EVOLUTIONARY DYNAMICS OF LARGE SYSTEMS

by

Nikhil Nayanar

A Dissertation

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the degree of

Doctor of Philosophy

School of Industrial Engineering

West Lafayette, Indiana

May 2021

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. Nagabhushana Prabhu, Chair
School of Industrial Engineering

Dr. Roshanak Nateghi
School of Industrial Engineering

Dr. Alok R. Chaturvedi
Krannert School of Management

Dr. Mark R. Lehto
School of Industrial Engineering

Approved by:
Dr. Abhijit Deshmukh

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Several socially and economically important real-world systems comprise large numbers of interacting constituent entities. Examples include the World Wide Web and Online Social Networks (OSNs). Developing the capability to forecast the macroscopic behavior of such systems based on the microscopic interactions of the constituent parts is of considerable economic importance.

Previous researchers have investigated phenomenological forecasting models in such contexts as the spread of diseases in the real world and the diffusion of innovations in the OSNs. The previous forecasting models work well in predicting future states of a system that are at equilibrium or near equilibrium. However, forecasting non-equilibrium states–such as the transient emergence of hotspots in web traffic–remains a challenging problem. In this thesis we investigate a hypothesis, rooted in Ludwing Boltzmann's celebrated H-theorem, that the evolutionary dynamics of a large system–such as the World Wide Web–is driven by the system's innate tendency to evolve towards a state of maximum entropy.

Whereas closed systems may be expected to evolve towards a state of maximum entropy, most real-world systems are not closed. However, the stipulation that if a system is closed then it should asymptotically approach a state of maximum entropy provides a strong constraint on the inverse problem of formulating the microscopic interaction rules that give rise to the observed macroscopic behavior. We make the constraint stronger by insisting that, if closed, a system should evolve monotonically towards a state of maximum entropy and formulate microscopic interaction rules consistent with the stronger constraint.

We test the microscopic interaction rules that we formulate by applying them to two real world phenomena: *the flow of web traffic in the gaming forums on Reddit* and *the spread of Covid-19 virus.* We show that our hypothesis leads to a statistically significant improvement over the existing models in predicting the traffic flow in gaming forums on Reddit. Our interaction rules are also able to qualitatively reproduce the heterogeneity in

the number of COVID-19 cases across the cities around the globe. The above experiments provide supporting evidence for our hypothesis, suggesting that our approach is worthy of further investigation.

In addition to the above stochastic model, we also study a deterministic model of attention flow over a network and establish sufficient conditions that, when met, signal imminent parabolic accretion of attention at a node.

1. INTRODUCTION

In the emerging hyper-connected world, systems comprising large numbers of interacting entities have become ubiquitous. The World Wide Web and social networks are prototypical examples of such large systems. As diverse large systems of increasing size and complexity emerge there is considerable economic interest in predicting–even if over a short time horizon–the evolution of such systems. For example, metrics such as the number of DAUs (Daily Active Users), which quantify the amount of user-attention accreted at a website, are being used to evaluate online enterprises.

In Section 3 we investigate the following question: *is it possible to detect an imminent surge in the accretion of user-attention (eg., number of DAUs) at a website solely based on an analysis of the flow of the user-attention across a network?* We establish a set of sufficient conditions that, when met, signal an imminent parabolic surge in the accretion of user-attention at a website. Although the model we study is deterministic it is instructive in that it shows that signals about imminent parabolic surge can be gleaned from a purely kinematic analysis of the flow.

Most large real-world systems are not deterministic but inherently stochastic. The key question about large stochastic systems then is: *what are the dynamical principles underlying the evolution of large real-world stochastic systems?* The question was studied in the context of systems comprising large numbers of interacting microscopic particles by Ludwig Boltzmann.

In his famous H-Theorem Ludwig Boltzmann established, under very general conditions, that a closed system with a large number of interacting degrees of freedom eventually reaches a state of maximum entropy [1]. Hence it is reasonable to hypothesize that the evolution of modern real-world systems with large numbers of interacting degrees of freedom is driven by an innate tendency of the system to reach a configuration of maximum entropy.

Whereas Boltzmann's result pertains to closed systems, the modern large systems of interest are not closed systems. Nevertheless, the innate tendency of a system to progress towards configurations of increasing entropy, discovered by Boltzmann in the context of closed systems, we hypothesize, provides the dynamical principle that underlies the evolution of open real-world systems as well. In fact, we use the constraint that the entropy of a real-world system should not only increase with time, but that it should increase monotonically with time, to guide our formulation of microscopic interaction rules. The numerical experiments we present appear to support our hypothesis.

To the best of our knowledge, the present models of evolution of large social and economic systems are not based on the principle of entropy maximization. Rather they are based on phenomenological hypotheses (see Section 2). Our hypothesis enables us to derive the stochastic evolution equations of large systems from first principles. We validate our hypothesis by applying it to two real-world problems: *predicting the flow of web traffic into pages in Reddit*, and on *capturing the variation in the number of infections due to COVID-19 in cities across the globe.* We present results from numerical experiments to demonstrate the predictive value of our stochastic evolution equations for large systems.

The ensuing discussion is organized as follows. In Chapter 2, we review existing models of evolution of complex networks in general. In Chapter 3, we describe and analyze a deterministic model of user-base flow within a social network. We analytically derive the sufficient conditions that, when met, signal an imminent parabolic surge in the accretion of user-base at a node in the network. In Chapter 4 we present the principles of statistical mechanics that provide the necessary background for our discussion. We also formulate our hypothesis and present preliminary empirical evidence to demonstrate the plausibility of the hypothesis.

In Chapter 5, we develop a collision-based model for the evolution of large systems. Specifically, following the philosophy of statistical mechanics, which seeks to explain the bulk properties of a system based on the microscopic interactions among the constituent particles in the system, we seek to explain the macroscopic evolution of a large infrastructure—such the World Wide Web—based on the microscopic interactions—

called ‘collisions’—among the constituent interacting ‘entities’ (such as websites) in the infrastructure. Following a closer examination of the conditions that must be met in order for the entropy of a large infrastructure to increase monotonically, we devise some candidate interaction rules that drive a closed infrastructure monotonically towards configurations of maximize entropy. As shown by Fluctuation Dissipation Theorem, entropy may not increase monotonically in real physical systems. Rather, real physical systems approach configurations of maximum entropy only asymptotically. For computational ease, however, we formulate a stronger criterion that the entropy should increase monotonically, and restrict attention to only those microscopic interaction rules that conform to the stronger criterion. We also clarify how we model stochastic perturbations to infrastructures and heterogeneity of interactions.

In Chapter 6 we describe the details of the simulation process and the datasets that we use to numerically validate our hypothesis. We also present and discuss the results of the simulation, comparing our results with the results from other models reported in the literature. Chapter 7 contains the concluding remarks.

2. LITERATURE REVIEW

A marketplace of user-bases is an increasingly important notion in this digital age since users are constantly bombarded with information about products, ideas through mediums that are equally available to all competing entities. One could even argue this makes the users in charge to decide where they divert their user-base [2]. By attracting user-base, products build a social capital, off of which they generate income , primarily through advertisements. A large body of work has used this facet of user-base scarcity from the perspective of entities themselves; to model their growth online. The spread of popularity is the same as a gain in the user-base directed towards the entity.

A relevant work here is that of Ratkiewicz et al.[3],in which the popularity of online systems like Wikipedia is modelled using existing preferential popularity mechanisms [4], [5] coupled with random bursts in popularity of its webpages due to external factors. The randomness is introduced by way of a uniform distribution to rerank a webpage, such that it is suddenly exposed to greater user user-base. The authors claim that the sustained popularity of webpages and sites is explained by these bursts of user-base they receive rather than a more intuitive notion of gradually accumulating user-base. In [6],the impact of social influence on the propagation of products is analysed. Using the ecosystem of Facebook [7], the popularity of applications is studied as an effect of the behaviour of the friends/acquaintances of the user and behaviour of the population as a whole. Interestingly, the Facebook ecosystem is one where the drivers of popularity are internal to the system ; the model avoids dealing with exogenous drivers

Ribeiro et al. [8] utilise the principle of a marketplace of user-base to model the success of Facebook coinciding with the simultaneous fall of other social networks. Using a set of catalytic reactions and diffusion equations, the interaction between the active, non-active and non-members constrained by their finite user-base is studied to capture the trajectory of total Daily Active Users (DAU) in the competing websites. Hood et al. [9], utilized a Yelp Dataset in predicting future user-base a business receives by learning features of existing successful business. The model does not explain the why businesses show an uneven amount of success, rather uses the features extracted for prediction of number of reviews(user-base) received in the future, using machine learning models.

The increase in user-base towards an entity can also be thought of as a rise in popularity. Existing work in the propagation of competing products, ideas and information in graph topologies have been predominantly inspired by biological process like the spread of an epidemic [10]–[12]. Prakash et al. [13], model two competing products and 'the word of mouth' adoption factor amongst them, using the principles of propagation of viruses. Attributing a rate of virus attack and healing to each product, the paper evaluates the stationary points of this dynamic system to show that once a product crosses an epidemic threshold it will not only have an increased market share but also complete domination of the market i.e attract all of the available user user-base.

In [14] , the author analyses the observed inequality in the distribution of user user-base across the contents of web , using tools in the study of economic inequality. The concentration of user-base towards a small fraction of the web is empirically shown. An interesting conjecture put forward in the paper is that the effective size of the Web remains bounded despite increasing number of web-based entities. Goldhaber details the impact of the scarcity of user-base in [15]. The economy that arises as a result of this , is claimed to potentially replace the existing market economy , just as capitalism replaced the feudal system. The author claims that user-base will gain dominance over monetary itself. Since people are spending a significant amount of time on the web with a finite amount of user-base, the monetary value generated from exploiting where the user-base is directed to, is in itself gaining importance. A rigourous investigation of these claims are in order.

A large body of work has been aimed at studying the diffusion and adoption of innovations through online networks of persons , [16]–[19]. Threshold models [20] have been widely used in comprehending the dissemination of information/ideas among networks of users. These models work on the principle that a user in a network graph adopts an idea/product if sufficient number of his/her friends adopt it. Acemoglu et al. [21], explored interesting results in the process of innovation diffusion in social networks; they studied the widely used linear threshold model with an interesting addition of the notion of path dependence; by modelling the adoption of an innovation as a stochastic process rather than a purely deterministic one, they show that a such behaviour can vastly alter the course of the diffusion process. Another popular model is the cascade model of spread of innovations. Unlike the threshold models, the principle here is that for every node v that has adopted an innovation, it tries to influence its neighbours w into adopting the innovation as well at time t with some probability p that may/ may not be dependent on the neighbours of w that have already tried to influence it into adopting the innovation. The in/dependence on the history of attemtped influence lead to independent/general cascade models.

An interesting associated problem here is one that pertains to marketing. Assuming that diffusion of innovations follows the above prescribed models, how many seed nodes (nodes that have initially adopted the product without a prior diffusion) are required to maximise the expected number of adopters. Classic compartmental models of disease spreading, the SIR AND SIRS models have also been widely studied to explain the diffusion of innovations [22]–[24]. The SIS model of propagation defines a set of users as either being Susceptible (S) to infection or being Infective (I). A susceptible individual can become infected with some probability by interaction with an infective individual. Similarly with some non-zero probability they can be healed from being infected to being susceptible again. The SIRS model in addition to the features of the SIS model has an addition compartment of Removed (R) users who essentially are considered removed from the system once they are healed (i.e cannot be infected again). Obviously the two models may be used in modelling the propagation of different types of innovations.These epidemic models have been fairly successful in predicting the spread of innovations across arbitrary networks. In [13], the authors have extended the SIS model to a system of two competing viruses and shown that the steady state populations infected by each of the viruses can have different scenarios depending on certain constraints the infection transmission and recovery rates for each virus.

