enemy (friend) of my friend (enemy) is my enemy'. In order to test the balance, triangles of connected nodes (closed triangles) are selected in random and tested to update the states of the edges. A triangle is considered balanced when no change in the states of the edges occurs, e.g. all three nodes are friends (enemies); otherwise it is considered unbalanced and a random change towards balance will occur [81,82]. In the case of a fully connected graph, mean-field based analytics can capture accurately the dynamics [82], yet not as good for sparse graphs [80]. A recent data/model analysis [83] indicates that 'online social networks are in general very poorly balanced'.

Reciprocity refers to the interaction between two nodes in a 'give and take' way. At a first glance, the concept of reciprocity can refer to the mutual connection between any two nodes, and of course it extends to mutual interactions between them [84]. Research has focused on the characterization of a network on its link reciprocity (referring to the mutual links of nodes) whether it is unweighted [85] or even more accurately, weighted [86, 87]. A large data analysis research on the

reciprocity of mobile phone calls suggests that social interactions are not entirely reciprocal [88]. Furthermore, empirical evidence indicates that money transfers for charity actually have a strong direction to reciprocity [89].

Homophily means, as the word in Greek indicates, 'love of the same'. It refers to individuals' drive to connect and interact with others that they have or they feel that they share similarities with [90]. Those similarities can be physical, cultural, class etc. characteristics [91]. There is a large number of research papers [92] exposing this strong social drive in humans. People with great similarities are more likely to be in agreement with each other on a new trend and opinion.

1.5 Social Influence

Here, we will focus on social influence, which can be described as behavioral contagion, conformity, compliance, peer pressure, product adoption etc. There have been large scale studies that demostrate the impact of Social Infleunce [93]. The aim is to model and quantify human complex interactions that are related to social influencing. A first framework of various cases of choices with externalities was made by Schelling back in 1973 [94] with a game theory based analysis, however his analysis did not include the impact of networks. Since then, more rigorous, systematic, and analytical methods have been developed. Here, we will address the methods that assume a microscopic rule of interaction. The macroscopic effects that we are interested in studying emerge from the spread of those microscopic interactions [6]. Macroscopically, in these models, the order parameter is typically the fraction of the nodes that have been influenced. By controlling the model's initial conditions, and network parameters we can track their macroscopic effects either by simulations or analytics. The most interesting cases are those for which tipping points or non-monotonic behaviors are observed. There is a number of different microscopic rules of adoption. The most extensively studied being the Linear Threshold Model (LTM), Majority rule, the Voter Model, and the Naming Game (NG) [95].

1.5.1 Linear Threshold Model

In his seminal work in collective behavior, Granovetter [96] introduced the Linear Threshold Model (LTM) to capture the social pressure that leads individuals to follow their group and adopt their behavior, regardless if it's beneficial for the individual or not. As introduced, the model is deterministic and binary. Each node i has an intrinsic resistance to adoption of the new behavior/opinion/product, that resistance is given by its threshold ϕ_i. Node i adopts the new opinion only when the fraction of its neighbors possessing that opinion is higher than its intrinsic value, that is $\phi_i \leq$ (number of active neighbors)$/k$. Granovetter, by introducing a fractional threshold, aimed to capture the impact of the group on a person, since on a larger group a person would have smaller probability of being apprehended [97]. Interestingly, the activation threshold in the LTM share similarities with the integrate-and-fire neuron model [98–101], with the addition that the thresholds are not hard, and a return on initial inactive state is allowed.

We will be studying extensively the LTM in Chapters 2, and 3. In Chapter 2 we will focus more on the model basic parameters such as the impact of the threshold distribution and the impact of the system size in the cascade size. In Chapter 3 we will focus on the importance of selection strategies for maximizing the cascade size.

1.5.2 Majority Rule

Conformity in Social Influence can be driven by normative social influence. According to social psychology it is defined as "the influence of other people that leads us to conform in order to be liked and accepted by them" [102]. Thus conforming and siding with the majority can lead to the benefits of being accepted within a social group, not because they want to consult/follow others who may have taken the correct action/choice as informational social influence recommends. The impact of laying with the majority is such that as indicated by conformity experiments [103], people are willing to change their opinion on a binary topic from an obvious correct answer, when the majority of a group supports the blindly false opinion.

The majority rule has been used to capture the dynamics of individuals in social networks with a great success [104–106] given its simplicity. The model is not

necessarily binary, and nodes can switch back and forth their opinions according to the rule. Finally, it doesn't obey disproportional impact, since all nodes are treated equally, or non-additivity since the changes are proportional, nor non-locality, since all the updates are local by default and Markovian.

1.5.3 Voter Model

The concept of the voter model was introduced back in 1970's [107,108] and it is a stochastic opinion spread model [109]. A node i adopts a new opinion by randomly selecting the opinion of one of its first neighbors k at each step. Assuming two states in the system $\sigma_i = \pm 1$, the probability of state change for node i is $P\left(\sigma_i \to -\sigma_i\right) = \frac{1}{2}\left(1 - \frac{\sigma_i}{k_i}\sum_{j\in N_i}\sigma_j\right)$. "A standard order parameter to measure the ordering process in the voter model dynamics is the average interface density ρ, defined as the density of links connecting sites with different spin values" (from [111]):

$$\rho = \frac{\sum_{i=1}^{N}\sum_{j\in N_i}\frac{1-\sigma_i\sigma_j}{2}}{\sum_{i=1}^{N}k_i} \tag{1.7}$$

The first analytical analysis on it was done for two (or higher) dimensional lattices [107, 108], but more recently it was expanded to small world networks [110], and Scale Free Networks (SF) [111, 112], where the authors studied the consensus time T_N in relationship to the degree distribution exponent of the SF and the system size N. Later, T_N was also studied in relationship to the average degree of the graph [113], and an exact solution has been given for complete graphs, also extended for sparse graphs by spectral analysis [114]. Finally, the model has been expanded for a q arbitrary number of opinions [115]. Furthermore, the selection order of nodes with initial state $\sigma = \pm 1$, has an impact on the probability of reaching consensus $\left(\sum_{i=1}^{N}\sigma_i = \pm N\right)$ for all the nodes, especially in networks with heterogeneous degree distributions [112]. The voter model also does not obey disproportional impact, or non-additivity, nor non-locality and is Markovian.

1.5.4 Naming Game

The Naming Game (NG) model was introduced to study how global agreement raises with multiple starting options. In particular, the model is designed to capture the patterns emerging from starting with multiple words representing the same concept/object converges to a single word agreed upon by everyone [116] in the network [117]. According to the authors such a process eventually allows for a common language to emerge. The model follows the properties of locality, proportional impact and additivity and is Markovian. Each node starts with an empty vocabulary for a particular concept/object and invents its own word randomly. Thus, initially there are essentially $O(N)$ generated unique words, one for each node. A 'speaker' and a 'listener' are selected randomly in a pairwise fashion. The speaker sends randomly one of the synonyms in his vocabulary to the listener. If that word is not included in the vocabulary of the listener, then he adds it to the list of synonyms. If the word is included in the listener's vocabulary, then the speaker and listener drop all other words from their vocabulary and 'agree' using this one word. The authors show that there is a sharp transition point from disorder (multiple synonyms) to order (consensus).

The model has been studied for various network structures [6,118]. The common parameter studied is the time to reach consensus and number of words in the equilibrium. An interesting case raises when initially each node starts with only of only two synonyms [118,119]. This restriction allows to study the dynamics between two competing synonyms, which can also represent competing opinions, A and B. Thus each node can be in either of three states, A, B, and their intermediate AB. When the speaker communicates the opinion (assume A) the listener does not possess (B), the listener will move to the intermediate state AB. If both possess that state (A), then they will agree on using it, and drop the state AB if they happened to possess it. The binary NG shares many similarities with the binary voter model, yet with the addition of the intermediate state. This addition changes the dynamics and the time to reach consensus [6,120]. Recently, it has been shown that networks with strong community structure are less likely to reach global agreement [118], which suggests that diversity (lack of global agreement) is re-enforced by communi-

ties. Another parameter that controls global consensus is the number of the initial synonyms [121], were it was shown that a system higher number of synonyms has a higher consensus time T_N. Finally, the most interesting results rise from the existance of *tipping points* in the NG. The presence of a large enough size ($p_c \approx 10\%$) of committed agents (agents who do never switch their opinion) can force the system to a global consensus [122]. This result is evident in several historical events, such as the 'rise of the African civil-right shortly after the size of the African-American population crossed the 10% mark".

CHAPTER 2
The Impact of Heterogeneous Thresholds on Social Contagion with Multiple Initiators

2.1 Introduction

The technological breakthroughs of the 21st century have strongly contributed to the emergence of network science, a multidisciplinary science with applications in many scientific fields and technologies. As mentioned in Chapter 1, several sociological opinion diffusion models first introduced in the middle of 20th century are now being thoroughly studied, while variations of these classical models have been introduced. Most of these models are based on social reinforcement, where simple rules based on the interaction of individuals with their respective nearest neighbors govern individual opinion evolution. The macroscopic outcome of these rules is a cascade of nodes switching opinions [94, 96, 124–128].

We focus our study on one of the classic models of social influencing, the Linear Threshold Model (LTM). The LTM is a binary opinion spread model first introduced by Granovetter [96] to model collective behavior socially driven by peer pressure. Under the LTM a node adopts a new opinion only when the fraction of its nearest neighbors possessing that opinion is larger than an assigned threshold, which represents the resistance of the node to peer pressure. The threshold of each node is considered an intrinsic value. Thus, it is safe to assume that it is not a unique for all the nodes, but rather heterogeneous.

Although the microscopic rule of opinion adoption in the LTM is simple, the collective behavior that arises is complex and non-linear. The resulting spread size depends on a large set of parameters, such as the network structure (e.g., clustering) [128–132], the size of the initially active nodes (initiators), the selection strategy

Portions of this chapter previously appeared as: P. Karampourniotis, S. Sreenivasan, B. K. Szymanski, and G. Korniss, "The impact of heterogeneous thresholds on social contagion with multiple initiators," PLOS ONE **3,** 2330 (2015).

of the initiators, and the distribution of threshold values among nodes of the network. In fact, Watts and Dodds [126] showed through simulations on various types of diffusion mechanisms that the cascade size is governed not by superspreaders, but by a small critical size of nodes with low resistance to influence. Hence, the importance of thresholds is critical for the dynamics.

The first thorough investigation of the LTM was made by Watts [125], who examined the effect of one randomly selected initiator on the cascade size. Gleeson and Cahalane [133–135], on the other hand, determined analytically the cascade size for varying initiator sizes (or fractions) for the infinite system size. Recent investigations of the LTM by Karimi and Holme [137] and Michalski et al. [138] also considered the impact of temporal networks on contagion cascades. Recently, Ruan et al. [139] studied the effects of "immune" individuals (those who resist adopting the new idea indefinitely) and external influencing (e.g., by mass media or advertisements) in the LTM. Furthermore, an extension of the LTM includes a persuasion mechanism, where the combination of adoption and persuasion giving rise to new dynamics [140, 141].

Watts [125], proposed the first analytic solution for the LTM, using percolation theory and generating functions to measure the size of the largest cluster of nodes requiring only one active neighbor to turn active (largest vulnerable cluster). The model applies to unweighted, undirected graphs with small clustering coefficient. In the infinite system size, when the vulnerable cluster percolates, there is a non-zero probability that a cascade will take over a large portion of the network (global cascade). A randomly selected initiator will activate the largest vulnerable cluster, if it is a part of the cluster or is one of its neighbors. Using this analytic method, Watts studied the regime for which global cascades are possible for one initiator, for different values of identical thresholds ϕ_0 and average degree $\langle k \rangle$ of synthetic graphs. He found that, for ER graphs with $\mathcal{O}(1)$ initiator the criterion for global cascades is $\langle k \rangle < 1/\phi_0$.

Gleeson and Cahalane [133] formulated an analytic approach for the LTM with varying initiator sizes. Their work was inspired by the zero-temperature Random-Field Ising Model (RFIM) [142,143], where the cascade size, the initiator size and the

threshold distribution correspond to the magnetization, the external uniform field and the local quenched random fields of the RFIM. The main difference between the two models is that in the LTM the activated nodes remain activated, while in the RFIM the spins may flip back to an inactive state. The analytic approach to the LTM model is applicable to locally tree-like structures [133], such as ER graphs. The graph is considered an infinite-level tree with a level-by-level updating of the spread size, starting from the bottom of the tree (for more see Appendix B.1).

An important problem in generalized models for social and biological contagion [73, 144, 145] is to optimize the set of initiators, i.e., for a fixed cost (seed size), find the set of initiators giving rise to the largest cascade, or alternatively, find the minimum size seed set required to activate the entire network [146]. As far as selection strategies are concerned, Kempe et al. [147] showed that the optimization problem of selecting the most influential nodes in any directed weighted graph with uniform random selection of thresholds is NP-hard. They also suggested a greedy algorithm [147], where each new initiator is selected based on the maximum spread it can cause, which unfortunately resulted in low efficiency of the algorithm. Chen et al. [148] designed a scalable algorithm (LDAG) which is based on the properties of directed acyclic graphs. Recently, Lim et al. [186] introduced a new node-level measure of influence, called cascade centrality (based on the size of the cascade resulting from the node being the only initiator), which may guide the selection of multiple initiators. Closely related to these studies and of practical interest is to find a set of initiators (not necessarily the smallest) in a scalable fashion that guarantees that the entire network will ultimately turn active, triggered by these initiators [149]. Their method was inspired by the k-shell decomposition of the network [150], which itself can be an effective heuristic for selecting initiators in a broad class of models for the spreading of social or biological contagion [73].

Singh et al. [129] studied the effect in the LTM of varying the fraction of initiators on the cascade size for various basic heuristic selection strategies when each node has identical threshold in the network. They showed [Fig. 2.1(a)] that there is a critical fraction of initiators ("tipping point") at which a sharp (discontinuous) phase transition occurs from small to large cascades in Erdős-Rényi (ER) graphs [21]. This

phase transition is apparent for the random, k-shell, and degree-ranked selection strategies, which are listed in the increasing order of their performance. These findings, in particular, the emergence of the discontinuous transition, were analogous to those found by Baxter et al. [151,152] for bootstrap percolation (there, activation of a node requires k active neighbors). Furthermore, Singh et al. looked at the impact of the network's average degree $\langle k \rangle$ to the cascade size [Fig. 2.1(b)].

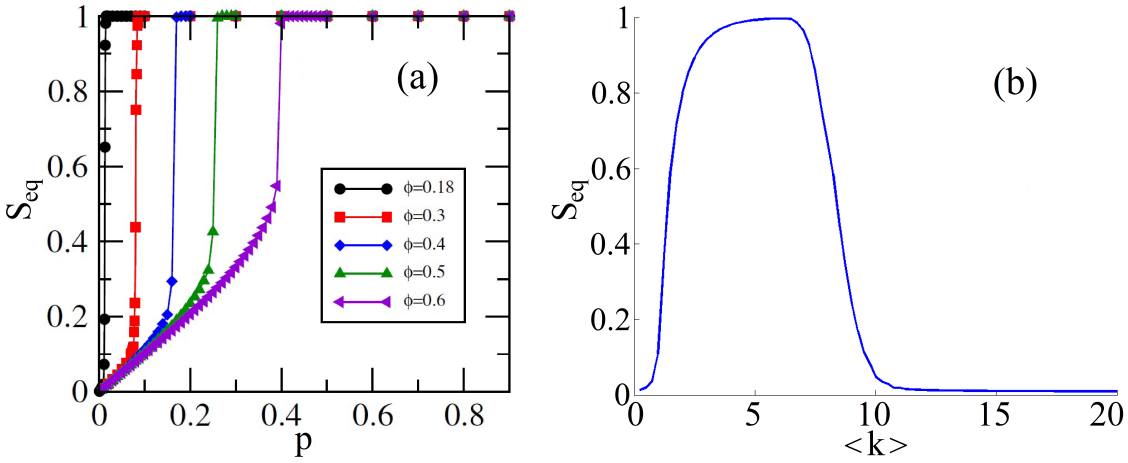

Figure 2.1: Tipping points and non-monotonic behavior. (a) Appearance of a discontinuous transition for various cases of identical thresholds on ER graphs with N=10000 and $\langle k \rangle$=10, plot taken from [129]. (b) Impact of the network's average degree $\langle k \rangle$ (ER, N=1000) in the cascade size S_{eq} for an initiator fraction of p=0.01, with identical thresholds.

In most of the past research, the cascade size has been thoroughly investigated for a identical threshold in the network [125, 129–132], or for a uniformly random threshold for each node [147,148]. However, a model with identical thresholds does not capture the complex nature of social influencing when multiple initiators are present. The small scale experiment conducted by Latane [153] and more recently an online experiment by Centola [154] as well as a large study on a Skype dataset [160] suggest that individuals have diverse thresholds for adopting a newly introduced opinion. Here, to capture the diversity of opinion adoption thresholds in a social influence context, we study the effect of heterogeneous thresholds on the cascade size under the LTM for empirical and synthetic unweighted and undirected networks

for randomly selected initiators.

2.1.1 Empirical Evidence of Linear Threshold Model-like Contagion

An increasing number of controlled experiments [153–155] and empirical data analysis on Twitter [156–159] and skype [160] datasets shine light on the existence of LTM like spread in social networks and on the existence of a distribution of susceptibility (thresholds). The works [155–158] support the existence of a version probabilistic threshold model of opinion diffusion, namely complex contagion. In 1996 Latané and L'Herrou contacted a controlled experiment on the impact of the choices of others in our choices, namely Conformity Game on an Email Experiment. Individuals were asked to predict the choice of the *majority* of their group on a bipolar issue, with the knowledge that they will receive a reward if their choice is in the majority and will receive nothing if it is not. The only goal is for the individuals to win the reward by siding with majority, hence holding their opinion has no other social benefit or any importance. There were 192 participants, all of which being undergraduate students, separated in 24 groups.

The participants would interact asynchronously through emails to their 4 neighbors. The computers were local and the duration of the experiments was 2.5 weeks. The only pre-game information the participants had is the size of the group (24) and the number of their neighbors, no network information was given to participants. To begin the experiment each participant would be given an general neutral and non-provocative question, such as " Now, see if you can predict the mathematician (Euler or Hilbert) which the majority of your groupmates will choose". The participants would pick their choice. On each following round, the participants would be notified on the choices of their nearest neighbors, and would be given the option to change their previous selection. The experiments revealed that individuals typically decide to change their answer when three out of their four neighbors have a different opinion. More importantly, it revealed the *heterogeneous threshold distribution* of the number of neighbors required for the participant to change his opinion.

A recent empirical study on a Skype dataset [160] further supports the LTM

based spread and the existence of heterogeneous threshold distribution. The authors contacted data analysis and modeling of a service adoption in on a OSN, Skype. Skype is the world's leading voice over internet service. The authors tracked the nodes which adopted the 'Skype paid service' "buy credit" for nearly 7.5 years since 2004, and only used the largest connected component of the aggregated free Skype service as the underlying structure. A node would either spontaneously adopt the service or would be influenced by its neighbors. Using the LTM model the authors found out that the threshold distribution follows a lognormal relationship to the degree of the node.

2.2 Materials and Methods

2.2.1 Simulations of the Linear Threshold Model

We assume that the thresholds are drawn from a truncated normal distribution with mean ϕ_0 and standard deviation σ. The threshold ϕ of each node is limited to the interval $[0, 1]$, thus the mean threshold ϕ_0 is also within this interval, and σ is in the range of $[0, 0.2887]$, boundaries of which correspond to the identical threshold and to the uniformly random threshold, respectively. The truncated threshold distribution $P(\phi, \sigma)$ is given by $P(\phi, \sigma) = N(\mu, \tau)/(1 - \int_{-\infty}^{0} N(\mu, \tau)d\mu - \int_{1}^{\infty} N(\mu, \tau)d\mu)$ for $0 \leq \phi \leq 1$, and $P(\phi, \sigma) = 0$ anywhere else [123]. Where, $N(\mu, \tau)$ is the normal distribution with mean μ and standard deviation τ, which take values $0 \leq \mu \leq 1$ and $0 \leq \sigma \leq \infty$ respectively. Unlike, in the formulation of the LTM in [133, 134], where thresholds drawn can be negative, allowing nodes to get spontaneously activated as innovators, and as a result randomizing the set of initiators, we are interested in the case where spread is initiated only with the insertion of randomly selected initiators in the network (Fig. 2.2).

Once a threshold for each node is set, for the simulations, we randomly assign initiators one by one and measure the cascade size. We repeat this process by drawing thresholds from the same distribution. The final cascade size for each threshold distribution is obtained by averaging one thousand times on different threshold distribution draws and, for the synthetic graphs, different network realizations. Detailed information on the networks used is located on Appendix (A.1).

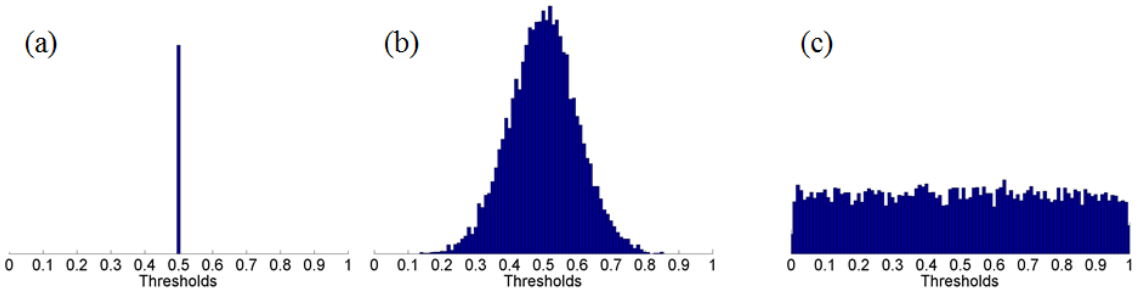

Figure 2.2: Visualization of the truncated Normal Distribution for mean threshold ϕ_0=0.5, and (a) σ=0 (Dirac distribution), (b) σ=0.2 (typical normal distribution), and (c) σ=0.2887 (uniform distribution).

2.3 Results

2.3.1 Multiple Initiators with Truncated Normal Threshold Distribution

First, we examine the effect of the standard deviation σ on the cascade size S_{eq} (averaged) for a constant initiator fraction and constant mean threshold ϕ_0 (Fig. 2.3). As σ increases so does a fraction of nodes whose threshold is far from the average causing a twofold effect. Of nodes far from average, the ones with thresholds below average are easily activated while those with thresholds above average are increasingly difficult to activate. Thus, when the initiator fraction is small, the cascade size S_{eq} is monotonically increasing since the presence of larger fraction of low threshold nodes facilitate the spread. However, when the initiator fraction are large, the increase in low threshold nodes helps a little since they are likely to be already activated without the increase in σ, but presence of additional high threshold nodes arrest the spread. This trade-off gives rise to the non-monotonic behavior seen in Fig. 2.3, which is apparent for different types of networks. Depending on the network structure and size of the initiators, the standard deviation σ for which the spread is optimal varies. The networks we use are Erdős-Rényi (ER) graphs [21], Scale Free (SF) networks [29, 161], a Facebook ego-network (FB) [162], and a high school friendship network (HS) [163][1] (for more information on the networks used

[1]We use the network-structure data sets from *Add Health*, a program project designed by J. Richard Udry, Peter S. Bearman, and Kathleen Mullan Harris, and funded by a grant P01-HD31921 from the National Institute of Child Health and Human Development, with cooper-

see Appendix A.3).

Interestingly, in the vicinity of $\sigma \approx 0$ the sharp decrease in the cascade size S_{eq} occurs because with non-zero σ, approximately half of the nodes acquire a threshold higher than $\phi_0 = 0.50$. For all the nodes with threshold $\phi > \phi_0$ with even degree, even the slightest non-zero σ value will increase the number of active neighbors by one, thus making cascades less likely to occur. Finally, for ER graphs [Fig. 2.3(a)] and SF networks [Fig. 2.3(b)] the analytic estimates are in good agreement with the simulations.

A visualization (Fig. 2.4) shows time steps of the spread on a random selection of initiators with $p = 0.20$ in the FB network. For the same set of initiators, the spread for large sigma ($\sigma = 0.20$) is much higher than for identical thresholds ($\sigma = 0.00$). Furthermore, the network visualization is such to identify different communities. Interestingly, we observe how the cascade is moving towards specific communities, a more analytical discussion of the impact of communities appears in [132].

In Fig 2.5, the cascade size S_{eq} is plotted for varying initiator sizes p for the same networks as in Fig. 2.3. As the initiator fraction increases, for small enough σ there is a transition from small local cascades to large global cascades, which, for synthetic graphs is a discontinuous phase transition [Fig. 2.5 (a) and (b)]. However, the line of the average cascade size S_{eq} appears smooth even in the presence of a discontinuous phase transition, because for each repetition the point of the discontinuous phase transition varies slightly. With increasing σ the initiator fraction for which the transition occurs is reduced, while for the synthetic graphs the spread size still exhibits a discontinuous phase transition. With largely diverse thresholds we find that a critical initiator size beyond which cascades become global ceases to exist and the tipping-point behavior of the social influencing process disappears and is replaced by a smooth crossover governed by the size of initiator set. This property can be important, for example, for a company's marketing strategy of a new product. If the threshold distribution is narrow enough, unless a critical

ative funding from 17 other agencies. For data files contact Add Health, Carolina Population Center, 123 W. Franklin Street, Chapel Hill, NC 27516-2524, (addhealth@unc.edu, available at http://www.cpc.unc.edu/projects/addhealth/ (Accessed September 27, 2015)

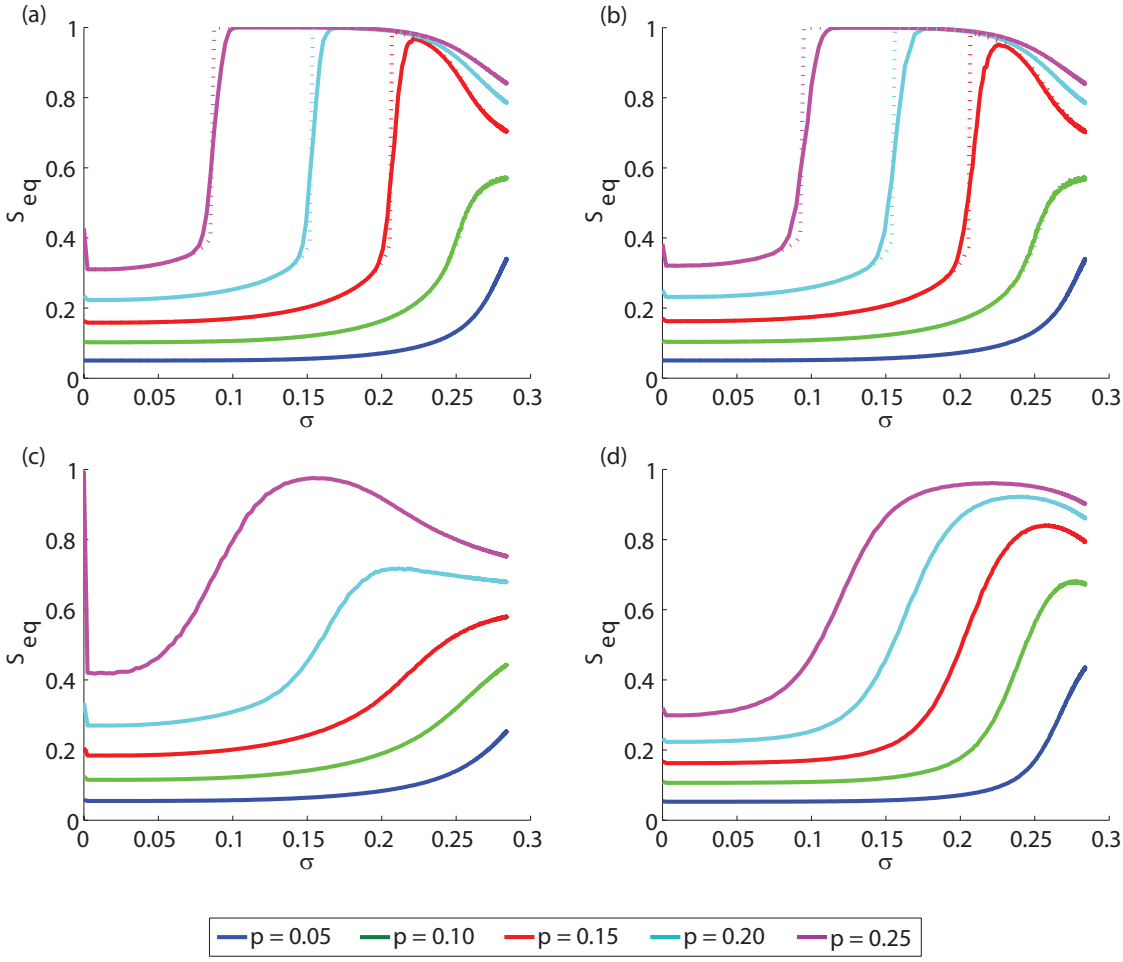

Figure 2.3: Behavior of the cascade size S_{eq} at equilibrium for varying standard deviation σ. (a) ER graphs with $\langle k\rangle$=10 and N=10^4; (b) SF networks with $\langle k\rangle$=10, γ=3, and N=10^4; (c) HS network with $\langle k\rangle$=5.96 and N=921; (d) FB network with $\langle k\rangle$=43 and N=4039. The mean threshold is ϕ_0=0.50. The simulations are averaged over one thousand repetitions. (a) and (b) also show the analytic estimates (dotted lines) based on the tree-like approximation (see Materials and Methods) [133].

initiator fraction is reached, there is a marginal local spread on a few of the first or second neighbor friends of the initiators. On the other hand, if the threshold distribution is wide, there is a significant spread.

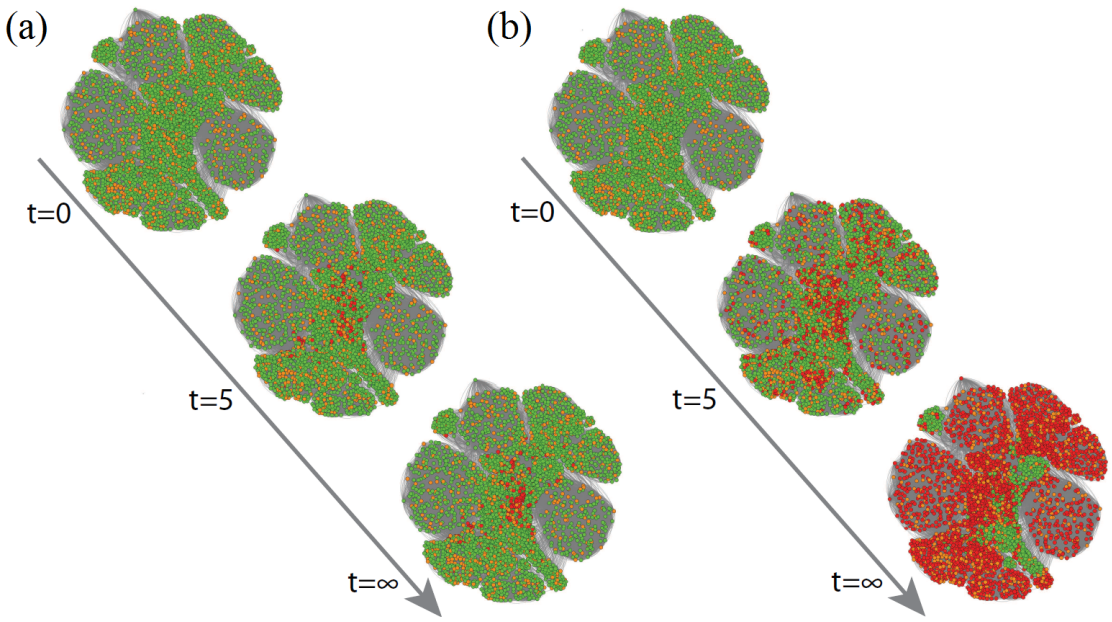

Figure 2.4: Visualization of the spread of opinion in the LTM model on a FB network with $\langle k \rangle$=43 and N=4039. The fraction of the randomly selected initiators is p=0.20. The mean threshold is ϕ_0=0.50 while the standard deviation of the threshold is (a) σ=0, (b) σ=0.20. Inactive nodes, initiators, and active nodes (through spreading) are marked with green, orange, and red, respectively.

2.3.2 Multiple Initiators with Truncated Lognormal Threshold Distribution

As we mentioned above, a recent empirical analysis on the skype dataset [160] suggests a lognormal threshold distribution. The threshold distribution was given by

$$P(\phi) = \frac{1}{\phi \sigma_T \sqrt{2\pi}} e^{-\left(\frac{\ln \phi - \mu}{2\sigma_T^2}\right)^2} \tag{2.1}$$

with $\mu_T = -2$ and $\sigma_T = 1$. The authors considered in their model nodes with negative threshold as self-initiators. We are not interested in this case, hence considering that the thresholds are bound $0 \leq \phi \leq 1$, we obtain $\phi_0 = 0.189$ and $\sigma = 0.233$. For this case the average threshold is small, and so a small initiator fraction can lead to global cascades. Nontheless, we use their threshold probability function combined with the truncation process to test the cascade size response. In Fig. 2.6 we show

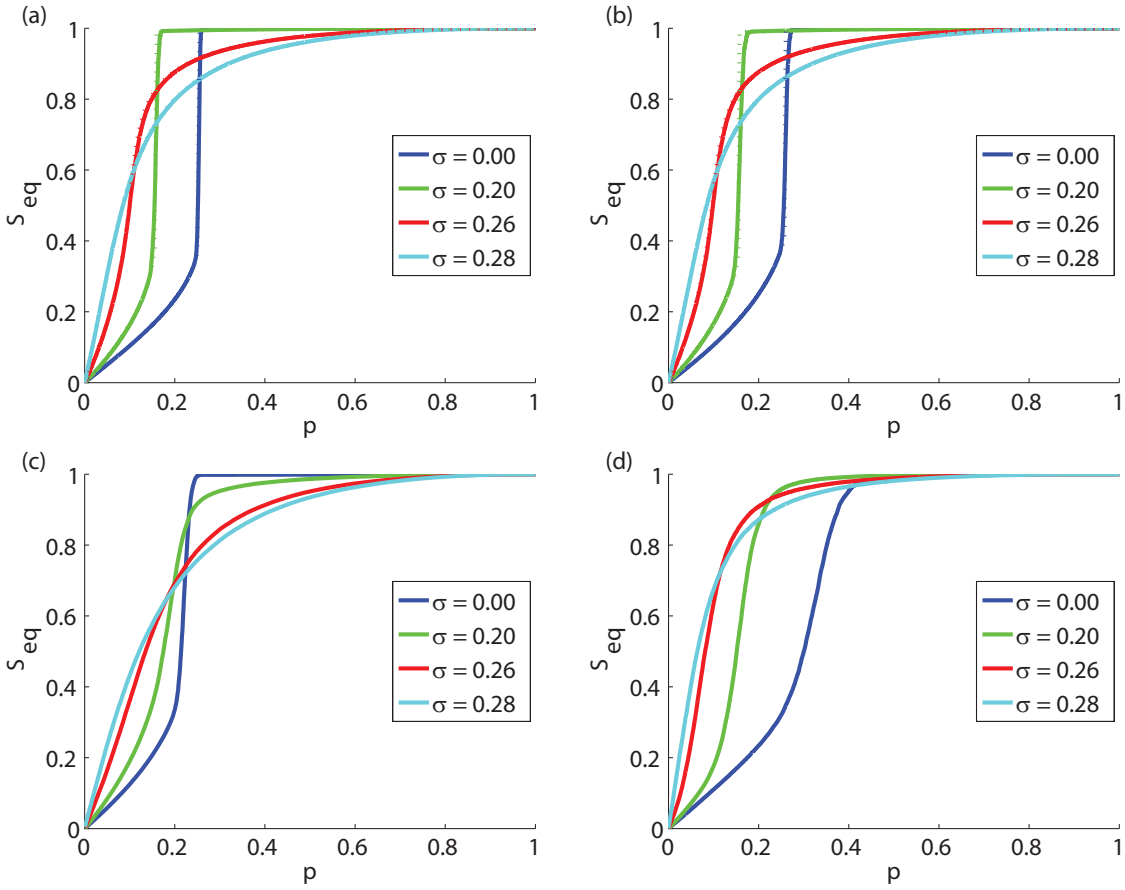

Figure 2.5: Behavior of the cascade size S_{eq} at equilibrium vs. the initiator fraction p. The networks are the same as in Fig. 2.3: (a) ER graphs with $\langle k\rangle$=10 and N=10^4; (b) SF networks with $\langle k\rangle$=10, γ=3, and N=10^4; (c) HS network with $\langle k\rangle$=5.96 and N=921; (d) FB network with $\langle k\rangle$=43 and N=4039. The mean threshold is ϕ_0=0.50. (a) and (b) also shows the analytic estimates (dotted lines) based on the tree-like approximation (see Materials and Methods) [133].

the shape of truncated lognormal distribution. For both cases of ϕ_0 with non-zero σ the positive skew leads to very large cascades, since there is a large fraction of nodes with low threshold, and fewer with high threshold. The minor decrease of the cascade size we observe vs. σ in the plots (c) and (d) is due to the small number of nodes with high threshold. Their size is significantly small that they do not happen to be connected often and block the spread.

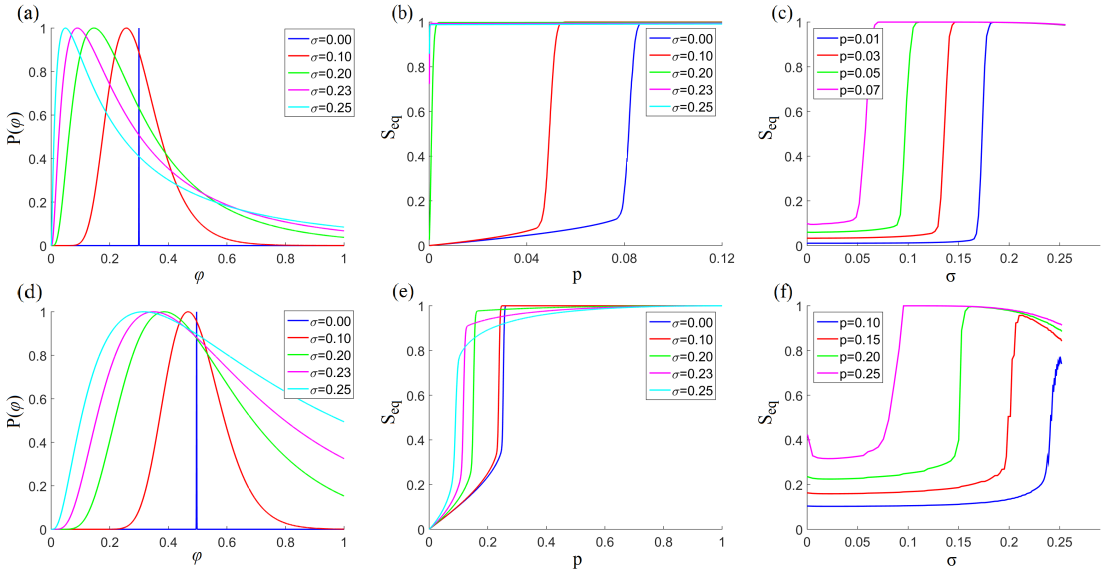

Figure 2.6: The impact of truncated lognormal threshold distribution Eq. (2.1) to the cascade size for ER graphs with N=10000, $\langle k\rangle$=10. (a,d) Visualization of the truncated lognormal distribution for mean threshold ϕ_0=0.3, and ϕ_0=0.5 respectively, (b,e) cascade size S_{eq} vs. inititator fraction p for ϕ_0=0.3, and ϕ_0=0.5 respectively, and (c,g) cascade size S_{eq} vs. standard deviation σ for ϕ_0 = 0.3, and ϕ_0=0.5 respectively. The input values of Eq. (2.1) for ϕ=0.3 and σ=0, 0.1, 0.2, 0.23, 0.25 are μ_T=−1.205, −1.2575, −1.36, −1.3325, −1.17 and σ_T=0.001 , 0.321, 0751, 1.041, 1.361 respectively. The input values of Eq. (2.1) for ϕ=0.5 and σ = 0, 0.1, 0.2, 0.23, 0.25 are μ_T=−0.7, −0.72, −0.71, −0.56, −0.21 and σ_T=0.001, 0.201, 0.491, 0.701, 0.971 respectively.

2.3.3 Impact of System Size

In Figs 2.7 and 2.8 we show that the behavior of the cascade size is largely independent of the system size N for any threshold distribution with the same degree distribution, for ER graphs and SF networks, respectively. We observe that with increasing system size N the cascade size S_{eq} is asymptotically converging.

2.3.4 Critical Initiator Fraction

We record the critical initiator fraction p_c for which a discontinuous phase transition occurs for varying mean threshold ϕ_0 (Fig. 2.9). For the measurement

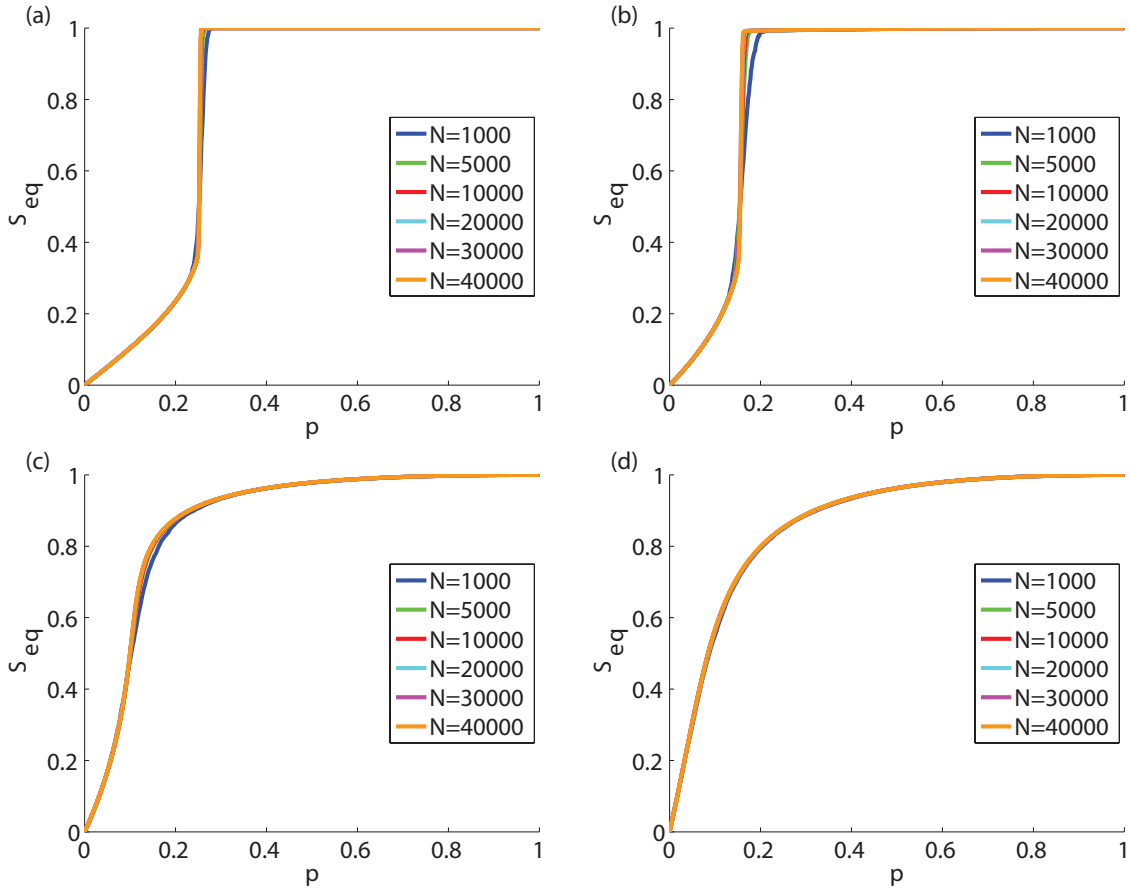

Figure 2.7: Finite-size behavior of the final cascade size S_{eq} vs. the initiator fraction p for ER graphs with average degree $\langle k\rangle$=10. The mean threshold is ϕ_0=0.50 while the standard deviation of the threshold is (a) σ=0.00, (b) σ=0.20, (c) σ=0.26 and (d) σ=0.28.

of p_c, first we calculated the derivative of the S_{eq} from Fig. 2.5 with respect to the initiator fraction p. The position of maximum of the derivative yields the p_c, in other words, $p_c = \arg\max_p \left(dS_{eq}\left(p\right)/dp\right)$. We used the same method for the calculation of the respective analytic estimates. We confine the threshold distribution for up to $\sigma = 0.15$ to assess if there is a discontinuous phase transition with increasing initiators. Above each p_c line global cascades occur. The value of p_c decreases with increasing σ. For identical thresholds ϕ_0 (in blue), the p_c line has some sharp jumps, for example at ϕ_0 equal to 0.50, 0.33, and 0.25 (Fig. 2.9). These jumps are artifacts of the discrete steps of the degree distribution in the presence of a

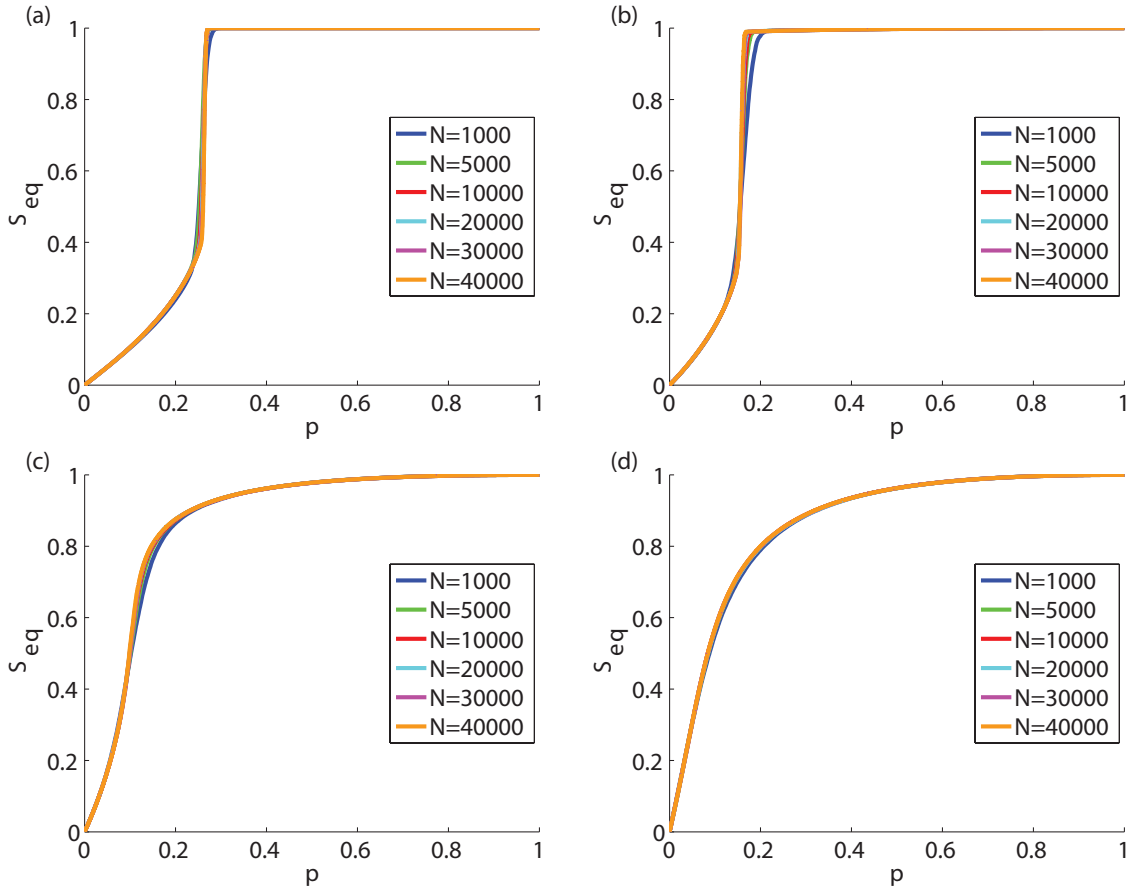

Figure 2.8: Finite-size behavior of the final cascade size S_{eq} at vs. the initiator fraction p for SF networks with $\langle k\rangle$=10 and γ=3. The mean threshold is ϕ_0=0.50 while the standard deviation of the threshold is (a) σ=0.00, (b) σ=0.20, (c) σ=0.26, (d) σ=0.28.

unique threshold for all the nodes. In particular, microscopically, the number of active neighbors required for a node to turn active increases by integer values. For example, for a node with degree 10 and $0.40 < \phi \leq 0.50$, that number is 5. For identical thresholds in the network, the cumulative effect of these integer steps gives rise to the jumps exhibited by the $p_c(\phi_0)$ curves (Fig. 2.9). Interestingly, this effect also shows in Fig. 2.3, where for large enough initiator fractions (i.e., $p = 0.25$ or higher) the cascade size drops abruptly as σ is increased from zero to small values. For nodes with mean threshold $\phi_0 = 0.50$, even the smallest non zero increase on the standard deviation σ results in approximately half of the nodes having threshold larger than $\phi_0 = 0.50$. The p_c lines are lower for the ER graph compared to the SF

networks because of the importance of a randomly selected very high degree node in SF networks can have on the spread. Our results obtained from simulations are in agreement with the analytic estimates.

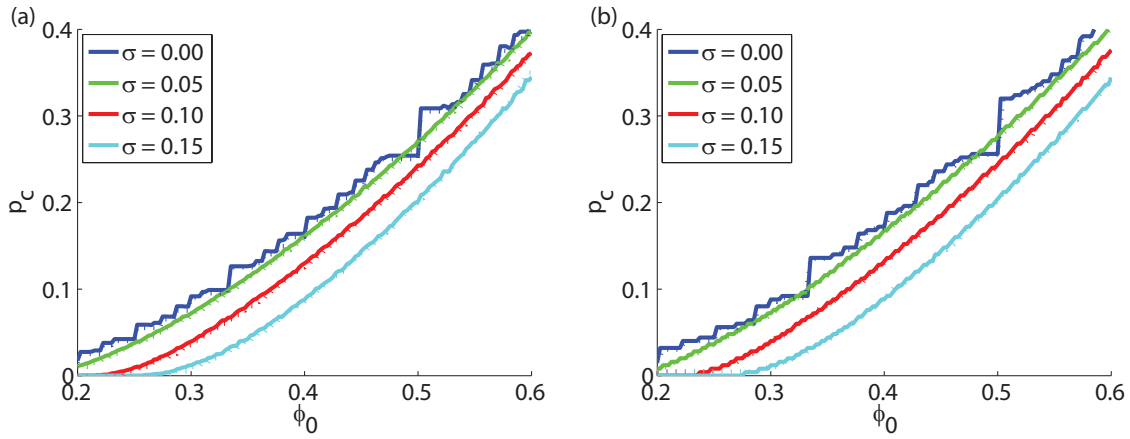

Figure 2.9: Critical initiator fraction p_c vs. mean threshold ϕ_0. (a) ER graphs and (b) SF networks with γ=3 with average degree $\langle k\rangle$=10 and system size N=10^4. An initiator size above the p_c line leads to global cascades. The analytic estimates (dotted lines) are based on the tree-like approximation [133] (see Materials and Methods).

2.3.5 Synchronous Updating and Cascade Size

To further understand the effect of the standard deviation σ, we study the dynamics of the spread for synchronous updating of the nodes. In phase-space, as shown in Fig. 2.10, the difference $\Delta S(n+1) - \Delta S(n)$ defines the number of nodes activated from time step n to $n+1$. The dynamic spread in the LTM is deterministic and evolves in one direction, hence, the spread stops when the change on the cascade size (vertical axis) reaches zero. Accordingly, the value of the cascade size in the steady state is indicated on the horizontal axis. When cascades are not possible, the spread rate decreases monotonically. However, when cascades are possible then for up to some σ the change is non-monotonic and the fractions of nodes in cascades reach almost one. But as σ's grow larger and larger, these fractions stop growing farther and stay farther from one. When σ approaches the standard deviation of uniform distribution the shape of the lines decreases linearly. Interestingly, similar

behavior is observed for the FB and HS networks as well.

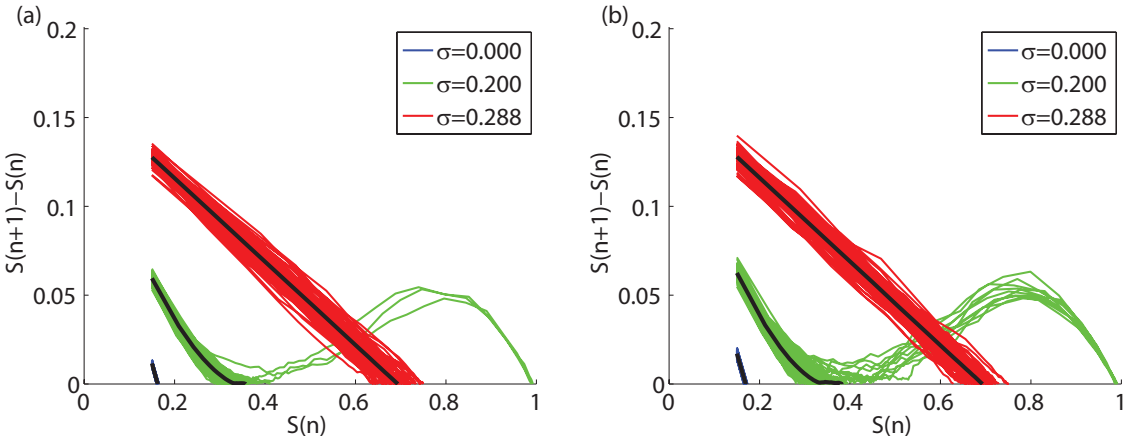

Figure 2.10: Phase-space diagrams for a constant initiator fraction p=0.15, and various standard deviations σ=0 (blue), σ=0.2 (green), σ=0.288 (red) for (a) ER graphs and (b) SF networks with γ=3, with $\langle k\rangle$=10 and N=10^4. The colored lines refer to a hundred independent repetitions, while the black lines are their averages.

2.3.6 Closed-form Analytic Estimate for the Uniform Threshold Distribution

For a uniform threshold distribution the phase-space line decreases linearly for any initiator fraction for synthetic graphs and almost linearly for the empirical networks (Fig. 2.11).

In addition, we show for this threshold distribution, using Gleeson and Cahalane's analytical methods, that the phase-space line has a closed form and is linearly decreasing. The extended proof of this is shown in Appendix B.2). For a uniform threshold distribution the iterative formula in Eq. (B.2) of the analytic approximation yields the following closed-form solution

$$q_{n+1} = p + bq_n, \tag{2.2}$$

with $b = (1-p)\,\frac{1}{\langle k\rangle}\,(\langle k\rangle - 1 + P_0)$. The solution of the above iterative equation

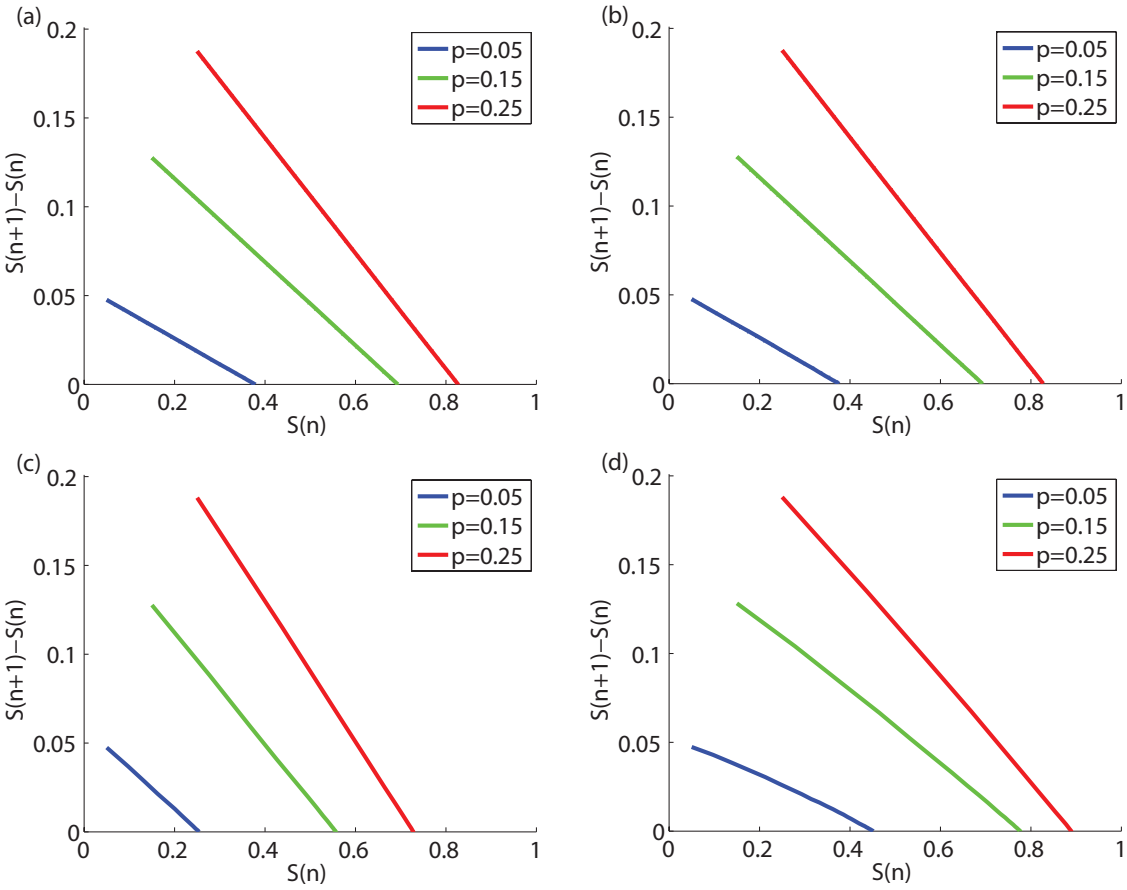

Figure 2.11: Phase-space diagrams for the uniform random threshold distribution (σ=0.288), for various initiator fractions p=0.05 (blue), p=0.15 (red) and p=0.25 (green) for (a) ER graphs, (b) SF networks, (c) HS network, and (d) FB network as in Fig. 2.3. The solid lines and dotted lines (complete overlap) correspond to the simulations and to the closed-form analytic estimates [Eq. (B.2)], respectively.

with the initial condition $q_0 = p$, is

$$q_n = p\frac{1-b^{n+1}}{1-b}. \tag{2.3}$$

According to [134], the spread at level $n+1$ is given by

$$S_{n+1} = h(q_n) = p + (1-p)\sum_{k=1}^{\infty} P_k \sum_{m=1}^{k} \binom{k}{m} q_n^m \left(1-q_n\right)^{k-m} F\left(\frac{m}{k}\right), \tag{2.4}$$

which, in the case of a uniform distribution of thresholds (A.1.3) simplifies to

$$S_{n+1} = p + cq_n, \tag{2.5}$$

with $c = (1-p)(1-P_0)$, where the initial spread is $S_0 = p$. Using the above Eq. and Eq. (B.13) we can calculate (A.1.3) the formula for the phase-space diagram

$$S_{n+1} - S_n = cp - (1-b)p - (1-b)S_n \tag{2.6}$$

The above Eq. is the closed form phase-space line of Fig. 2.11. On the other hand, at the equilibrium (as $n \to \infty$) the spread size in Eq. 2.5 becomes

$$S_{eq} = p + cq_\infty, \tag{2.7}$$

with $q_\infty = p\frac{1}{1-b}$(A.1.3). Note that in this approximation for uniform threshold distribution, the size of the final cascade for uncorrelated networks does not dependent on the details of the degree distribution, it only depends on the average degree $\langle k \rangle$. In addition, it is easy to show that the derivative of the final cascade size [Eq. (2.7)] with respect to the initiator size p is monotonically decreasing, in agreement with the submodularity property of the LTM for the uniform threshold distribution [147].

2.3.7 Discontinuous Phase Transitions in the Linear Threshold Model

To further understand the final cascade size behavior at the critical point for synthetic graphs, we are examining the system size dependence. The spread size at the equilibrium is independent of the method of the insertion of initiators, e.g., it does not matter whether the addition occurs in fractions or by individual addition of initiators. Using Monte-Carlo simulations, Singh [129] showed that the average cascade size is largely independent of the system size for the same initiator fraction for an identical threshold for ER graphs with unique degree distribution. We use the same approach to show that this is true for other threshold distributions for ER graphs (Fig. 2.7) and SF networks (Fig. 2.8). These results indicate that given an initiator fraction p_0 and an average cascade size $S_{eq}(p_0)$, the addition of another

initiator fraction p_1 will cause the same change $\Delta S = S_{eq}(p_0 + p_1) - S_{eq}(p_0)$ in the average cascade size S_{eq}, largely independently of the system size, for large system sizes, for the same input degree and threshold distributions.

Our analysis so far focused on the cascade size at the steady state S_{eq} averaged over many realizations of networks, threshold values and assignment of initiators (Figs. 2.7 and 2.8). To verify the presence and nature of phase transitions, we follow the approach presented in [151]. We start by measuring the increase of the cascade size of each sample in response to the one-by-one addition of initiators. If a discontinuous phase transition arises, at the critical point, the increase of the cascade size should remain constant and independent of the system size. To investigate this, let v be the current size of initiator set. For a given sample i, let $\Delta S_i = S_i(\frac{v+1}{N}) - S_i(\frac{v}{N})$ denote the increase in the cascade size caused by the addition of a single randomly selected initiator to the current initiator set. Let $(\Delta S_i)_{\max}(N)$ be the maximum value of $\Delta S_i(N)$ for all initiator sets of size $\frac{v}{N}$. Then, varying σ, we study how $(\Delta S_i)_{\max}(N)$ averaged over one thousand repetitions depends on the system size N (Fig. 2.12) (solid lines). We observe that for the plotted cases with $\sigma = 0.00$ and $\sigma = 0.24$, $\langle(\Delta S_i)_{\max}\rangle(N)$ is independent of the system size. Moreover, the contribution of the rest of the initiators to the cascade tends to zero in the limit of infinite system size. However, for $\sigma = 0.26$, $\langle(\Delta S_i)_{\max}\rangle(N)$ decreases with the system size, indicating the absence of a discontinuous phase transition in the infinite system-size limit. Thus, there appears a qualitative change somewhere between $\sigma = 0.24 - 0.26$.

A similar analysis can be applied to the analytical estimation, with the tree-like approximation, of the increase in the cascade size $(\Delta S_{TL})_{max}(\delta p)$ with a marginal addition of initiators. However, since the analytical estimation is set for an infinite system size, the one-by-one addition of initiators on larger and larger system sizes is not possible. Hence, we insert smaller and smaller fractions of initiators δp. In Fig. 2.12 the top horizontal axis is the fractional step increase of the number of initiators. For consistency, we include the corresponding increase in the cascade size $\langle(\Delta S_i)_{\max}\rangle(\delta p)$ that δp, a fractional step increase of the number of initiators, measured through simulations. In this case, the minimum possible fraction of initiators

is $\delta p = 1/N$. We observe, that the results for the one-by-one addition of initiators with varying systems through simulations, agree with those for the fractional increase of an infinite system size with varying δp. We conclude that it is between $\sigma = 0.24 - 0.26$ (for $\phi_0 = 0.50$) where the discontinuous phase transitions cease to emerge in the thermodynamic limit.

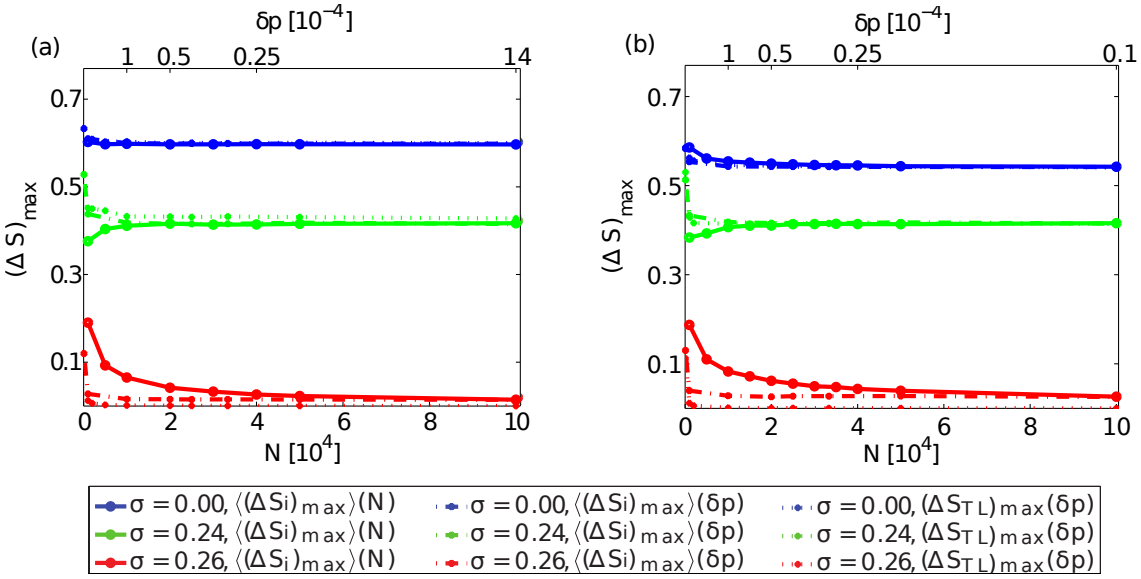

Figure 2.12: Maximum contribution of initiators to the cascade size for various σ values. (a) for ER graphs and (b) for SF networks with γ=3, for $\langle k\rangle$=10. Solid lines: $\langle(\Delta S_i)_{\mathbf{max}}\rangle(N)$ of O(1) initiator with one-by-one addition of initiators for varying system sizes (bottom horizontal axis). Dashed lines: $\langle(\Delta S_i)_{\mathbf{max}}\rangle(\delta p)$ for various initiator fractions (top horizontal axis) for a constant system size N=10^5. Dotted lines: $(\Delta S_{TL})_{\mathbf{max}}(\delta p)$ for various initiator fractions (top horizontal axis) for the TL approximation. The mean threshold is kept at ϕ_0=0.50 in all cases.

2.4 Discussion

Past experimental online studies [153,154,159] indicate the existence of diverse adoption thresholds of individuals in social networks. Prompted by this observation, we studied the impact of diversity of thresholds in spreading a new opinion, by intuitively assuming that the adoption thresholds are drawn from a truncated

normal distribution. We explored this impact by using the linear threshold model, a reinforcement model which has lately drawn significant attention in the scientific community. We showed that in the presence of a small spread (standard deviation) of the threshold distribution in a network, unless a critical initiator fraction is reached, the impact of the randomly selected initiators is small. Furthermore, we showed that, when discontinuous transitions in cascade size are possible for synthetic graphs, the addition of a single randomly-selected initiator can have a significant (global) impact on the final cascade size, i.e., the manifestation of the tipping point. However, with a sufficiently large spread in the individual thresholds (with the same mean), the cascade size exhibits a smooth transition, where the impact of each added initiator is reduced by the current size of the initiator set. Finally, we showed that in the case of a uniform threshold distribution, the spreading rate is linearly decreasing with the spread size for synthetic graphs and close to linearly decreasing for empirical graphs. In summary, our results indicate that information on the diversity of the thresholds is critically important for the understanding of the behavior of cascades in threshold-limited social contagion with multiple initiators. Most importantly, sufficiently large spread in the individual thresholds can change not only the quantitative aspects of triggering global cascades, but also the qualitative behavior of the system: the cascade size exhibits a smooth change (as opposed to a discontinuous jump) as a function of the fraction of initiators.

CHAPTER 3
Influence Maximization for Fixed Heterogeneous Thresholds

3.1 Introduction

Cascading processes emerge naturally through the interactions of nodes in different states in natural and human-made networks. Microscopic processes can potentially have large macroscopic impact on the networks. In the case of human-made networks, their ever increasing size and interconnectedness exponentially increases the uncanny impact of cascades processes. For instance, in financial or power grid networks, small initial perturbations or failures can result in cascades in the network causing tremendous disasters of global impact [49, 58, 60]. In social networks, contact processes, namely social influence (or contagion), enable the spread of new behaviors, opinions and products, driving politics, social movements and norms, and Viral Marketing.