In [25] the adoption/spread of an innovation is modelled taking into consideration the interdependencies between multiple exposures a node has to the innovation. In effect a memory feature was introduced for every node such that the probability that the node itself is infected is dependent on the number of past exposures and the dose of the exposures in a short period of time. The probability of being exposed to an infected node and the dose of the exposure themselves are tunable distributions in the model. Its interesting to observe that the authors make no accommodation for the structure of the network; this plays a crucial role in the diffusion of user-base [21]. In [26], the authors explain the propagation of recommendations and the observed power-law distribution of their cascade sizes, using a massive person-to-person network. Using the temporal nature of this network several interesting observations have been put forward. [27], [28] present an exhaustive survey of the literature on complex networks. It is interesting to note here that a principle guideline for the research on random networks has been to identify critical probabilities when certain properties of the graph undergo a sudden transition. In the context of such critical phenomena , the percolation theory holds importance in networks ; for a graph with a defined degree sequence , there exists a constraint on the degree distribution which if satisfied almost surely guarantees the existence of a large connected component in the graph [29]. These results are extended to scale free directed and non-directed networks in [30]

From empirical studies on real-world networks , a characteristic of several networks is the power law distribution their node degree follows[27]. An extensive body of work covers an explanation for this distribution. A widely utilized growth mechanism of networks to explain the previously stated distribution has been the 'preferential attachment' model [4], [31]. Intuitive in principle, the model posits that a node i attaches itself to a node j with a probability proportional to the degree of node j. In [32], the user-base inequality in the Twitter social network, is studied. The paper uses the preferential attachment model of gaining user-base. The model is able to capture the skewed nature of the user-base distribution among users into the future. While this model does explain the power law distribution of node degree, it is highly questionable to assume this as a de facto model of network growth[33]. In [34], the authors build on this preferential attachment model of growth by adding a local feature to the growth; rather than having the ability to process the entire network to decide upon the nodes to connect to, this capability is limited to a fraction of the nodes in the network. Using this paradigm, the authors show that the power-law distribution is in fact refined to an exponentially truncated one.

Other observed features of real world networks worth mentioning here are the ’small world’ and clustering property. The ’small world’ property essentially means that the average path length between two nodes in a network is relatively much small compared to the size of the graph in itself [35]. Clustering is the tendency of nodes in a network to form cliques measured using a clustering coefficient [36]. A good amount of work has been focused on studying the evolution of networks, specifically the temporal changes in the network structure [26], [37]–[39] . In [40], the author provides the definition of a class of models(stochastic) that explains the power-law distribution observed in a wide range of empirical data such as word frequencies,scientific publications,city sizes. An alternate approach to understanding the spread of innovations has been through the agent-based modelling paradigm [41], [42]. In [33], the authors study network characteristics as emergent properties of an ensemble of agents interacting through specific rules.

The field of statistical mechanics essentially uses statistical methods to deal with systems composed of a large number of particles (read as large degree of freedom). Perhaps the most well known use of statistical mechanics has been to explain the thermodynamics of large systems [43]. There have been some previous attempts at linking statistical mechanics and other existing physical phenomena to explain some of the observed characteristics of networks and their growth. In [44] the authors show that an evolving network is similar to a Bose gas [45] at equilibrium. The model treats each node as an energy level and an edge between two nodes as a set of two particles one on each energy level corresponding to the two nodes. Using a continuum approximation of the rate of change of the degree of the nodes the authors show that under specific conditions imposed on the system, a preferential attachment mechanism in the growth of the network can be observed. In [46], the authors use a simple stochastic process used previously to explain scaling phenomena in copy growth processes. In this model, the existing nodes in the network are divided into classes depending on their existing connectivity. At any time step in the evolution of the network, two possibilities are described; (i) With a non zero probability, a new node joins the network and can randomly connect itself to any other node , (ii) if an existing node forms a new edge in the network , the probability of connecting to a node belonging to a class k is a function of the product of its connectivity and cardinality.

Thermodynamics essentially gives us the basis for understanding how heat and work are related and rules the macroscopic properties of systems follow at equilibrium [47]. There is a natural extension of the thermodynamics at equilibrium to systems away from but close to equilibrium. It is based on the Local Equilibrium Hypothesis. If the constraints with respect to which a system is in equilibrium are relaxed , in the (linear)vicinity of this relaxation , the system can be thought of as being in local equilibrium.In other words, if the system is in a steady state in terms of the flux of energy and particles [48], it can be divided into small volumes where the Gibbs equation holds and changes in variables are not infinitely slow [49]. Because of the usual disparity between macroscopic and microscopic scales, most steady state systems can be included in this category.

3. A DETERMINISTIC MODEL OF INTERACTION

The spectacular rise of Facebook with the simultaneous fall of MySpace,Orkut has been an interesting subject of discussion and research (Section 2) from a system dynamics perspective. Is it possible to detect the imminent growth of the user-base position of an entity solely through the kinematics of user-base flow amongst all entities ?. We try to answer this question by arriving at a set of detectable conditions which when met signal an entity is on the verge of a parabolic growth in terms of user-base. We regard the online enterprises vying for user user-base as an interconnected system of nodes. The amount of user-base–measured using any metric of choice (for example, DAU)–is regarded as a fluid that accretes at and flows among nodes, subject to the constraint that the amount of fluid is finite. We present a simple model that allows for continuous influx of user-base into the network and describes the competition for available user-base. The significance of our analysis is that it shows the possibility of formulating analytical sufficient conditions for predicting parabolic growth of user-base position.

3.1 Model of user-base flow

We model the online enterprises vying for user-base–hereafter called just user-base–as a complete graph. Specifically, we assume that the enterprises are numbered 1, 2, . . . , N. The associated graph has N nodes with an undirected edge between every pair of nodes. We model user-base as a fluid, using the continuum approximation. Time is discretized and labeled by an integer variable n.

The kinematics of user-base in the above network are governed by three equations, namely (3.3), (3.6) and (3.9), which describe respectively the *inflow of new user user-base into the network*, the *redistribution of user-base among the nodes* and the *leakage of user-base from each node*, in each step. Deferring a more precise description of the equations to later discussion we summarize their contents below.

In each time step, we assume a fixed nonnegative amount of new user-base–denoted u–flows into the network. For example, if we take a time step to be a day, then a fixed amount of new user-base flows into the network daily. We do not insist that $u > 0$, allowing for the possibility that no new user-base flows into the network. Equation (3.3) specifies that the inflow is distributed among the enterprises in proportion to the user-base position of the nodes in the previous instant of time.

The contention for user-base among enterprises is modeled using (3.6). In each time step, an enterprise attempts to siphon user-base towards itself from the other enterprises. The amount of user-base an enterprise k siphons from another enterprise j depends on j's user-base position as well as the ratio of the user-base positions of the two enterprises j and k at the previous instant of time. An enterprise k siphons a greater fraction of user-base from j than an enterprise i, if the user-base position of k in the previous time instant is greater than that of i. The net flow of user-base from j to k is the difference between the user-base-flow from j to k and the user-base-flow from k to j. If the user-base position of k is greater than that of j, in the previous time instant, then the net flow from j to k in the current time step is positive, and nonpositive otherwise.

Finally, equation (3.9) models the leakage of user-base from a node. In every time step, a fixed fraction, $1 - \delta$, of the user-base position of a node in the previous instant is assumed to leak away from the node with a part of it flowing to the other nodes in the network and the remaining part disappearing altogether from the network. The accumulation of user-base in a node is thus determined by the net inflow of new user-base, the net inflow of user-base from other nodes, and the net loss of user-base from the node due to leakage. The model is described more precisely by the following definitions.

$\mathbf{T}(n)$: denotes the time interval $(n-1, n]$.

$\mathbf{A}_{\mathrm{i}}(n),\ 1 \leq \mathrm{i} \leq N$: denotes the **a**ttention position of (the amount of user-base vested in) node i at time $t = n$.

$\mathbf{A}_\Sigma(n)$: denotes $\sum_{\mathrm{i}=1}^{N} A_\mathrm{i}(n)$. In general, we follow the convention of replacing a subscript by a summation symbol Σ to indicate the sum over all the values of the subscript.

$\bar{\mathbf{A}}_\mathrm{i}(n)$: denotes the sum the user-base positions of all the nodes except node i. Specifically,

$$\bar{A}_\mathrm{i}(n) \quad = \quad A_\Sigma(n) - A_\mathrm{i}(n)$$

$\rho_\mathrm{i}(n),\ 1 \le \mathrm{i} \le N$**:** denotes the fraction of total user-base resident in the network that is vested in node i, at $t = n$. Specifically,

$$\rho_\mathrm{i}(n) \quad = \quad \frac{A_\mathrm{i}(n)}{A_\Sigma(n)} \tag{3.1}$$

$\mathbf{R}_\mathrm{ij}(n),\ 1 \le \mathrm{i},\mathrm{j} \le N$: denotes the **r**atio of the user-base positions of nodes j and i at $t = n$. Specifically,

$$R_\mathrm{ij}(n) \quad = \quad \frac{A_\mathrm{j}(n)}{A_\mathrm{i}(n)} \tag{3.2}$$

$\mathbf{U}_\mathrm{i}(n),\ 1 \le \mathrm{i} \le N$: denotes the amount of external user-base flowing into node i in the interval $T(n)$. We assume that a constant amount of new user-base, denoted u, flows into the network at every time step. We assume that the new user-base is distributed among the nodes in proportion of their user-base positions. Specifically,

$$U_\mathrm{i}(n) \quad = \quad u \cdot \frac{A_\mathrm{i}(n-1)}{A_\Sigma(n-1)} \tag{3.3}$$

Equation (3.3) provides the first kinematic equation of our model.

$\bar{\mathbf{U}}_\mathrm{i}(n),\ 1 \le \mathrm{i} \le N$: denotes the sum of the total of external user-base flowing into the network excluding the external user-base flowing into node i. Specifically,

$$\bar{U}_\mathrm{i}(n) \quad = \quad \sum_{\mathrm{j}\neq\mathrm{i}} U_\mathrm{j}(n)$$

$\mathbf{O}_{\mathrm{ij}}(n),\ 1 \leq \mathrm{i}, \mathrm{j} \leq N,\ \mathrm{i} \neq \mathrm{j}$: denotes the net

amount of user-base that flows from node j to node i in time interval $T(n)$. $O_{\mathrm{ij}}(n)$ is the amount of user-base, initially vested in node j at $t = n - 1$ that users divert towards node i in the interval $(n - 1, n]$. We assume that the amount of user-base flowing from node j to node i is proportional to $A_{\mathrm{j}}(n-1)$, the user-base position of node j at $t = n - 1$. Further, we assume that a node with larger position siphons user-base from nodes that have smaller user-base positions. Therefore

$$O_{\mathrm{ij}}(n) = -O_{\mathrm{ji}}(n). \tag{3.4}$$

We assume that O_{ij} satisfies the following two limits:

$$\begin{aligned} \lim_{R_{\mathrm{ij}}(n-1)\to 0} O_{\mathrm{ij}}(n) &\propto A_{\mathrm{j}}(n-1); \\ \lim_{R_{\mathrm{ij}}(n-1)\to \infty} O_{\mathrm{ij}}(n) &\propto -A_{\mathrm{i}}(n-1) \end{aligned} \tag{3.5}$$

Noting that $R_{\mathrm{ij}} = 1/R_{\mathrm{ji}}$, it is easily verified that the following functional form of $O_{\mathrm{ij}}(n)$ satisfies the antisymmetry property (3.4) as well as the limits in (3.5).

$$\begin{aligned} O_{\mathrm{ij}}(n) = \ & \alpha \Big\{ A_{\mathrm{j}}(n) \left(1 - \mathrm{e}^{-R_{\mathrm{ji}}(n-1)}\right) \\ & - A_{\mathrm{i}}(n-1) \left(1 - \mathrm{e}^{-R_{\mathrm{ij}}(n-1)}\right) \Big\} \end{aligned} \tag{3.6}$$

Equation (3.6) is the second kinematic equation of our model. The factor α is included to ensure that the amount of flow from node j to node i over the interval $T(n)$ does not exceed the user-base position of node j at $t = n - 1$. In Lemma 2 we show that the outflow from node j will be constrained to be less than the user-base position at node j if α is chosen so that it does not exceed an upper bound.

Using the definition of R_{ij}, in (3.2), we can rewrite (3.6) as

$$\begin{aligned} O_{\mathrm{ij}}(n) &= \alpha A_{\mathrm{i}}(n-1) \\ &\quad \left[R_{\mathrm{ij}}(n-1)\left(1-\mathrm{e}^{-1/R_{\mathrm{ij}}(n-1)}\right)\right. \\ &\quad \left.-\left(1-\mathrm{e}^{-R_{\mathrm{ij}}(n-1)}\right)\right] \\ &= \alpha A_{\mathrm{i}}(n-1)\cdot f\left(R_{\mathrm{ij}}(n-1)\right) \qquad (3.7) \end{aligned}$$

where $f(x) := x\left(1-\mathrm{e}^{-1/x}\right)-(1-\mathrm{e}^{-x})$.