The identification of nodes whose change of state can potentially affect large portions of the network becomes a key challenge. It is a computationally hard problem, and as such, multiple heuristics, theoretical analyses and algorithms have been introduced to solve it [165–167]. Some are designed to address the specific nature of the cascade process, while others are based on more general algorithmic approaches or network based centrality measures. Such algorithms can be used to minimize disasters by, for example, re-enforcing weak nodes in power-grid nodes [54, 61, 62], or placing sensors to detect the contamination of water pipe network [168]. Likewise, to arrest spread of infectious disease requires a global sense of awareness [64, 73, 74, 169]. Understanding cascades is also important for optimizing Viral Marketing [170–172]. Yet, it is challenging to find the set of initiators (also called seeds) which when put into a new state (opinion/idea/product), will

Portions of this chapter to appear in: P. D. Karampourniotis, B. K. Szymanski, and G. Korniss, "Influence maximization for fixed heterogeneous thresholds," (unpublished, 2017).

maximize the spread of this state [146–149, 173–179].

Here, we study the problem of IM on a classical opinion contagion model, namely the Linear Threshold Model (LTM) [94, 96], although our methods can be used for any percolation based model. The LTM is designed to capture the peer pressure dynamics that lead an individual to accept a new state being propagated. It is a binary state model, where a node i has either adopted a new product/state/opinion, $n_i = 1$, or not, $n_i = 0$. According to the LTM, each node in the network has a fractional threshold, an intrinsic parameter representing the node's resistance to peer pressure. The spreading rule is that an inactive node ($n_i = 0$), with in-degree k_i^{in} and threshold ϕ_i, adopts a new opinion only when the fraction of its neighbors $j \in \partial i$ possessing that new opinion is higher than the node's threshold, that is $\sum_{j\partial i} n_j \geq \phi_i k_i^{\text{in}}$. The process is deterministic and nodes cannot return to their previous state. The integer number of active neigbhors required for node i to get active is given by its resistance $r_i = \lceil \phi_i k_i^{\text{in}} \rceil$. A node gets activated through spread when its resistance drops to zero, $r_i = 0$, with the maximum resistance of a node being k^{in}. Bootstrap percolation [151,152] is an alternative formulation of the LTM where the thresholds are not fractional, but integer (resistance). Interestingly, the activation threshold in the LTM and bootstrap percolation conceptually share similarities with the integrate-and-fire neuron model [98–101], with the difference being that there is a probability distribution describing the probability of activation of a node, rather than a fixed threshold value, and return on the initial (inactive) state $n_i = 0$ is allowed. The size of cascades in the LTM is governed by the thresholds of the nodes [123, 126], the size of the initiator set [129], the strategies for selecting initiators [179], and of course the structure of the underlying network [130–132, 135, 136].

Examples of an LTM type spread mechanism and of the heterogeneity of the thresholds are provided through a number of controlled experiments [153–155] and empirical data analysis [156–160]. Watts and Dodds [126] showed through simulations on various types of spread mechanisms that the cascade size is governed not by superspreaders, but by a small critical set of nodes with low resistance to influence. Karampourniotis et al. showed that the threshold distribution is important for the overall dynamics [123]. In particular, with an increasing standard variation

σ of thresholds (while holding the average of the thresholds fixed) a smaller initiator fraction is required for global cascades. Furthermore, they showed that in the vicinity of $\sigma \approx 0$ a tipping point appears as the fraction of randomly selected initiators gradually increases. Yet, with gradually increasing variance σ eventually the tipping point is replaced with smooth transition (See Chapter 2). In addition, Watts and Dodds [126] showed that a critical size of nodes with high susceptibility contribute to social influence much more than initiators with high network centrality.

And so, the knowledge of the thresholds or the threshold distribution is important for the IM algorithms. In the case of zero information on the thresholds or the threshold distribution, a good assumption to make is the threshold distribution is uniform. This is a very interesting case since then, the spread function is submodular, that is, it follows a diminishing returns property [147], which we can be used for maximizing the influence [147, 148, 168, 173–175, 180, 181]. Even though there could be specific cases of thresholds distributions where a weakly submodularity property [182] could be used, in general submodularity does not hold when the thresholds are known and fixed or for any threshold distribution other than the uniform. Such a case is when the threshold of each node is known with some uncertainty [183]. When the thresholds are known and fixed the influence of any seed set can be computed in polynomial time [184]. In the special case of a threshold identical for all nodes, a first-order transition appears [125, 129, 133]. Then a powerful algorithm, namely CI-TM with complexity $\mathcal{O}\left(\langle k\rangle N \log N\right)$, provides the best performance [179]. In Ref. [185] the authors express the IM problem as a constraint-satisfaction problem and use belief propagation to solve it, yet it does not perform as well as the CI-TM for the case of identical thresholds that it was compared with [179]. Other analytical based metrics show the importance of the network structure, but only for a small number of initiators [186]. Furthermore, Ref. [187] proposes the use of an evolutionary algorithm implemented with general-purpose computing on graphics processing units (GPGPU) to tackle the challenge of combinatorics, at the additional cost of higher time and memory complexity. The authors show that their approach clearly outperforms the greedy algorithm for known thresholds both in cascade size and time, but it is currently limited to graphs

of size of the order of $N = 10^3$.

Here, we study the efficiency of known selection strategies for fixed (and known) heterogeneous thresholds generated from different threshold distributions, and a range of assortativities, and we introduce two new selection strategies for the LTM with fixed thresholds, and compare them in terms of their performance with a number of other strategies, including the CI-TM, and greedy. Since we focus on fixed (and known) thresholds we do not include the performance of various network centrality measures like the Page Rank and k-core [129], which do not take into account the provided threshold information and thus are outperformed by the strategies that do.

3.2 Selection Strategies

We use a number of simple and fast heuristics, which take advantage of the knowledge of thresholds. Since the thresholds are fixed and known, the cascade size caused by an initiator is deterministic. Hence, we sequentially introduce initiators *on the inactive subgraph* of the original network. First, the node with the highest dynamic fractional threshold ($thres$) is a reasonable choice. Likewise, a natural selection is the node with the highest dynamic out-degree k_i^{out} (deg) at each step. Another possible heuristic for the LTM is the selection of the node with the highest resistance (res) at each step. Resistance r_i is the current (dynamic) integer threshold of node i, that is the number of active neighbors required for the node to get activated. Accordingly, when a node is activated by a cascade or by being selected as an initiator, its resistance turns zero, and so a fully activated network has total resistance of zero.

The selection of any inactive node i as an initiator results to the decrease of the resistance of all its inactive neighbor nodes by one, for a total k_i^{out}. Due to the spread process taking place, if any neighbors of i had resistance equal to one, they will also get activated, further resulting to other nodes reducing their resistance. In addition, node i is the initiator, hence its resistance is reduced to zero, $r_i = 0$. Therefore, the total resistance drop of the entire graph is reduced by at least $r_i + k_i^{\text{out}}$. To capture the direct resistance drop we introduce the heuristic strategy RD, with

metric

$$\mathrm{RD}_i = r_i + k_i^{\mathrm{out}}. \tag{3.1}$$

In addition, we introduce the heuristic method red total resistance (RT), by adding on the RD metric the drop of the resistance of the inactive subgraph caused by the neighbors of the node i which are indirectly activated by the selection of i as seed. That is

$$\mathrm{RD}_i = v_i + k_i^{\mathrm{out}} + \sum_{j \in \partial i | r_j = 1} (k_j^{\mathrm{out}} - 1). \tag{3.2}$$

The drop of the network's total resistance caused by choosing i as an initiator is at least as high as RT_i; it is be more if the spread expands to the 2nd neighborhood (and so on). This heuristic is equivalent to the CI-TM algorithm [179] for a sphere of influence $L = 1$ with the addition of the resistance r_i of the selected node i in the metric. For very large L, a node's i CI-TM score is essentially (assuming a tree-like approximation of the network) equal to the drop of the network's total resistance if i was the seed, minus its resistance r_i. The metric of CI-TM is governed by the out-degree of the nodes surrounding the target node ignoring the challenge of activating nodes with high resistance and/or low in-degree. For comparison to our methods, we apply the CI-TM algorithm itself (for $L = 6$), and the greedy algorithm for fixed thresholds, where at each step the node which would cause the maximum cascade size is selected.

3.2.1 Balanced Index Strategy

Constructing a selection strategy mainly based on the network structure or just on the resistance of the nodes, is not ideal, since useful information is not being utilized. On one hand, selecting nodes solely based on some network centrality metric, leads to many easily susceptible nodes being selected as initiators, nodes that could potentially be activated through spread. On the other hand, aiming on selecting high resistance nodes, that is nodes with high resistance, does not guarantee they will be great influencers. The RD and RT strategies aim to address this weakness, by using intuitive heuristics. To quantify on the interplay of importance between low resistance vs. high centrality nodes, we introduce the Balanced Index

(BI) selection strategy. For this strategy, we essentially introduce weights (a, b, c) on each term of RD_i to capture the importance of each feature, that is

$$\mathrm{BI}_i = ar_i + bk_i^{\mathrm{out}} + c \sum_{j \in \partial i | r_j = 1} \left(k_j^{\mathrm{out}} - 1 \right), \tag{3.3}$$

where $a + b + c = 1$ and $a, b, c \geq 0$. The optimal weights for influence maximization are determined by scanning for the allowed range of weights in the ensemble of graphs and for various threshold distributions. In this case, the degree (deg) strategy corresponds to $(a, b, c) = (0, 1, 0)$, res to $(a, b, c) = (1, 0, 0)$, the CI-TM for $L = 1$ to $(a, b, c) = (0, 1/2, 1/2)$, and the two heuristics we introduced, RD_i and RT_i, correspond to weights $(a, b, c) = (1/2, 1/2, 0)$ and $(a, b, c) = (1/3, 1/3, 1/3)$ respectively. Interestingly, the weighted metric BI_i can be viewed as a measure (units) of resistance, however in general, it does not correspond to the network's total resistance drop when i is the seed. As far as the time complexity of each method is concerned, the computation of a seed's induced spread takes $\mathcal{O}(\langle k \rangle N)$ time. Yet, (similar to [179]) when computing the spreading process, we can place a stopping condition on the algorithm L levels away from the seed node, reducing the complexity by $\mathcal{O}(N)$. In addition, using a heap structure, re-ordering the highest BI nodes takes $\mathcal{O}(\log N)$.

3.2.2 Group Performance Index Algorithm

All of the above strategies are essentially local in nature, since they aim to maximize the number of activated nodes or to reduce the total resistance of the system caused by one initiator at a time. They lack in their metrics the impact of the combination of initiators on influence maximization, which by default limits their performance. Algorithms that use combinations of nodes in their metrics can lead to approximate global solutions. However, look-ahead algorithms suffer from the potentially prohibitive computational costs. For instance, to measure the total impact of g from N nodes, a deterministic greedy algorithm would require $\binom{N}{g}$ selections of possible initiator sets. The algorithm would choose iteratively at each step $(t = 1, 2 \ldots, g)$ the highest impact node from the set of the inactive ones,

therefore its complexity would be $\mathcal{O}(\langle k \rangle N^g)$, feasible only for very small g and moderate N. In addition, to compute the cascade size for each possible initiator set would require $\mathcal{O}(\langle k \rangle N^2)$.

Instead, a probabilistic greedy algorithm would aim to reduce the number of combinations by randomly selecting initiators. That is, at each step t, in order to measure the impact of a node i in the presence of other nodes, i would have to be selected as an initiator. Then, the remaining initiators would be randomly selected. This process would be repeated v times, each time recording the cascade size. We would have essentially measured the impact of node i as an initiator in the presence of any randomly selected set of initiators. Then, we would repeat this process for all other inactive nodes, and finally select the node with the highest impact. Since we would have to measure the impact of each inactive node, and run v simulations to do so, the time complexity per step t is $\mathcal{O}(vN)$. Typically, g is comparable to the total number of nodes N, adding an additional $\mathcal{O}(N)$. Finally, computing the simulations takes another $\mathcal{O}(\langle k \rangle N^2)$, and so the total complexity of the probabilistic greedy algorithm for g initiators is $\mathcal{O}(v\langle k \rangle N^4)$, which is still very expensive.

Here, we introduce the Group Performance Index algorithm (GPI). With GPI we target the nodes which, when included in any randomly selected initiator set (group), the group has the highest in average performance. GPI shares similarities with the probabilistic greedy, however it is much more efficient. First, we take advantage of the property that permutations of any set of initiators do not impact the total cascade size returned by that set for the LTM. By not having to scan each node individually when computing its impact in the presence of other initiators, we reduce the number of computations by $\mathcal{O}(N)$. Moreover, for the probabilistic greedy algorithm we would be selecting each initiator one-by-one, each time having to update the impact of each node by re-running simulations. Here, we select q initiators (instead of one-by-one, where $q = \lceil sN \rceil$, and s is a fraction of N), thus reducing the complexity by another $\mathcal{O}(N)$. Also, when randomly picking nodes as test-initiators for our metric, we do not predefine the order of selecting initiators, but we pick them randomly one by one after ensuring that they did not get activated. Finally, typically the quantity we wish to maximize is the cascade size for a specific

number of initiators, which is essentially our cost budget. However, the impact of even a small fraction of initiators can potentially have a large impact on the cascade size, especially near a tipping point. That means, that we could potentially be near a tipping point, but since we have a limited budget, we missed it. To address this, we aim to minimize the size of the initiator set in order for the cascade size to be at least as large as a specific predefined size S_{goal}. However in general, GPI can also be used when constraining on the cost budget, or computational time.

Let us start with the initial Graph $G(V, E, r)$, where $V(G)$ is the set of nodes, $E(G)$ is the set of edges of the graph, while r is the resistance of each node. Our goal is to find the initiator set Y such that $S(Y) \geq S_{\text{goal}}$ [Alg. 1]. To do so, at each step t we select $q = |Q|$ nodes as initiators placing them in Y, with $q = \lceil sN \rceil$, where Q_t is the set the initiators selected at step t, that is $Y = \cup_t Q_t$. The number of active nodes required to get activated is $d = \lceil S_{\text{goal}} N \rceil$. In every step t, we need to find the q nodes with the highest GPI-ranking (we will define it below), which we place in Q. Then, we include Q in the initiator set Y, compute the cascade induced from that set, update the spread size, and update H (reduce the resistance, and remove all activated nodes and their edges). We define the function $f(Q_t|H_t)$ to express the number of nodes that got activated at step t from the current bunch of initiators Q_t at the inactive subgraph H_t.

Now, at any step t we look for the nodes with the highest GPI value, that is, the nodes which when present in the initiator set, the desired cascade size is (on average) reached faster (smaller initiator set size). To measure the expected GPI for each step t we run simulations till the desired cascade size is reached, for a total of v times. The simulations are run on the graph H_{test}. In particular, at the beginning of every step j, we set $H_{\text{test}} = H_t$ and start with the empty set X_j. We keep placing randomly selected inactive nodes as test-initiators one-by-one on X_j and run the simulations on H_{test} (which is updated for the spread caused by every test-initiator), until the desired spread size has been reached, that is until $f(X_j|H_{\text{test}_j}) \geq d - \sum_t f(Q_{t-1|H_{t-1}})$. The metric can be expressed analytically as

$$\text{GPI}_i = \frac{\sum_{j=1}^{r} |X_j| \, x_{j,i}}{\sum_{j=1}^{r} x_{j,i}} \tag{3.4}$$

where, $x_{j,i}$ represents whether a node i is included in the set X_j; if $x_{j,i} = 1$ then $i \in X_j$, else $x_{j,i} = 0$. Since at the beginning of each step j, X_j is empty, $x_{j,i}$ for each node is zero. Once we have the GPI for each node, we select the $q = |Q_t|$ highest ranked nodes and place them in Y, that is $Y = Y \cup Q_t$.

The nominator of GPI_i is the cumulative of the sizes of the randomly selected initiator sets $|X_j|$ in the presence of node i. Since we select inactive nodes uniformly at random, nodes do not appear in the initiator set equal number of times. If that was not the case (that is if all nodes will be equally frequently chosen), just like for the probabilistic greedy we mentioned above, the nominator of the fraction of Eq. (3.2) would be enough to be used as a metric, where the smaller the cumulative the larger the impact of node i would be. The presence of the denominator is necessary to normalize the number of times node i is selected as an initiator. In addition, because we only select inactive nodes; nodes which are likely to be activated through spread, that is typically nodes with low resistance and high in-degree, are going to be part of the initiator set less frequently than other nodes. And so, nodes with a large number of appearances are nodes are less likely to be activated than others.

A node may on average contribute to the best performance over any set of randomly selected initiators, yet it may not be a part of the optimal initiator set. Since GPI deals with the expected impact of nodes, it is by default slower than the rest of the strategies but can potentially find much better initiator sets aiming for the global optimum (instead of local one). Estimating the time complexity of GPI, it takes $\mathcal{O}(N)$ steps to go through all the nodes and it takes $\mathcal{O}(\langle k \rangle N)$ to measure the cascade size caused by any of them. It further takes v times to repeat this process and approach the expected GPI values, adding a factor $\mathcal{O}(N)$. Furthermore, the number of initiators selected adds another $\mathcal{O}(N)$ factor when selecting them one-by-one, or it takes just an additional constant when selecting a fraction of nodes as initiators. Thus, the total complexity of GPI is $\mathcal{O}(v\langle k \rangle N^2)$ for fractional addition and $\mathcal{O}(v\langle k \rangle N^3)$ for one-by-one addition of initiators. However, for the case of identical threshold for all nodes, since there is a sharp phase transition point (and small spread otherwise), the total complexity drops by an $\mathcal{O}(N)$ factor [179]. For the remaining of the paper, unless otherwise specified, the control parameters of

Algorithm 1 Group Performance Index Algorithm

procedure Group Performance Index Algorithm

Input Graph $G(V, E)$ and thresholds ϕ_i for each node i

Input the desired cascade size (fraction) S_{goal}, the fraction s, the number of randomizations r

Initialize the initiator set $Y = \emptyset$, the integer cascade size induced by the current Y, $S = 0$

Get the resistance of each node, $r_i = \lceil \phi_i k_i \rceil$, the number of initiators selected at step t, $q = \lceil sN \rceil$, and desired integer cascade size $d = \lceil S_{\text{goal}} N \rceil$

Initiate step counter $t = 0$

we start the graph reduction from G, that is $H = G$

while $S < d$ **do**

$t \leftarrow t + 1$

Initialize GPI_i ▷ average impact GPI for node i

Initialize no_i ▷ nominator of Eq. (3.4) for node i

Initialize de_i ▷ denominator of Eq. (3.4) for node i

$Q = \emptyset$ ▷ set of top q GPI-ranking nodes

for $j = 1 : v$ **do**

$H_{\text{test}} = H$

Initialize $X = \emptyset$ ▷ set of test-initiators

Initiate the local integer test cascade size $S_l = 0$

while $S_l < d - S$ **do**

Randomly select an inactive node i as test initiator

$X \leftarrow X \cup \{i\}$ ▷ Include i to the initiator set X

Run the cascade on H_{test} induced by i:

compute the additional cascade size $f(i)$, and update S_l, $S_l \leftarrow S_l + f(i)$

insert all newly activated nodes to X

$de_{i \in X} \leftarrow de_{i \in X} + 1$, $no_{i \in X} \leftarrow no_{i \in X} + |X|$

reduce the resistance of all affected nodes and remove the inactive nodes and their edges (that is, update H_{test})

$\text{GPI}_i = no_i / de_i$, for all $i \in H$

The top q GPI-ranking nodes are inserted in Q, and also in Y, $Y \leftarrow Y \cup Q$

Run the cascade on H induced by the Q:

compute the additional cascade size $f(Q)$, and update S, $S \leftarrow S + f(Q)$

reduce the resistance of all affected nodes and remove the inactive nodes and their edges (that is, update H)

GPI we are using are $s = 10^{-3}$, $v = 10^5$, and $S_g = 0.5$.

3.3 Results

We are comparing the performance of the strategies for the entire parameter space of network assortativity ρ, and threshold distribution with fixed average threshold $\overline{\phi} = 0.50$ and varying standard deviation σ. For more information on the methods for the generation of ER graphs, and graphs with controlled assortativity see the Appendix B.2. In Fig. 3.1 we present our main results for the ensemble of ER graphs for the extreme cases of high positive ($\rho = 0.9$) and high negative ($\rho = -0.9$), and $\rho = 0$ assortativity, measured with Spearmans ρ [188]. Furthermore, we examine the cases of a identical thresholds ($\sigma = 0$), a uniform threshold distribution ($\sigma = 0.287$), and for some truncated normal distribution in between ($\sigma = 0.2$). For GPI strategy, we present for which critical initiator fraction the S_{goal}. First, focusing especially on ER graphs ($\rho = 0$) we notice that as we move from a threshold distribution with standard deviation $\sigma = 0$ to larger σ, there is change from a first-order phase transition to a smooth crossover also seen for randomly selected initiators in [123]. Interestingly, in the case of the uniform threshold distribution ($\sigma = 0.287$) at the ensemble level, we observe that all the direct methods appear to have diminishing returns with increasing cascade size, that is, the contribution in the cascade size of any additional initiator in an initiator set is diminishing as the initiator set is increasing larger.

As far as the performance of the strategies is concerned, the degree (deg) strategy's relative performance is decreasing for larger σ's, while the resistance strategy's performance is increasing. In addition, CI-TM which incorporates a network structure decomposition using the neighboring nodes with the information about resistance $v = 1$ in it as metric, is out-performing the degree strategy for the case of $\sigma = 0$ and $\rho = 0$ but it is not performing as well in the rest of cases. On the other hand, the RT approach is outperforming the degree, the resistance and the CI-TM strategy for all cases of ρ and σ. Naturally, the introduced weighted strategy is outperforming in all cases the strategies that incorporates their ranking metrics (deg, res, CI-TM, RD, RT), especially for $S_{eq} = 0.5$ which is what we optimized it for

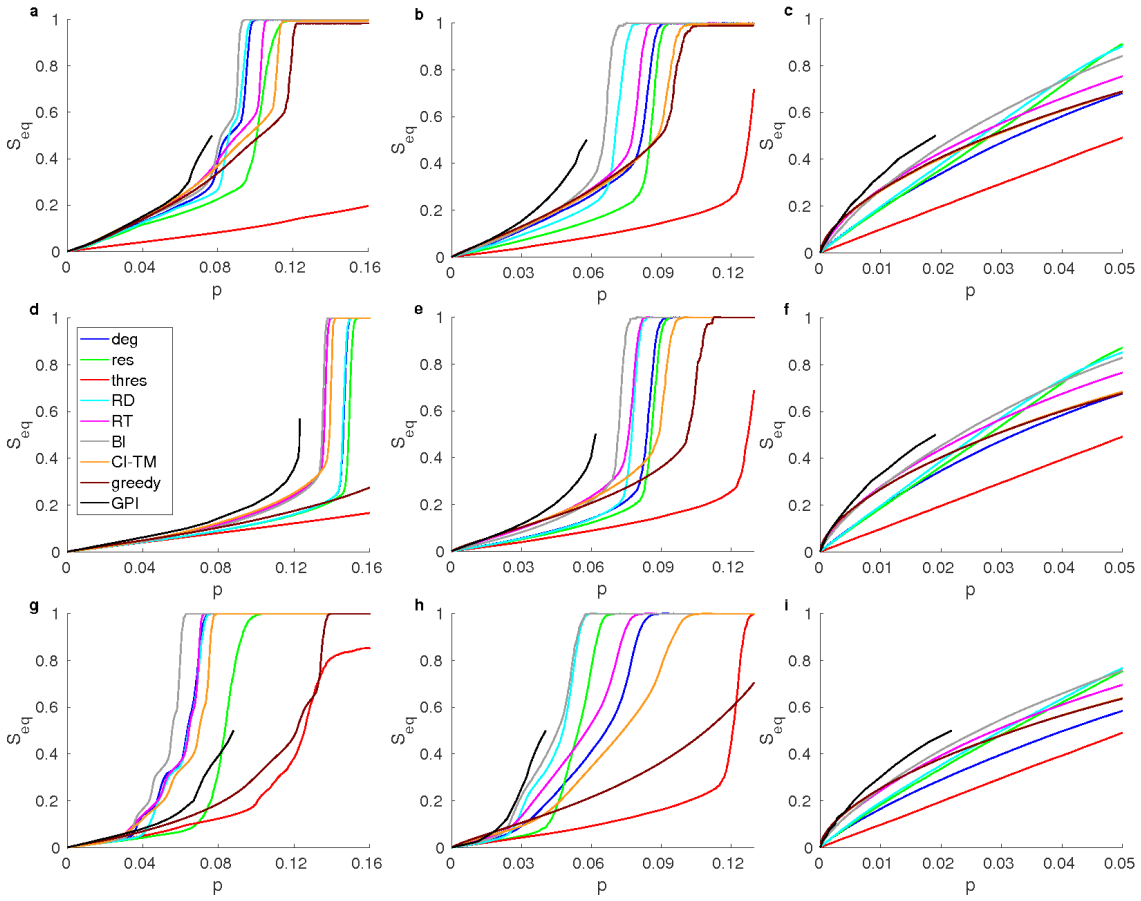

Figure 3.1: Comparison of the BI and GPI (for different S_{goal}) selection strategies in terms of cascade performance S_{eq} for (a-b-c) ρ=−0.9, for (d-e-f) ρ=0, for (g-h-i) ρ=0.9, for (a-d-g) σ=0, for (b-e-h) $\sigma = 0.20$, for (c-f-i) σ=0.2887, with $\overline{\phi}$=0.5, averaged for 500 different network realizations (except for the GPI which is 20) each with a different threshold generation, applied on ER graphs with N=10000 and $\langle k \rangle$=10.

here.

The GPI strategy largely outperforms all other strategies in all cases except for the case of very high assortativity ρ with identical thresholds ($\sigma = 0$). Yet, with even lower s and higher v the performance of GPI improves. Since, this method is computationally expensive, we have used a $s = 10^{-3}$ fraction of initiator addition with $v = 10^5$, the best resolution we could achieve for a graph with size N would be $s = 1/N$, by inserting the initiators one-by-one.

To further study the direct methods, we evaluate their performance in the

assortativity space ρ (Fig. 3.2) and the threshold distribution space σ (Fig. 3.3) for $S_{eq} = 0.5$. Evaluating through ρ for $\sigma = 0$, we observe a highly non-monotonic behavior for all the strategies. Between the range of approximately $-0.8 \leq \rho \leq 0.4$, CI-TM, is outperforming the degree strategy, which is approximately the same regime, in which RT is outperforming all other direct strategies. Yet, for $\rho \geq 0.5$ RD is the best strategy. With increased deviation of the thresholds $\sigma = 0.20$ (Fig. 3.2b), the performance of the strategies which depend more on the network structure, like *deg* and CI-TM, is getting worse, while strategies which give higher importance to the resistance of the node, like *res*, RD and RT are performing better. Finally, for a uniform threshold distribution $\sigma = 0.2887$ (Fig. 3.2c), we observe that strategies show a convex response to ρ, with RT being the best strategy (for a desired cascade size $S_{\text{goal}} = 0.5$). On the other hand, the threshold (*thres*) strategy appears to be independent of ρ for thresholds generated randomly with $\sigma = 0.2887$. Finally, from Fig. 3.3 we observe that RT has the best performance compared to all the direct strategies over nearly all of the σ range, making it the best overall strategy (excluding the weighted and GPI strategies).

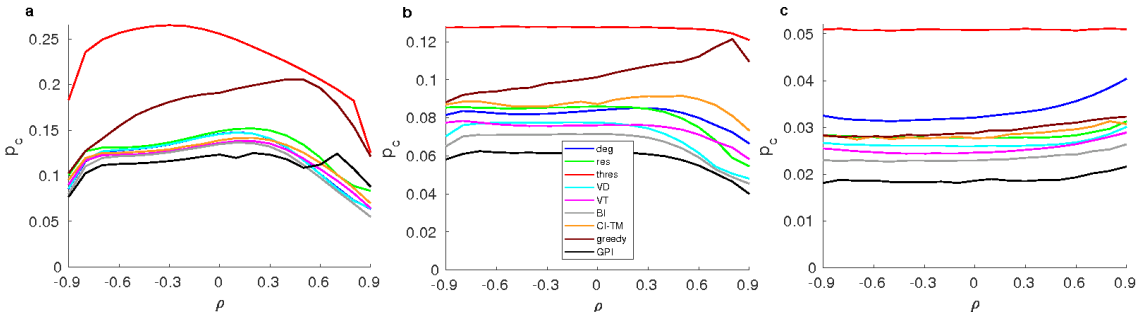

Figure 3.2: Initiator fraction p_c required to reach spread S_{eq}=0.5 vs. degree assortativity ρ for graphs with N=10000 and $\langle k \rangle$=10 for (a) σ=0, for (b) σ=0.20, and for (c) σ=0.2887 with $\overline{\phi}$=0.5, averaged for 500 different network realizations (except for the GPI which is 20) each with a different threshold generation.

Furthermore, we are interested in studying not just the average performance of each strategy but also the probability of each being the best strategy for any initiator fraction p (Fig. 3.4) for a fixed network, and for various assortativity and threshold distribution cases. The actual cascade size S_{eq} vs. initiator fraction p can be seen in Fig. 3.5 (average) and Fig. 3.6 (50 runs). Since the GPI strategy

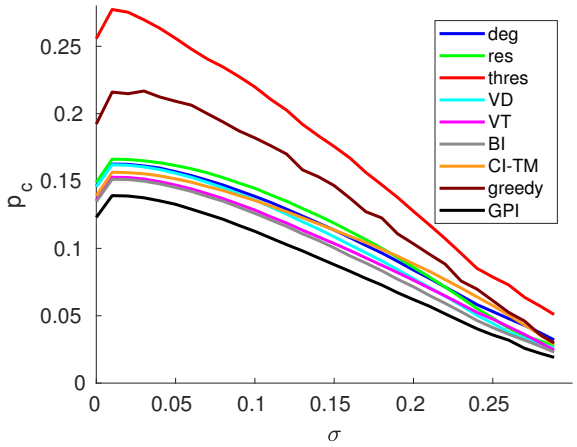

Figure 3.3: Initiator fraction p_c required to reach spread $S_{eq} = 0.5$ vs. the standard deviation of the generating threshold distribution σ with $\overline{\phi}$=0.5 for graphs with $N = 10000$, $\langle k \rangle = 10$ and a ρ=−0.9, b ρ=0 c, ρ=0.9, averaged for 500 different network realizations (except for the GPI which is 20) each with a different threshold generation.

is optimized for a specific S_{goal} each time, we have not included it here. First, we notice that the greedy algorithm is in all cases outperforming all other strategies for a small initiator fraction, while for a very large initiator fraction all strategies have the same probability, because all of the network is activated. For an ER graph ($\rho = 0$) with $\sigma = 0$ the spread is minimal until the phase transition point has been reached, and CI-TM leads until the tipping point for RT has been reached. Since the metric of CI-TM takes into account the out-degree of the nodes, and does not consider the effort required to activate nodes with high resistance and low in degree, it is outperforming other methods for smaller initiator sizes, but eventually gets surpassed, when the nodes with high resistance with a structural importance have not been activated. However, for very low ($\rho = -0.9$) and high ($\rho = 0.9$) degree assortativity, there are multiple initiator fractions for which a large spread occurs, which vary for the strategies, allowing for a sudden change of 'lead' between the four network depend strategies (CI-TM, RT, deg, and RD), with RD and RT performing the best. For larger σ the importance of resistance increases while the importance of the network structure declines, making the strategies more depended on resistance to take the 'lead', while also reducing the number of sharp 'lead' changes.

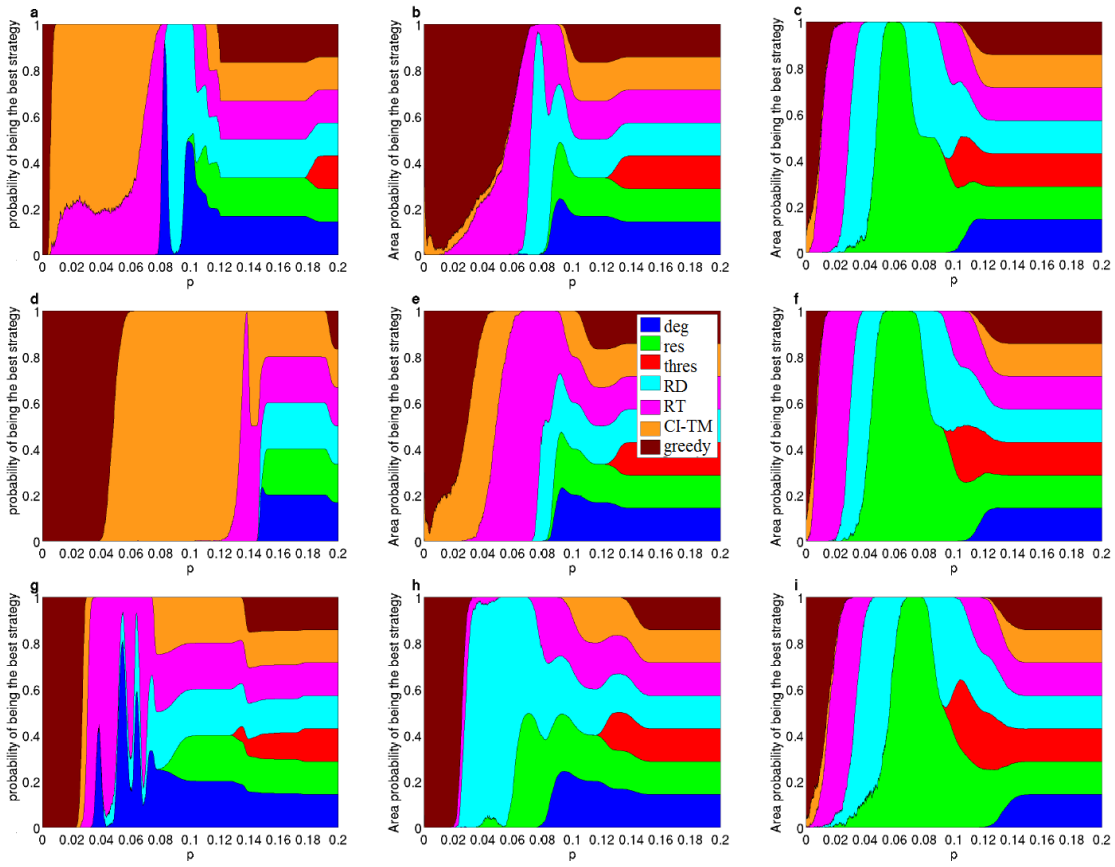

Figure 3.4: Probability of each strategy being the best strategy for one network with N=10000, $\langle k\rangle$=10 and for (a-b-c) ρ=−0.9, for (d-e-f) $\rho = 0$, for (g-h-i) ρ=0.9, for (a-d-g) σ=0, for (b-e-h) $\sigma = 0.20$, for c-f-i σ=0.2887, with $\overline{\phi}$=0.5, for 500 threshold generations (same for each strategy).