$\mathbf{O}_{\mathrm{i}\Sigma}(n),\ 1 \le \mathrm{i} \le N:$ denotes the sum of O_{ij} over $\mathrm{j} \neq \mathrm{i}$. Specifically,

$$O_{\mathrm{i}\Sigma}(n) = \sum_{\mathrm{j}\neq\mathrm{i}} O_{\mathrm{ij}}(n) \qquad (3.8)$$

$\mathbf{I}_{\mathrm{i}}(n),\ 1 \le \mathrm{i} \le N:$ denotes the net change in the user-base position of node i over the interval $T(n)$. Specifically,

$$I_{\mathrm{i}}(n) = A_{\mathrm{i}}(n) - A_{\mathrm{i}}(n-1)$$

$\bar{\mathbf{I}}_{\mathrm{i}}(n),\ 1 \le \mathrm{i} \le N:$ denotes the net change in the user-base at all the nodes except node i over the interval $T(n)$. Specifically,

$$\bar{I}_{\mathrm{i}}(n) = \sum_{\mathrm{j}\neq\mathrm{i}} I_{\mathrm{i}}(n)$$

Finally, we state the third kinematic equation for the flow of user-base among the nodes of the network.

$$A_{\mathrm{i}}(n) = \delta A_{\mathrm{i}}(n-1) + O_{\mathrm{i}\Sigma}(n) + U_{\mathrm{i}}(n), \qquad (3.9)$$

where $0 < \delta < 1, \quad 1 \le \mathrm{i} \le N$.

That is, we assume that a fraction $\delta A_{\mathrm{i}}(n-1)$ of the user-base position at node i at the beginning of the interval $T(n)$ remains at the node at the end of the interval, while the remaining fraction $(1-\delta)$ $A_{\mathrm{i}}(n-1)$ leaks away from the node–a part of $(1-\delta)$ $A_{\mathrm{i}}(n-1)$ flowing to the other nodes in the network and the remaining part disappearing altogether from the network. Besides the user-base that a node retains, it also receives $O_{\mathrm{i}\Sigma}(n)$ amount of user-base from other nodes and $U_{\mathrm{i}}(n)$ amount of external user-base from users. Equations (3.3), (3.6) and (3.9) together provide the three kinematic equations that govern the flow of user-base in our model.

We complete the definitions by proving the following two Lemmas that establish the required upper bound on the factor α in (3.6).

Lemma 1. *Given the kinematic equation (3.9) the total user-base in the network at time $t=n$, namely $A_\Sigma(n)$ can be written as*

$$A_\Sigma(n) \quad = \quad u \cdot \frac{1-\delta^n}{1-\delta}$$

Proof: Recalling the antisymmetry of $O_{\mathrm{ij}}(n)$ mentioned in (3.4) we note that

$$\sum_{\mathrm{i}=1}^{N}\sum_{\mathrm{j}=1}^{N} O_{\mathrm{ij}}(n) \quad = \quad \sum_{\mathrm{i}=1}^{N} O_{\mathrm{i}\Sigma}(n) \quad = \quad 0 \tag{3.10}$$

Summing (3.9) and using (3.3) and (3.10) we get

$$\begin{aligned} A_\Sigma(n) &= \delta A_\Sigma(n-1) + u \quad \Longrightarrow \\ A_\Sigma(n) &= u\left(1+\delta+\ldots+\delta^{n-1}+\delta^n A_\Sigma(0)\right) \end{aligned} \tag{3.11}$$

Assuming that there is no user-base in the network at $t=0$, $A_\Sigma(0)=0$. Summing the geometric series in (3.11) yields the result.

■

Lemma 1 is used to establish an upper bound on α in Lemma 2.

Lemma 2. *The total outflow of user-base from node* j *over an interval* $T(n)$ *to all of the other nodes is less than the total amount of user-base at node* j *available for redistribution, that is,*

$$\sum_{\mathrm{i} \neq \mathrm{j}} O_{\mathrm{ij}}(n) \quad < \quad (1 - \delta) A_{\mathrm{j}}(n-1)$$

if

$$\alpha \quad < \quad \frac{1 - \delta}{N - 1}$$

Proof: The net flow of user-base from node j to node i, namely O_{ij} over an interval $T(n)$ is the difference of the flow from j to i and the flow from i to j over the interval $T(n)$. The flow from node j to node i over the interval $T(n)$ is F_{ij}, given by

$$F_{\mathrm{ij}}(n) \quad = \quad \alpha A_{\mathrm{j}}(n-1) \left(1 - \mathrm{e}^{R_{\mathrm{ij}}(n-1)}\right)$$

We want

$$\begin{aligned} \sum_{\mathrm{i} \neq \mathrm{j}} F_{\mathrm{ij}}(n) \quad &= \quad \alpha \sum_{\mathrm{i} \neq \mathrm{j}} A_{\mathrm{j}}(n-1) \left(1 - \mathrm{e}^{-R_{\mathrm{ij}}(n-1)}\right) \\ &< \quad (1 - \delta) A_{\mathrm{j}}(n-1) \end{aligned} \tag{3.12}$$

Noting that $1 - \mathrm{e}^{-R_{\mathrm{ij}}(n-1)} < 1$ we conclude that in order to satisfy (3.12) it is sufficient if α satisfies

$$\begin{aligned} \alpha \sum_{\mathrm{i} \neq \mathrm{j}} A_{\mathrm{j}}(n-1) \left(1 - \mathrm{e}^{-R_{\mathrm{ij}}(n-1)}\right) < \alpha A_{\mathrm{j}}(N-1) \\ < \quad (1 - \delta) A_{\mathrm{j}}(n-1) \quad \Longrightarrow \quad \alpha < \frac{1 - \delta}{N - 1} \end{aligned}$$

as claimed. ■

3.2 Accelerated accretion of user-base

Our main result–which establishes sufficient conditions for accelerated accretion of user-base–is contained in Theorem 1. The proof of the theorem requires a few intermediate results, which are established in Lemmas 3 to 9. We begin by stating Theorem 1, but defer its proof until after the required intermediate results are stated and proved.

The following remarks about the notation we use will likely make the discussion more readable. Our interest is in determining if a chosen node–which we call node k in this section–is about to experience a surge in the user-base flowing towards it. Thus, it is helpful to remember that among the many subscripts and node labels that will be used in the following discussion, the subscript or label k has a special status. It is also helpful to remember that the parabolic surge, if it occurs, is assumed to start at time n^*. Since we work with discretized time, we use the symbol n to denote time.

Theorem 1. *Assume that*

1. *for* $1 \leq \mathrm{i}, \mathrm{j} \leq N$, *the user-base flow into and within the network are governed by the kinematic equations (3.3), (3.6) and (3.9)*

$$\begin{aligned} U_{\mathrm{i}}(n) &= u \cdot \frac{A_{\mathrm{i}}(n-1)}{A_{\Sigma}(n-1)} \\ O_{\mathrm{ij}}(n) &= \alpha \Big[A_{\mathrm{j}}(n) \left(1 - \mathrm{e}^{-R_{\mathrm{ji}}(n-1)}\right) \\ &\quad - A_{\mathrm{i}}(n-1) \left(1 - \mathrm{e}^{-R_{\mathrm{ij}}(n-1)}\right) \Big] \\ A_{\mathrm{i}}(n) &= \delta A_{\mathrm{i}}(n-1) + O_{\mathrm{i}\Sigma}(n) + U_{\mathrm{i}}(n), \end{aligned}$$

   where $0 < \delta < 1, \alpha < \dfrac{1-\delta}{N-1}$,

2. *at some* $n^* > 0$, *the internode flows satisfy the following conditions for some node labeled* k,

$$O_{k\Sigma}(n^*) \quad > \quad 0 \tag{3.13}$$

$$\begin{aligned} O_{\mathrm{i}\Sigma}(n^*) \quad &< \quad 0,\ 1 \leq \mathrm{i} \leq N, \\ &\qquad \mathrm{i} \neq k \end{aligned} \tag{3.14}$$

$$\frac{O_{k\Sigma}(n^*)}{O_{k\Sigma}(n^*-1)} \quad > \quad \frac{1-\delta^{n^*}}{1-\delta^{n^*-1}} \tag{3.15}$$

3. and that

$$R_{k\mathrm{i}}(n) \quad > \quad R^*, \quad 1 \leq \mathrm{i} \leq N,\ \mathrm{i} \neq k \tag{3.16}$$

for all $n \in [n^*, n^*+r]$*, where* $r \geq 0$*.*

Then, the fraction of the total user-base vested in node k*, undergoes accelerated (parabolic) growth in the time interval* $[n^*, n^*+r]$*. Specifically,*

$$\begin{aligned} \Delta_k^2(n)\mathrm{j} := \mathrm{j}\rho_k(n) - 2\rho_k(n-1)) + \rho_k(n-2) & \\ > \quad 0, \qquad & n \in [n^*, n^*+r] \end{aligned}$$

∼∼ ◦ ∼∼

The intermediate results needed to prove Theorem 1 are established in the following sequence of lemmas. Lemma 3 shows that the condition for accelerated growth of user-base position–that is the condition that the second time derivative of user-base position is positive–can be reformulated in terms of inequalities involving the inter-node flows (the $O_{\mathrm{i}\Sigma}$).

Lemma 3. *For* $1 \leq \mathrm{i} \leq N$ *and* $n \geq 2$*, the following two inequalities are equivalent.*

$$\Delta_k^2(n) > 0 \qquad \text{and} \qquad \frac{O_{\mathrm{i}\Sigma}(n)}{O_{\mathrm{i}\Sigma}(n-1)} > \frac{1-\delta^n}{1-\delta^{n-1}}$$

Proof: Noting that $A_\Sigma(n) = \delta A_\Sigma(n-1) + u$, and $U_k(n) = u\rho_k(n-1)$, we have

$$\begin{aligned}
&\rho_k(n) - \rho_k(n-1) \\
&= \frac{\delta A_k(n-1) + U_k(n) + O_{k\Sigma}(n)}{A_\Sigma(n)} - \frac{A_k(n-1)}{A_\Sigma(n-1)} \\
&= \frac{\delta A_k(n-1) + u \cdot \rho_k(n-1) + O_{k\Sigma}(n)}{\delta A_\Sigma(n-1) + u} \\
&\quad - \frac{A_k(n-1)}{A_\Sigma(n-1)} \\
&= \frac{O_{k\Sigma}(n)}{A_\Sigma(n)} \qquad (3.17)
\end{aligned}$$

Using (3.17) we can write $\Delta_k^2(n)$ as,

$$\begin{aligned}
\Delta_k^2(n)\mathrm{j} &:= \mathrm{j}\rho_k(n) - 2\rho_k(n-1) + \rho_k(n-2) \\
&= \frac{O_{k\Sigma}(n)}{A_\Sigma(n)} - \frac{O_{k\Sigma}(n-1)}{A_\Sigma(n-1)}
\end{aligned}$$

If $O_{k\Sigma}(n-1), A_\Sigma(n) > 0$, then $\Delta_k^2(n) > 0$ can be rewritten as

$$\frac{O_{k\Sigma}(n)}{O_{k\Sigma}(n-1)} > \frac{A_\Sigma(n)}{A_\Sigma(n-1)} = \frac{1-\delta^n}{1-\delta^{n-1}} \qquad (3.18)$$

■

Lemma 4. *For all* $1 \leq \mathrm{i} \leq N$*, and* $n \geq 1$*, the following two inequalities are equivalent:*

$$\rho_\mathrm{i}(n) > \rho_\mathrm{i}(n-1) \qquad \text{and} \qquad \frac{O_{\mathrm{i}\Sigma}(n)}{1-\rho_\mathrm{i}(n-1)} > 0.$$

Proof:

$$\langle A_{\mathrm{i}}(n)\rangle = \delta\langle A_{\mathrm{i}}(n-1)\rangle + \langle O_{\mathrm{i}\Sigma}(n)\rangle + \langle U_{\mathrm{i}}(n)\rangle$$

$$\begin{aligned}
\langle I_{\mathrm{i}}(n)\rangle &= A_{\mathrm{i}}(n) - A_{\mathrm{i}}(n-1) \\
&= (\delta-1)\langle A_{\mathrm{i}}(n-1)\rangle + \langle U_{\mathrm{i}}(n)\rangle + \langle O_{\mathrm{i}\Sigma}(n)\rangle \\
\left\langle I_{\mathrm{i}}(n) + \bar{I}_{\mathrm{i}}(n)\right\rangle &= (\delta-1)\langle A_{\Sigma}(n-1)\rangle + \langle U_{\Sigma}(n)\rangle \qquad (3.19)
\end{aligned}$$