Next we focus on the BI and GPI strategies and their performance for their different parameters. For the BI strategy, we scan the parameters space $a \times b$ and record the average minimum p_c at which the cascade size is $S_{\text{goal}} = 0.5$ (Fig. 3.8). The second and third feature of this weighted method Eq. (3.3) correspond to the first two terms of the CI-TM metric. The three features are interdependent, e.g. before the cascade begins with $\sigma = 0$, a node's i resistance r_i is nearly linearly proportional to its degree k_i, which is why we observe those linear contours on the plots a, d, e. Moreover, it is clear that any strategy that would exclude the resistance ($a = 0$) from its metric, such as the CI-TM, will have inferior performance. The contours indicate that the importance of the resistance and degree of a node

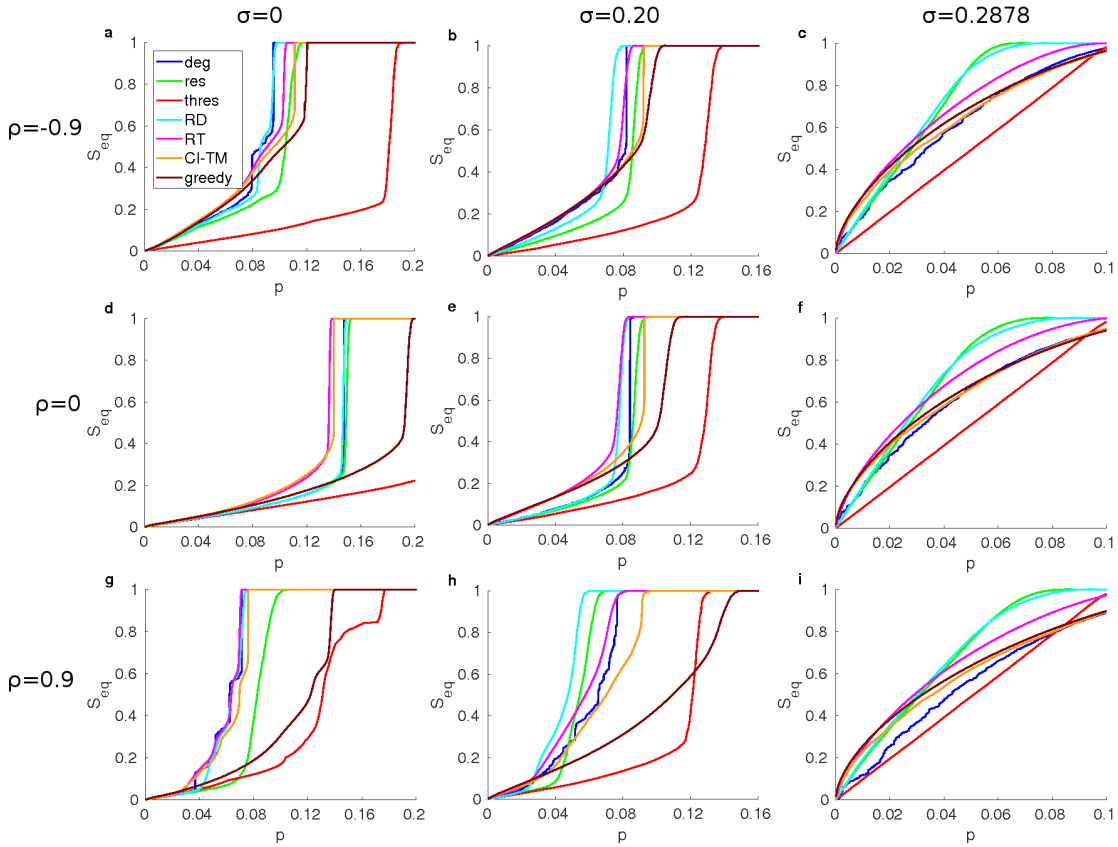

Figure 3.5: Comparison of average cascade performance S_{eq} for one network with N=10000, $\langle k\rangle$=10 and for (a-b-c) ρ=−0.9, for (d-e-f) ρ=0, for (g-h-i) ρ=0.9, for (a-d-g) σ=0, for (b-e-h) σ=0.20, for c-f-i σ=0.2887, with $\overline{\phi}$=0.5, for 500 threshold generations (same for each strategy).

is much higher that the third feature, which in addition is computationally most costly to obtain. Furthermore, we have recorded the impact of the standard deviation σ on the optimal weights (Fig. 3.7). As σ increases, the optimal c coefficient decreases. Interestingly, the most important feature is the resistance ($a \approx 0.53$), then the degree ($b \approx 0.32$), and the smallest importance is left for $c \approx 0.15$. This result is especially important since other strategies do not fully utilize the resistance information combined with other network centrality measures.

For the GPI strategy, on Fig. 3.9 and Fig. 3.10 we present the cascade size behavior for various numbers of randomizations v and step sizes s respectively. On average, with increasing v and decreasing s we always minimize p_c for obtaining

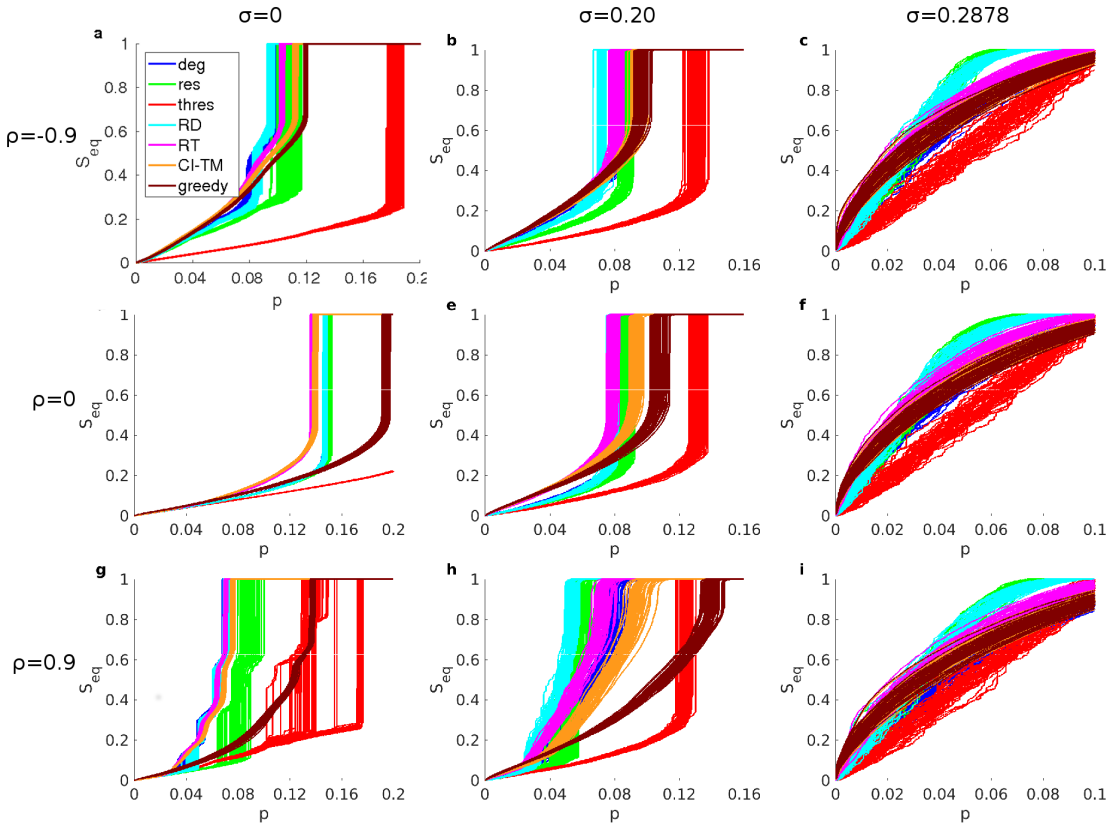

Figure 3.6: Cascade performance S_{eq} for 50 threshold randomizations for one network with N=10000, $\langle k\rangle$=10 and for (a-b-c) ρ=−0.9, for (d-e-f) ρ=0, for (g-h-i) ρ=0.9, for (a-d-g) σ=0, for (b-e-h) σ=0.20, for (c-f-i) σ=0.2887, with $\overline{\phi}$=0.5, for 50 threshold generations (same for each strategy).

a cascade size (here we aim at $S_{\text{goal}} = 0.5$). On average, we expect and observe an asymptotic return with increasing v. For computational efficiency, we fix the s and v when the additional performance is minimal. Further investigation is required in order to find the interplay between the two control parameters in order to optimize the performance of the algorithm for the smallest computational time possible. Interestingly, for $\sigma = 0.2878$, in contrast to the direct methods, there is a large transition to cascade size as we reach S_{goal}.

3.4 Discussion

The challenge of Influence Maximization for the LTM or other diffusion processes is finding low complexity, yet well performing algorithms for the discovery

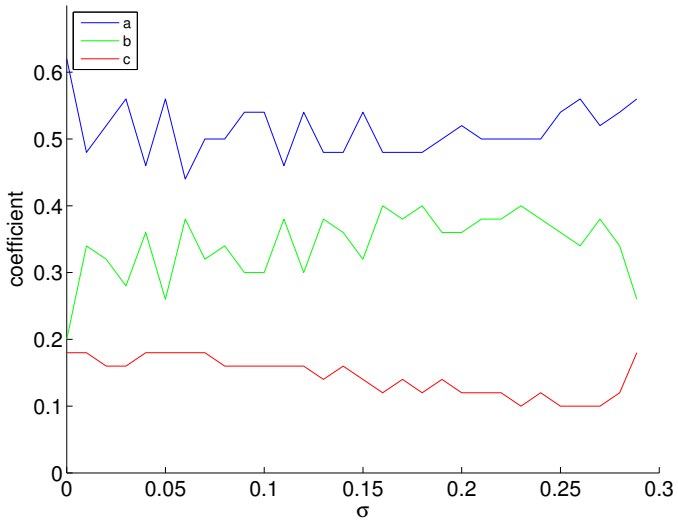

Figure 3.7: Impact of standard deviation σ on the optimal weights (from Eq. (3.3), with a+b+c=1) for desired cascade S_{goal}=0.5 for ER graphs with N=10000, $\langle k\rangle$=10, $\rho = 0$, with $\overline{\phi} = 0.5$, averaged for 500 different network realizations (except for the GPI which is 20) each with a different threshold generation. The resolution in the a and b weight space is 0.02.

of superspreaders. Most strategies do not directly consider the combination of initiators but rather use some heuristic/analytical metric, leading to local solutions of reduced time complexity. Among the strategies discussed here, by default our weighted strategy is outperforming strategies that take into account the same features as the weighted one. Our weighted strategy was inspired by the combination of analytics (CI-TM [179]) and computational methods. The great performance of our weighted strategy demonstrated that that mixing of methods can be very successful. Furthermore, it brings a connection between different selection strategies discussed here. In addition, by scanning the weights space we discovered that between resistance, degree, and 2nd neighborhood of actived nodes, resistance is the most important feature. Hence, the cascade size is governed not by initiators with high network centrality measures but by low resistance nodes, a result also supported by Watts and Dotts [126]. Finally, we show that even when the weights are not optimized, but are equal among the features (RT strategy), we can still obtain better results than the CI-TM (for at least the cases of bidirectional random graphs that we explored) for the same time complexity $\mathcal{O}\left(\langle k\rangle N \log N\right)$ (which is reduced

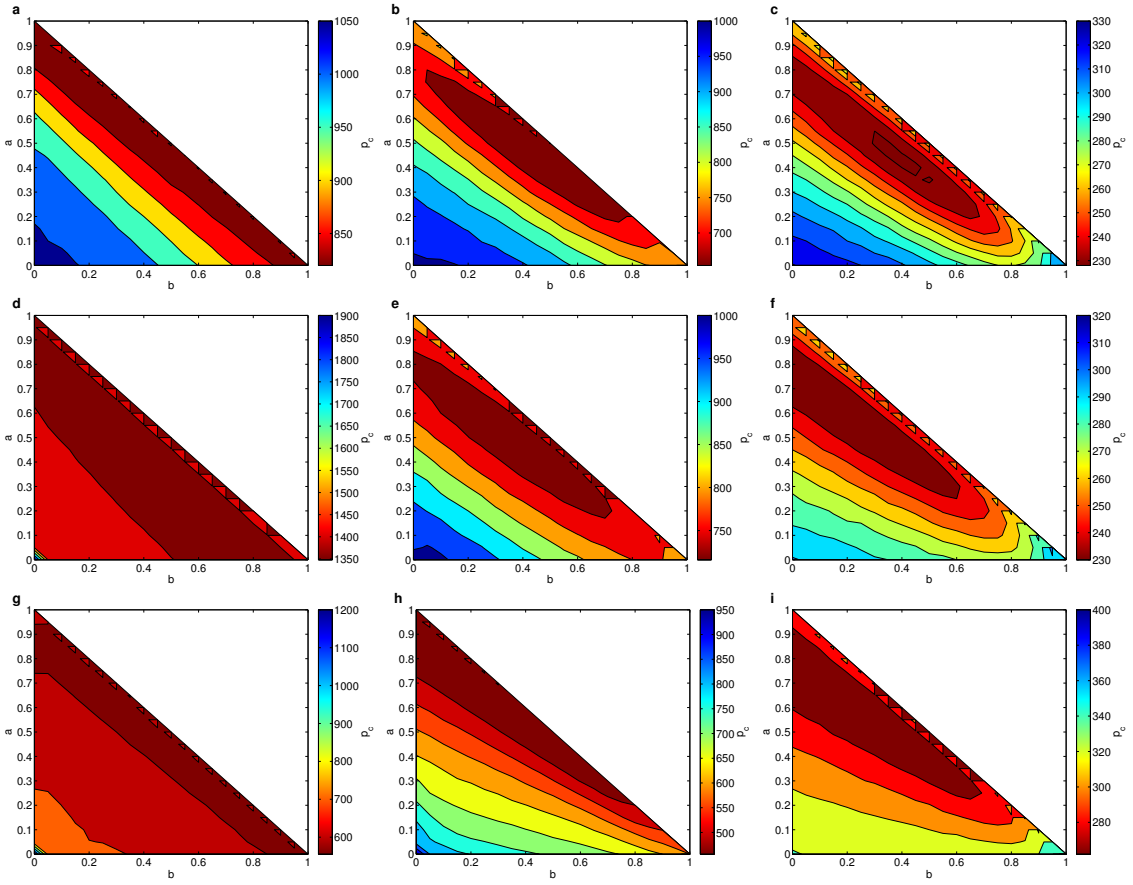

Figure 3.8: Contours of p_c for reaching S_{eq}=0.5 by controlling parameters a and b (from Eq. (3.3), with a+b+c=1) for graphs with N=10000, $\langle k\rangle = 10$ and for (a-b-c) ρ=−0.9, for (d-e-f) ρ=0, for (g-h-i) ρ=0.9, for (a-d-g) σ=0, for (b-e-h) $\sigma = 0.20$, for (c-f-i) σ=0.2887, with mean threshold $\overline{\phi}$=0.5, for 500 repetitions, averaged for 500 different network realizations (except for the GPI which is 20) each with a different threshold generation. The resolution in the a and b weight space is 0.05.

by a factor $\mathcal{O}(\langle k\rangle)$ for sparse graphs).

On the other hand, strategies using combinations of initiators, such as GPI, can have higher performance by targeting a global optimum, at the expense of higher complexity, $\mathcal{O}(\langle k\rangle N^2)$ (which is reduced by a factor $\mathcal{O}(\langle k\rangle)$ for sparse graphs). Those methods can be further improved by utilizing the network and model specific properties (LTM in this case), analytics or even learning methods. More research is required to elegantly mix those methods to deal with the challenge of Influence Maximization. For those investigations, GPI serves as a benchmark for (synthetic)

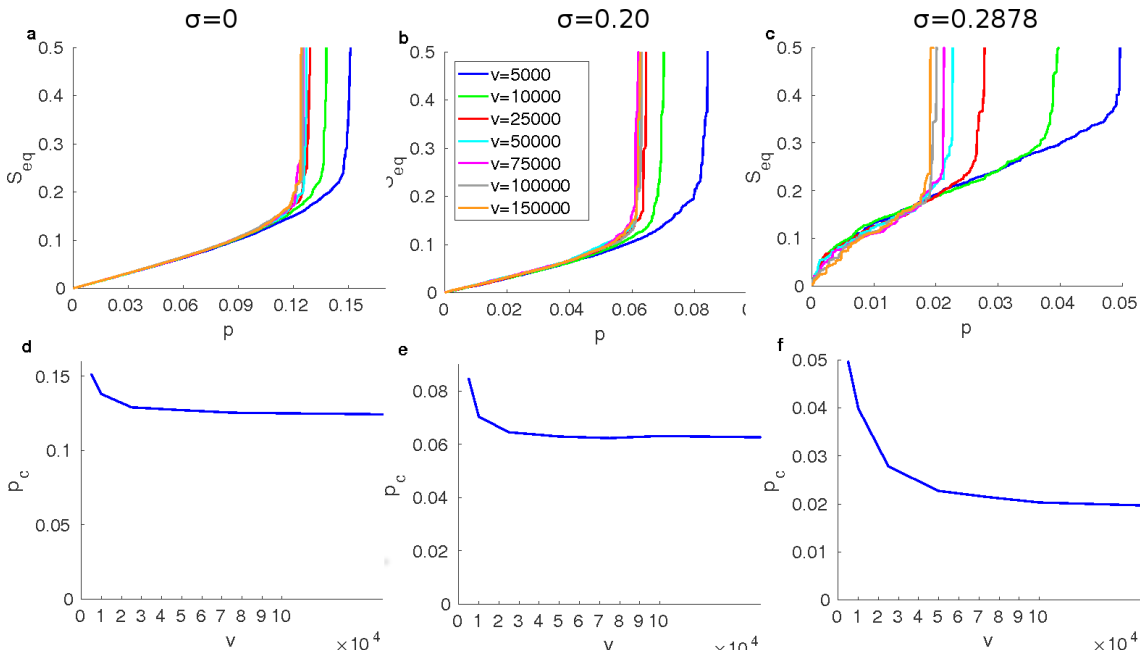

Figure 3.9: Impact of the number of randomizations v on the performance of the GPI strategy for s=0.0025 for desired cascade S_{goal}=0.5 for ER graphs with N=10000, $\langle k\rangle$=10, ρ=0. (a-b-c) cascade S_{eq} vs. the initiator fraction p for σ =0, 0.2, 0.2887 respectively (for one realization). (e-f-g) initiator fraction p_c required for desired cascade S_{goal}=0.5 vs. randomizations v for σ=0, 0.2, 0.2887 respectively (for one realization).

graphs and sets a minimum bound for the optimal initiator set. Finally, in terms of applicability, both strategies can be used for directed and weighted graphs as well, although a few adjustments would have to be made for the weighted strategy in case of weighted graphs.

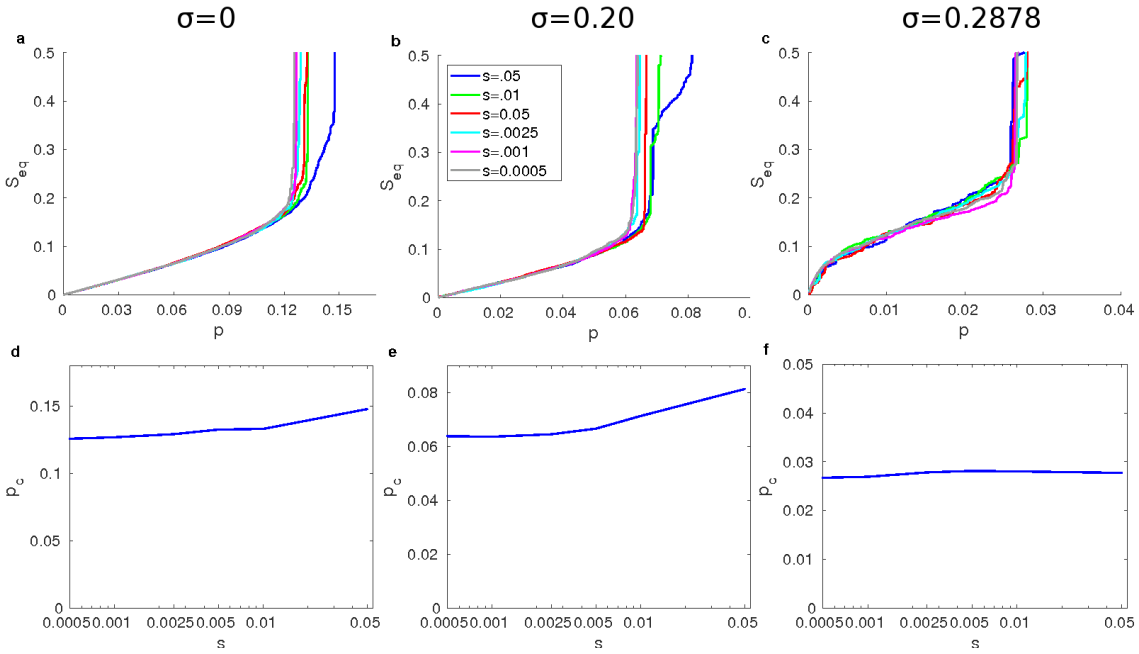

Figure 3.10: Impact of the size of initiator fraction s on the performance of the GPI strategy for randomizations v=25000 for desired cascade S_{goal}=0.5 for ER graphs with $N = 10000$, $\langle k\rangle$=10, ρ=0. (a-b-c) cascade S_{eq} vs. the initiator fraction p for σ=0, 0.2, 0.2887 respectively (for one realization). (e-f-g) initiator fraction s required for desired cascade $S_{\text{goal}} = 0.5$ vs. randomizations v for σ=0, 0.2, 0.2887 respectively (for one realization).

CHAPTER 4
Peer-to-Peer Lending and Bias in Crowd Decision-Making

4.1 Introduction

The idea of a new level playing field where global economic equality gradually improves is seductive [190]. Models of financial markets suggest that international capital flows are reaching more countries [191] and dominating national institutional policies [192], thereby laying a groundwork for global equality in access to capital that can promote new possibilities for prosperity among the worlds poor [193–202]. However, others have countered that outside of a handful of cities/countries the vast majority of economic activities (e.g., institution and government investment, web traffic, and telecommunications) have remained domestic over time [203, 204]. Such critiques of flat world can be explained by the Lucas Paradox [205], which states that "capital does not flow in relatively large amounts from developed countries to developing countries despite the fact that developing countries have lower levels of capital per worker". As crowdfinancing grows, is it a flat-world mechanism for creating opportunities for the worlds poor, or is it becoming biased similar to other established economic activities? Nevertheless, the Lucas Paradox indicates that counterintuitively the liberalization of international capital regimes has not produced an open club, but a rich club – that is, a group of countries with similarly well-developed monetary institutions, cultures, and wealth that display in-group preferences [206] in lending and borrowing, thus restricting capital to poor nations [207, 208].

New data on global crowdfinancing allows questions to be asked about the role of peer-to-peer lending networks in leveling global capital financial flows and development. Crowdfinancing is a recent innovation. It enables private lenders and borrowers to find and directly interact with one another through a website. Private

Portions of this chapter to appear in: P. Singh, J. Uparna, P. D. Karampourniotis, E.-A. Horvat, B. K. Szymanski, G. Korniss, J. Z. Bakdash, and B. Uzzi, "Peer-to-peer lending and bias in crowd decision-making," (submitted to Manag. Sci., 2017).

individuals on the website, from theoretically anywhere around the world, can lend or borrow capital directly from each other. Borrowers put forth their reasons (see Appendix Fig. C1) and make requests for capital directly to lenders; in turn, lenders make their lending decisions free of institutional constraints. In this way, peer-to-peer lending sidesteps the long-standing institutional arrangements and cultural norms that have up to this point characterized lending [209] (see Appentix C 3.4.1 for a comparison between Kiva and government aid between countries). Crowdfinance offers an alternative and/or supplemental mechanism to more institutionalized forms of foreign aid. The flow of such aid is associated with increased stability, such as reductions in terrorism [210]. However, the success of foreign aid is marred by corruption, political changes, and other factors (e.g., see [211]). Patterns in crowdfinance is associated with corruption in the country [212]. Thus, crowdfinance provides a potential mechanism for unmediated, direct aid especially if it is flat in terms of opportunities for the poor.

Despite the possibility for crowdfinancing to level the playing field in capital flows, its potential is debated [213] and empirical patterns are largely unknown [214]. One critical association between peer-to-peer lending and global financial flows concerns the flat-world hypothesis [190]. The flat-world hypothesis holds that crowdfinancing counter-balances lending biases by acting as a functional substitute for capital from traditional lenders and lending institutions ([190], see pp. 492–493). However, the increased interconnectedness may also potentially make the world less flat by reinforcing the existing global or individual level biases. If the flat-world hypothesis is correct, peer-to-peer lending systems should display no preferential attachment of capital flows between lender-borrower pairs [215].

To examine the flat-world hypothesis, we analyzed the total aggregate lending of over half a billion dollars in over 600,000 peer-to-peer loans made on one of the largest and well-regarded crowdfinancing websites in the world, Kiva from its inception in 2005 to 2013 [216]. Loans from private individual lenders in more than 220 countries were made to private borrowers in 80 nations. Kiva is philanthropic in nature and lenders receive their capital back without interest and borrowers receive loans without paying interest. By comparison, the aggregated (2005–2013) gov-

ernment aid for the same time period involves 48 donor countries making interest-based loans (data from AidData [217] see Appentix C). Our study examines three related questions about crowdfinancing. First, to test whether crowdfunding loans are associated with a flatter world, we measure the degree of flatness in the lending system. A flat world has capital flows that display no preferential attachment between lender–borrower pairs [218]. To quantify flatness, we randomly rewire the observed co-country network of loans, which creates a hypothetical Kiva network wherein the propensity for any lender–borrower transactions is no greater than expected by chance. Deviation from the expected null network of flows reflects choice in lending and hence a less flat world [218]. Second, we investigate the potential susceptibility of the Kiva network to shocks that could change the systems ability potential for flatness. Shocks to lending systems include national policy changes, market collapse, climate change, health or security risks [219], and have been shown to dramatically alter capital flows [225]. We represent these hypothetical changes in the system as the disappearance of network nodes or links [45, 48, 220] and then observe their simulated effects on the network structure and its flatness. Third (this part was done by our collaborators, for more information see Appendix C), we use regression analysis to predict bias in country–pair transactions based on variables such as GDP, geographical distance that are typically used in gravity models of trade [226, 238]. Although previous studies [214, 221] have investigated the biases associated with lending on Kiva, our study presents a longitudinal analysis for a longer observation window (2005-2013) (see Apendix C). Since the number of participating borrower countries as well as the transactions have grown significantly in the later years, it becomes important to account for yearly changes in the network as opposed to treating it in a cross-sectional fashion (see Appendix C). Nevertheless, some of the factors that we find associated with lending bias are qualitatively consistent with the findings of Burtch et al. [214].

4.2 Data

Crowdfinancing networks differ in orientation. Some lending systems provide funds in exchange for equity in an investment (e.g., Equitynet.com, CrowdCube.com,

Seedrs.com) versus financial return (e.g., Prosper.com) versus interest-free loans made for the developmental aid of the borrower (e.g., Kiva.com). Our dataset of lenders, borrowers, and loans includes all transactions made on Kiva.com, 2005–2013. The vast majority of loan contributions are made in multiples of $25.00 and most loans are for $25.00 and $50.00. These loans typically support purchases of machinery for petty entrepreneurs, livestock for farmers, or domestic items such as water purification systems that improve living conditions (see Appendix Fig C.2). For each loan we know the time of effectuation, size of the loan, the location of the lender and borrower as well as the specific Kiva field partner, that is, a representative of Kiva who provides access to computers to potential borrowers, helps them translate or edit their requests for a loan into English, and manages lender-borrower transactions. We constructed a yearly co-country (multi-edge) network aggregated from the country-to-country transactions (an example is shown in Fig. 4.1). Loans to compatriots (i.e., self-loops in the network), are allowed. Fig. 4.2 (A–G) summarizes the growth of the co-country network and shows that money lent in the form of loan contributions and the number of participating borrower and lender nations grew dramatically on Kiva between 2005 and 2013. A few lender countries account for a large portion of the loan transactions. Fig. 4.2(F) shows the top 5 lender countries and their share of transaction volumes by year. It can be seen that these 5 countries together account for about 80% of all observed contributions with the US alone being responsible for more than half of the contributions. The top 5 borrower countries benefit from a large portion of the total contribution, but there is no clear outlier and there are many countries with a similar share of received contributions (Fig. 4.2 [G]). The same trend can be seen in terms of the degree distribution of the network. The in-degree (out-degree) of a country is the sum of transactions made to (by) that country. Fig. 4.2 (D) and (E) show that both in-degree and out-degree distributions are skewed (log scale), but the out-degree distribution is highly skewed (i.e., a few lender countries provide a very large portion of the observed transactions).

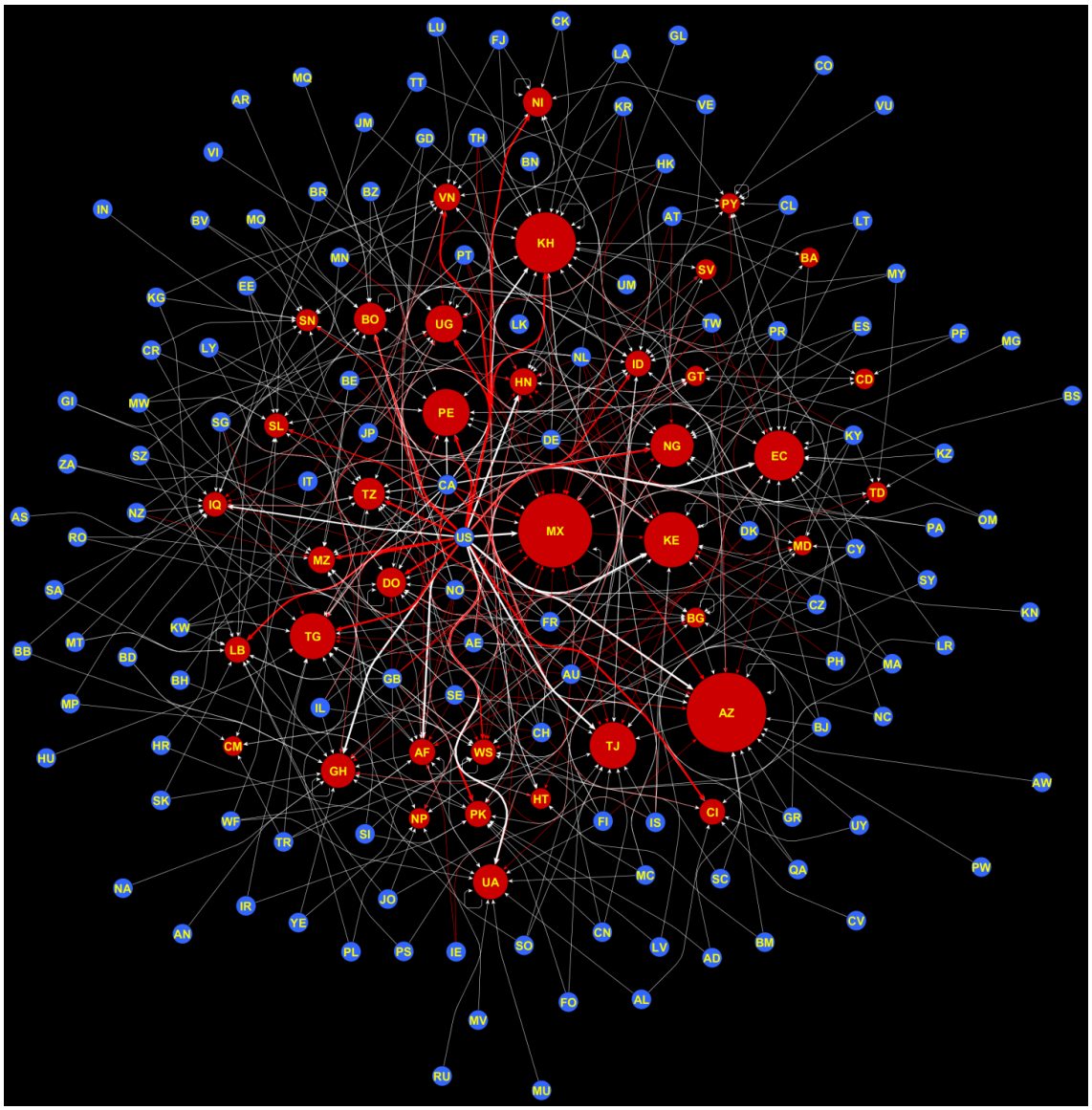

Figure 4.1: Biased links in the Kiva network. Visualization of positively (colored white) and negatively (colored red) biased links in the Kiva co-country network for 2007. Borrower countries (nodes) are shown in red with size proportional to the total transactions received by that country; whereas, lender countries are shown in blue and all nodes are of the same size. The link thickness corresponds to the actual number of transactions made between the country–pairs.

4.3 Results

To analyze the structural property of the network, we used degree-preserving network randomization, a common technique for assessing the statistical significance

of observed network properties, including biased links between nodes [1,199,222,224]. Using the randomization method for weighted (multiedge) networks, we generate many synthetic networks by randomly rewiring the loan transactions in the observed network [223] while preserving the total transactions made to and from, for each country (i.e., in- and out-degree of every node). Many synthetic networks provide a distribution of every bilateral exchange, giving an expected mean and standard deviation across all links in the network, which are used to determine how far observed relationships are from expected values. A comparison between the null model and the observed data enables us to identify country-level lending biases in this network – that is, which countries have a lending–borrowing relationship that is greater or smaller than expected by chance, where chance theoretically reflects a system without bias [215]. To measure the flatness of the lending network, we count the number of country–pairs (positive as well as negative) where the observed links are statistically different from what is expected using a z-score for each pair of countries. The z-score z_{ij} of any link ij is given by

$$z_{ij} = \frac{O_{ij} - E_{ij}}{\sigma_{ij}} \tag{4.1}$$

where O_{ij} is the observed number of transactions from a country i to country j. $_{ij}$ and σ_{ij} are the expected number of transactions and the associated standard deviation according to the null model. For a country–pair, the z-score provides a normalized and relative measure of how far away the observed number of transactions is from what is expected by chance Fig.4.3. A pair is classified as biased if its observed number of transactions is 2 standard deviations above or below the null model ($p < .05$). Here p refers to the p-value used in statistical hypothesis testing. For a probability distribution, the p-value is the cummalitive probability of a particular value or any value above not being generated by the probability distribution. For the Normal distribution a (double sided) p-value of 0.05 corresponds to $z = 1.96$ standard deviations. The flatness is then given by the fraction of unbiased links:

$$\text{flatness} = 1 - \frac{\text{number of biased links}}{\text{total number of links}} \tag{4.2}$$

The measured flatness in the year range 2006–2013 is shown in Fig. 4.4 and is systematically decreasing with time. This indicates a statistically significant trend of less rather than more flatness. Between 2006 and 2013 (we drop the year 2005 from this analysis due to the small number of transactions made in that year), the flatness dropped by nearly 10% from its initial value. A detailed comparison of z-score distributions is shown in Fig. 4.3.

An examination of country–pairs reveals that some pairs show persistent bias (positive as well as negative), whereas others remain unbiased through time. Fig. 4.5 shows the time evolution of z-scores of a few of these country pairs. An example of positive bias (over-lending relative to the null model expectations) in the network is illustrated by loans from the US to Mexico. In the year 2012 there were $59k$ transactions made from the US to Mexico, about $5k$ more than expected by the null model ($54k$), which corresponds to a z-score of $+32$. Loan contributions made to borrowers in US and lenders from other countries usually show a negative bias. For example, transactions from Australia to the US in the same year (2012) were only 639. This observation is much lower than expected, 1,962 transactions with a z-score of -31. However, this is compensated by US-to-US over-lending (self-loop) as shown in Fig. 4.5. Interestingly, within country lending and borrowing (positive bias associated with self-loops) is seen consistently across the whole network and over time.

4.4 Network Robustness

World events have the potential to significantly change the Kiva network and lending systems like it. For example, events can impact the nodes or links in the network at random with events being precipitated by unpredictable financial collapses, coups, or natural disasters [225]. Events that drop nations out of the system can be strategically determined by new regulations, policies, or relationship failures. For example, the construction of a wall between the US and Mexico, an embargo, or a Brexit event could reduce or shut down flows in country–pairs [45, 48, 220].

To take a first step in trying to capture these network events in an abstract way, we explore key what-if scenarios of how the Kiva network responds to events

that disrupt capital flows. Our what-if shocks occur at the country level (affecting a node) or the country-pairs link level. For country/node level effects we remove nodes and all their links in four scenarios: (i) random removal of borrower nations, (ii) random removal of lender nations, (iii) removal of nations according to their lending volume (out-degree), and (iv) the removal of nations according to the borrowing volume (in-degree). For link removal, we remove links (i) at random, (ii) with minimal z-score, (iii) maximal z-score, and (iv) maximal transaction volume. Node removal is equivalent to a total edge removal when all the edges of a specific node are removed at the same time. For each reshaped network topology, we compare the new network to its corresponding null model distribution.

Fig. 4.6 shows the change in flatness as a result of node removal, broken down by year. The vertical axis represents the percentage of nodes removed for each of our four scenarios and the vertical axis shows the flatness. Our results indicate that the systems flatness responds differently to random and targeted removal of nodes. The system is remarkably stable when lender or borrower nodes are removed at random. This suggests that shocks that might impact nodes in the network at random are unlikely to change the system properties in regard to flatness. By contrast, the removal of just 10% of nodes targeted by their ranked out- or in-degree rapidly change system dynamics. The removal of only a few big lenders increases flatness quickly in all years. This makes intuitive sense as the big lenders correspond to pairs with larger per capita GDP difference, and therefore, are associated with bias (Fig. 4.5). This increase reaches saturation when the network attains an almost flat configuration. The trend in the elimination of the big borrowers is similar, but not as pronounced. This can potentially be attributed to the difference in out-degree and in-degree distributions. Since the out-degree distribution is more skewed (Fig. 4.2[D]), a few high-degree lender nodes account for a significantly larger portion of observed transactions. Hence, their removal results in the disappearance of more biased links than a high-degree borrower.

Since the in-degrees and out-degrees of nodes are preserved, presence of highly biased connection to a node may force other connections to that node to be biased as well (e.g., under-lending to a country from one or more lender countries balanced by

over-lending by others). Due to this interdependency of link biases, a local disruptive change in the network may have cascading effects causing a larger number of links to become biased.

The systems flatness is robust against random removal of edges and increases in flatness with removal of high transaction links (Fig. 4.7). In addition, we investigate the effect of edge removal according to the positivity or negativity of bias. Gradually removing links with strong positive bias causes flatness to increase comparatively to targeting maximum transaction links. This change is more drastic for small fraction of removed links and holds especially for earlier years (when the network was small). Targeting links with strong negative bias results in a weaker increase in flatness. This difference can be understood qualitatively in terms of the slight asymmetry in the z-score distribution. There are more positively biased links than negatively biased links. We also observe that the selection order of link removal based on the highest number of transactions increases most the flatness of the networks for later years.

This sensitivity analysis about system responses to different kinds of removals (nodes or links, random or targeted) reveals that random removal of nodes or links causes little-to-no change in overall flatness of the lending system. However, the flatness increases rapidly as big-players are removed from the network or few important channels of capital flow are blocked. We find that most of the bias in the system is accounted for by these few key countries or country–pairs.

4.5 Discussions and Conclusions

Global interconnectedness has raised the possibility that the world is becoming flatter and offering more equality of opportunity worldwide. Online crowdfinancing platforms like Kiva provide alternative channels of capital flow to traditional institutions raising the question as to whether peer-to-peer financing is making a flatter world – that is, one with fewer institutional and cultural biases in lending. To the contrary, we find continued and increasing bias in an inter-country, peer-to-peer crowdfinancing network. This drift towards a less-flat world may arise from individual level preferences or global factors. Although crowdfinancing provides a lending platform that connects lenders with borrowers and eliminates conventional

intermediaries such as banks, it is the individual lenders who decide whom they give loan to and can often be biased in their decisions. These biases are reinforced and made even stronger by the rapid growth of the crowdfinancing platform itself (rich get richer effect). An example of this growing bias in the crowdfinancing network is seen in the form of self-loops (lenders lending to borrowers in the same country), which are consistently biased in the positive direction. Nonetheless, whether or not these biases will continue to persist in the long run, remains an open question. We explored the effects of hypothetical disruptive events on system-level flatness with simulations and found that the lending network is not vulnerable to random losses of countries or bilateral ties. However, the targeted removal of a few high-volume lenders or high-transaction links could cause the networks flatness to increase significantly. This implies that the decreasing flatness is not centered on all lending, but on the lending bias of a few giant lenders that skew the overall system. In this way, the flatness of the system is directly linked with the dominance of a few big players.

Using regression analysis, we identified a few factors associated with preferential lending on this platform. One of the factors that significantly affect lending is economic disparity. Lenders in high-GDP per capita countries show a preference to provide money for low-GDP per capita countries – facilitating capital flow from developed to developing nations. This is important from the point of view of equality as it suggests that Kiva favors links that allow capital to flow from rich to poor countries (a counterexample of Lucas paradox). Other factors effecting lending are migration and colonial past, which are positively associated with lending, along with geographical distance, which has a negative association. Interestingly, these factors also effect other forms of international capital flows in the same manner (the effect size may vary from one system to another), as revealed by analyzing the government aid and shown by previous studies on international trade [226], thus reflecting the embeddedness of crowdfinancing in a larger ecosystem [227, 228]. The association of these factors with trade flow and government aid have to do with reasons that may be logistic (e.g., in trade flows, distance adds to the cost for supplying goods) or sociopolitical (e.g., a colonizing power providing development aid to its past colonies).

The same factors that determine the level of bilateral trade or aid are also associated with biasing the capital flows in an online crowdfinancing platform where loan transactions have zero logistic costs. This suggests that while crowdfinancing holds promise to add flatness to the world system of finance, it is embedded in a larger system of stable inequities that limit its effects and influences its development.

4.6 Methods

4.6.1 Node Removal

Starting from the original observed network, we remove a node (or a set of nodes), and all their edges, either randomly or in a particular order. Then we are interested in comparing the flatness of the remaining network with a null model generated from it. For measuring the flatness, we need the expected number of transactions of all links, as well as their standard deviation. We use the following analytical approximation to estimate the null model distribution. Let k_i^{out} denote the out-degree of node i. Similarly, k_j^{in} is the in-degree of node j. Assuming that the probability of observing a link is independent of all other links, the probability of appearance of an edge from node i to j is independent of the connectivity of the rest of the edges, and it is given by

$$p_{ij} = \frac{k_i^{\text{out}} k_j^{\text{in}}}{N_E^2} \tag{4.3}$$

where N_E corresponds to the total number of edges in the network. Using the above probability, the expected number of transactions from i to j is

$$E_{ij} = N_E p_{ij} = \frac{k_i^{\text{out}} k_j^{\text{in}}}{N_E} \tag{4.4}$$

with standard deviation (since the distribution is binomial)

$$\sigma_{ij} = \sqrt{N_E p_{ij} \left(1 - p_{ij}\right)}. \tag{4.5}$$

4.6.2 Edge Removal

Starting from the original observed network, we now remove links, according to the selected removal order. Similar to the case of node removal, we compare the remaining network with a null model. Here however, due to the eliminated links, which now have forbidden flows, both analytical approximations and simulations are challenging. Therefore, to obtain the desired distribution for the null model, we use the algorithm MaxEnt [229–232] to find the probability distribution that maximizes the Shannon entropy of the system given the node-level constraints (in- and out-degree) and the imposed edge-level constraints (no flow across certain edges). The distribution corresponding to maximum Shannon entropy is the least informative distribution, which in our case corresponds to the distribution of the null model [229, 230]. The Shannon Entropy is given by

$$H = -\sum_{ij} p_{ij} \log p_{ij}, \tag{4.6}$$

and is a non-linear, convex function. We use non-linear programming to

$$\begin{aligned} \underset{p_{ij}}{\text{maximize}} \quad & H\left(p_{ij}\right) \\ \text{subject to} \quad & \sum_{j} p_{ij} = \frac{k_i^{\text{out}}}{N_E} \\ & \sum_{i} p_{ij} = \frac{k_j^{\text{in}}}{N_E} \\ & 0 \leq p_{ij} \leq 1 \\ & p_{ij} = 0, \text{when } \{ij\} \in L \end{aligned}$$

where, L is the set of constrained links. The expected number of transactions E_{ij} is then given by $E_{ij} = p_{ij} * N_E$. Since MaxEnt cannot provide us with the standard deviation σ_{ij}, we approximate it using Eq. 4.5 and assuming that appearance of each edge is independent of other edges (thus it follows a binomial distribution).

4.6.3 Node and Link Removal Simulations

Because of the large number of nodes and links in the network, simulations for all of them were computationally infeasible; therefore, we applied a well-accepted numerical method to approximate the simulation results. Here, we show that the network estimated by simulation or analytical methods show close agreement with one another for 2006. A comparison between simulations and the numerical approximation (Figures 4.8 and 4.9) shows that the agreement between the approximation and degree preserving simulation is good. Since this approach overestimates the standard deviation slightly because of the small contribution from node degree not being preserved exactly but only on an average, we see that the flatness obtained by analytical approximation is larger in a systematic way even though the magnitude of the difference is small. Moreover, since the analytical method always overestimates the flatness (and this is true for all node/link removal methods), it only shifts the flatness measure by a small amount and does not affect the overall trend of flatness change with respect to removal (Figures 4.8 and 4.9).

4.6.4 Targeted Link Removal

Figure 4.10 shows that there are slightly more positively biased links than negatively biased for all years. This has an effect when links are removed by maximum and minimum z-scores. Since removal of biased links have a stronger effect on flatness, the curve corresponding to minimum z-score becomes flatter before the one corresponding to maximum z-score (Figure 4.7). The discrepancy is larger for the year 2006 where there are much fewer negatively biased links than positively biased links. The flatness continues to increase for link removal based on maximum transactions beyond maximum z-score and minimum z-score based removal. This can be attributed to the fact that the number of transactions are highly correlated with the absolute z-score as shown in Figure 4.11. Removal according to maximum or minimum z-score starts by targeting either positively or negatively biased links, respectively; whereas, removal by transaction has the advantage of potentially targeting positively as well as negatively biased links.

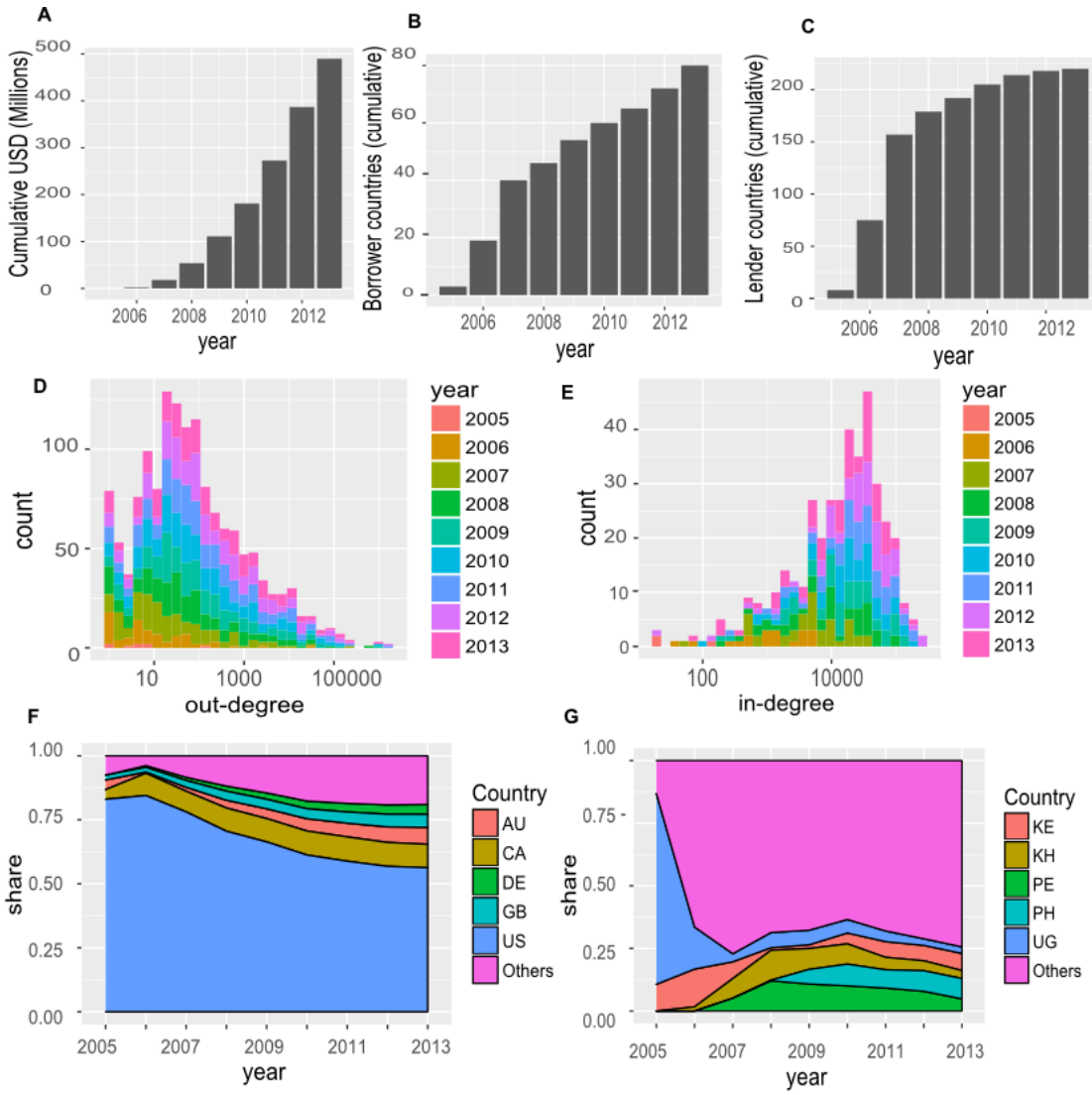

Figure 4.2: Evolution of the Kiva country network. (A) Total annual money lent through Kiva (cumulative). (B) Cumulative number of borrower and (C) lender countries. Plot (A)–(C) show the rapid growth of Kiva as platform for crowdfinancing both in terms of money lent and level of participation. (D) Histogram of total number of outgoing transactions from countries (out-degree) and (E) histogram of total number of incoming transactions to countries (in-degree), color stacked by year. The horizontal axis scale is logarithmic, thus, the histograms reflects the skewness of the distributions. (F) Top five lender countries and their share of loans given and (G) top five borrower countries for each year. The US accounts for a major share of the lending activity ($> 50\%$), however the US dominance is decreasing with time as we see increased participation levels from more countries.

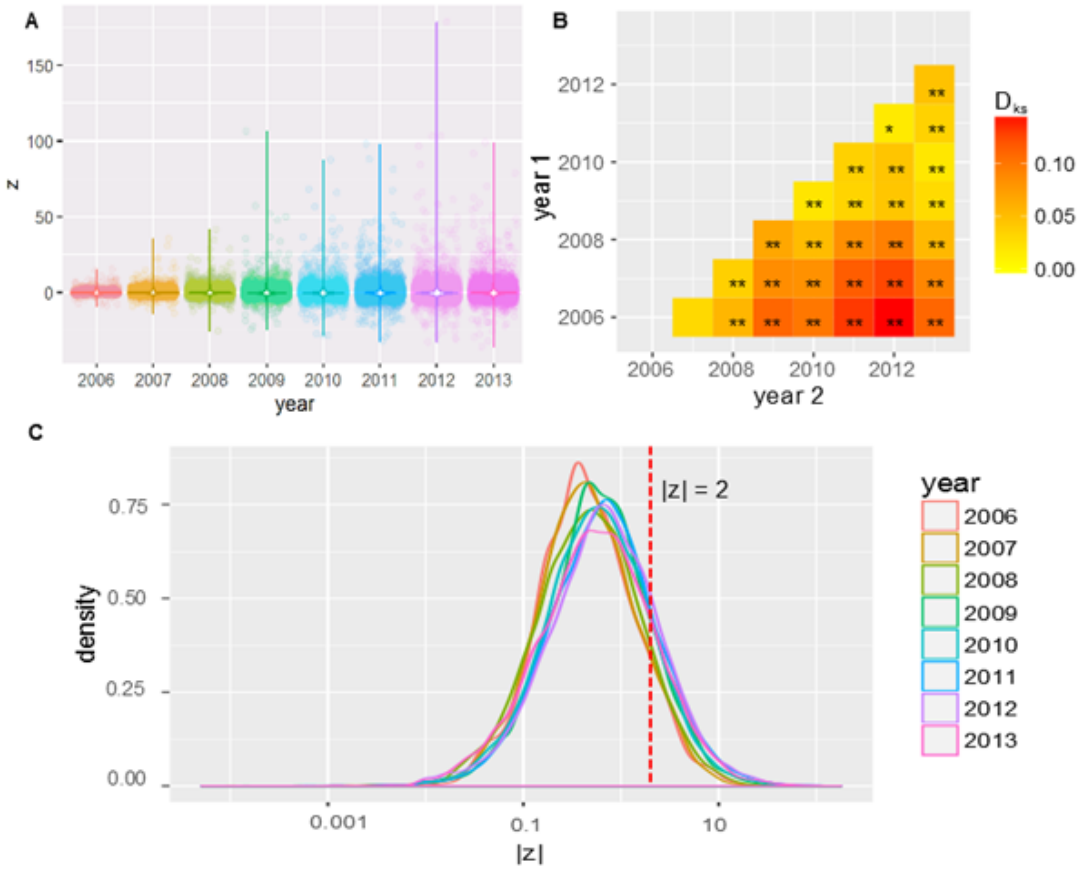

Figure 4.3: Broadening of z-score distributions. (A) Distribution of z-scores for each year shown by violin plots overlaid on the cloud of data points. It can be seen that the range of z-scores is becoming wider with time indicating a growing abundance of biased country–pairs. (B) KS test statistic D_{KS} of the z-score distributions for every pair of years. Significance levels are indicated by stars (*$p < 0.1$, **$p < 0.05$) in each cell. All pairs of years show a significant ($p < 0.1$ or $p < 0.05$) difference except one (2006–2007), which is not significant. The color of each cell corresponds to the value of the KolmogorovSmirnov (KS) statistic D_{KS}, which measures how far away the two distributions are. (C) Probability distribution function of $|z|$ for each year with the dashed line showing the cut-off $|z| = 2$. Each curve corresponds to a particular year. This density plot shows that with time the distribution is shifting right, which indicates that a larger fraction of links is becoming biased (fraction beyond the $|z| = 2$ cut-off shown by the dashed line).

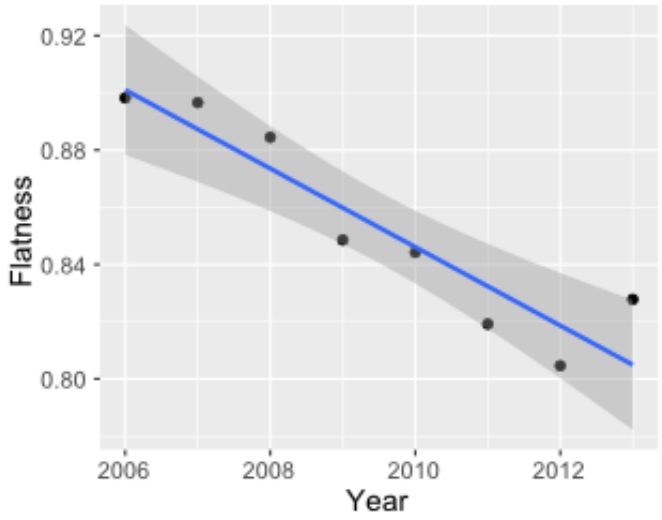

Figure 4.4: **Flatness of the Kiva network. Flatness is defined as the fraction of unbiased country–pairs ($|z| \leq 2$) under the null model. The flatness is measured by comparing the observed flow with the expected flow as described in the text. The line fit reveals a trend of decreasing flatness over the considered time frame (2006–2013).**

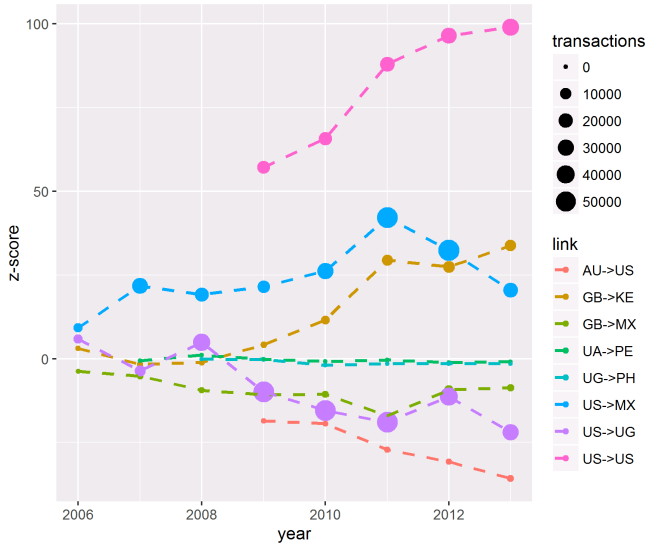

Figure 4.5: **Evolution of link biases. Time series of link level bias measured as z-scores for select links. The size of each dot corresponds to the number of transactions across the corresponding link. Few of these links (e.g., US to MX) are consistently biased while a few are unbiased (contributing to flatness) for all years.**

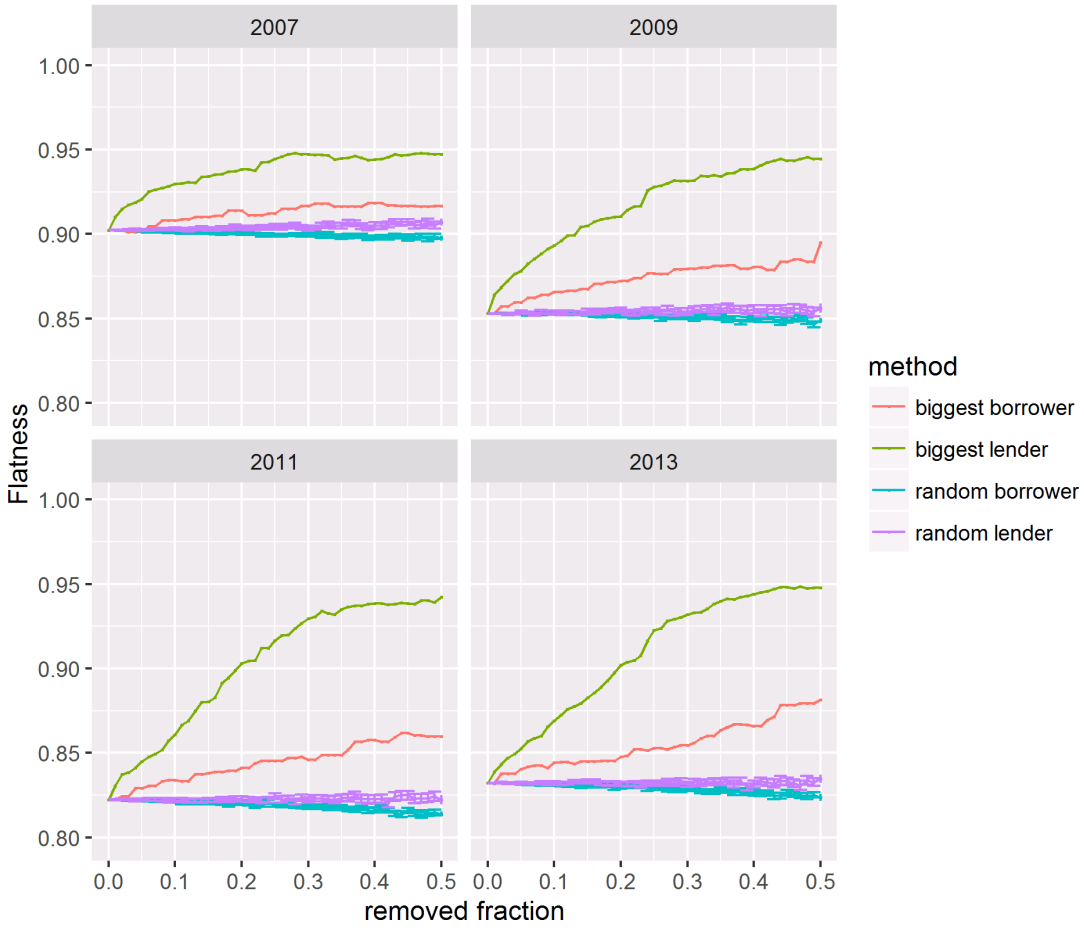

Figure 4.6: Simulated shocks: the effect of node removal. Change in flatness (defined as the fraction of unbiased links in the network) of the system as a function of removed fraction of nodes for different selection methods and for a few selected years (other years show a similar trend). The error bars correspond to $\pm$ 2 standard error for the random borrower and random lender case. The plots suggest that when nodes are removed randomly, the system flatness does not change; however, removing the biggest lenders or borrowers drives the system towards a more flat configuration.

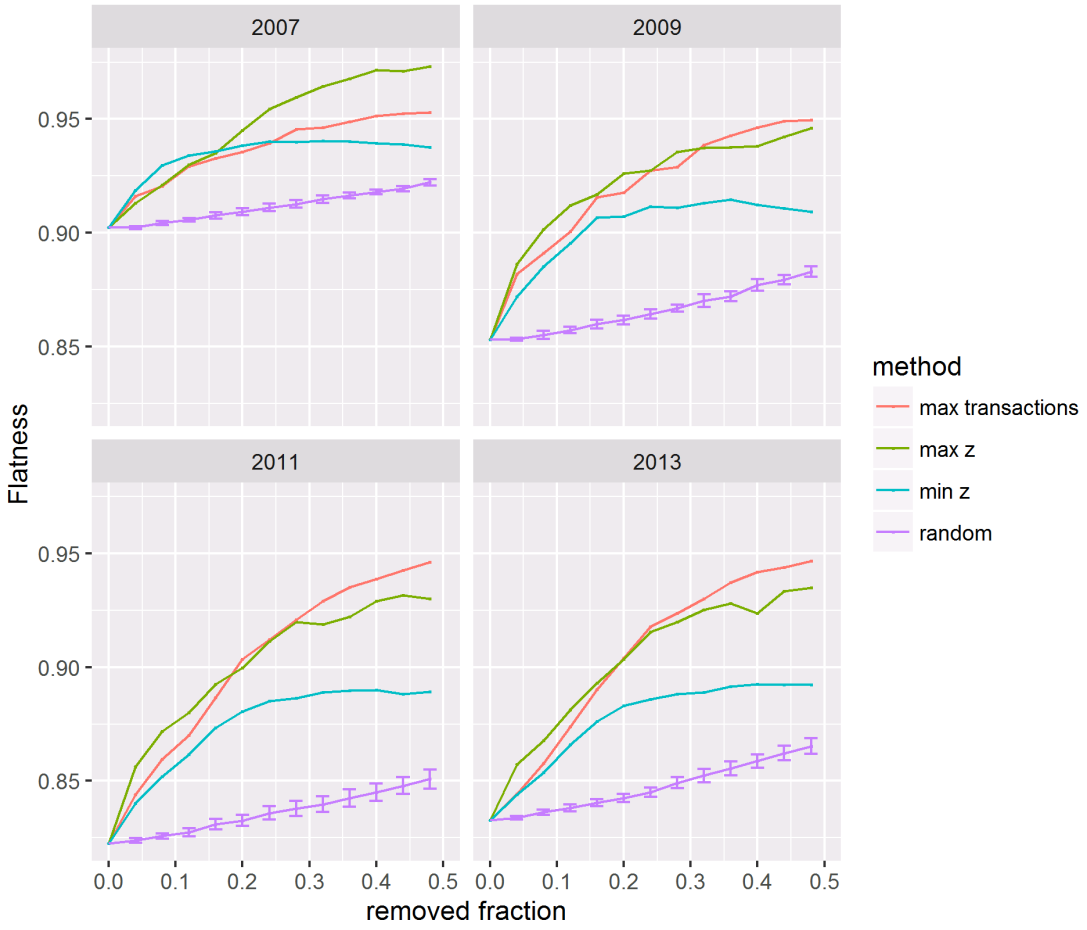

Figure 4.7: Simulated shocks: the effect of link removal. Change in flatness (defined as the fraction of unbiased links in the network) as a function of removed fraction of links for different selection methods and for a few selected years (other years show a similar trend). The error bars correspond to $\pm$ 2 standard error for the random link removal case. Similar to the node removal case, the system flatness does not change appreciably as links are removed randomly. Removing biased links (i.e., maximum or minimum z-scores) and links with maximum transactions makes the system flatter.

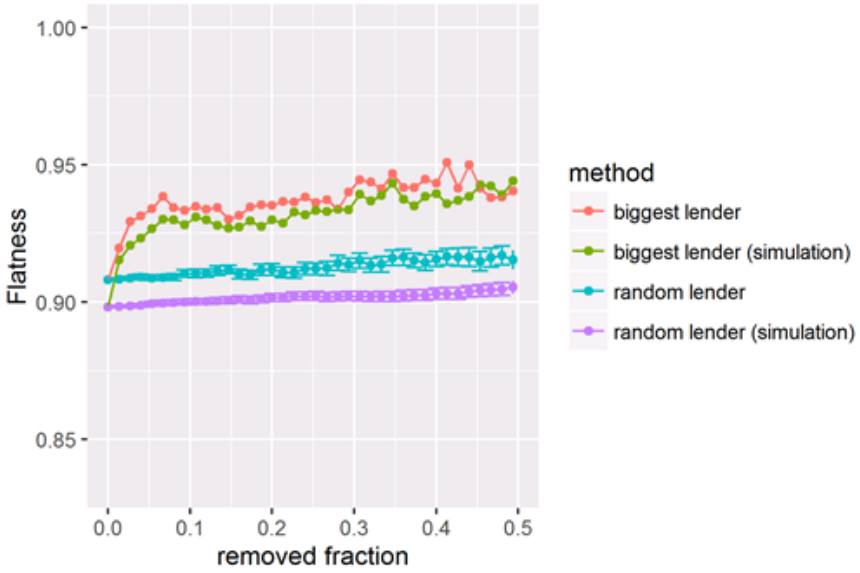

Figure 4.8: Comparison between simulation and analytical approximation for node removal for 2006 (for random and degree based removals). Results show good agreement between simulation and analytical approximation. The analytical approximation by construction overestimates the flatness as explained in the text. The error bars correspond to ± 2 standard error for the random removal case.

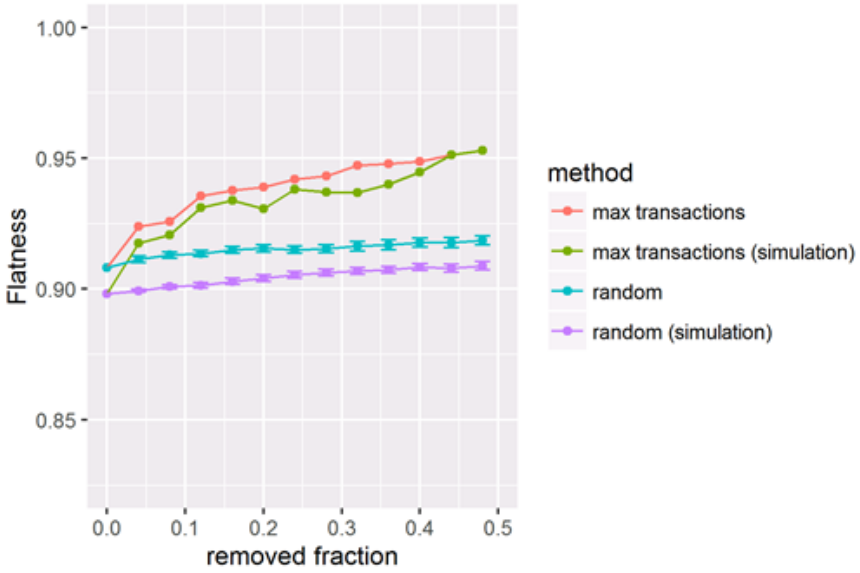

Figure 4.9: Comparison between simulation and maximum entropy method for link removal for 2006 (for random and transaction-based removals). Results show a good agreement in the trend between simulation and maximum entropy method. The maximum entropy method overestimates the flatness as explained in the text. The error bars correspond to $\pm$ 2 standard error for the random removal case.