We can write

$$\rho_{\mathrm{i}}(n) > \rho_{\mathrm{i}}(n-1),$$

as

$$\frac{\langle A_{\mathrm{i}}(n-1)\rangle + \langle I_{\mathrm{i}}(n)\rangle}{\langle A_{\Sigma}(n-1)\rangle + \langle I_{\mathrm{i}}(n)\rangle + \left\langle \bar{I}_{\mathrm{i}}(n)\right\rangle} > \frac{\langle A_{\mathrm{i}}(n-1)\rangle}{\langle A_{\Sigma}(n-1)\rangle} \qquad (3.20)$$

Assuming $\langle A_{\mathrm{i}}(n-1)\rangle, \langle A_{\Sigma}(n-1)\rangle, \left\langle I_{\mathrm{i}}(n) + \bar{I}_{\mathrm{i}}(n)\right\rangle > 0$, (3.20) can be rewritten as

$$\frac{\langle I_{\mathrm{i}}(n)\rangle}{\langle A_{\mathrm{i}}(n-1)\rangle} > \frac{\left\langle I_{\mathrm{i}}(n) + \bar{I}_{\mathrm{i}}(n)\right\rangle}{\langle A_{\Sigma}(n-1)\rangle} \qquad (3.21)$$

Using (3.19), (3.21) can be rewritten as

$$\begin{aligned}
&(\delta-1) + \frac{\langle U_{\mathrm{i}}(n)\rangle + \langle O_{\mathrm{i}\Sigma}(n)\rangle}{\langle A_{\mathrm{i}}(n-1)\rangle} \\
&\qquad > (\delta-1) + \frac{\langle U_{\Sigma}(n)\rangle}{\langle A_{\Sigma}(n-1)\rangle} \qquad (3.22)
\end{aligned}$$

Noting that $A_\Sigma(n-1) = A_\mathrm{i}(n-1) + \sum_{\mathrm{i}\neq k}^{N} A_\mathrm{i}(n-1) := A_\mathrm{i}(n-1) + \bar{A}_\mathrm{i}(n-1)$ and similarly $U_\Sigma(n) = U_\mathrm{i}(n) + \bar{U}_\mathrm{i}(n)$ (3.22) can be rewritten as

$$\begin{aligned} \langle O_{\mathrm{i}\Sigma}(n)\rangle \langle A_\Sigma(n-1)\rangle &> -\langle U_\mathrm{i}(n)\rangle \left\langle \bar{A}_\mathrm{i}(n-1)\right\rangle \\ &\quad + \left\langle \bar{U}_\mathrm{i}(n)\right\rangle \langle A_\mathrm{i}(n-1)\rangle \end{aligned} \tag{3.23}$$

Dividing (3.23) throughout by $A_\Sigma(n-1)$ and rearranging we get

$$\begin{aligned} &\langle O_{\mathrm{i}\Sigma}(n)\rangle + (1-\rho_\mathrm{i}(n-1))\langle U_\mathrm{i}(n)\rangle \\ &\qquad -\rho_\mathrm{i}(n-1)\left\langle \bar{U}_\mathrm{i}(n)\right\rangle > 0 \end{aligned} \tag{3.24}$$

We observe that

$$\sum_\mathrm{i} O_{\mathrm{i}\Sigma}(n) = O_{\mathrm{i}\Sigma}(n) + \bar{O}_{\mathrm{i}\Sigma}(n) = 0$$

Therefore,

$$\rho_\mathrm{i}(n-1)\langle O_{\mathrm{i}\Sigma}(n)\rangle + \rho_\mathrm{i}(n-1)\left\langle \bar{O}_{\mathrm{i}\Sigma}(n)\right\rangle = 0 \tag{3.25}$$

Subtracting (3.25) from (3.24) and dividing throughout by $1-\rho_\mathrm{i}(n-1)$, we get

$$\begin{aligned} &\langle O_{\mathrm{i}\Sigma}(n) + U_\mathrm{i}(n)\rangle \\ &\quad -\frac{\rho_\mathrm{i}(n-1)}{1-\rho_\mathrm{i}(n-1)}\left[\left\langle \bar{U}_\mathrm{i}(n) + \bar{O}_{\mathrm{i}\Sigma}(n)\right\rangle\right] > 0 \end{aligned} \tag{3.26}$$

Using (3.3) and noting that $\bar{O}_{\mathrm{i}\Sigma}(n) = -O_{\mathrm{i}\Sigma}(n)$ (see (3.25)) we get

$$\begin{aligned} &\langle u\cdot\rho_\mathrm{i}(n-1) + O_{\mathrm{i}\Sigma}(n)\rangle - \frac{\rho_\mathrm{i}(n-1)}{1-\rho_\mathrm{i}(n-1)} \\ &\qquad [u\cdot(1-\rho_\mathrm{i}(n-1)) - \langle O_{\mathrm{i}\Sigma}(n)\rangle\,] > 0 \end{aligned} \tag{3.27}$$

which can be rewritten as

$$\frac{\langle O_{\mathrm{i}\Sigma}(n)\rangle}{1-\rho_{\mathrm{i}}(n-1)} > 0$$

thereby establishing the equivalence of the inequalities as claimed. ■

Lemma 5. *If the conditions (3.13) and (3.14), listed in Theorem 1, are satisfied, then the following inequalities hold at n^*.*

$$\begin{aligned} \rho_k(n^*) &> \rho_k(n^*-1) && (3.28)\\ \rho_{\mathrm{i}}(n^*) &< \rho_{\mathrm{i}}(n^*-1),\ 1 \le \mathrm{i} \le N,\ \mathrm{i} \ne k && (3.29) \end{aligned}$$

Proof: The Lemma follows at once from Lemma 4 and inequalities (3.13) and (3.14) in the statement of the Theorem. ■

Lemma 6. *If the conditions (3.13) and (3.14), listed in Theorem 1, are satisfied, then the following inequalities hold at n^*:*

$$\begin{aligned} R_{k\mathrm{i}}(n^*) &< R_{k\mathrm{i}}(n^*-1) \qquad 1 \le \mathrm{i} \le N,\ \mathrm{i} \ne k\\ A_k(n^*) &> A_k(n^*-1) \end{aligned}$$

Proof: The first inequality follows from the two inequalities in Lemma 5. The second inequality follows from the first inequality in Lemma 5 by observing that

$$A_\Sigma(n^*) = u\frac{1-\delta^{n^*}}{1-\delta} > u\frac{1-\delta^{n^*-1}}{1-\delta} = A_\Sigma(n^*-1)$$

■

Lemmas 8 and 9 are based on certain properties of the function f (see (3.7)) that are established in Lemma 7. Before stating Lemma 7 it is helpful to illustrate the graph of f over the interval [0,1]; see Figure 3.1. The salient features of the graph are that f attains

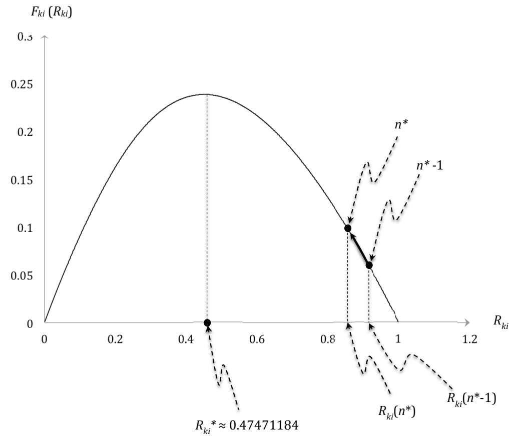

Figure 3.1. Graph of $f(x) := x\left(1 - \mathrm{e}^{-\frac{1}{x}}\right) - (1 - \mathrm{e}^{-x})$ over the interval $[0, 1]$

a maximum at $R^*_{k\mathrm{i}} \approx 0.47471184$ and is monotonically decreasing in the interval $(R^*_{k\mathrm{i}}, 1)$. Theorem 1 shows that the parabolic growth continues as long as $R_{k\mathrm{i}} > R^*$ for all $\mathrm{i} \neq k$.

Lemma 7. *In the interval* $[0.4, 1]$*, the function*

$$f(x) \;\; := \;\; x\left(1 - \mathrm{e}^{-\frac{1}{x}}\right) - \left(1 - \mathrm{e}^{-x}\right)$$

has a unique maximum at $x^* \approx 0.47471184$*. Further,* f *is a monotonically decreasing function in the interval* $(x^*, 1)$*. That is, if* $x^* < x < x < 1$*, then* $f(x) > f(x) > 0$*.*

Proof: Note that

$$f(x) := \frac{d^2 f(x)}{dx^2} \;\; = \;\; \mathrm{e}^{-x} - \frac{\mathrm{e}^{-1/x}}{x^3}$$

We show that $f(x) < 0$ for $0.4 < x < 1$. Since $x > 0$ in the interval (0.4, 1), we note that the following inequalities are equivalent.

$$f(x) < 0 \iff \mathrm{e}^{-x} < \frac{\mathrm{e}^{-1/x}}{x^3} \iff x^3 < \mathrm{e}^{x-1/x}$$

Since $log(y)$ is a monotonically increasing function of y for $0 < y < 1$, in order to establish that $x^3 < e^{x-1/x}$ for $0.4 < x < 1$, we take logarithms on both sides to conclude that the following two inequalities are equivalent.

$$
\begin{aligned}
f(x) < 0 \iff 3\log x < x - \frac{1}{x} \iff \\
x - \frac{1}{x} - 3\log x > 0,\ 0.4 < x < 1
\end{aligned}
$$

Change variables by setting $x = e^z$. For $0.4 < x < 1$, the range of z is $\log(0.4) < z < 0$. Using (3.30) we note that the following inequalities are equivalent

$$
f(x) < 0 \quad \iff \quad \sinh(z) > \frac{3z}{2} \tag{3.30}
$$

We note that the function $g(z) := \sinh(z) - 1.5z$, has exactly three real zeros at $z_0 = 0$, and at $z_1 \approx -1.62213$ and $z_2 \approx +1.62213$. For $z_1 < z < z_0$, $g(z) > 0$. Since $z_1 < \log(0.4)$, for $\log(0.4) < z < 0$ we conclude that $g(z) > 0$. From (3.30) we conclude that $f(x) < 0$ over $0.4 < x < 1$.

Since $f(x) < 0$ over $0.4 < x < 1$, $f(x) := \dfrac{df(x)}{dx}$ is a monotonically decreasing function of x over the interval. Therefore, $f(x)$ can have at most one zero in (0.4, 1). We note that

$$
f(0.45) \approx 0.0132, \qquad f(0.50) \approx -0.0125
$$

showing that $f(x)$ has a zero in the interval (0.45, 0.50). We denote the zero of $f(x)$, as x^*. Since $f(x) < 0$ in the interval (0.45,0.50), x^* is a maximum of $f(x)$.

Since $f(x^*) = 0$ and $f(x) < 0$ on the interval (0.4,1), we conclude that $f(x)$ is a monotonically decreasing function in the interval $(x^*, 1)$. Noting that $f(1) = 0$, completes the proof. ■

Lemma 8. *If assumptions (3.13), (3.14) and (3.15), listed in Theorem 1, are satisfied, then* $O_{k\Sigma}(n^* + 1) > 0$.

Proof: From Lemma 6, Lemma 7 and assumption (3.16) in Theorem 1, we have

$$\begin{aligned}
f(R_{k\mathrm{i}}(n^*)) &= \frac{O_{k\mathrm{i}}(n^*+1)}{A_k(n^*)} > \frac{O_{k\mathrm{i}}(n^*)}{A_k(n^*-1)} \\
&= f(R_{k\mathrm{i}}(n^*-1)) \qquad \Longrightarrow \\
O_{k\Sigma}(n^*+1) &> \frac{A_k(n^*)}{A_k(n^*-1)} \cdot O_{k\Sigma}(n^*) \\
&> O_{k\Sigma}(n^*) \\
&> 0
\end{aligned}$$

■

Lemma 9. *If assumption (3.16), listed in Theorem 1, is satisfied then* $O_{\mathrm{i}\Sigma}(n^*+1) < 0$.