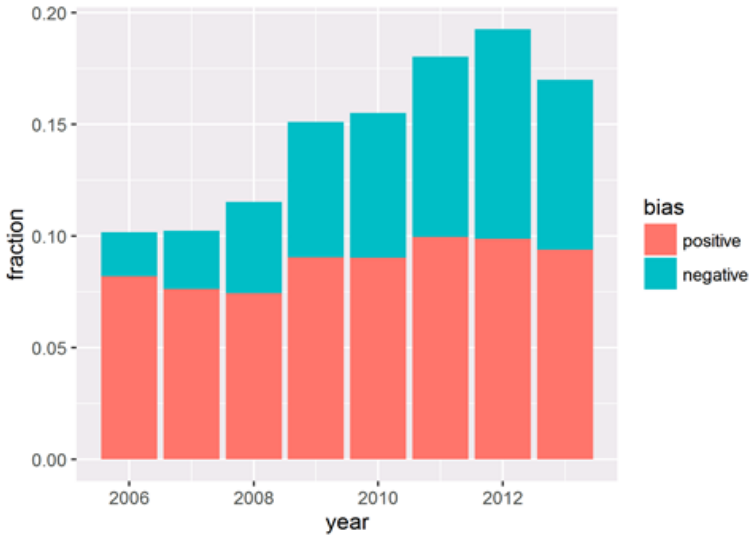

Figure 4.10: Positively vs. negatively biased fraction of links. Fraction of positively ($z > 2$) biased (red), and negatively ($z < 2$) biased links (blue) for the years 2006–2013. The figure shows a slightly larger proportion of positively than negatively biased links.

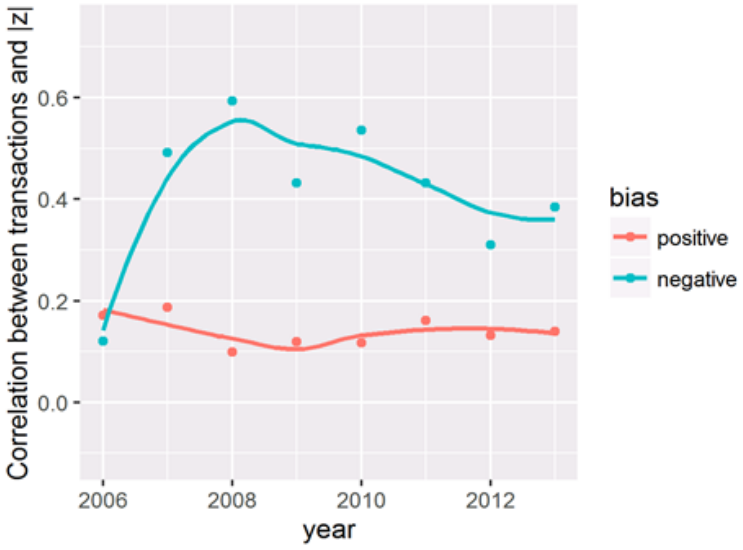

Figure 4.11: Correlation between the number of transactions and absolute value of z-scores. Linear correlation between absolute value of z-score for the biased pairs of countries in the network (computed separately for positively ($z > 2$) and negatively ($z < 2$) biased links) and the number of transactions between a pair of countries for years 2006–2013. Red and blue points correspond to positive and negative z-scores. The number of transactions seem to be correlated with both positively and negatively biased links. Thus, the removal of links with maximum transactions has a similar effect on the system flatness as removal of highly biased (positive or negative) links.

CHAPTER 5
Summary and Future Work

5.1 Summary

In this dissertation we have investigated the impact of heterogeneity in social networks in two different frameworks. In the first framework, we focus on the heterogeneity of a node's susceptibility to the adoption of new opinions. There, each node is the receiver of peer pressure and it has a finite capacity (threshold) of resisting to it. Hence, interesting dynamics occur when a node's resistance is suppressed. We study these dynamics using a model based on simplistic rules. In the second framework, using empirical data we study the impact of the heterogeneity of nodes when actively taking a decision, with no direct monetary benefit from their choice. Studying the node's choice reveals the direction a network growths. To this end, we have used Monte Carlo simulations, analytical and optimization methods, as well as methods from network and data science. A summary of our findings is given below.

In Chapter 2, we explored the impact of the heterogeneity of resistance (thresholds) in an opinion diffusion model (threshold model) with multiple initiators. The parameter we controlled is the standard deviation σ of the threshold distribution. In the case of $\sigma = 0$ all the nodes have the same threshold, and in the case of $\sigma = 0.2878$ the thresholds are selected uniformly at random. We found out that as the distribution of the thresholds increases the cascade's change from having a tipping point to cascades with a smooth transitions to reach consensus. By studying the response of the cascade size, we computationally detected the critical σ value for which (for a given fraction of randomly selected initiators) we move from cascades with a tipping point p_c to a smooth crossover, where p_c does not exist. Furthermore, in the case of thresholds selected uniformly at random we found a closed form analytical expression of the cascade size in relationship to the inserted initiator fraction. Interestingly, it is independent of the network structure. Finally, all of our above findings hold for synthetic and empirical networks.

In Chapter 3, following on our findings in the dynamics of the threshold model, we introduced two new strategies for selecting initiators in order to maximize the social influence. We compared the performance of our methods with other existing ones. We further study the impact of the heterogeneity of the thresholds and the network assortativity for all methods. To find a good set of initiators, our first method, namely Balanced Index (BI), looks for the optimal balance between the threshold of the nodes and their network centrality. Typically, nodes with high resistance and high network centrality are the better spreaders. Our second method, namely Group Performance Index, uses the power of the combination of nodes to reveal superspreaders. To do so, it selects the nodes for which any set of randomly selected nodes performs better in their presence. We show that our methods outperform any other methods for any initiator fraction, any threshold distribution, and any network assortativity.

In Chapter 4 we study the aggregated heterogeneity (biases) of individuals in the country level using the data of the lending patterns from a global charitable non-profit platform (KIVA). To this end, we defined a metric of the bias between each lender to borrower country (edge). The global bias was given as the fraction of biased edges, namely flatness. We showed that the world flatness is decreasing in time. In addition, we studied the response of the flatness in simulated scenarios of (borrower or lender) country or edge removals. Those removals simulated policy and event shocks (raising walls). We studied this response for both random shocks, and targeted shocks. We found that the high-transaction links could cause the networks flatness to increase significantly, while random shocks would insignificantly affect the flatness. These results, combined with regression analysis of our collaborators revealed that geographical distance and cultural biases are the leading factors of this bias.

Overall, our model and data based results indicate that node heterogeneity governs the world dynamics from social influencing to global patterns of country to country biases. We hope our results can be utilized for spreading an opinion or idea faster or considering local and global policies for balancing the global heterogeneity towards a less polarizing more interconnected and stable world.

5.2 Future Work

5.2.1 Analytical Model for Dynamical Selection of Inactive Nodes as Initiators

For the analytics of the Threshold Model (TM) we used Gleeson and Cahalane's tree like approximation [133], which is accurate for studying ensembles of for ER or SF networks. This analytical model has been expanded to capture different variations of the TM as well as the impact of specific network measures, such as the clustering or modularity of a network, in the final cascade size. The model can describe random (perhaps also degree or resistance based) selection of initiators, where the nodes are pre-selected before the spreading process is initiated. Hence, as designed, this analytical model cannot describe cases where the nodes are selected dynamically. For instance, it cannot model the case of a random selection of initiators which are dynamically selected from the pool of the inactive nodes.

A future plan is to create a variation of the tree-like approximation for dynamic selection of inactive initiators on the ensemble level of ER and SF networks. On our proposed model, on each step, a small fraction of inactive nodes is selected as initiators, and the cascade size is computed following the tree-like approximation. We repeat the process assuming part of the graph has been activated, and only focusing on the inactive part of the graph. One challenge on this proposed model is to compute the probability that any node with degree k has been activated. Another challenge is keeping track of the normalization factors necessary when applying this process on the inactive graph. In particular, Gleeson and Cahalane's analytical model is based on the assumption of an infinite size graph. Hence, in order to dynamically select only inactive nodes as initiators, we would have to assume that each time a small fraction of inactive nodes is selected as initiators and a cascade takes places, then the remaining inactive subgraph is always infinite in size. Yet, it will have to be smaller from the initial graph we started from. Hence, the contribution on the total cascade size of each dynamically selected initiator must be normalized to the size of the initial graph.

5.2.2 BI Algorithm: Improving on the Metric

Our current features for the BI algorithm is the resistance of a node, its dynamic degree, and the sum of the degrees of the second level vulnerable neighbors. The benefit of selecting those three features is that they are fast to compute while the performance of the algorithm is good. Yet, we have not systematically examined the performance of the higher order terms, such as the impact of a third level neighborhood etc. In the cases we have explored we found out that the resistance was the most important feature, followed by the dynamic degree. A possible question to examine is whether the drop on the weights of higher-order terms follows a specific pattern, such as exponential decay. That would reveal more on the dynamics that take place in the LTM and the number of higher-order neighborhoods that we should consider when computing the BI of each node. Furthermore, BI considers the resistance of the node in question, and only the neighboring nodes which are vulnerable (have resistance equal to one). Thus, nodes with resistance higher than one are disregarded on this metric. Perhaps, a variation of the BI metric where more of the information of the resistance of each node is utilized, would lead to better performance. For instance, we could be taking into account the average resistance or average dynamic degree of first level neighborhood of the node in question.

5.2.3 GPI Algorithm: Reducing the Time Complexity

The GPI algorithm's speed limits its usability. To tackle this, we propose a number of possible ways to reduce GPI's complexity and number of Monte Carlo simulations (v), while attempting to maintain the performance of the algorithm the same. Currently, the time complexity of GPI is $\mathcal{O}\left(v\langle k\rangle N^2\right)$ (for any graphs, reduced by $\langle k\rangle$ for sparse graphs), where $\mathcal{O}\left(v\right)$ is the number of runs required to compute with a good accuracy the expected GPI for all the nodes, $\mathcal{O}\left(N\right)$ is the number of test-initiators required to satisfy the goal (either looking to maximize the cascade size for a fixed number of initiators, or minimize the number of initiators for a fixed cascade size), and computing the spread induced by each initiator takes $\mathcal{O}\left(\langle k\rangle N\right)$. We should notice that on each step j after computing the GPI of each node, we need to rank the nodes based on their current GPI value, which takes $\mathcal{O}\left(N\log N\right)$ using

Heapsort. However, the time complexity per step j (coming from the computation of GPI) is larger than that, and so it is omitted.

Computing the impact (cascade size, reduced resistance) of one initiator in graphs of any size and average degree takes $\mathcal{O}(v\langle k\rangle N)$ time. However, similar to the process used in the CI-TM algorithm and our BI algorithm, we can compute the impact within a sphere of influence L, the size of which we control; the complexity of which is $\mathcal{O}(1)$ (assuming sparse graphs). Then, the leading complexity term per step j, becomes the time complexity of the Heapsort. Hence, the total time complexity of the algorithm could would become $\mathcal{O}(vN\log N)$.

To compute the GPI of all the nodes, we require a large number of runs v, which are independent from each other. Hence, it is possible to parallelize this process. Furthermore, the selection of test-initiators is completely random, thus a large number of runs v is required to distinguish the expected GPI of the most important nodes. To reduce the number of runs v, we can control the probability that specific nodes get selected for testing (so far, we have been selecting inactive nodes uniformly at random). One suggestion is to focus on nodes which more likely are superspreaders, for instance, nodes with high degree and resistance, or nodes with high BI ranking. Another suggestion is to focus on nodes which on the previous initiator insertion step had a high GPI ranking, yet not high enough to be selected as initiators. Unless the insertion of the initiators on the previous step has impacted their performance significantly, those nodes will still have a high GPI value. A final suggestion is to focus on nodes which have a high GPI ranking for the current initiator insertion step for a first small number of runs.

5.2.4 GPI Algorithm: Improving on the Metric

GPI measures the average impact of a node in the presence of any other set of initiators that satisfy a preset goal. This computational metric does not take into account other approaches such as analytics, learning algorithms, or meta-data heuristics. A possible future direction would be to combine those methods with GPI. For instance, by using a non-uniform random selection of test-initiators, emphasized on nodes with particular properties can potentially increase (not just the speed

of the algorithm as mentioned before but also) the performance of the algorithm. Those properties could be their network (first or/and second neighborhood) or their threshold/resistance. In addition, when testing the impact of each test-initiators, we only focus on the set of nodes which has not be activated yet either through being an initiator or through spread. Hence, nodes with high in-degree or low resistance are typically more likely to be activated through spread than other nodes. This additional information is not utilized. Taking into account how often a node gets activated through spread, or for how many inserted initiators it got activated, could lead to a better selection of initiators.

GPI is a metric of the expected number of initiators required to satisfy the goal. Yet, additional information on the statistics of the number of initiators could be collected and utilized for the improvement on the performance of the GPI. For instance, with no additional time and/or memory complexity cost the maximum and minimum test-initiator size as well as the standard deviation of GPI of each node could be recorded. Then, when selecting between two nodes with similar GPI values, perhaps the node higher standard deviation for the computation of GPI or the node with a lower minimum GPI would, or the node which was less times activated through spread would make for a better initiator.

5.2.5 GPI Algorithm: Network Destruction

As mentioned in Chapter 1, and Chapter 4, the robustness of a network in random failures and targeted attacks is an essential characteristic of a network. The detection of the set of nodes whose removal would cause the largest damage on a network is an NP-hard problem. The typical order parameter to quantify the damage of a network is the remaining largest connected component of a graph, namely the giant component. The larger the giant component is, the smaller the damage on the network. The classical case of a network destruction ignores any dynamical processes, and focuses only on the giant component given the removal of a set of nodes. We propose the use of an variation of the GPI algorithm to measure the size of the giant component. Here, the goal is to reduce the size of the giant component of any graph as much as possible for a given set of nodes. Thus, in this

case, GPI_i would measure the average size of the remaining giant component when node i is included in the set of randomly selected removed nodes.

5.2.6 Selection Strategies for Probabilistic Thresholds

In most cases, we do not know exactly the threshold of each node, but we will be able to infer it around some value, with some probability of it being different. A probabilistic threshold has a large impact on the performance of the algorithms. This is because now there is less certain information to relay on. A possible future project would be to first test the performance of the current algorithms we introduced, BI and GPI, as well as other successful algorithms (CITM) with increasing uncertainty on the threshold of each node. Then, we would have to examine possible variations on the introduced algorithms to make them more suitable for probabilistic thresholds. For instance, if the uncertainty of our inference of the probabilistic threshold of each node is relative small, then we can apply the BI and GPI algorithms as they are, but instead of using the expected probabilistic threshold of each node, assume each node's threshold is higher that, such as two standard deviations higher.

5.2.7 Long term Impact of Random Failures and Attacks on the Robustness of the Flatness of Empirical Directional Networks

Our analysis on Chapter 4 on the robustness of the flatness of empirical weighted directional networks (KIVA) focused on the removal of a link with multiedges from a lender to a borrower reducing the degree of both nodes. Removing both the demand and supply of the two nodes respectively can capture the transient impact on the flatness of the system, when both nodes have not had enough time to redirect their capacity (their demand and supply). To capture the *long term* impact of the removal of a (borrower or lender) node from the network or the removal of link we would have to assume that both the lender, and borrower nodes would redirect their supply and demand respectively. Thus, a new model has to be designed where now, the remaining nodes of the network would have to absorb this perturbation by increasing their transactions with those two nodes respectively. The challenge here would be to correctly redirect the multi-edge capacity. For instance, would

each non-constrained lender and borrower node increase their transactions with the respected two nodes proportionality to the empirical network or the null model?

In our future work we plan to address the above proposals and questions.

REFERENCES

[1] M. Newman, *Networks, an Introduction* (Oxford University Press, New York, NY, 2011).

[2] K. Coronges, A.-L. Barabási, and A. Vespignani, prepared by K. Klemic and J. Zeigler, "Future directions of Network Science: A workshop report on the emerging science of network," (2016), available at http://www.acq.osd.mil/rd/basic_research/references/docs/Network_Sciences.pdf (Accessed on May 25, 2017).

[3] M. Newman, "The physics of networks," Phys. Today **61**, 33 (2008).

[4] R. Albert, and A.-L. Barabási, "Statistical mechanics of complex networks," Rev. Mod. Phys. **74**, 47 (2002).

[5] E. Estrada, "Graph and network theory in physics," (2013), available at https://arxiv.org/abs/1302.4378 (Accessed on May 25, 2017).

[6] C. Castellano, S. Fortunato, and V. Loreto, "Statistical physics of social dynamics," Rev. Mod. Phys. **81**, 591 (2009).

[7] A.-L. Barabási, and Z. N. Oltvai, "Network biology: understanding the cell's functional organization," Nat. Rev. Genet. **5**, 101 (2004).

[8] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee, "Measurement and Analysis of Online Social Networks," *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement* (ACM, New York, NY, 2007), pp. 29–42.

[9] B. Wellman, J. Salaff, D. Dimitrova, L. Garton, M. Gulia, and C. Haythornthwaite, "Computer networks as social networks: Collaborative work, telework, and virtual community," Annu. Rev. Sociol. Vol. **22**, 213 (1996).

[10] M. Trusov, R. E. Bucklin, and K. Pauwels, "Effects of word-of-mouth versus traditional marketing: Findings from an Internet social networking Site," J. Marketing, **73**, 90 (2009).

[11] C. Hawn, "Take two aspirin and tweet me in the morning: How Twitter, Facebook, and other social media are reshaping health care," Health Aff. (Millwood) **28**, 361 (2009).

[12] M. Swan, "Emerging patient-driven health care models: an examination of health social networks, consumer personalized medicine and quantified self-tracking," Int. J. Environ. Res. Public Health **6**, 492 (2009).

[13] W. Bridewell, and A. K. Das, "Social network analysis of physician interactions: the effect of institutional boundaries on breast cancer care," AMIA Annu. Symp. Proc. **2011**, 152 (2011).

[14] N. A. Christakis, and J. H. Fowler, "The spread of obesity in a large social network over 32 years," N. Engl. J. Med. **357**, 370 (2007).

[15] H. Grassegger, and M. Krogerus, "The Data that turned the world upside down," Motherboard (2017), available at https://motherboard.vice.com/en_us/article/big-data-cambridge-analytica-brexit-trump (Accesed on March 23, 2017).

[16] T. Gohmann, "How Donald Trump won the election: A behavioral economics explanation," (2016), available at http://www.behavioralsciencelab.com/news/ (Accesed on March 23, 2017).

[17] M. Chau, and J. Xu, "Mining communities and their relationships in blogs: A study of online hate groups," Int. J. Hum. Comput. Stud. **65**, 57 (2007).

[18] M. J. Brzozowski, T. Hogg, and G. Szabo, "Friends and Foes: Ideological Social Networking," CHI '08 proceedings of the SIGCHI conference on human factors in computing systems, 817–820 (2008).

[19] O. Oh, . Agrawal, and H. R. Rao, "Information control and terrorism: Tracking the Mumbai terrorist attack through twitter," Inf. Syst. Front. **13**, 33 (2011).

[20] J. P. Farwell, "The media strategy of ISIS," Survival Global Politics and Strategy **56**, (2014).

[21] P. Erdős, and A. Rényi, "On random graphs," Publ. Math. Debrecen **6**, 290 (1959).

[22] A. J. Veraart, E. J. Faassen, V. Dakos, E. H. van Nes, M. Lurling, and M. Scheffer, "Recovery rates reflect distance to a tipping point in a living system," Nature **481**, 357 (2012).

[23] D. P. Redlawsk, A. J. W. Civettini, and K. M. Emmerson, "The affective tipping point: Do motivated reasoners ever Get it?," Polit. Psychol. **31**, 563 (2010).

[24] W. Zhang, C. Lim, and B. K. Szymanski, "Analytic treatment of tipping points for social consensus in large random networks," Phys. Rev. E **86**, 061134 (2012).

[25] M. E. J. Newman, "The structure and function of complex networks," SIAM Rev. **45**, 167 (2003).

[26] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang, "Complex networks: Structure and dynamics," Phys. Rep. **424**, 175 (2006).

[27] D. J. Watts, and S. H. Strogatz, "Collective dynamics of 'small-world' networks," Nature **393**, 440 (1998).

[28] S. H. Strogatz, "Exploring complex networks," Nature **410**, 268 (2001).

[29] A. L. Barabási, and R. Albert, "Emergence of scaling in random networks," Science **286**, 509 (2002).

[30] A. Derzsi, N. Derzsy, E. Káptalan, and Z. Néda, "Topology of the Erasmus student mobility network," Physica A **390**, 2601 (2011).

[31] N. Derzsy, Z. Néda, and M. A. Santos, "Income distribution patterns from a complete social security database," Physica A **391**, 5611 (2012).

[32] M. Faloutsos, P. Faloutsos, and C. Faloutsos, "On power-law relationships of the Internet topology," Comput. Commun. Rev. 29, 251 (1999).

[33] H. Ebel, L.-I. Mielsch, and S. Bornholdt, "Scale-free topology of e-mail networks," Phys. Rev. E **66**, 035103(R) (2002).

[34] V. M. Eguíluz, D. R. Chialvo, G. A. Cecchi, M. Baliki, and A. V. Apkarian, "Scale-free brain functional networks," Phys. Rev. Lett. **94**, 018102 (2005).

[35] S. Milgram, "The small world problem," Psychol. Today **2**, 60 (1967).

[36] J. Travers, and S. Milgram, "An experimental study of the small world problem," Sociometry **32**, 425 (1969).

[37] C. Korte, and S. Milgram, "Acquaintance networks between racial groups: Application of the small world method," J. Personality and Social Psych. **15**, 101 (1978).

[38] R. Albert, H. Jeong, and A.-L. Barabási, "The diameter of the World Wide Web," Nature **401**, 130 (1999).

[39] V. Latora, and M. Marchiori, "Economic small-world behavior in weighted networks," Eur. Phys. J. B **32**, 249 (2003).

[40] D. S. Bassett, and E. D. Bullmore, "Small-world brain networks," Neuroscientist **12**, 512 (2006).

[41] V. Latora, and M. Marchiori, "Is the Boston subway a small-world network?" Physica A **314**, 109 (2002).

[42] P. Sen, S. Dasgupta, A. Chatterjee, P. A. Sreeram, G. Mukherjee, and S. S. Manna, "Small-world properties of the Indian railway network," Phys. Rev. E **67**, 036106 (2003).

[43] V. Latora, and M. Marchiori, "Efficient behavior of small-world networks," Phys. Rev. Lett. **87**, 198701 (2001).

[44] M. E. J. Newman, and D. J. Watts, "Scaling and percolation in the small-world network model," Phys. Rev. E **60**, 7332 (1999).

[45] R. Albert, H. Jeong, and A.-L. Barabási, "Error and attack tolerance of complex networks," Nature **406**, 378 (2000).

[46] S. D. S. Reis, Y. Hu, A. Babino, J. S. Jr Andrade, S. Canals, M. Sigman and H. A. Makse, "Avoiding catastrophic failure in correlated networks of networks," Nat. Phys. **10**, 762 (2014).

[47] Jr. F. Molnár, N. Derzsy, B. K. Szymanski, and G. Korniss, "Building damage-resilient dominating sets in complex networks against random and targeted Attacks," Sci Rep. **5**, 8321 (2015).

[48] J. Gao, B. Barzel, and A.-L. Barabási, "Universal resilience patterns in complex networks," Natute **530**, 307 (2016).

[49] P. Crucitti, V. Latora, and M. Marchiori, "A topological analysis of the Italian electric power grid," Physica A **338**, 92 (2004).

[50] S. V. Buldyrev, R. Parshani, G. Paul, H. E. Stanley, and S. Havlin, "Catastrophic cascade of failures in interdependent networks," Nature **464**, 1025 (2010).

[51] A. Vespignani, "Complex networks: The fragility of interdependency," Nature **464**, 984 (2010).

[52] C. D. Brummitt, G. Barnett, and R. M. D'Souza, "Coupled catastrophes: sudden shifts cascade and hop among interdependent systems," J. R. Soc. Interface **12**, 20150712 (2015).

[53] A. Motter, and Y. Yang, "The unfolding and control of network cascades," (2017), available at https://arxiv.org/abs/1701.00578 (Accessed on May 25, 2017).

[54] A. Motter, and Y.-C. Lai, "Cascade-based attacks on complex networks," Phys. Rev. E **66**, 065102 (2002).

[55] A. G. Haldane, and R. M. May, "Systemic risk in banking ecosystems," Nature **469**, 351 (2001).

[56] M. Elliott, B. Golub, and M. O. Jackson, "Financial networks and contagion," Am. Econ. Rev. **104**, 3115-3153 (2014).

[57] Z. Su, L. Li, H. Peng, J. Kurths, J. Xiao, and Y. Yang, "Robustness of interrelated traffic networks to cascading failures," Sci. Rep. **4**, 5413 (2014).

[58] R. Kinney, P. Crucitti, R. Albert, and V. Latora, "Modeling cascading failures in the North American power grid," *Eur. Phys. J. B* **46**, 101–107 (2005).

[59] T. Nagatani, "The physics of traffic jams," Rep. Prog. Phys, **65**, 1331 (2002).

[60] P. Hines, K. Balasubramaniam, and E. C. Sanchez, "Cascading failures in power grids," IEEE Potentials, **28**, 101 (2009).

[61] J.-W. Wang, and L.-L. Rong, "Cascade-based attack vulnerability on the US power grid," Saf. Sci. **47**, 1332 (2009).

[62] A. Asztalos, S. Sreenivasan, B. K. Szymanski, and G. Korniss, "Cascading failures in spatially-embedded random networks," *PLoS ONE* **9**, e84563 (2014).

[63] A. Moussawi, N. Derzsy, X. Lin, B. K. Szymanski, and G. Korniss, "Limits of predictability of cascading overload failures in spatially-embedded networks with distributed flows," (2017), available at https://arxiv.org/abs/1706.04579 (Accessed on May 25, 2017).

[64] R. Pastor-Satorras, C. Castellano, P. van Mieghem, and A. Vespignani, "Epidemic processes in complex networks," Rev. Mod. Phys. **87**, 925 (2015).

[65] A. S. Klovdahl, "Social networks and the spread of infectious diseases: The AIDS example," Soc. Sci. Med. **21**, 1203 (1985).

[66] A. J. Tatem, D. J. Rogers, and S. I. Hay, "Global transport networks and infectious disease spread," Adv. Parasitol. **62**, 293 (2006).

[67] D. Balcan, V. Colizza, B. Goncalves, H. Hu, J. Ramasco, and A. Vespignani, "Multiscale mobility networks and the spatial spreading of infectious diseases," Proc. Natl. Acad. Sci. U.S.A. **106**, 21484 (2009).

[68] R. Pastor-Satorras, and A. Vespignani, "Epidemic spreading in scale-free networks," Phys. Rev. Lett. **86**, 3200 (2001).

[69] F. Liljeros, C. R. Edling, L. A. N. Amaral, H. E. Stanley, and Y. Åberg, "The web of human sexual contacts," Nature **411**, 6840 (2001).

[70] B. Rothenberg, C. Sterk, K. Toomey, J. Potterat, D. Johnson, M. Schrader, and S. Hatch, "Using social network and ethnographic tools to evaluate Syphilis transmission," Sex. Transm. Dis. **25**, 154 (1998).

[71] W. O. Kermack, and A. G.McKendrick, "A contribution to the mathematical theory of epidemics," Proc. of Royal Soc. **115**, 700 (1927).

[72] C. Castellano and R. Pastor-Satorras, "Thresholds for epidemic spreading in networks," Phys. Rev. Lett. **105**, 218701 (2010).

[73] M. Kitsak, L. K. Gallos, S. Havlin, F. Liljeros, L. Muchnik, H. E. Stanley, and H. Makse, "Identification of influential spreaders in complex networks," Nat. Phys. **6**, 888 (2010).

[74] F. Altarelli, A. Braunstein, L. DallAsta, J. R. Wakeling, and R. Zecchina, "Containing epidemic outbreaks by message-passing techniques," *Phys. Rev. X* **4**, 021024 (2014).

[75] A. Guille, H. Hacid, C. Favre, and D. Zighed, "Information diffusion in online social networks: A survey," ACM SIGMOD Record **42**, 17 (2013).

[76] S. Aral, L. Muchnika, and A. Sundararajana, "Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks," Proc. Natl. Acad. Sci. U.S.A. **106**, 21544 (2009).

[77] C. R. Shalizi, and A. C. Thomas, "Homophily and contagion are generically confounded in observational social network studies," Socio. Meth. Res. **40**, 211 (2011).

[78] F. Heider, "Attitudes and cognitive organization," J. Psychol. **21**, 107 (1946).

[79] D. Cartwright, and F. Harary, "Structural balance: a generalization of Heider's theory," Psychol. Rev. 63, 277 (1956).

[80] P. Singh, S. Sreenivasan, B. K. Szymanski, and G. Korniss, "Competing effects of social balance and influence," Phys. Rev. E **93**, 042306 (2016).

[81] T. Antal, P. L. Krapivsky, and S. Redner, "Dynamics of social balance on networks," Phys. Rev. E **72**, 036121 (2005).

[82] S. A. Marvel, S. H. Strogatz, and J. M. Kleinberg, "Energy landscape of social balance," Phys. Rev. Lett. **103**, 198701 (2009).

[83] E. Estrada, and M. Benzi, "Are social networks really balanced?," (2014), available at https://arxiv.org/abs/1406.2132 (Accessed on May 25, 2017).

[84] S. Wasserman, and K. Faust, *Social Network Analysis: Methods and Applications* (Cambridge University Press, Cambridge, UK, 1994).

[85] D. Garlaschelli, and M. I. Loffredo, "Patterns of link reciprocity in directed networks," Phys. Rev. Lett. **93**, 268701 (2004).

[86] C. Wang, O. Lizardo, C. Hachen, A. Strathman , Z. Toroczkai, and N. Chawla, "A dyadic reciprocity index for repeated interaction networks," Netw. Sci. **1**, 31 (2013).

[87] T. Squartini, F. Picciolo, F. Ruzzenenti, and D. Garlaschelli, "Reciprocity of weighted networks," Sci Rep. **3**, 2729 (2013).

[88] L. Kovanen, J. Saramaki, and K. Kaski, "Reciprocity of mobile phone calls," (2010), available at https://arxiv.org/abs/1002.0763 (Accessed on May 25, 2017).

[89] J. Blumenstock, M. Fafchamps, and N. Eagle, "Risk and reciprocity over the mobile phone network: Evidence from Rwanda," NET Institute Working Paper No. 11–25 (2011).

[90] P. F. Lazarsfeld, and R. K. Merton, *Freedom and Control in Modern Society* (Van Nostrand, New York, 1954), Vol. 18, pp. 18–66.

[91] M. McPherson, L. S. Lovin, and J. M. Cook, "Birds of a feather: Homophily in social networks," Ann. Rev. Sociol. **27**, 415 (2001).

[92] M. Dehghani, K. Johnson, J. Hoover, E. Sagi, J. Garten, N. J. Parmar, S. Vaisey, R. Iliev, and J. Graham, "Purity homophily in social networks," J. Exp. Psychol.-Gen. **145**, 366 (2016).

[93] R. M. Bond, C. J. Fariss, J. J. Jones, A. D. I. Kramer, C. Marlow, J. E. Settle, and J. H. Fowler, "A 61-million-person experiment in social influence and political mobilization," Nature **489**, 295 (2012).

[94] T. C. Schelling, "Hockey helmets, concealed weapons, and daylight saving: A study of binary choices with externalities," J. Confl. Resolut. **17**, 381 (1973).

[95] B. K. Szymanski, O. Lizardo, C. Doyle, P. D. Karampourniotis, P. Singh, G. Korniss, and J. Z. Bakdash, "The Spread of Opinions in Societies," (2016) in *Modeling Sociocultural Influences on Decision Making: Understanding Conflict, Enabling Stability*, edited by J. V. Cohn, S. Schatz, H. Freeman, D. J. Y. Combs, (CRC Press, Boca Raton, FL, 2016), pp. 61–84.

[96] M. Granovetter, "Threshold models of collective behavior," Am. J. Sociol. **83**, 1420 (1978).

[97] R. A. Berk, "A gaming approach to crowd behavior," Am. Sociol. Rev. **39**, 355 (1974).

[98] R. Kozma, M. Puljic, P. Balister, B. Bollobás, and W. J. Freeman, "Phase transitions in the neuropercolation model of neural populations with mixed local and non-local interactions," Biol. Cybern. **92**, 367 (2005).

[99] J.-P. Eckmanna, O. Feinermanb, L. Gruendlingerc, E. Mosesb, J. Sorianob, and T. Tlustyb, "The physics of living neural networks," Phys. Rep. **449**, 54 (2007).

[100] H. Amini, "Bootstrap percolation in living neural networks," J. Stat. Phys. **141**, 459 (2010).

[101] I. N. Lymperopoulos, and G. D. Ioannou, "Online social contagion modeling through the dynamics of Integrate-and-Fire neurons," Inf. Sc. **320**, 26 (2015).

[102] E. Aronson, T. D. Wilson, and A. M. Akert, *Social Psychology (5th ed.)* (Upper Saddle River, NJ: Prentice Hal, 2005).

[103] A. Solomon, "Opinions and social pressure," Sci. Am. **193**, 31 (1955).

[104] J. Zhang, C. K. Hsee, and Z. Xiao, "The majority rule in individual decision making," Organ. Behav. Hum. Decis. Process. **99**, 102 (2006).

[105] A. A. Hung, and C. R. Plott, "Information Cascades: Replication and an Extension to Majority Rule and Conformity-Rewarding Institutions," Am. Econ. Rev. **91**, 1508 (2001).

[106] D. B. Bahr, R. C. Browning, H. R. Wyatt, and J. O. Hill, "Exploiting social networks to mitigate the obesity epidemic," Obesity **17**, 723 (2009).