Proof: We note that for all $n \geq 1$,

$$\begin{aligned}
O_{\mathrm{ij}}(n) &= \alpha \left[A_{\mathrm{j}}(n-1) \left(1 - \mathrm{e}^{-R_{\mathrm{ji}}(n-1)}\right) \right. \\
&\qquad \left. - A_{\mathrm{i}}(n-1) \left(1 - \mathrm{e}^{-R_{\mathrm{ij}}(n-1)}\right) \right] \\
&= \alpha \; A_{\mathrm{i}}(n-1) \\
&\quad \left[R_{\mathrm{ij}}(n-1) \left(1 - \mathrm{e}^{-R_{\mathrm{ji}}(n-1)}\right) \right. \\
&\qquad \qquad \left. - \left(1 - \mathrm{e}^{-R_{\mathrm{ij}}(n-1)}\right) \right] \\
&= \alpha \; A_{\mathrm{i}}(n-1) \cdot f(R_{\mathrm{ij}}(n-1)) \\
\frac{O_{\mathrm{ij}}(n)}{A_{\mathrm{i}}(n-1)} &= \alpha f(R_{\mathrm{ij}}(n-1)) \qquad (3.31)
\end{aligned}$$

Lemma 6 and assumption (3.16) imply that $R^* < R_{k\mathrm{i}}(n^*) < R_{k\mathrm{i}}(n^*-1)$. Therefore, from Lemma 7 we get

$$\alpha \; f(R_{k\mathrm{i}}(n^*)) > \alpha \; f(R_{k\mathrm{i}}(n^*-1)) > 0 \tag{3.32}$$

Since $f(R_{k\mathrm{i}}) = -f(R_{\mathrm{i}k})$, using (3.31) and (3.32) we get

$$\begin{aligned} \alpha\ f(R_{\mathrm{i}k}(n^*)) < \alpha\ f(R_{\mathrm{i}k}(n^*-1)) < 0 \quad &\Longleftrightarrow \\ \frac{O_{\mathrm{i}k}(n^*+1)}{A_{\mathrm{i}}(n^*)} < \frac{O_{\mathrm{i}k}(n^*)}{A_{\mathrm{i}}(n^*-1)} &< 0 \end{aligned} \tag{3.33}$$

Since k is the label of the largest node at time n^*, $O_{\mathrm{i}k}(n^*) < 0$. From (3.33), and noting $A_{\mathrm{i}}(n^*), A_{\mathrm{i}}(n^*-1) > 0$, the claim follows.

■

Proof of Theorem 1: From Lemma 3 and assumption 3.15 we conclude that

$$\Delta_k^2(n^*) \quad > \quad 0$$

In order to prove Theorem 1, it is sufficient to show that if inequalities (3.13), (3.14), (3.15) and (3.16) are satisfied at n^* and inequality (3.16) is satisfied at $n = n^*+1$, that is,

$$R_{k\mathrm{i}}(n^*+1) \quad > \quad R^*, \quad 1 \leq \mathrm{i} \leq N,\ \mathrm{i} \neq k \tag{3.34}$$

then analogous forms of inequalities (3.13), (3.14) and (3.15) will remain valid at n^*+1; that is, the following inequalities will be satisfied at n^*+1.

$$O_{k\Sigma}(n^*+1) \quad > \quad 0 \tag{3.35}$$

$$O_{\mathrm{i}\Sigma}(n^*+1) \quad < \quad 0, \quad 1 \leq \mathrm{i} \leq N, \mathrm{i} \neq k \tag{3.36}$$

$$\frac{O_{k\Sigma}(n^*+1)}{O_{k\Sigma}(n^*)} \quad > \quad \frac{1-\delta^{n^*+1}}{1-\delta^{n^*}} \tag{3.37}$$

Establishing inequalities (3.35), (3.36) and (3.37), assuming the validity of inequality (3.34), will prove by induction that

$$\Delta_k^2(n) \quad > \quad 0, \qquad n^* \leq n \leq n^*+r \tag{3.38}$$

as claimed in the Theorem.

Inequality (3.35) is established in Lemma 8, which assumes the validity of inequalities (3.13), (3.14) and (3.15). Inequality (3.36) is established in Lemma 9, which assumes the validity of inequalities (3.13), (3.14) and (3.16) . In the following we will establish inequality (3.37).

From Lemma 6 and the validity of inequality (3.16) at $n = n^*$, we know that

$$R^* < R_{k\mathrm{i}}(n^*) < R_{k\mathrm{i}}(n^* - 1) \; 1 \leq \mathrm{i} \leq N, \mathrm{i} \neq k \tag{3.39}$$

Using Lemma 7 and (3.39) we can then conclude that for $1 \leq \mathrm{i} \leq N, \;\mathrm{i} \neq k$,

$$f(R_{k\mathrm{i}}(n^*)) > f(R_{k\mathrm{i}}(n^* - 1)) > 0 \tag{3.40}$$

Summing inequality (3.40) over all i with $\mathrm{i} \neq k$, we obtain

$$\sum_{\mathrm{i} \neq k} f(R_{k\mathrm{i}}(n^*)) > \sum_{\mathrm{i} \neq k} f(R_{k\mathrm{i}}(n^* - 1)) > 0 \tag{3.41}$$

Using (3.31) we can write the left hand side of (3.37) as

$$\begin{aligned} \frac{O_{k\Sigma}(n^* + 1)}{O_{k\Sigma}(n^*)} &= \frac{A_k(n^* + 1)}{A_k(n^*)} \\ &\quad \cdot \frac{\sum_{\mathrm{i} \neq} f(R_{k\mathrm{i}}(n^*))}{\sum_{\mathrm{i} \neq k} f(R_{k\mathrm{i}}(n^* - 1))} \end{aligned} \tag{3.42}$$

Using (3.40), (3.30) and (3.13), and noting that $A_k(n^*) > 0$, we obtain

$$\begin{aligned}
\frac{A_k(n^*+1)}{A_k(n^*)} &= \frac{\delta A_k(n^*) + u \cdot \rho_k(n^*) + O_{k\Sigma}(n^*)}{A_k(n^*)} \\
&= \delta + \frac{u}{A_\Sigma(n^*)} + \frac{O_{k\Sigma}(n^*)}{A_k(n^*)} \\
&> \delta + \frac{u}{A_\Sigma(n^*)} \\
&= \frac{1-\delta^{n^*+1}}{1-\delta^{n^*}} \qquad (3.43)
\end{aligned}$$

Using (3.41), (3.42) and (3.43) we obtain

$$\frac{O_{k\Sigma}(n^*+1)}{O_{k\Sigma}(n^*)} > \frac{1-\delta^{n^*+1}}{1-\delta^{n^*}}$$

thereby establishing inequality (3.37). ■

To summarize, we have presented a simple model in which the flow of user-base across the network of enterprises depends nonlinearly on the state of the network. It is somewhat surprising that the model allows us to formulate a set of sufficient conditions to detect the onset of parabolic surge in the user-base position of a specific node. Note that the sufficient conditions do not require that a specific node (node k in the above discussion) garner a singularly large user-base position before it undergoes parabolic growth. In fact, the condition that $R_{k\mathrm{i}} > R^* \approx 0.48$ shows that the user-base position of node k is required to be less than about twice (≈ 2.1 times) the size of every other node.The signal of imminent parabolic surge, hence, appears to be rather subtle, and cannot be inferred by looking at merely the user-base position of the most dominant node. Our model is based on a deterministic dynamical system. Clearly, a realistic network of enterprises involves stochastic events that are not captured by a deterministic model. If the deviations from the deterministic model due to stochastic events are sufficiently small, then a deterministic model could provide useful insights.

4. BACKGROUND

A large real world system, such as the World Wide Web, may be thought of as a macroscopic system of particles (websites/webpages) interacting with each other. The sheer size of the system makes it very difficult to concentrate on the behaviour of every particle; we have to rely heavily on probabilistic arguments to understand the behaviour of large systems in general. In this section we present a brief overview of statistical mechanics, a discipline that has enabled us to gain an understanding of the bulk behavior of large systems, *ab initio*, by looking at the microscopic interactions of the constituent particles. We also present our hypothesis and preliminary evidence in support of it, before undertaking a deeper investigation of the hypothesis in later chapters.

4.1 Overview of Statistical Mechanics

Statistical mechanics in general concerns with the connection between the microscopic and macroscopic dynamics of systems comprising large numbers of particles. We begin with a few preliminary remarks.

We assume a large system is characterized by a certain set Ω_{sys} of configurations also called microstates. For example, in case of the World Wide Web, a microstate would be a vector of the user-base position of each and every entity on the World Wide Web. A system is assumed to evolve through random transitions between its microstates.

An equilibrium state is characterized by uniformity of properties in an average sense. The significance of the equilibrium state is that in it the system attains its maximum possible entropy.

Entropy maybe thought of as the amount of information required to describe a system. Entropy of a macrostate–a state that is fully described by bulk parameters such as temperature and pressure–may understood as logarithm of the number of microstates associated with the macrostate. The second law of thermodynamics states that an isolated system will tend to move towards the state of maximum entropy.

Let us take the example of the World Wide Web. We'll derive the theoretical distribution of user-base amongst it's entities in an equilibrium state.

Considering a total of N entities, let a_1 entities have user-base level ϵ_1, a_2 entities user-base level ϵ_2 and so on. We can define p_{j} as the probability of finding an entity in an user-base level ϵ_{j}. We use the Gibbs definition of entropy

$$S = -\sum_{\mathrm{i}} p_{\mathrm{i}} \ln p_{\mathrm{i}}$$

In a state of equilibrium, the entropy is maximized subject to the constraints that the total user-base E and the total number of particles N are conserved. These constraints may be written as :

$$\sum_{\mathrm{i}} p_{\mathrm{i}}\epsilon_{\mathrm{i}} = \frac{E}{N}$$

$$\sum_{\mathrm{i}} p_{\mathrm{i}} = 1$$

Hence the maximization problem can be formulated as

$$P = -\sum_{\mathrm{i}} p_{\mathrm{i}} \ln p_{\mathrm{i}} + \alpha(\sum_{\mathrm{i}} p_{\mathrm{i}} - 1) + \beta(\sum_{\mathrm{i}} p_{\mathrm{i}}\epsilon_{\mathrm{i}} - \frac{E}{N})$$

$$\frac{dP}{dp_{\mathrm{i}}} = -\ln p_{\mathrm{i}} - 1 + \alpha + \beta\epsilon_{\mathrm{i}} = 0$$

$$p_{\mathrm{i}} = \mathbf{e}^{-\beta\epsilon_{\mathrm{i}} - \alpha + 1} = C\mathbf{e}^{-\beta\epsilon_{\mathrm{i}}} \tag{4.1}$$

It may be shown that the density of states between ϵ and $\epsilon + d\epsilon$ is of the form $g(\epsilon)d\epsilon = C_2\epsilon^{\theta}d\epsilon$. So the number of particles $n(\epsilon)$ found in user-base states between ϵ and $\epsilon + d\epsilon$ is of the form

$$n(\epsilon)d\epsilon = f(\epsilon)g(\epsilon)d\epsilon \tag{4.2}$$

We know that $\int_0^\infty n(\epsilon)d\epsilon = N$. So it follows that

$$\int_0^\infty f(\epsilon)g(\epsilon)d\epsilon = N$$
$$M\int \mathbf{e}^{-\beta\epsilon}\epsilon^\theta d\epsilon = N$$
$$M\int \mathbf{e}^{-\beta\epsilon}\frac{(\beta\epsilon)^\theta}{\beta^\theta}\frac{d(\beta\epsilon)}{\beta} = N$$
$$M\beta^{-\theta-1}\int \mathbf{e}^{-\beta\epsilon}(\beta\epsilon)^{(\theta+1)-1}d(\beta\epsilon) = N$$

The integral term here is the gamma function $\Gamma(\theta+1)$, so that $M = \frac{N}{\beta^{-\theta-1}\cdot\Gamma(\theta+1)}$.Using this in (4.2), we have

$$n(\epsilon)d\epsilon = \frac{N}{\beta^{-\theta-1}\cdot\Gamma(\theta+1)}\cdot\mathbf{e}^{-\beta\epsilon}\cdot\epsilon^\theta d\epsilon \tag{4.3}$$

4.2 Empirical evidence

We evaluate the distribution of user-base amongst entities where the change in number of users and content is ~ 0 and no perturbations exist i.e approximately an isolated system. We focus on the popular open-source social-media/aggregation website **Reddit**. Users post content on subreddits which is then 'upvoted' or 'downvoted' by other users. The website has numerous user-defined categories of content. We identified collections of such subreddits in a category with an almost constant amount of online users and entities; they may be approximated as an isolated system so that the user-base amongst the webpages should be distributed as a gamma distribution (general form of Maxwell-Boltzmann distribution) as per our hypothesis. We arrive at the same conclusion from the empirical data. The distribution of user-base is recorded against several timestamps (the distribution at two timestamps separated by $\sim$ 6 months are shown in Figure 4.1). In each instance, the histogram has been fitted against three gamma curves. The expected value of this gamma distribution is a measure of the 'temperature' of the system since the expected value of the distribution is essentially the average user-base per entity in

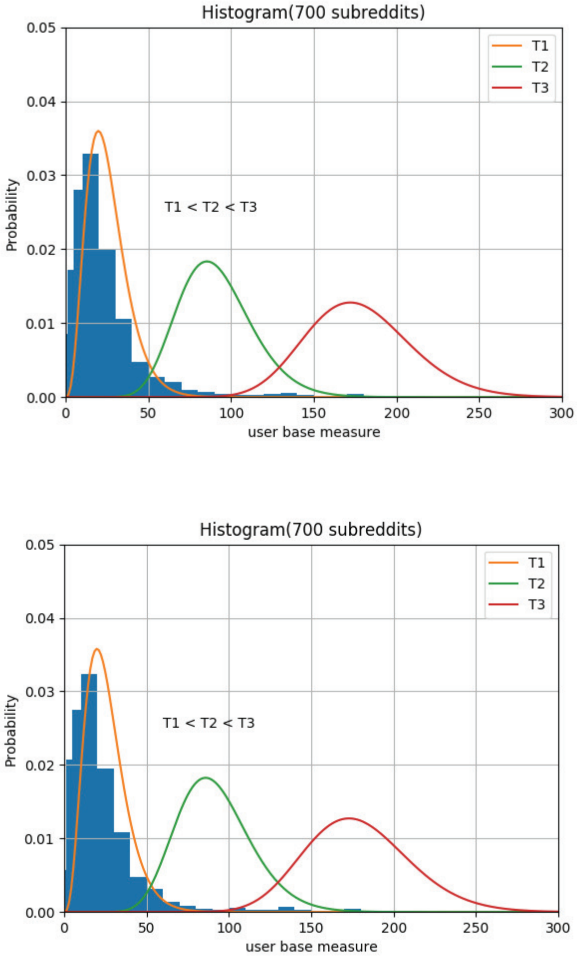

Figure 4.1. The user-base distribution at 2 instances

the website. Notice that the fit curve is of the least temperature denoted by $T1$ in all cases.

4.3 Proposed hypothesis

The empirical data suggests that in pockets of the Web where the user-base and entity population have approached approximately constant values, the distribution of user-base amongst the entities follows a gamma distribution. In section 4.1, we have shown that an isolated system (fixed user-base, entity count and no new perturbations) should evolve to such a state of maximum entropy.

This leads us to posit that such macroscopic systems of interacting entities *are infact trying to move towards a state of maximum entropy; however, perturbations such as the ever changing number of users and entities implies that this state of maximum entropy is also ever-changing.* The rules of interaction amongst the entities must consequently embody such an entropic force. In the ensuing discussion we develop such an interaction rule based model.

5. BINARY INTERACTION MODEL

We approximate the interactions among entities of a system as a series of binary ’collisions’ in which the colliding entitites exchange some amount of a finite conserved resource. For example, in the case of colliding particles, the resource is energy or angular momentum, and in the case of a ‘collision’ between websites the resource could be number of of DAUs (Daily Active Users). The collision is a stochastic process subject to a conservation law. In the World Wide Web consider a collision between two websites labeled i and j having user-base levels (for eg., DAUs) A_{i} and A_{j}. The exact metric used to quantify user-base levels or the time scale of the collision depend on the context. An example of such a collision on the web could result when a user follows a link on, say, a news website i to a different news website j and subsequently shifts his/her attention gradually towards website j. The user-base levels of the websites i and j after the collision will be denoted A_{i} and A_{j}. The conservation law then stipulates that

$$A_{\mathrm{i}} + A_{\mathrm{j}} \;=\; A_{\mathrm{i}} + A_{\mathrm{j}}$$

We model the evolution of a system as a Markov process, and seek to understand the properties that the interactions among entities must have in order for the entropy of the system to be maximized asymptotically.

5.1 Microscopic interaction rules and entropy maximization

When the system of entities (like the World Wide Web) is isolated, we expect the system to evolve to a state of equilibrium (signified by maximum entropy). In this section, we take a closer look at the evolution of the entropy of the isolated system, with the objective of understanding the conditions that interactions among the constituent entities in a closed system must satisfy in order for the system to reach a configuration of maximum entropy asymptotically. Our discussion is based on Boltzmann’s argument [1]

The Gibbs entropy of the system is given by

$$S(t) = -\sum_{\mathrm{i}} p(\mathrm{i}, t) \ln p(\mathrm{i}, t)$$

where $p(\mathrm{i}, t)$ is the probability of an entity with user-base of magnitude i . Recalling that

$$\sum_{\mathrm{i}} p(\mathrm{i}, t) \quad = \quad 1$$

it is easily verified that

$$\frac{dS}{dt} = -\sum_{\mathrm{i}} \frac{dp(\mathrm{i}, t)}{dt} \ln p(\mathrm{i}, t)$$

We can express the evolution of the probability as

$$\frac{dp(\mathrm{i}, t)}{dt} = \sum_{\mathrm{j}} p(\mathrm{j}, t) w(\mathrm{j} \to \mathrm{i}) - p(\mathrm{i}, t) w(\mathrm{i} \to \mathrm{j}) \tag{5.1}$$

$w(\mathrm{i} \to \mathrm{j})$ is the probability of transitioning from a state with user-base i to a state with user-base j. Therefore,

$$w(\mathrm{i} \to \mathrm{j}) \quad = \quad \sum_{k} \sum_{m} p(k, t) \ p_c(\mathrm{i}, k \to \mathrm{j}, m) \, \delta_{\mathrm{i}+k-\mathrm{j}-m,0}$$

where

$$\delta_{\mathrm{i}+k-\mathrm{j}-m,0} \quad = \quad \begin{cases} 1, \mathrm{ji} + k - \mathrm{j} - m = 0 \\ 0\mathrm{ji} + k - \mathrm{j} - m \neq 0 \end{cases}$$

and $p_c(\mathrm{i}, k \to \mathrm{j}, m)$ is the probability that the collision of a website A with user-level i with another website B with user-level k, will result in the website A having user-level j and website B having user-level m after the collision. For brevity, in the following we will use the abbreviation

$$\Delta\mathrm{j} := \mathrm{j}\delta_{\mathrm{i}+k-\mathrm{j}-m,0}$$

Equation (5.1), can be rewritten as follows

$$\begin{aligned}
\frac{dp(\mathrm{i},t)}{dt} &= \Delta\sum_{\mathrm{j}} p(\mathrm{j},t)\sum_m\sum_k p(m,t)\,p_c(\mathrm{j},m\to\mathrm{i},k) - p(\mathrm{i},t)\sum_k\sum_m p(k,t)\,p_c(\mathrm{i},k\to\mathrm{j},m) \\
&= \Delta\sum_{\mathrm{j}}\sum_m\sum_k p(\mathrm{j},t)\,p(m,t)\,p_c(\mathrm{j},m\to\mathrm{i},k) - p(\mathrm{i},t)\,p(k,t)\,p_c(\mathrm{i},k\to\mathrm{j},m)
\end{aligned}$$

Using this in the expression $\frac{dS}{dt}$, we have

$$\frac{dS}{dt} = -\Delta\sum_{\mathrm{i}}\sum_{\mathrm{j}}\sum_m\sum_k [p(\mathrm{j},t)p(m,t)\,p_c(\mathrm{j},m\to\mathrm{i},k) - p(\mathrm{i},t)p(k,t)\,p_c(\mathrm{i},k\to\mathrm{j},m)]\ln p(\mathrm{i},t)$$

Noting that $\mathrm{i},\mathrm{j},k,m$ are dummy indices, and relabeling them we get

$$\begin{aligned}
\frac{dS}{dt} &= -\Delta\sum_{\mathrm{i}}\sum_{\mathrm{j}}\sum_m\sum_k [p(\mathrm{j},t)p(m,t)\,p_c(\mathrm{j},m\to\mathrm{i},k) - p(\mathrm{i},t)p(k,t)\,p_c(\mathrm{i},k\to\mathrm{j},m)]\ln p(k,t) \\
\frac{dS}{dt} &= \Delta\sum_{\mathrm{i}}\sum_{\mathrm{j}}\sum_m\sum_k [p(\mathrm{j},t)p(m,t)\,p_c(\mathrm{j},m\to\mathrm{i},k) - p(\mathrm{i},t)p(k,t)\,p_c(\mathrm{i},k\to\mathrm{j},m)]\ln p(\mathrm{j},t) \\
\frac{dS}{dt} &= \Delta\sum_{\mathrm{i}}\sum_{\mathrm{j}}\sum_m\sum_k [p(\mathrm{j},t)p(m,t)\,p_c(\mathrm{j},m\to\mathrm{i},k) - p(\mathrm{i},t)p(k,t)\,p_c(\mathrm{i},k\to\mathrm{j},m)]\ln p(m,t)
\end{aligned}$$

Adding the above equations we get

$$4\left[\frac{dS}{dt}\right] = -\Delta\sum_{\mathrm{i}}\sum_{\mathrm{j}}\sum_m\sum_k$$

$$\mathrm{jj}[\underbrace{p(\mathrm{j},t)\,p(m,t)\,p_c(\mathrm{j},m\to\mathrm{i},k) - p(\mathrm{i},t)\,p(k,t)\,p_c(\mathrm{i},k\to\mathrm{j},m)]\ln\frac{p(\mathrm{i},t)\,p(k,t)}{p(m,t)\,p(\mathrm{j},t)}}_{a(t)} \tag{5.2}$$

Using the abbreviation

$$f_{\mathrm{j}}\mathrm{j} := \mathrm{j}p(\mathrm{j}, t)$$

we have

$$a(t) \quad = \quad [f_{\mathrm{j}}\, f_m\, p_c(\mathrm{j}, m \to \mathrm{i}, k) - f_{\mathrm{i}}\, f_k\, p_c(\mathrm{i}, k \to \mathrm{j}, m)] \ln \left(\frac{f_{\mathrm{i}} f_k}{f_m f_{\mathrm{j}}} \right)$$

For entropy to increase monotonically, the following condition must be satisfied:

$$a(t) < 0 \tag{5.3}$$

If $\frac{f_{\mathrm{i}} f_k}{f_m f_{\mathrm{j}}} > 1$, then condition (5.3) reduces to

$$f_{\mathrm{j}} f_m\, p_c(\mathrm{j}, m \to \mathrm{i}, k) - f_{\mathrm{i}} f_k\, p_c(\mathrm{i}, k \to \mathrm{j}, m) < 0 \tag{5.4}$$

Similarly if $\frac{f_{\mathrm{i}} f_k}{f_m f_{\mathrm{j}}} < 1$, then condition (5.3) reduces to

$$f_{\mathrm{j}} f_m\, p_c(\mathrm{j}, m \to \mathrm{i}, k) - f_{\mathrm{i}} f_k\, p_c(\mathrm{i}, k \to \mathrm{j}, m) > 0 \tag{5.5}$$

We consider the two cases $p_c(\mathrm{j}, m \to \mathrm{i}, k) \neq p_c(\mathrm{i}, k \to \mathrm{j}, m)$, and $p_c(\mathrm{j}, m \to \mathrm{i}, k) = p_c(\mathrm{i}, k \to \mathrm{j}, m)$ separately below.

Case 1: $p_c(\mathrm{j}, m \to \mathrm{i}, k) \neq p_c(\mathrm{i}, k \to \mathrm{j}, m)$

If $\frac{f_{\mathrm{i}} f_k}{f_m f_{\mathrm{j}}} < 1$, then (5.5) can be written as

$$\frac{f_{\mathrm{j}} f_m}{f_{\mathrm{i}} f_k} > \frac{p_c(\mathrm{i}, k \to \mathrm{j}, m)}{p_c(\mathrm{j}, m \to \mathrm{i}, k)} \tag{5.6}$$

We note that $p_c(\mathrm{i}, k \to \mathrm{j}, m)$ and $p_c(\mathrm{j}, m \to \mathrm{i}, k)$, the scattering probabilities for collisions between two websites are independent of the $f_{\mathrm{i}}, f_{\mathrm{j}}, f_k$ and f_m, which are proportional to the numbers of websites with user-bases of magnitude $\mathrm{i}, \mathrm{j}, k$ and m.

Hence, (5.6) does not hold in general. The same reasoning holds for $\frac{f_\mathrm{i} f_k}{f_m f_\mathrm{j}} > 1$, and we conclude that in this case, entropy does not increase monotonically with time, in general.