[107] P. Clifford, and A. Sudbury, "A model for spatial conflict," Biometrika **60**, 581 (1973).

[108] R. A. Holley, and T. M. Liggett, "Ergodic theorems for weakly interacting infinite systems and the Voter Model," Ann. Probab. **3**, 643 (1975).

[109] T. M. Liggett, *Stochastic Interacting Systems: Contact, Voter, and Exclusion Processes* (Springer-Verlag, New York, NY, 1999).

[110] C. Castellano, D. Vilone, A. Vespignani, "Incomplete ordering of the voter model on small-world networks," Europhys. Lett. **63**, 153 (2003).

[111] K. Suchecki, V. M. Eguíluz, and M. S. Miguel, "Voter model dynamics in complex networks: Role of dimensionality, disorder, and degree distribution," Phys. Rev. E **72**, 036132 (2005).

[112] V. Sood, and S. Redner, "Voter model on heterogeneous graphs," Phys. Rev. Let. **94**, 178701 (2005).

[113] F. Vazquez, and V. M. Eguíluz, "Analytical solution of the voter model on uncorrelated networks," New J. Phys. **10**, 063011 (2008).

[114] W. Pickering, and C. Lim, "Solution of the voter model by spectral analysis," Phys. Rev. E **91**, 012812 (2015).

[115] C. Castellano, M.A. Munoz, and R. Pastor-Satorras, "Nonlinear q-voter model," Phys. Rev. E **80**, 041129 (2009).

[116] L. Steels, "A self-organizing spatial vocabulary," Artif. Life, **2**, 319 (1995).

[117] A. Baronchelli, M. Felici, V. Loreto, E. Caglioti, and L. Steels, "Sharp transition towards shared vocabularies in multi-agent systems," J. Stat. Mech.: Theory Exp. **2006**, P06014 (2006).

[118] Q. Lu, G. Korniss, and B. K. Szymanski, "The naming game in social networks: community formation and consensus engineering," J. Econ. Interact. Coord. **4**, 221 (2009).

[119] J. Xie, S. Sreenivasan, G. Korniss, W. Zhang, C. Lim, and B. K. Szymanski, "Social consensus through the influence of committed minorities," Phys. Rev. E **84**, 011130 (2011).

[120] X. Castello, A. Baronchelli, and V. Loreto, "Consensus and ordering in language dynamics," Eur. Phys. J. B **71**, 557 (2009).

[121] W. Pickering, B. K. Szymanski, and C. Lim, "Analysis of the high-dimensional naming game with committed minorities," Phys. Rev. E **93**, 052311 (2016).

[122] A. M. Thompson, B. K. Szymanski, and C. Lim, "Propensity and stickiness in the naming game: Tipping fractions of minorities," Phys. Rev. E **90**, 042809 (2014).

[123] P. D. Karampourniotis, S. Sreenivasan, B. K. Szymanski, and G. Korniss, "The impact of heterogeneous thresholds on social contagion with multiple initiators," PLoS ONE **10**, e0143020 (2015).

[124] A. Nowak, J. Szamrej, and B. Latané, "From private attitude to public opinion: A dynamic theory of social impact," Psychol. Rev. **97**, 362 (1990).

[125] D. J. Watts, "A simple model of global cascades on random networks," Proc. Natl. Acad. Sci. U.S.A. **99**, 5766 (2002).

[126] D. J. Watts, and P. S. Dodds, "Influentials, networks, and public opinion formation," J. Consum. Res. **34**, 441 (2007).

[127] J. Xie, S. Sreenivasan, G. Korniss, W. Zhang, C. Lim, and B. K. Szymanski, "Social consensus through the influence of committed minorities," Phys. Rev. E **84**, 011130 (2011).

[128] D. Centola, V. M. Eguíluz, and M. W. Macy, "Cascade dynamics of complex propagation," Physica A **374**, 449 (2007).

[129] P. Singh, S. Sreenivasan, B. K. Szymanski, and G. Korniss, "Threshold-limited spreading in social networks with multiple initiators," Sci. Rep. **3**, 2330 (2013).

[130] Y. Ikeda, T. Hasegawa, and K. Nemoto, "Cascade dynamics on clustered network," J. Phys.: Conf. Ser. **221**, 012005 (2010).

[131] K. M. Lee, C. D. Brummitt, and K. I. Goh, "Threshold cascades with response heterogeneity in multiplex networks," Phys. Rev. E **90**, 062816 (2014).

[132] A. Nematzadeh, E. Ferrara, A. Flammini, and Y. Y. Ahn, "Optimal network modularity for information diffusion," Phys. Rev. Lett. **113**, 088701 (2014).

[133] J. P. Gleeson and D. J. Cahalane, "Seed size strongly affects cascades on random networks," Phys. Rev. E **75**, 056103 (2007).

[134] J. P. Gleeson, and D. J. Cahalane, "An Analytical Approach to Cascades on Random Networks," Proc. SPIE 6601, Noise and Stochastics in Complex Systems and Finance, 66010W (2007).

[135] J. P. Gleeson, "Cascades on correlated and modular random networks," Phys, Rev. E **77**, 046117 (2008).

[136] G. Curato, and F. Lillo, "Optimal information diffusion in stochastic block models," Phys. Rev. E **94**, 032310 2016.

[137] F. Karimi, and P. Holme, "Threshold model of cascades in empirical temporal networks," Physica A **392**, 3476 (2013).

[138] R. Michalski, T. Kajdanowicz, P. Bródka, and P. Kazienko, "Seed Selection for spread of influence in social networks: Temporal vs. static approach," New Generat. Comput. **32**, 213 (2014).

[139] Z. Ruan, G. Iniguez, M. Karsai, and J. Kertesz, "Kinetics of social contagion," Phys. Rev. Lett. 115, 218702 (2015).

[140] W. M. Huang, L. J. Zhang, X. J. Xu, X. Fu, "Contagion on complex networks with persuasion," Sci. Rep. **6**, 23766 (2016).

[141] P. Wang, L.-J. Zhang, X.-J. Xu, G. Xiao, "Heuristic strategies for persuader selection in contagions on complex networks," PLoS ONE, **12**, e0169771 (2017).

[142] D. Dhar, P. Shukla, J. P. Sethna, "Zero-temperature hysteresis in the random-field Ising model on a Bethe lattice," J. Phys. A: Math. Gen. **30**, 5259 (1997).

[143] J. P. Sethna, K. Dahmen, S. Kartha, J. A. Krumhansl, B. W. Roberts, J. D. Shore, "Hysteresis and hierarchies: Dynamics of disorder-driven first-order phase transformations," Phys. Rev. Lett. **70**, 3347 (1993).

[144] P. S. Dodds, D. J. Watts, "A generalized model of social and biological contagion," J. Theor. Biol. **232**, 587 (2005).

[145] M. Karsai, G. Iñiguez, K. Kaski, and J. Kertész, "Complex contagion process in spreading of online innovation," J. R. Soc. Interface **11**, (2014).

[146] F. Morone, and H. A. Makse, "Collective influence optimization uncovers the strength of weak nodes in complex networks," Nature **524**, 65 (2015).

[147] D. Kempe, J. Kleinberg, and E. Tardos, "Maximizing the Spread of Influence Through a Social Network," in *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, New York, NY, 2003), pp. 137–146.

[148] W. Chen, Y. Yuan, and L. Zhang, "Scalable Influence Maximization in Social Networks Under the Linear Threshold Model," in *Proceedings of the 2010 IEEE International Conference on Data Mining* (IEEE Computer Society, Washington, DC, 2010), pp. 88–97.

[149] P. Shakarian, S. Eyre, and D. Paulo, "A scalable heuristic for viral marketing under the tipping model," Soc. Netw. Anal. Min. **3**, 1225 (2003).

[150] S. Carmi, S. Havlin, S. Kirkpatrick, Y. Shavitt, and E. Shir, "A model of Internet topology using k-shell decomposition," Proc. Natl. Acad. Sci. U.S.A. **104**, 11150 (2007).

[151] G. J. Baxter, S. N. Dorogovtsev, A. V. Goltsev, and J. F. F. Mendes, "Bootstrap percolation on complex networks," Phys. Rev. E **82**, 011103 (2010).

[152] G. J. Baxter, S. N. Dorogovtsev, A. V. Goltsev, and J. F. F. Mendes, "Heterogeneous k-core versus bootstrap percolation on complex networks," *Phys. Rev. E* **83**, 051134 (2011).

[153] B. Latané, and T. L'Herrou, "Spatial clustering in the conformity game: dynamic social impact electronic groups," J. Pers. Soc. Psychol. **70**, 1218 (1996).

[154] D. Centola, "The spread of behavior in an online social network experiment," Science **329**, 1194 (2010).

[155] B. Monsted, P. Sapiezynski, E. Ferrara, and S. Lehmann, "Evidence of complex contagion of information in social media: An experiment using Twitter bots," (2017), available at https://arxiv.org/abs/1703.06027 (Accessed on May 25, 2017).

[156] D. M. Romero, B. Meeder, and J. Kleinberg, "Differences in the Mechanics of Information Diffusion Across Topics: Idioms, Political Hashtags, and Complex Contagion on Twitter," *Proceedings of the Twentieth International Conference on World Wide Web* (ACM, New York, NY, 2011), pp. 695–704.

[157] C. Fink, A. Schmidt, V. Barash, C. Cameron, and M. Macy, "Complex contagions and the diffusion of popular Twitter hashtags in Nigeria," Soc. Netw. Anal. Min. **6**, 1 (2016).

[158] C. Fink, A. C. Schmidt, V. Barash, J. Kelly, C. Cameron, and M. Macy, "Investigating the Observability of Complex Contagion in Empirical Social Networks," *Proceedings of the Tenth International AAAI Conference on Web and Social Media*, (AAAI Press, Palo Alto, CA, 2016), pp 121–130.

[159] D. State, and L. Adamic, "The Diffusion of Support in an Online Social Movement: Evidence from the Adoption of Equal-Sign Profile Pictures," *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing* (ACM, New York, NY, 2015), pp. 1741–1750.

[160] M. Karsai, G. Iñiguez, R. Kikas, K. Kaski, and J. Kertész, "Local cascades induced global contagion: How heterogeneous thresholds, exogenous effects, and unconcerned behaviour govern online adoption spreading," Sci. Rep. **6**, 27178 (2016).

[161] M. Catanzaro, M. Boguna, and R. Pastor-Satorras, "Generation of uncorrelated random scale-free networks," Phys. Rev. E **71**, 027103 (2005).

[162] Stanford Network Analysis Project (SNAP), available at http://snap.stanford.edu/data (Accessed April 23, 2015).

[163] Add Health, available at http://www.cpc.unc.edu/projects/addhealth (Accessed September 27, 2015).

[164] P. D. Karampourniotis, B. K. Szymanski, and G. Korniss, "Influence Maximization for fixed heterogeneous thresholds," (unpublished, 2017).

[165] S. Mugisha, and H. J. Zhou, "Identifying optimal targets of network attack by belief propagation," Phys. Rev. E **94**, 012305 (2016).

[166] H. J. Zhou, "Structural resilience of directed networks," (2016), available at https://arxiv.org/abs/1701.03404 (Accessed on May 25, 2017).

[167] A. Braunstein, L. DallAsta, G. Semerjian, and L. Zdeborová, "Network dismantling," *Proc. Natl. Acad. Sc. U.S.A.* **113**, 12368 (2016).

[168] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance, "Cost-Effective Outbreak Detection in Networks," *Proceedings of the Thirtieth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, New York, NY, 2007), pp. 420–429.

[169] F. D. Malliaros, M.-E. G. Rossi, and M. Vazirgiannis, "Locating influential nodes in complex networks," Sci. Rep. **87**, 925 (2016).

[170] P. Domingos, and M. Richardson, "Mining the Network Value of Customers," *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, New York, NY, 2001), pp. 57–66.

[171] M. Richardson, and P. Domingos, "Mining Knowledge-Sharing Sites for Viral Marketing," *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, New York, NY, 2002), pp. 61–70.

[172] N. Barbieri, and F. Bonchi, "Influence Maximization with Viral Product Design," *Proceedings of the 2014 SIAM International Conference on Data Mining* (ACM, New York, NY, 2014), pp. 55–63.

[173] A. Goyal, W. Lu, and L. V. S. Lakshamanan, "CELF++: Optimizing the Greedy Algorithm for Influence Maximization in Social Networks," *Proceedings of the Twentieth International Conference Companion on World Wide Web* (ACM, New York, NY, 2011), pp. 47–48.

[174] Q. Jiang, G. Song, G. Cong, Y. Wang, W. Si, and K. Xie, "Simulated Annealing Based Influence Maximization in Social Networks," *Proceedings of the Twenty-fifth AAAI Conference on Artificial Intelligence* (AAAI, Palo Alto, CA, 2011), pp. 127–132.

[175] A. Goyal, W. Lu, and L. V. S. Lakshamanan, "SIMPATH: An Efficient Algorithm for Influence Maximization under the Linear Threshold Model," *2011 IEEE Eleventh International Conference on Data Mining (ICDM)* (IEEE Computer Society, Washington, DC, 2011), pp. 211–220.

[176] C. Wang, W. Chen, and Y. Wang, "Scalable influence maximization for independent cascade model in large-scale social networks," Data Min. Knowl. Disc. **25**, 545 (2012).

[177] L. Lü, D. Chen, X.-L. Ren, Q.-M. Zhang, Y. C. Zhang, and T. Zhou, "Vital nodes identification in complex networks," Phys. Rep. **650**, 1 (2016).

[178] J. Jankowski, P. Bródka, P. Kazienko, B. K. Szymanski, R. Michalski, and T. Kajdanowicz, "Balancing speed and coverage by sequential seeding in complex networks," Sci Rep. **7**, 891 (2017).

[179] S. Pei, X. Teng, J. Shaman, F. Morone, and H. Makse, "Efficient collective influence maximization in threshold models of behavior cascading with first-order transitions," Sci. Rep. **7**, 45240 (2017).

[180] Y. Tang, X. Xiao, and Y. Shi, "Influence Maximization: Near-Optimal Time Complexity Meets Practical Efficiency," *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data* (ACM, New York, NY, 2014), pp. 75–86.

[181] X. Liu, M. Li, S. Li, S. Peng, X. Liao, and X. Lu, "IMGPU: GPU-accelerated influence maximization in large-scale social networks," IEEE Trans. Parallel. Distrib. Syst. **25**, 136–145 (2014).

[182] C. Boutsidis, E. Liberty, and M. Sviridenko, "Greedy minimization of weakly supermodular set functions," (2015), available at https://arxiv.org/abs/1502.06528v1 (Accessed on May 25, 2017).

[183] A. Alshamsi F. L. Pinheiro, and C. A. Hidalgo, "When to target hubs? Strategic diffusion in complex networks," (2017), available at https://arxiv.org/abs/1705.00232v1 (Accessed on May 25, 2017).

[184] Z. Lu, W. Zhang, W. Wu, J. Kim, and B. Fu, "The complexity of influence maximization problem in the deterministic linear threshold model," J. Comb. Optim. **24**, 374–378 (2012).

[185] F. Altarelli, A. Braunstein, L. Dall'Asta and R. Zecchina, "Optimizing spread dynamics on graphs by message passing," J. Stat. Mech.: Theory Exp. **9**, P09011 (2013).

[186] Y. Lim, A. Ozdaglary, and A. Teytelboymz, "A Simple model of cascades in networks," (2015). Available at http://t8el.com/wp-content/uploads/2015/08/SimpleCascades.pdf (Accessed on May 20, 2017).

[187] M. Weskida, and R. Michalski, "Evolutionary Algorithm for Seed Selection in Social Influence Process," *2016 IEEE/ACM International Conference on*

Advances in Social Networks Analysis and Mining (ASONAM) (IEEE Computer Society, Washington, DC, 2016), pp. 1189–1196.

[188] R. Hovstad, and N. Litvak, "Degree-degree dependencies in random graphs with heavy-tailed degrees," Internet. Math. **10**, 287 (2014).

[189] P. Singh, J. Uparna, P. D. Karampourniotis, E.-A. Horvat, B. K. Szymanski, G. Korniss, J. Z. Bakdash, and B. Uzzi, "Peer-to-peer lending and bias in crowd decision-making," (submitted to Manag. Sci., 2017).

[190] T. Friedman, *The World is Flat* (New York, NY: Farrar, Straus and Giroux, 2005).

[191] K. J. Forbes, and F. E. Warnock, "Capital flow waves: Surges, stops, flight, and retrenchment," J. Int. Econ. **88**, 235 (2012).

[192] V. Bruno, and H. S. Shin, "Cross-border banking and global liquidity," Rev. Econ. Stud. **82**,535 (2015).

[193] A. Dreher, "Does globalization affect growth? Evidence from a new index of globalization," Appl. Econ. **38**, 1091 (2006).

[194] D. Dollar, "Globalization, poverty, and inequality since 1980," World Bank Res. Obs. **20**, 145 (2005).

[195] A. Harrison, "Globalization and poverty: An introduction," NBER, 1–32 (2007).

[196] M. D. Litonjua, "The socio-political construction of globalization," Int. Rev. Mod. Sociol. **34**, 253 (2008).

[197] B. Milanovic, "Global income inequality in numbers: in history and now," Global Policy **4**, 198 (2013).

[198] N. Yazdani, and D. Mamoon, "The economics and philosophy of globalization," Economics and Philosophy of Globalization (January 20, 2012), available at SSRN: http://ssrn.com/abstract=2001063 (Accessed November 7, 2016).

[199] J. Reichardt, and S. Bornholdt, "Statistical mechanics of community detection," Phys. Rev. E **74**, 016110 (2006).

[200] C. R. Reinhart, and V. R. Reinhart, "Capital flow bonanzas: an encompassing view of the past and present," NBER Working Paper No. 14321 (2008).

[201] A. Sen, *Inequality Reexamined* (Harvard University Press, Cambridge, MA, 1992).

[202] A. Nishi, N. A. Christakis, "Human behavior under economic inequality shapes inequality," Proc. Natl. Acad. Sci. U.S.A. **112**, 15781 (2015).

[203] R. Florida, "The World is spiky," The Atlantic Monthly **48**, (2005), available at www.theatlantic.com/past/docs/images/issues/200510/world-is-spiky.pdf (Accessed Nov 7, 2016).

[204] P. Ghemawat, "Why the world isn't flat," Foreign Policy, (2009), available at foreignpolicy.com/2009/10/14/why-the-world-isnt-flat/ (Accessed Nov 7, 2016).

[205] R. E. Lucas, "Why doesnt capital flow from rich to poor countries?" Am. Econ. Rev. **80**, 92 (1990).

[206] A. Banerjee, A. G . Chandrasekhar, E. Duflo, and M. O. Jackson, "The diffusion of microfinance," Science **341**, 6144 (2003).

[207] L. Alfaro, S. Kalemli-Ozcan, and V. Volosovych, "Why doesnt capital flow from rich to poor countries? An empirical investigation," Rev. Econ. Stat. **90**, 347 (2008).

[208] M. Chinazzi, G. Fagiolo, J. A. Reyes, and S. Schiavo, "Post-mortem examination of the international financial network," J. Econ. Dyn. Control. **37**, 1692 (2013).

[209] T. Bruett, "Cows, Kiva, and Prosper.Com: How disintermediation and the internet are changing microfinance," Comm. Dev. Inv. Rev. **3**, 44 (2007).

[210] S. Bandyopadhyay, T. Sandler, and J. Younas, "Foreign Aid as Counterterrorism Policy," Oxford Economic Papers **63**, 423 (2011).

[211] S. Mallaby, "The reluctant imperialist: Terrorism, failed states, and the case for american empire," Foreign Affairs **81**, 2 (2002).

[212] H. Roy, S. Kase, "The Relation Between Microfinacing and Corruption by Country: An Analysis of an Open Source Dataset," Intelligence and Security Informatics (ISI), 2015 IEEE International Conference, (2015), available at http://ieeexplore.ieee.org/document/7165959/ (Accessed Nov 7, 2016).

[213] E. Mollick, "The dynamics of crowdfunding: An exploratory study," J. Bus. Venturing **29**, 1 (2014).

[214] G. Burtch, A. Ghose, and S. Wattal, "Cultural differences and geography as determinants of online pro-social lending," MIS Quarterly **38**, 773 (2014).

[215] S. Itzkovitz, R. Milo, N. Kashtan, G. Ziv, and U. Alon, "Subgraphs in random networks," Phys. Rev. E **68**, 026127 (2003).

[216] Kiva, "Loans that change lives," (2016), available at www.kiva.org (Accessed Nov 7, 2016).

[217] AidData, "Open data for international development," (2016), available at http://aiddata.org (Accessed Nov 7, 2016).

[218] D. Gefen, and E. Carmel, "Is the world really flat? A look at offshoring at an online programming marketplace," MIS Quarterly **32**, 1 (2008).

[219] B. K. Szymanski, X. Lin, A. Asztalos, and S. Sreenivasan, "Failure dynamics of the global risk network," Sci. Rep. **5**, 10998 (2015).

[220] R. Cohen, K. Erez, D. Ben-Avraham, and S. Havlin, "Resilience of the internet to random breakdowns," Phys. Rev. Lett. **85**, 4626 (2008).

[221] J. Galak, D. Small, and A. Stephen, "Micro-finance decision making: A field study of prosocial lending," J. Mark. Res. **48**, S130 (2011).

[222] N. Gotelli, and G. Graves, "Null Models in Ecology," Smithsonian Institution Press, (1996).

[223] O. Sagarra, C. J. Perez Vicente, and A. Diaz-Guilera, "Statistical mechanics of multiedge networks," Phys. Rev. E **88**, 062806 (2013).

[224] M. E. J. Newman, and J. Park, "Why social networks are different from other types of networks," Phys. Rev. E **68**, 036122 (2003).

[225] R. M. May, S. A. Levin, and G. Sugihara, "Complex systems: Ecology for bankers," Nature **451**, 893 (2008).

[226] J. E. Anderson, and E. van Wincoop, "Gravity with gravitas: A solution to the border puzzle," Am. Econ. Rev. **93**, 170 (2003).

[227] B. Uzzi, "Embeddedness in the making of financial capital: How social relations and networks benefit firms seeking financing," Am. Sociol. Rev. **64**, 481 (1999).

[228] P. Ingram, "The intergovernmental network of world trade: IGO connectedness, governance, and embeddedness," Am. J. Sociol. **111**, 824 (2005).

[229] R. C. Dewar, and A. Porte, "Statistical mechanics unifies different ecological patterns," J. Theor. Biol. **251**, 389 (2008).

[230] R. J. Williams, "Biology, methodology or chance? The degree distributions of bipartite ecological networks," PLoS ONE **6**, e17645 (2011).

[231] J. Park, and M. E. J. Newman, "Statistical mechanics of networks," Phys. Rev. E **70**, 066117 (2004).

[232] E. T. Jaynes, "Information theory and statistical mechanics," Phys. Rev. **106**, 620 (1957).

[233] T. Britton, M. Deijfen, and A. Martin-Lőf, "Generating simple random graphs with prescribed degree distribution," J. Stat. Math. Phys. **124**, 1377 (2005).

[234] F. Molnár, S. Sreenivasan, B. K. Szymanski, and G. Korniss, "Minimum dominating sets in scale-free network ensembles," Sci Rep. **3**, 1736 (2003).

[235] M. E. J. Newman, "Assortative mixing in networks," Phys. Rev. Lett. **89**, 208701 (2002).

[236] F. Jr. Molnár, N. Derzsy, É. Czabarka, L. Székely, B. K. Szymanski, and G. Korniss, "Dominating scale-free networks using generalized probabilistic methods," Sci. Rep. **4**, 6308 (2014).

[237] G. Grimmett, and D. Welsh, *Probability: An Introduction* (Oxford: Oxford University Press, 2014).

[238] P. Egger, "A note on the proper econometric specification of the gravity equation," Econ. Lett. **66**, 25 (2000).

[239] J. V. Tu, P. C. Austin, and B. B. Chan , "Relationship between annual volume of patients treated by admitting physician and mortality after acute myocardial infarction," J. Am. Med. Assoc. **285**, 3116 (2001).

[240] T. Mayer, and S. Zignago, "Notes on CEPII's distances measures: The GeoDist database," CEPII Working Paper **25**, (2011).

[241] D. Ratha, and W. Shaw, *South-South Migration and Remittances* (The World Bank, Washington, D.C. 2007), World Bank Working Paper No. **102**.

[242] K. Head, T. Mayer, and J. Ries, "The erosion of colonial trade linkages after independence," J. Int. Econ. **81**, 1 (2010).

[243] K. P. Burnham, and D. R. Anderson, *Model Selection and Multimodel Inference* (Springer-Verlag, New York, 2002).

[244] A. Agrawal, C. Catalini, and A. Goldfarb, "Crowdfunding: Geography, social networks, and the timing of investment decisions," J. Econ. Manag. Strateg. **24**, 253 (2015).

[245] R. B. OHara and D. J. Kotze, "Do not log-transform count data," Methods Ecol. Evol. **1**, 118 (2010).

[246] A. C. Cameron, and P. K. Trivedi, *Regression Analysis of Count Data* (Cambridge University Press, Cambridge UK, 1998).

[247] A. E. Raftery, "Bayesian model selection in social research," Sociol. Methodol. **25**, 111 (1995).

[248] World development indicators, The World Bank (2016), available at http://data.worldbank.org/data-catalog/world-development-indicators (Accessed Nov 7, 2016).

APPENDIX A
Synthetic and Empirical Networks

A.1 Generation of Synthetic Networks

The networks we use are undirected and unweighted. The synthetic networks used are Erdős-Rényi (ER) graphs and scale-free (SF) networks. For the generation of ER graphs [21] we used the $G(N, p_{\mathrm{ER}})$ model with N being the system size and p_{ER} the probability that a random node will be connected to any node in the graph. The probability p_{ER} is given by $p_{ER} = z/(N-1)$, where z is the nominal average degree in the network. We keep the average degree $z = 10$. For the generation of uncorrelated SF networks [29, 161] ($N = 10^4$, $z = 10$, with power law constant $\gamma = 3$) we employ the configuration model [161, 233] with a structural cut-off, and a maximum possible node degree set to $\sqrt{N}$, using a high accuracy look-up table from [234].

A.2 Generation of Synthetic Networks with Controlled Assortativity

The degree assortativity was first introduced by Newman [235] to describe the connectivity between neighboring nodes with different degrees. To measure it we use Spearman's ρ [188]. ER graphs have degree assortativity measured with Spearman's $\rho = 0$. To control the degree assortativity we use the method applied in [236].

A.3 Empirical Networks

The empirical networks used are a connected ego-network from a Facebook (FB) dataset, available from the Stanford Network Analysis Project (SNAP) [162] (system size $N = 4048$, average degree $z = 43$), and a high-school (HS) friendship network [163]. For the HS network, we only used the giant connected component of that network, with $N = 921$ and $z = 5.96$. The network contains two communities

which are roughly equal in size (Table A.1). Although SF, FB, and HS networks are connected networks, the generated ER graphs may have a disconnected component with probably e^{-z}, which for $z = 10$ is approximately 0.000045.

Table A.1: Basic statistics of the two empirical networks used. The properties measured are: the type of network (directed or undirected), total number of nodes N, total number of edges m, average degree z, power law coefficient α, network diameter d, fraction of closed triangles C_1, average clustering coefficient C_2, assortativity (Spearman's) ρ.

Network	Type	N	m	z	α	d	C_1	C_2	ρ
FB	Undir.	4039	88234	43.691	1.72	8	0.2647	0.6055	0.5432
HS	Undir.	921	2745	5.9674	3.30	12	0.0521	0.1254	0.2817

APPENDIX B
Analytical Approximation for the Linear Threshold Model

For analytic methods, we apply Gleeson and Cahalane's tree-like approximation for synthetic networks [133, 134]. The approximation is given by the following set of equations

$$S_{eq} = p + (1-p)\sum_{k=1}^{\infty} P_k \sum_{m=1}^{k} \binom{k}{m} q_{\infty}^{m} \left(1-q_{\infty}\right)^{k-m} F\left(\frac{m}{k}\right) \tag{B.1}$$

$$q_{n+1} = p + (1-p)\sum_{k=1}^{\infty} \frac{k}{z} P_k \sum_{m=1}^{k-1} \binom{k-1}{m} q_{n}^{m} \left(1-q_{n}\right)^{k-m-1} F\left(\frac{m}{k}\right). \tag{B.2}$$

In this approximation the graph is considered an infinite level tree. The spread diffuses level-by-level starting from the bottom of the tree. q_n is defined as the conditional probability that a node on level n is active, conditioned on its parent on level $n+1$ being inactive" and it is given by Eq. (B.2). The final spread S_{eq} is given by Eq. (B.1), and is measured at the top of the tree. The fraction of initially active nodes is given by p. In the bottom of the tree at level $n = 0$, the fraction of active nodes is only based on the initiators, thus $q_0 = p$. The graph degree distribution is given by P_k, which for an infinite size ER graph is given by $P_k = \left(z^k e^{-z}\right)/k!$, where z is the average degree, while for SF networks it's given by $P_k \sim k^{-\gamma}$. $F\left(\frac{m}{k}\right)$ is the cumulative probability that a node requires m or less active neighbors to get active, which depends on the assigned threshold distribution.

B.1 Closed-form Analytical Estimate for Uniform Thresholds

Here, we show explicitly the derivation of the closed form equation of the treelike approximation [133, 134] of the fraction S_n of active nodes at level n on Eq. (6) in the main text. According to [134] the level (or time) dependent evolution of the fraction q_{n+1} of nodes with inactive parents at level $n+1$ for synchronous

updating of the nodes is given by

$$q_{n+1} = g(q_n) = p + (1-p)\sum_{k=1}^{\infty}\frac{k}{z}P_k\sum_{m=1}^{k-1}\binom{k-1}{m}q_n^m\left(1-q_n\right)^{k-m-1}F\left(\frac{m}{k}\right), \quad \text{(B.3)}$$

and the fraction of active nodes at level $n+1$ is given by

$$S_{n+1} = h(q_n) = p + (1-p)\sum_{k=1}^{\infty}P_k\sum_{m=1}^{k}\binom{k}{m}q_n^m\left(1-q_n\right)^{k-m}F\left(\frac{m}{k}\right). \quad \text{(B.4)}$$

The replacement of the cumulative probability function $F\left(\frac{m}{k}\right)$ in the particular case of a uniform distribution of thresholds in the above two equations yields the closed form solution. Let a node i have degree k and an assigned threshold ϕ. Resistance l is the absolute number of active neighbors required for node i to get activated, and it is given by $l = \text{ceil}(\phi \times k)$. The cumulative probability distribution $F\left(\frac{m}{k}\right)$ of nodes with degree k, having resistance less or equal to m, is given by $F\left(\frac{m}{k}\right) = \sum_{k=1}^{m} r_{l,k}$, where $r_{l,k}$ is the probability that a node has resistance l, conditioned that it has degree k. For a uniform threshold distribution the probability that a node has resistance l, conditioned that it has degree k, is $r_{(l,k)} = 1/k$. For example, a node with degree $k = 2$ will have resistance $l = 1$, with probability $r_{(1,2)} = 1/2$ and resistance $l = 2$ with probability $r_{(2,2)} = 1/2$. Thus, the fraction $F\left(\frac{m}{k}\right)$ of nodes that have resistance m or less conditioned that they have degree k for the uniform random threshold distribution is given by

$$F\left(\frac{m}{k}\right) = \sum_{k=1}^{m} r_{l,k} = \sum_{k=1}^{m}\frac{1}{k} = \frac{m}{k}. \quad \text{(B.5)}$$

Now, replacing Eq. B.5 in Eq. B.3 we show the linear relationship between the fraction q_{n+1} of nodes with inactive parents at level $n+1$ with the fraction q_n at the previous level n of the approximated tree for networks with uniform distribution of thresholds (see Eq (3) in the main text). So,

$$q_{n+1} = p + (1-p)\sum_{k=1}^{\infty}\frac{k}{z}P_k\sum_{m=1}^{k-1}\binom{k-1}{m}q_n^m\left(1-q_n\right)^{k-1-m}\frac{m}{k}, \quad \text{(B.6)}$$

which simplifies to

$$q_{n+1} = p + (1-p)\,\frac{1}{z}\sum_{k=1}^{\infty} P_k \sum_{m=1}^{k-1} \binom{k-1}{m} q_n^m \,(1-q_n)^{k-1-m}\, m. \tag{B.7}$$

However,

$$\sum_{m=1}^{k} \binom{k}{m} q_n^m \,(1-q_n)^{k-m}\, m = \sum_{m=0}^{k} \binom{k}{m} q_n^m \,(1-q_n)^{k-m}\, m, \tag{B.8}$$

where the right hand of the equation is the mean of the binomial distribution, and it is given by kq_n [237], thus

$$\sum_{m=1}^{k-1} \binom{k-1}{m} q_n^m \,(1-q_n)^{k-1-m}\, m = (k-1)\,q_n \tag{B.9}$$

Using the above equation in Eq. (B.7) yields

$$q_{n+1} = p + (1-p)\,\frac{1}{z}\sum_{k=1}^{\infty} P_k\,(k-1)\,q_n, \tag{B.10}$$

which can be rewritten as

$$q_{n+1} = p + (1-p)\,\frac{1}{z}\left(\sum_{k=0}^{\infty} P_k\,(k-1) + P_0\right) q_n. \tag{B.11}$$

Since the average degree is given by $z = \sum_{k=0}^{\infty} kP_k$, the above equation becomes

$$q_{n+1} = p + (1-p)\,\frac{1}{z}\,(z - 1 + P_0)\,q_n. \tag{B.12}$$

which can be rewritten as

$$q_{n+1} = p + bq_n, \tag{B.13}$$

with $b = (1-p)\,\frac{1}{z}\,(z-1+P_0)$. The solution of the above equation with initial condition $q_0 = p$ is

$$q_n = p\frac{1-b^{n+1}}{1-b} \tag{B.14}$$

Similarly, replacing $F\left(\frac{m}{k}\right)$ in B.4 by the right hand side of Eq. (B.5), the analytic approximation yields

$$S_{n+1} = p + (1-p)\sum_{k=1}^{\infty} P_k \sum_{m=1}^{k} \binom{k}{m} q_n^m \left(1-q_n\right)^{k-m} \frac{m}{k}. \tag{B.15}$$

Using again the property of the mean of the binomial distribution the above equation reduces to

$$S_{n+1} = p + (1-p)\sum_{k=1}^{\infty} P_k \frac{1}{k}\left(kq_n\right), \tag{B.16}$$

which yields

$$S_{n+1} = p + (1-p)\, q_n \sum_{k=1}^{\infty} P_k. \tag{B.17}$$

Thus, the closed form solution of cascade size at level $n+1$ is given by

$$S_{n+1} = p + cq_n, \tag{B.18}$$

with $c = (1-p)\left(1-P_0\right)$. Subtracting S_n from both parts of the above equation and combining it with Eq. B.13 we get

$$S_{n+1} - S_n = c\left(q_n - q_{n-1}\right). \tag{B.19}$$

Substituting $q_n = p + bq_{n-1}$ from Eq. B.13 into the above equation yields

$$S_{n+1} - S_n = c\left(p + (b-1)q_{n-1}\right). \tag{B.20}$$

Solving Eq. B.18 for q_{n-1} at level $n-1$ and substituting to the above equation yields

$$S_{n+1} - S_n = c\left(p + (b-1)\left(\frac{S_n - p}{c}\right)\right). \tag{B.21}$$

Expansion of the above equation yields to the closed form phase-space equation at Eq. (6) in the main text

$$S_{n+1} - S_n = cp - (1-b)p - (1-b)S_n. \tag{B.22}$$

Now, going back to the calculation of S_{n+1} at Eq. B.18, substituting q_n with the right part of Eq. B.14 yields

$$S_{n+1} = p + cp\frac{b^{n+1} - 1}{b - 1}, \tag{B.23}$$

where the cascade size S_0 at level $n = 0$ is just the fraction of the initiators, $S_0 = p$. On the other hand, in the equilibrium state (as $n \to \infty$) the cascade size S_{eq} is given by

$$S_{eq} = p + cp\frac{1}{1 - b}, \tag{B.24}$$

since $0 \leq b < 1$. Interestingly, the final cascade size doesn't depend for uncorrelated networks on the degree distribution, but only on the average degree $< k >$.