Case 2: $p_c(\mathrm{j}, m \to \mathrm{i}, k) = p_c(\mathrm{i}, k \to \mathrm{j}, m)$

If $\frac{f_\mathrm{i} f_k}{f_m f_\mathrm{j}} < 1$, then (5.5) reduces to

$$\frac{f_\mathrm{i} f_k}{f_\mathrm{j} f_m} \mathrm{j} < \mathrm{j} \frac{p_c(\mathrm{j}, m \to \mathrm{i}, k)}{p_c(\mathrm{i}, k \to \mathrm{j}, m)} \quad = \quad 1$$

which is clearly satisfied. The same reasoning holds for $\frac{f_\mathrm{i} f_k}{f_m f_\mathrm{j}} > 1$ and we conclude that the entropy will increase monotonically regardless of the initial configuration of the system, as Boltzmann observed. In this case, it is easy to see that at equilibrium—that is, when $\frac{dS}{dt} = 0$—the system is described by the Boltzmann distribution. At some t

$$\frac{dS}{dt} = 0 \implies \frac{f_\mathrm{j} f_m}{f_\mathrm{i} f_k} = 1 \implies \ln f_\mathrm{j} + \ln f_m = \ln f_\mathrm{i} + \ln f_k$$

So $\ln f$ is summational invariant in the binary collision. However , as per our definition of the collision, the only summational invariant in the binary collision is the total user-base. Hence, $\ln f$ has to be a linear function of the user-base amount. That is,

$$\ln f_\mathrm{j} \quad = \quad \ln C - \beta \mathrm{j} \implies f_\mathrm{j} = C \mathbf{e}^{-\beta \mathrm{j}}$$

yielding the Boltzmann distribution, which characterizes the equilibrium.

5.2 Boltzmann interaction rules

In Section 5.1, we saw that if

$$p_c(\mathrm{i},\mathrm{j} \to m,k) \quad = \quad p_c(m,k \to \mathrm{i},\mathrm{j}) \qquad \forall \mathrm{i},\mathrm{j},m,k \tag{5.7}$$

then a closed system, governed by an interaction rule satisfying the above property evolves monotonically to a state of maximum entropy irrespective of the initial configuration of the isolated system. We call an interaction rule satisfying (5.7) a *Boltzmann Interaction Rule*. In section 5.2, we focus on the nontrivial task of constructing Boltzmann interaction rules. We start by showing, through two examples, that not all collision rules satisfy (5.7). The first is a simple rich-get-richer rule and the second involves a Gaussian damping factor. In both the rules described below, we consider a collision in which two colliding entities have user-base positions A and B before the collision and A and B after the collision.

- **Rule 1**: The rule stipulates that

$$A = A + min(B, |A - B|) \;\; if \;\; A > B$$

$$A = A - min(A, |A - B|) \quad if \;\; A < B$$

$$B = B + min(A, |A - B|) \;\; if \;\; B > A$$

$$B = B - min(B, |A - B|) \quad if \;\; B < A$$

The rule ensures conservation of total user-base amount and implies that the rich get richer. Simulation results for a system with 1000 interacting entities (particles) are shown in Figure Figure 5.1b. The simulation results show that in a system governed by the above interaction rule all of the user-base eventually accrues at a single entity. Thus a system governed by the above rule does not asymptote to a configuration of maximum entropy.

- **Rule 2**: The second interaction rule we consider is given by the following joint probability density function

  $$f(A, B, A, B)\mathrm{j} \propto \mathrm{j}e^{-(A-A)^2}\,\delta(A + B - A - B)$$

  where $\delta(x)$ is the Dirac delta function. We see that Rule 2 satisfies the Boltzmann condition. With this rule, it is more likely that the interacting entities will remain at or close to their user-base levels prior to their interaction. Simulation results for this interaction rule are shown in Figure 5.1c.

We simulate system with 1000 interacting entities. The system is assumed to be closed, that is we assume that the total user-base and number of entities are constant. The Initially the system of entities is randomly allocated a user-base amount (epoch = 0) as shown in Figure 5.1a. Notice that in case of **Rule 1** (Figure 5.1b), all of the user-base is eventually accrued by a single entity. In case of **Rule 2** (Figure 5.1c), the user-base is distributed in a gamma distribution (shape = 1.5591, scale = 2, R^2_{adj} =0.91) across the entities.

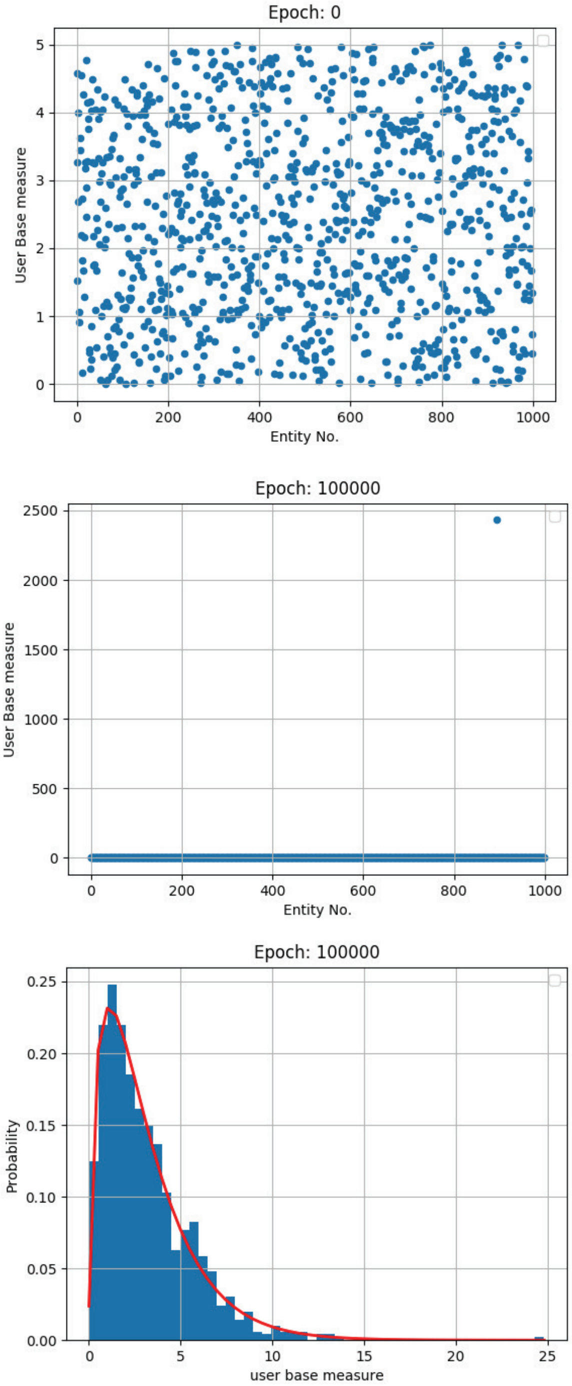

Figure 5.1. a) Initial user-base positions, b) user-base position after 100000 epochs with Rule 1, c) user-base distribution after 100000 epochs with Rule 2

5.3 Enhancements of the microscopic model of interactions

The real-world large systems embody two complex features that are not present in systems comprising distinguishable, but identical, interacting particles. First, a real-world system, such as the World Wide Web, endures unforeseen, stochastic perturbations, such as the emergence of an innovation or a novel social phenomenon, that affect the subsequent evolution of the system. Secondly, unlike the simple system of interacting identical particles, a real-world system is heterogeneous—it comprises different species of constituent entities and consequently different types of interactions among the entities. In order to be able to predict the behavior of such large real-world systems, the microscopic interaction rule that we have discussed above needs to be enhanced to allow for stochastic perturbations, and heterogeneity. The significant challenge that arises due to such enhancements is *to devise interaction rules that, in the face of heterogeneity and perturbations, still drive a closed system (modulo perturbations) towards a configuration of maximum entropy.* We address the above challenging task in this section.

In section 5.3.1, we quantify perturbations and their temporal nature. In section 5.3.2, we define a candidate rule of interaction that will drive the atomic-level interactions in the system. Section 5.3.3 is a discussion on establishing that with the given candidate rules we can maintain the probability mass function (p.m.f) of the interaction model. In section 5.3.4, we show that the collision model will ensure that the necessary and sufficient condition for entropy maximization follows from our interaction rule when the system is isolated, except for the perturbations. In section 5.3.5, we reformulate our collision model and interaction rule to accommodate for the heterogeneous nature of real world n-body systems. Section 5.3.6 shows the probability mass function of the heterogeneous model. Finally in section 5.3.7 we show that even in the heterogeneous case, the system reaches a state of maximum entropy when its isolated.

5.3.1 Perturbations

There is almost no restriction on the mathematical modelling of perturbations. The only requirement is that the perturbations should vanish within some time $t < \infty$. We'll detail how we've modelled the perturbations for the experiments that we have performed on Reddit.

Our objective is to predict the amount of DAU (Daily Active Users) on gaming forums in Reddit. The perturbations affecting this system are a) updates to existing games, b) new games. These perturbations are introduced into the system at fixed points in time. A record of these perturbations are available in [50]. In order to simulate perturbations, we need to predict their impact using historical data; by *impact*, we mean the magnitude of perturbations. The two available and measurable independent variables in this regard are number of pre-release responses for a game/game-update and the average rating of these responses (also extracted from [50]).

Our intention is to *model the perturbation as an exponentially decaying signal.* The magnitude of the perturbation is a probabilistic estimate. Based on the discretised levels of current user-base position of the entity and predictive parameters (number of pre-release responses and average rating/score of these responses), the perturbation is sampled from empirical data (a histogram). For example: let the predictive parameters (pre-release responses average score and number of pre-release responses) be in ranges 0 - 10 and 1000 - 1100 respectively and the current user-base position be in the range 10000 - 11000. The magnitude of perturbation is sampled from a histogram of historical observations satisfying these criterion.

We observed that on an average, the time taken for a perturbation to die out (we take 'die out' to mean that the mentions regarding the perturbation dies down to $\sim$ 1% of initial amount on the reddit forums) is 100 days (Figure A.1a). We estimate a perturbation's magnitude by measuring the change in Daily Active Users over the next 365 days (from the date that the perturbation was introduced).

In the case of Reddit, let an entity E denote a gaming forum, $t_{0\mathrm{j}}$ be the instance when the j^{th} perturbation is imposed on it and $t_{1\mathrm{j}}$ be set to $t_{0\mathrm{j}} + 365$ days. Then τ_{j} is estimated by approximating that the perturbation should decay to 1 percent of initial value within next 100 days (as per the empirical data). Let $K_{E,\mathrm{j}}$ be the initial value of the exponential function. Then for a perturbation of magnitude M, the value of $K_{E,\mathrm{j}}$ is calculated by solving

$$M \;=\; \int_0^{365} K_{E,\mathrm{j}} \cdot \mathrm{e}^{-t\cdot\tau_{\mathrm{j}}} dt$$

Then the magnitude of the j^{th} perturbation on entity E at time t as

$$s(E,\mathrm{j},t) = [K_{E,\mathrm{j}} \cdot \mathrm{e}^{-(t-t_0)\cdot\tau_{\mathrm{j}}}](u(t-t_{0\mathrm{j}}) - u(t-t_{1\mathrm{j}})),\ t \geq t_{0\mathrm{j}} \tag{5.8}$$

where $u(t)$ denotes the unit step function.

5.3.2 Candidate rule of interaction

We have already defined the magnitude of the j^{th} perturbation impacting entity E at time t as $s(E,\mathrm{j},t)$. We use $s(E,\mathrm{j})$ as a shorthand for $s(E,\mathrm{j},t)$. When two entities interact, they exchange some quantity (e.g. user-base in the World Wide Web, virus load in disease transmission etc.).

Let the two entities colliding (interacting) be E_1 and E_2 such that E_1 is at user-base level A_1, and E_2 is at B_1. Let the perturbations impacting E_1 be referenced by i and similarly j for E_2. Hence, the magnitude of the i^{th} perturbation on E_1 is $s(E_1,\mathrm{i})$ at time t. Similarly, $s(E_2,\mathrm{j})$ for E_2.

At time t, let the set of all perturbations impacting E_2 be denoted as $\{N_{E_2,t}\}$. Then the total impact of these perturbations on E_2 at time t is merely $\sum_{\mathrm{i}\in\{N_{E_2,t}\}} s(E_2,\mathrm{i})$. Notice that $\{N_{E_2,t}\}$ will vary with time as not only does the magnitude of perturbations change over time, but also new perturbations might enter the system and old ones may die out.