APPENDIX C
Further Information and Analysis on the Kiva Data

C.1 Extended Introduction

Example of Kiva Borrower and Lender Narratives (Fig. C.1).

Reasons Asking for Loan Text

The following description was written by Moses O., a volunteer with Village Enterprise Fund and partner representative for Kiva in Uganda:Justine O. is among the most successful small-scale business leaders. He got a grant of 100 US dollars from VEF. He started by buying and selling of goats right away. Most butchers have known his business, so a number of them come to him to buy goats. He can now buy and sell between 15 and 20 goats in a month -- a great amount. The market is very open since most of Ugandans rear goats and cows for raising money whenever needs arise like school fees, sickness, death, journey etc.O. has attended business training and he is capable of handling the loan and paying it back. Given a loan of 500 US dollars, he is targeting to introduce bulls for slaughter and opening up a butcher shop himself.

Reasons for Lending Text

"If everybody did the world may just be a better place."

"I can. Good ideas and ambitious people deserve a chance."

"As Mother Teresa so wisely said, "We can do no great things, only small things with great love."

Figure C.1: A sample of representative borrowers and lenders' images and reasons for asking taken from Kiva's webpage [216].

A visualization of the Kiva sub-networks from the year 2007 is shown in Figure C.2

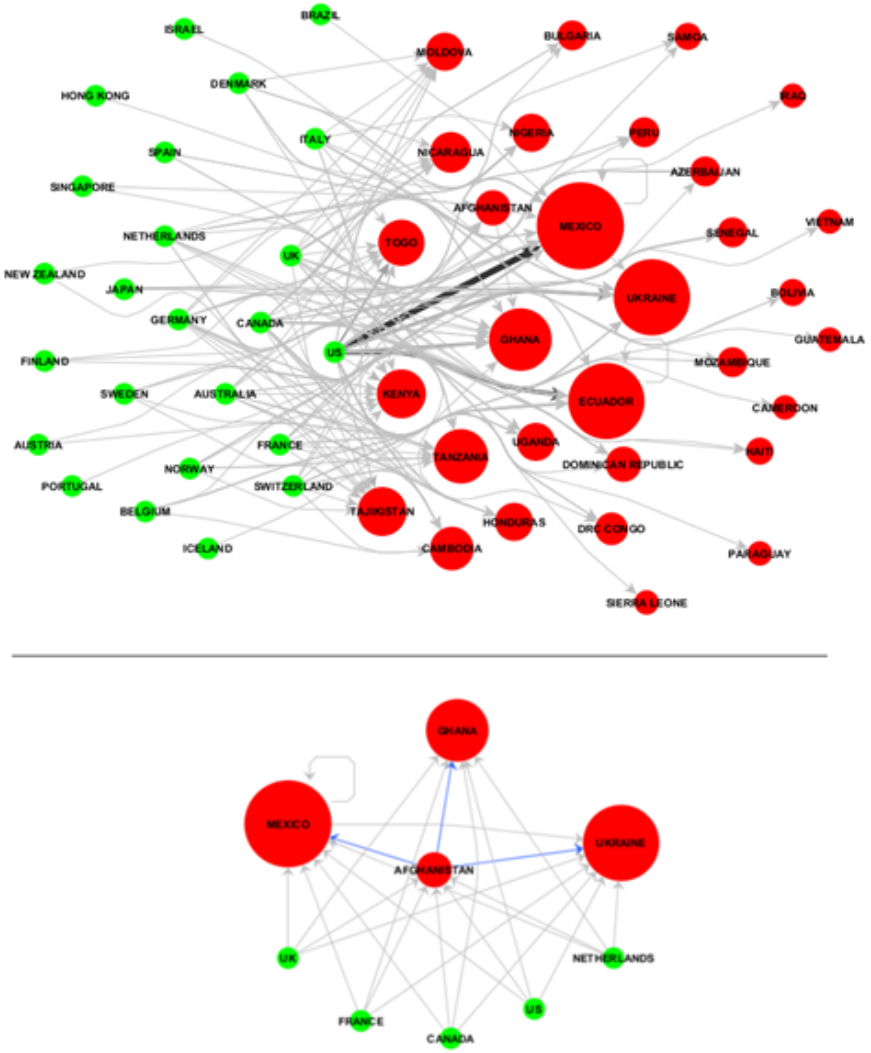

Figure C.2: Examples of Kiva sub-networks. Top 200 links by number of transactions in the 2007 Kiva network (top). The borrower (lender) countries are colored red (green). The size of borrower country nodes is proportional to the received transactions; whereas, the lender countries are shown to be of the same size. Edge thickness is related to the number of transactions from lender country to borrower country. The figure contains only a subset of country–pairs for clarity. The ego-network of Afghanistan is for the same year (bottom). The outgoing links from Afghanistan have been colored differently following the same convention for node size and link thickness.

C.2 Gravity Model and Regression Analysis

To further investigate the factors associated with lending bias between nation pairs, we regress the level of lending between nations on factors effecting bilateral international trade with the widely used fixed-effect gravity model [226,238]. In this model, the level of trade from country i to country j, Y_{ij}, is modeled as

$$Y_{ij} = 1 - G\frac{M_i^{\alpha} M_j^{\beta}}{M_{ij}^{\gamma}} \tag{C.1}$$

where M_i and M_j are the economic masses (e.g., GDP) of i and j, d_{ij} is the geographical distance between i and j, and G is a constant. The parameters to be estimated are α, β, and γ, respectively. We aggregate transactions such that each observation Y_{ijfy} denotes the number of transactions from the lender country i to the borrower country j involving the Kiva field partner f for given year y. Field partners are microfinance institutions (e.g., NGOs, schools, or social enterprises) operating in the borrower country and are responsible for connecting borrowers with Kiva, screening them, posting their loan requests online, and disbursing and collecting repayments. Since many country–pairs in our data show zero transactions, the log transformation of the level of bilateral trade typically used in the gravity model is not feasible in our setting. Thus, we ran a second model that appropriately accounts for the skewness in the level of loans between countries by discretizing the dependent variable Y_{ijfy} into four categories (denoted by Q_{ijfy}) that correspond to zero, low, medium, or high levels of lending [239]. We performed a fixed-effects ordered logistic regression on the transformed variable to control for unobserved heterogeneity related to the lender country, borrower country, Kiva field partner, or year. Zero transactions category is the omitted category. The ordered logit and the gravity model produce qualitatively similar results (see Table C.5 and C.6).

Per the gravity model, we include four explanatory variables in our regression: (i) the difference of per capita GDP between lender and borrower countries (World Bank data [248]); (ii) the geographical distance between the country–pairs [240]; (iii) the size of the migrant population of borrower country living in the lender country [241]; and (iv) an indicator variable showing that lender country colonized

borrower country (1 = yes) , which captures common culture and institutional structures [240, 242] (see Tables C.2 and C.3. Our model is as follows:

$$Q_{ijfy} = \beta_1 \text{GDP difference}_{ij} + \beta_2 \text{Distance}_{ij} + \beta_3 \text{Migration}_{ji} + \beta_4 \text{Colony}_{ij} + \epsilon_{ijfy} \quad \text{(C.2)}$$

This model (Model 4) unequivocally had the best fit, with an evidence/likelihood ratio of 12.05 105 over the next best fit model (Model 3) [243]. The regression findings reported in Table C.1 suggest that bilateral transaction volumes in this peer-to-peer lending system reflect general patterns of trade between nations rather than unique peer-to-peer patterns. The per capita GDP difference between countries, migration between county pairs, and the historical presence of a colonial relationship are all positively (odds ratio > 1) and significantly associated with lending volumes, while geographical distance is negatively and significantly associated with the level of lending (odds ratio < 1). These findings suggest that the greater global context within which peer-to-peer lending is embedded impacts crowdfinancing in much the same way that it does other forms of global trade. We also apply the same model on AidData using four categories of country-to-country government aid money (zero, low, medium, high) as the outcome variable (see Appendix C.4). The results shown in the last column of Table C.1 imply that distance, migration, and colonial tie are associated with level of aid in the same manner. However, much higher odds ratio for migration and colony (compared to Kiva) indicate that these variables have a much stronger association with flow of government aid. Surprisingly the effect of per capita GDP difference is not found to be significant, which is positive and significant for Kiva.

To depict these effects in Kiva over the range of the variables, we plot the relationship between transaction flows, GDP difference, and migration from an ordered logistic regression (ologit) using quantiles of GDP difference and high and low migration (split at the median). Fig. C.3 shows the probability of high transaction volumes (8 to 54,136 transactions) at different quantiles of GDP difference for different levels of migration. The plot shows an increasing trend in lending associated with growing per capita GDP for country–pairs that share a large (above the median) and no significant change for small (below the median) immigrant population.

Table C.1: Fixed-effect ologit estimates of levels of lending between countries. Odds ratio reported for 4 levels of transactions (4 levels of commitment amount in the case of government aid). **$p < 0.05$

Variable	Odds ratio Model 1	Odds ratio Model 2	Odds ratio Model 3	Odds ratio Model 4	Odds ratio (Govt. Aid)
GDP (pc) difference	1.76**	1.73**	1.73**	1.74**	0.99
Distance	...	0.94**	0.94**	0.94**	0.77**
Migration	...	...	2.47**	2.24**	2.52**
Colony	...	...	...	1.41**	12.65**
AIC:	181781.5	176346.9	162648.1	162624.7	...
BIC:	181950.9	176555.7	162855	162624.7	...
Fixed effects:					
Year	Yes	Yes	Yes	Yes	Yes
Partner	Yes	Yes	Yes	Yes	-
Lender (donor) country	Yes	Yes	Yes	Yes	Yes
Borrower (recipient) country	Yes	Yes	Yes	Yes	Yes

We observe that the effect of GDP difference is weak up to its 60th percentile after which it shows a much stronger impact on loan levels. This suggests that much of the source of bias in the system is keyed to high GDP lenders. Specifically, for lower than 60th percentile, the probability of observing biasedly high-volume transactions is quite small (< 0.2) but grows rapidly for higher percentiles of GDP difference (0.75 at 90th percentile, in the case where migration level is also high).

Interestingly, the results show that migration from borrower to lender country only plays a role when the per capita GDP of the lender country is sufficiently higher than that of the borrower country (otherwise migration shows a slight negative association). It can also be seen that higher GDP difference with high migration has a strong positive effect on the transaction volumes, suggesting that the deeply embedded structures that characterize relationships among nations continue to impact the

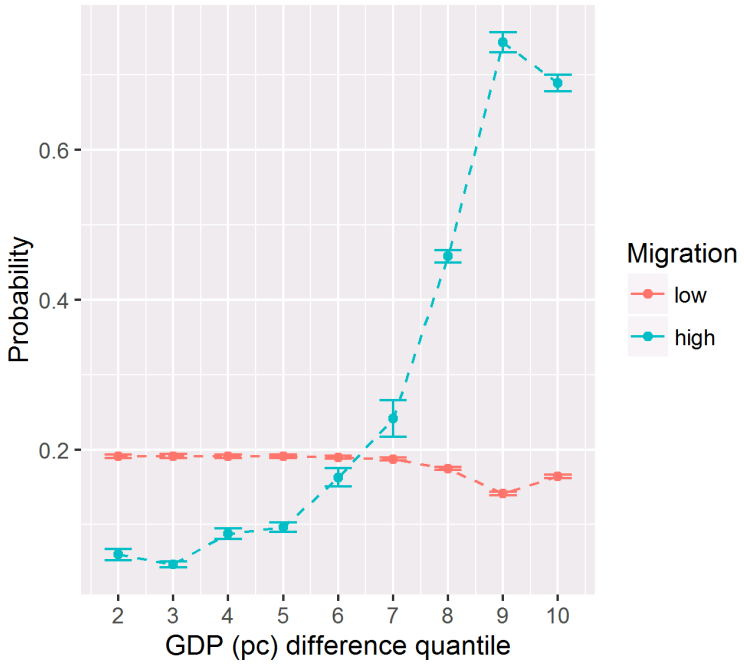

Figure C.3: Marginal effects of GDP per capita difference and level of migration. The vertical axis measures the probability of observing large numbers of transactions (i.e., the outcome Q_{ijfy} falling into a high category), as a function of GDP difference quantile and for different levels of migration (low vs. high). For low migration the probability shows no increase with GDP difference quantile, but for high migration the probability shows a significant increase – specifically beyond the 50th percentile of GDP difference. The plot shows that migration is only effective when it moves migrants from a low GDP to a high GDP country (which corresponds to direction across a large and positive GDP difference).

networked systems such as Kiva. These findings indicate that while crowdfinancing may have reduced some biases [244] in the lending system, the greater global context within which peer-to-peer lending is embedded impacts crowdfinancing in much the same way it does other forms of global trade. Factors associated with the magnitude of bias continue to be correlated with lending pair relationships that deviate from flatness.

C.2.1 Regression Specification

The ordered logit is a non-linear model where the dependent variable Y_{ijfy} (defined as the aggregated number of transactions from the lender country i to the borrower country j and involving the Kiva field partner f for a given year y) is converted from a continuous variable to quantiles of transaction count between countries (amount of aid between countries in the case of government aid) as the dependent variable with outcomes zero (1), low (2), medium (3), and high (4) based on natural break points in the distribution. This conversion is done to deal with the non-normality of count data that makes up the dependent variable, the problem caused by log transforming the variables with zero values [245], and also because of the limitation of Poisson models for dealing with this type of data (skewed distribution and containing a large number of zero observations) [246]. We supplement the Kiva data with our explanatory variables: per capita GDP difference (averaged over 2005–2013), inter-country distance, migration, and a categorical variable indicating whether the lender country was a colonizer of the borrower country in the past (a colonial tie). Data on distance between lending and borrowing countries and the presence of absence of colonial past relationships between countries were obtained from the GeoDist data of CEPII, Research and Expertise on the World Economy [240]. Country per capita GDP data were obtained from the World Banks World Development Indicators. Finally, data about the number of immigrants between countries came from 2010 estimates of the International Migrant Stocks of the United Nations population division [241]. Since the data are obtained from different sources, after merging, our number of observations is reduced from 174,468 to 140,418 due to availability of data. In addition, the model considers the fixed effects of lender country, borrower country, field partner, and year. (See Appendix for a summary of the dependent and the independent variables (Table C.2) and correlations among them (Table C.2).) We check the robustness of our model by comparing it to other models that use a subset of explanatory variables. The model we use corresponds to the optimal set of Akaike information criterion (AIC) and Bayesian information criterion (BIC) statistics [247] (Table C.1). To test for multicollinearity among explanatory variables, VIF statistics were checked and found to

be satisfactorily low.

C.3 Extended Information on Regression Analysis

C.3.1 Categorical Dependent Variable

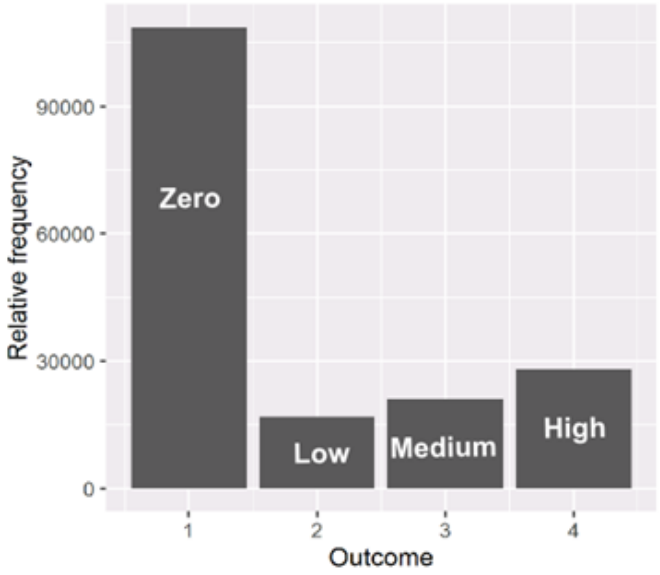

Figure C.4: Outcome variable for Kiva loans. Quantiles of Y_{ijfy}. Outcomes (Q) represents zero (0 transactions), low (1 transaction), medium (2–7 transactions), and high volume (8–54,136 transactions) of transactions, respectively.

Table C.2: Descriptive statistics.

Variable	Obs.	Mean	SD	Min	Max
Q (outcome)	174,468	1.82	1.16	1	4
GDP (pc) difference (thousands USD)	157,609	11.24	21.62	-47.96	122.16
Distance (thousands of kms)	164,803	8.55	4.55	0.010	19.95
Migration (Millions)	155,558	0.01	0.18	0	11.63
Colony	164,803	0.01	0.09	0	1

Table C.3: Correlation matrix.

Variable	Q (outcome)	GDP (pc) difference	Distance	Migration	Colony
Q (outcome)	1	...	...	...	...
GDP (pc) difference	0.42	1	...	...	...
Distance	-0.05	-0.03	1	...	...
Migration	0.06	0.03	-0.04	1	...
Colony	0.11	0.02	-0.03	0	1

C.3.2 Gravity Model

The results shown from the gravity model are qualitatively consistent with the ologit model in section C.2. They show a positive and significant association of transaction with economic disparity, migration and colony, and a negative and significant association with geographical distance. Here we model the number of transactions Y_{ijfy} from country i (lender) to country j (borrower) through the field partner f and in a given year y, using the gravity equation in the following way:

$$\log\left(Y_{ijfy}\right) = \log\left(G\right), \tag{C.3}$$

where G is a constant, GDP_i and GDP_j are the per capita GDP of the lender and the borrower countries, *distance* is the geographical distance between i and j, *migration* is the migrant population of borrower country in the lender country, *colony* represents a colonial link between i and j (i being colonizer of j) and ϵ_{ijfy} is the error term. The model coefficient to be estimated is α, β, γ, and δ. We also include the fixed effects of lender country, borrower country, field partner, and year. Equation C.3 is the log transformed gravity equation (with fixed effects) where we included terms that capture the economic disparity between lender and borrower country (as the ratio of their per capita GDPs), distance, migration, and colonial past. The associated coefficients are estimated by performing a linear regression (see Tables C.4, C.5 and C.6).

The results in Table C.5 indicate that 1 unit increase in GDP ratio is associated with a 408% increase (fractional change $= e^a - 1$) in number of transactions, 1

Table C.4: Descriptive statistics for gravity model.

Variable	Obs.	Mean	SD	Min	Max
log (transactions)	65896	2.09	2.04	0	10.90
log (GDP ratio)	157609	0.78	1.52	-4.44	5.36
log (distance)	164803	1.93	0.77	-4.56	2.99
log (migration)	47081	-7.61	3.10	-13.81	2.45
Colony	164.803	0.01	0.09	0	1

Table C.5: Correlation matrix for gravity model.

	log (transactions)	log (GDP ratio)	log (distance)	log (migration)	Colony
log (transactions)	1	...	...	...	...
log (GDP ratio)	0.3575	1	...	...	...
log (distance)	0.20	0.23	1	...	...
log (migration)	0.35	-0.10	-0.28	1	...
Colony	0.08	0.01	0.08	0.27	1

Table C.6: Gravity model regression. Gravity model regression with number of transactions as the dependent variable. $N = 30216$**, goodness of fit** $R^2 = 0.85$**.**

log (transactions)	Coefficient	Robust SE	t	$P > \|t\|$
log (GDP ratio)	1.63	0.019957	81.53	0
log (distance)	-0.11	0.013652	-7.89	0
log (migration)	0.02	0.004096	4.84	0.001
Colony	0.12	0.017699	7.09	0
Fixed effects	...	...	...	...
Lender country	Yes	...	...	...
Borrower country	Yes	...	...	...
Field partner	Yes	...	...	...
Year	Yes	...	...	...

unit increase in distance is associated with a 10% decrease in transactions, 1 unit increase in migration is associated a 2% increase in transactions, and a presence of a colonial tie is associated with a 13% increase in transactions. Although the numerical estimates shown above cannot be compared exactly with the results obtained by the ologit model, their relationship to the bilateral transaction levels is similar.

We also look at the interaction between per capita GDP difference and migration by considering 10 quantiles of per capita GDP difference and migration (high

= above median, low = below median) and modeling the number of transactions by the fixed-effect gravity model discussed above. The trend shown in Figure C.5 is found to be qualitatively consistent with the ologit regression discussed in Appendix C.2.

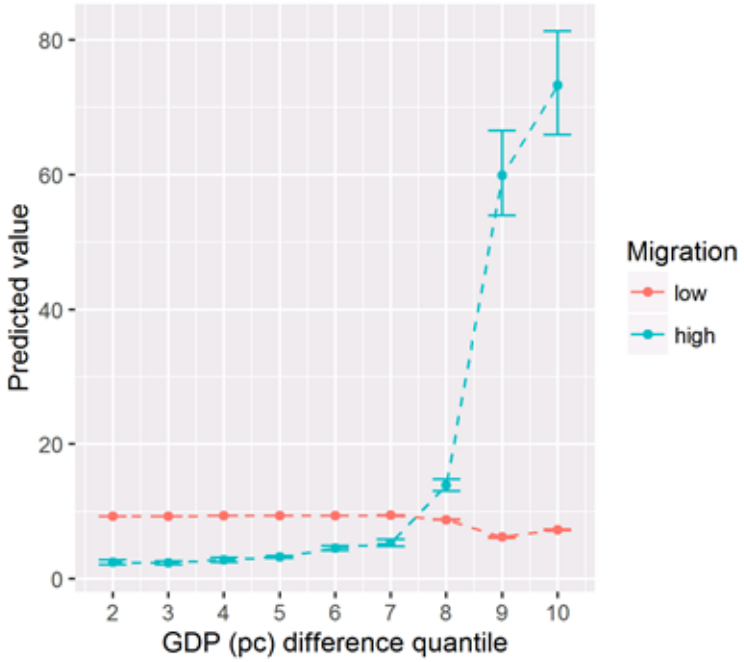

Figure C.5: Predicted transactions. Predicted number of bilateral transactions as a function of per capita GDP difference quantile and level of migration. Error bars indicate ± 2 standard error (i.e., 95% confidence interval).

C.4 Aid Data

C.4.1 Global Financial Lending Flows: Kiva vs. Government Aid

We compare the participation level of countries on Kiva and aggregated aid using data from AidData, (available at: http://aiddata.org/) from one country to another (only looking at country-to-country aid) for the same time period as Kiva (2005–2013). Figure C.6 (A) and (C) show the sum of commitment aid money (USD) given and received, respectively, by each donor country; and Figure C.6 (B) and (D) show the total loan contributions made by the lenders in a lending country and total contributions made to the loans and borrower country, respectively.

The distribution of receivers of money through bilateral aid and through Kiva (by individual lenders) looks quite different. The Aid is distributed among recipient countries more uniformly whereas Kiva focuses mostly on fewer developing regions. The other distinguishing feature of Kiva is the global presence of individual lenders. The donor countries providing aid in the AidData are fewer in numbers (48) in comparison to Kiva lenders contributing from almost every country in the world (i.e., capital flow from few-to-many vs. many-to-few). Thus the Kiva dataset accounts for a much larger number of inter-country links that reach developing regions from developed regions.

C.4.2 Analysis of Government Aid Data

We construct a null model for the co-country aid network using data on international development aid [217] and extracting the yearly flow of country-to-country government aid (for the years 2005–2012). The null model is constructed by randomly rewiring the multi-edges in the network, where each (directional) edge represents an aid commitment made between a pair of countries. We preserve the total number of incoming edges and outgoing edges for each node (country). As in the case of the Kiva network (described in Chapter 3), by comparing the observed network with the null model we identify the biased links and compute the yearly flatness (as fraction of unbiased links in the given year). The flatness of the aid network is shown in Figure C.7. We observe that the level of flatness in this network is lower than Kiva and does not follow a systematic trend. It can be inferred from Figure C.7 that lending in the form of developmental aid by governments on an average is more biased than Kiva.

Next, to identify the potential factors associated with the observed bias, we model the level of aid using the fixed-effect ordered logistic regression given as follows:

$$Q^{\text{aid}}_{ijy} = \beta_1(\text{GDP difference})_{\text{ijy}} + \beta_2\text{Distance}_{\text{ij}} + \beta_3\text{Migration}_{\text{ji}} + \beta_4\text{Colony}_{\text{ij}} + \epsilon_{\text{ijy}} \quad \text{(C.4)}$$

The fixed-effects of donor country, borrower country, and year were included in the model. The categorical outcome variable is constructed by using four quantiles

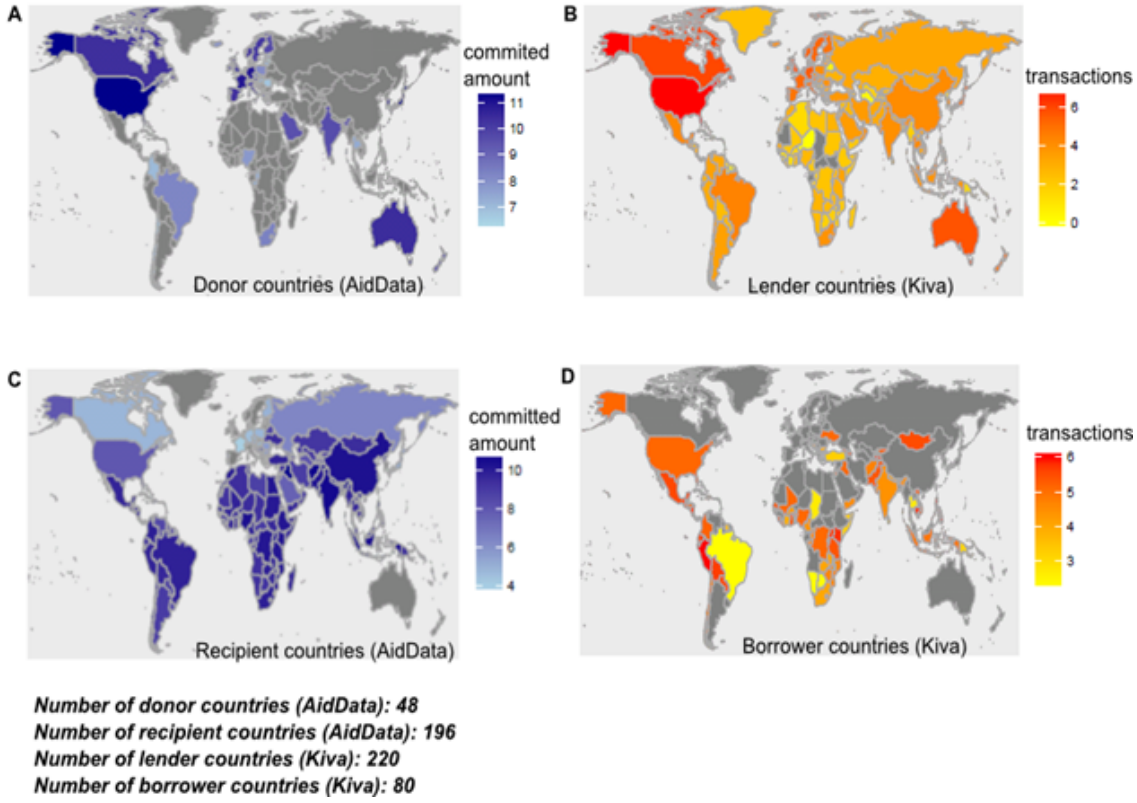

Figure C.6: Geographical coverage of Kiva and government aid. (A) Donor countries by their total commitment amounts (USD), (B) lender countries in Kiva by the total number of contributions made, (C) recipient countries by total commitment amount (USD), and (D) borrower countries in Kiva by the total number of contributions received. All values are aggregated sum from 2005–2013. The scale shown is logarithmic with a base of 10. The coverage patterns show a difference in the potential channels for capital flow. There are more participating lender countries on Kiva compared to number of donor countries from AidData in the same time period.

(zero, low, medium, and high) of aid amount (shown in Fig. C.8). The description of variables and their correlations are presented in Tables C.7 and C.8.

The results of our ologit regression are reported in Table C.9 ($N = 39031$; *pseudo* $R^2 = 0.4068$ for the final model) and show that similar to lending in Kiva, government aid is also driven by the same exogenous variables with the exception of GDP difference, which in the case of government aid was not found to be significant. The effect of migration and colonial past, as reflected by very high odds ratios, are

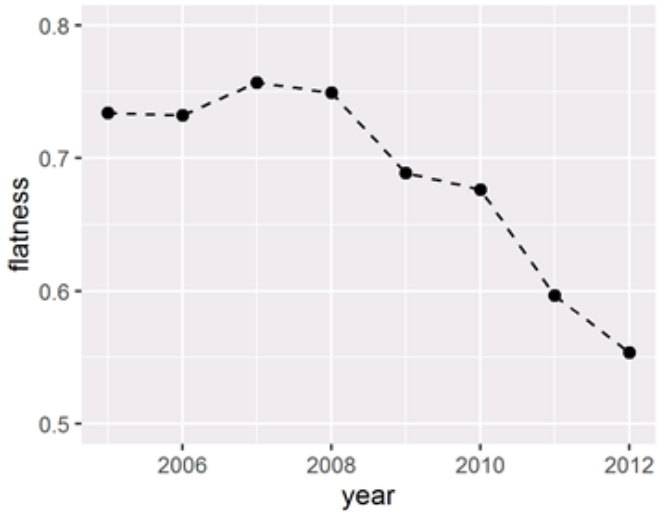

Figure C.7: Flatness of government aid network. The level of flatness is low (compared to Kiva, which is between 90% and 80%) and increases between 2006 and 2007 and shows a decrease afterwards.

Table C.7: Descriptive statistics for government aid data.

Variable	Obs.	Mean	SD	Min	Max
Outcome	46456	2.89	2.105959	1	4
GDP (pc) difference *(thousands USD)*	39938	2766	192136	-44.72371	115.852
Distance *(thousands of kms)*	44094	8.07	4.181782	0.0361766	19.95116
Migration *(Millions)*	42094	0.02	0.1857959	0	11.63599
Colony	44094	0.03	0.1620495	0	1

Table C.8: Correlation matrix for government aid data.

Variable	Q (outcome)	GDP (pc) difference	Distance	Migration	Colony
Outcome	1	...	...	...	...
GDP (pc) difference	0.26	1	...	...	...
Distance	-0.14	-0.03	1	...	...
Migration	0.08	0.02	-0.05	1	...
Colony	0.16	-0.01	-0.02	0.06	1

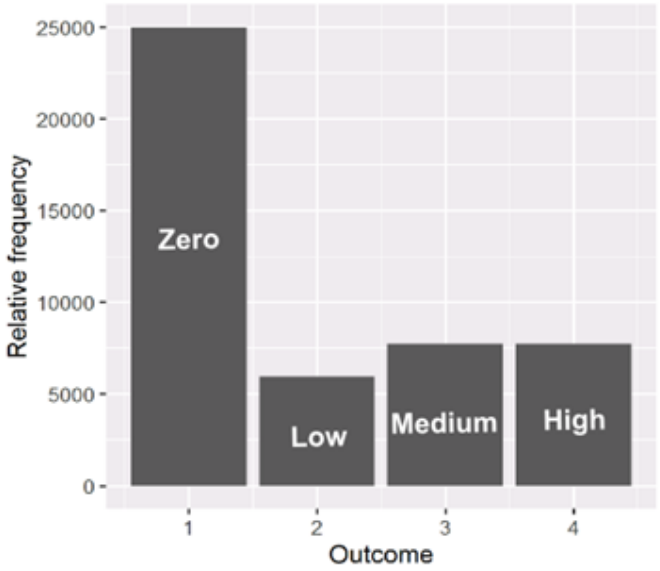

Figure C.8: Relative frequency of levels of commitment amount (Zero: 0 USD; low: 8 USD–0.3 Million USD; medium: 0.3 Million USD–6.5 Million USD; high: 6.5 Million USD–11.1 Billion USD).

much stronger in this case.

Table C.9: Fixed-effect ologit estimates of levels of lending between countries. Odds ratio reported for 4 levels of transactions (4 levels of commitment amount in the case of government aid). **$p < 0.05$

Variable	Odds ratio Model 1	Odds ratio Model 2	Odds ratio Model 3	Odds ratio Model 4
GDP (pc) difference	0.99	0.99	0.99	0.99
Distance	...	0.77**	0.77**	0.77**
Migration	...	...	5.21**	2.52**
Colony	...	...	...	12.65**
AIC:	62907.5	58361.99	58284.72	5737125
BIC:	62976.26	58430.56	58353.3	57439.83
Fixed effects:	...	...	...	...
Year	Yes	Yes	Yes	Yes
Donor country	Yes	Yes	Yes	Yes
Rrecipient country	Yes	Yes	Yes	Yes