When E_1, E_2 collide, let the two levels they arrive at after collision be A_2, B_2. Let the total amount of user-base before collision be T. Given that the total user-base level is conserved, the total amount after collision is also T. Then at time t we have

$$\begin{aligned}
&\left|A_2 - (A_1 + \sum_{\mathrm{i}\in\{N_{E_1,t}\}} s(E_1,\mathrm{i},t) - \sum_{\mathrm{j}\in\{N_{E_2,t}\}} s(E_2,\mathrm{j},t))\right| \\
&= \left|(T - B_2) - ((T - B_1) + \sum_{\mathrm{i}\in\{N_{E_1,t}\}} s(E_1,\mathrm{i},t) - \sum_{\mathrm{j}\in\{N_{E_2,t}\}} s(E_2,\mathrm{j},t))\right| \\
&= \left|-B_2 - (-B_1 + \sum_{\mathrm{i}\in\{N_{E_1,t}\}} s(E_1,\mathrm{i},t) - \sum_{\mathrm{j}\in\{N_{E_2,t}\}} s(E_2,\mathrm{j},t))\right| \\
&= \left|-B_2 + B_1 - \sum_{\mathrm{i}\in\{N_{E_1,t}\}} s(E_1,\mathrm{i},t) + \sum_{\mathrm{j}\in\{N_{E_2,t}\}} s(E_2,\mathrm{j},t)\right| \\
&= \left|B_2 - B_1 + \sum_{\mathrm{i}\in\{N_{E_1,t}\}} s(E_1,\mathrm{i},t) - \sum_{\mathrm{j}\in\{N_{E_2,t}\}} s(E_2,\mathrm{j},t)\right| \\
&= \left|B_2 - (B_1 - \sum_{\mathrm{i}\in\{N_{E_1,t}\}} s(E_1,\mathrm{i},t) + \sum_{\mathrm{j}\in\{N_{E_2,t}\}} s(E_2,\mathrm{j},t))\right|
\end{aligned}$$

Let's consolidate the above expression into

$$\begin{aligned}
\left|A_2 - (A_1 + \sum_{\mathrm{i}\in\{N_{E_1,t}\}} s(E_1,\mathrm{i},t) - \sum_{\mathrm{j}\in\{N_{E_2,t}\}} s(E_2,\mathrm{j},t))\right| &= \\
\left|B_2 - (B_1 + \sum_{\mathrm{j}\in\{N_{E_2,t}\}} s(E_2,\mathrm{j},t) - \sum_{\mathrm{i}\in\{N_{E_1,t}\}} s(E_1,\mathrm{i},t))\right| &= \\
\beta_{A_1\to A_2,(E_1,E_2),T,t} &= \\
\beta_{B_1\to B_2,(E_1,E_2),T,t}
\end{aligned} \tag{5.9}$$

Let the entities in the system be the set $\{E^*\}$. Lets define a quantity $N_{cp} = \sum_{E_\mathrm{i} \in \{E^*\}} 1$ We now define the probability distribution of E_1 at user base A_1 colliding with E_2 at user base B_1 to arrive at user base levels A_2, B_2 at time t as

$$p_c([A_1, B_1 \to A_2, B_2], E_1, E_2, t) \quad =$$

$$\begin{cases} \frac{\mathrm{e}^{-\beta^2_{A_1 \to A_2,(E_1,E_2),T,t}}}{N_{cp}}, \mathrm{j} A_1 + B_1 = A_2 + B_2 \ , \quad A_1 \neq A_2 \\ 0 \mathrm{j} A_1 + B_1 \neq A_2 + B_2 \end{cases} \tag{5.10}$$

From (5.10), we have

$$p_c([A_1, B_1 \to A_1, B_1], E_1, E_2, t) \quad = \quad 1 - \sum_{\mathrm{j} \neq 1} p_c([A_1, B_1 \to A_\mathrm{j}, A_1 + B_1 - A_\mathrm{j}], E_1, E_2, t),$$

When entities E_1, E_2 at user base levels A_1, B_1 collide, they can arrive at several user-base levels, restricted only by the conservation of the total user-base amount before and after collision. For an entity E_1, let the possible user base levels it can take be $\{A^*\}$. From the expression (5.9) we see that at some time t if we no longer have any perturbations impacting the system (i.e. $\sum_{\mathrm{i} \in \{N_{E_1,t}\}} s(E_1, \mathrm{i}, t) = 0, \sum_{\mathrm{j} \in \{N_{E_2,t}\}} s(E_2, \mathrm{j}, t) = 0$) , we have

$$\beta_{A_1 \to A_2,(E_1,E_2),T,t} = \beta_{A_2 \to A_1,(E_2,E_1),T,t}$$

5.3.3 Probability mass function of transition probability

Once again considering the example of the World Wide Web, let us verify that the transition probability sums to 1 for a transition of entity E_1 from user-base level A_1 to all possible levels. Let's consider the case where an entity E_1 at user-base level A_i has to transition to level A_j. We denote this transition probability at time t as

$$w(A_\mathrm{i} \to A_\mathrm{j}, E_1, t)$$

This can happen only through an instantaneous collision of E_1 with some other entity E_2 at user-base level B_1 possibly leading to two new user-base levels A_2, B_2. We use the following notation:

- $p(E_1, t)$: the probability of randomly selecting an entity E_1 at time t.
- $p_2(A_k, E_2, t)$: the probability of finding an entity E_2 at user-base level A_k at time t. Let the set of all possible user-base levels be $\{A^*\}$. It follows that for an entity E_j, $\sum_{A_\mathrm{i} \in \{A^*\}} p(A_\mathrm{i}, E_\mathrm{j}, t) = 1$.
- $p_c([A_\mathrm{i}, A_k \rightarrow A_\mathrm{j}, A_m], E_1, E_2, t)$: the probability of entities E_1, E_2 at user base levels A_i, A_k colliding to end up at levels A_j, A_m respectively, at time t. This is already defined in (5.10).

Let the entities in the system be the set $\{E^*\}$. Let the set of all possible user-base levels be $\{A^*\}$. We can now express the transition probability $w(A_\mathrm{i} \rightarrow A_\mathrm{j}, E_1, t)$ at some time t as

$$w(A_\mathrm{i} \rightarrow A_\mathrm{j}, E_1, t) = \sum_{E_2 \in \{E^*\}} p(E_2, t) \sum_{A_k \in \{A^*\}} p_2(A_k, E_2, t) \sum_{A_m \in \{A^*\}} p_c([A_\mathrm{i}, A_k \rightarrow A_\mathrm{j}, A_m], E_1, E_2, t) \tag{5.11}$$

Let's analyze the sum of transition probabilities to all possible user-base levels A_j,

$$\sum_{A_\mathrm{j} \in \{A^*\}} w(A_\mathrm{i} \rightarrow A_\mathrm{j}, E_1, t) =$$

$$\sum_{A_\mathrm{j} \in \{A^*\}} \sum_{E_2 \in \{E^*\}} p(E_2, t) \sum_{A_k \in \{A^*\}} p_2(A_k, E_2, t) \sum_{A_m \in \{A^*\}} p_c([A_\mathrm{i}, A_k \rightarrow A_\mathrm{j}, A_m], E_1, E_2, t)$$

On the R.H.S, let's push the newly introduced summation term inwards so that the R.H.S now translates to

$$\sum_{E_2 \in \{E^*\}} p(E_2, t) \sum_{A_k \in \{A^*\}} p_2(A_k, E_2, t) \sum_{A_\mathrm{j} \in \{A^*\}} \sum_{A_m \in \{A^*\}} p_c([A_\mathrm{i}, A_k \rightarrow A_\mathrm{j}, A_m], E_1, E_2, t) \tag{5.12}$$

We make a few observations:

- From (5.10), we have a probability distribution for $p_c([A_\mathrm{i}, A_k \rightarrow A_\mathrm{j}, A_m], E_1, E_2, t)$. For any A_i, A_k, it follows from this probability distribution definition that

$$\sum_{A_\mathrm{j} \in \{A^*\}} \sum_{A_m \in \{A^*\}} p_c([A_\mathrm{i}, A_k \rightarrow A_\mathrm{j}, A_m], E_1, E_2, t) = 1$$

- Since an entity E_2 has to be at some user-base level at time t, it follows that $\sum_{A_k \in \{A^*\}} p_2(A_k, E_2, t) = 1$
- $\sum_{E_2 \in \{E^*\}} p(E_2, t) = 1$

Using these observations in 5.12, we have

$$\sum_{A_\mathrm{j} \in \{A^*\}} w(A_\mathrm{i} \rightarrow A_\mathrm{j}, E_1, t) = 1$$

5.3.4 Verifying entropy maximization when system is isolated

Extending the discussion from Section 5.3.2, let E be some entity and i be an event/perturbation impacting it. We have defined $s(E, \mathrm{i}, t)$ in Equation (5.8). Notice that in the definition, we have enveloped the exponentially decaying function with a rectangular wave $u(t - t_{\mathrm{i}0}) - u(t - t_{\mathrm{i}1})$, where $t_{\mathrm{i}0}$ is the time when the perturbation was introduced and $t_{\mathrm{i}1} = t_{\mathrm{i}0} + 365$. This ensures that any perturbation affecting the system will only have a finite time of impact.

Let the system be isolated at t^*. Then it is obvious there is some $\delta \geq t^*$ such that we can guarantee $s(E_{\mathrm{j}}, \mathrm{i}, t) = 0 \ \forall t \geq \delta, \quad \forall E_{\mathrm{j}} \in \{E^*\}, \quad \forall \mathrm{i} \in \{N_{E_{\mathrm{j}}, t^*}\}$. All that we are saying here is, after some finite time after the system is isolated, all the perturbations affecting entities within the system will cease to exist. Using Equation (5.9), for $t \geq \delta$, the collision operation for entities E_1, E_2 from user base levels A_1, B_1 to A_2, B_2 may be rewritten as

$$|A_2 - A_1| = |B_2 - B_1| = \beta_{A_1 \to A_2, (E_1, E_2), T, t} \text{ or } \beta_{B_1 \to B_2, (E_1, E_2), T, t} \ , \quad t \geq \delta$$

Similarly, the collision operation for E_1, E_2 from user base levels A_2, B_2 to A_1, B_1 may be rewritten as

$$|A_1 - A_2| = |B_1 - B_2| = \beta_{A_2 \to A_1, (E_1, E_2), T, t} \ = \ \beta_{B_2 \to B_1, (E_1, E_2), T, t} \ , \quad t \geq \delta \tag{5.13}$$

Combining the relation from (5.13) and the probability distribution defined in (5.10), we see that

$$p_c([A_1, B_1 \to A_2, B_2], E_1, E_2, t) = p_c([A_2, B_2 \to A_1, B_1], E_1, E_2, t) \ \ \forall t \geq \delta \geq t^* \geq 0$$

In Section 5.1, we have already shown that such a symmetry in the collision probability (i.e. the probability of two entities at user base levels A_1, B_1 colliding to arrive at levels A_2, B_2) will ensure a monotonic growth in entropy.

5.3.5 Accomodating for the heterogeneity of entities

Let us take the example of the World Wide Web. In the previous discussion, we have implicitly assumed that the entities in the system are of the same 'type' i.e. they have same properties. However, empirically we can easily observe the heterogeneity of the Web (e.g Amazon.com , cnn.com are two entities on the web with almost no similarity and consequently there is a very low probability that they compete with each other for user-base). We have to accommodate for such heterogeneity. One possibility is to tag the entities by a feature vector and develop a measure of similarity that should be incorporated into the probabilistic interaction rule.

Let s_{i} be the feature vector for an entity i, s_{j} be the feature vector for entity j. We define the similarity measure as the cosine of the angle between them. The 'coupling' between two entities with feature vectors i, j colliding is $s_{\mathrm{ij}} = cos\,(\theta_{\mathrm{ij}})$. We add this component to the definition of interaction rule.

Consider E_1 at user base A_1 with feature vector m colliding with E_2 at user base B_1 with feature vector n to arrive at user base levels A_2, B_2 respectively, at time t. We update the probability distribution in (5.10) to incorporate the heterogeneity as follows:

$$p_c^{(h)}([A_1, B_1 \to A_2, B_2], E_1, E_2, t) \quad =$$

$$\begin{cases} \dfrac{\mathrm{e}^{-\left(\frac{\beta^2_{A_1 \to A_2, (E_1, E_2), T, t}}{s_{mn}}\right)}}{N_{cp}}, \mathrm{j} A_1 + B_1 = A_2 + B_2\ , \quad A_1 \neq A_2 \\ 0 \mathrm{j} A_1 + B_1 \neq A_2 + B_2 \end{cases} \tag{5.14}$$

From (5.14), we have

$$p_c^{(h)}([A_1, B_1 \to A_1, B_1], E_1, E_2, t) \quad = \quad 1 - \sum_{\mathrm{j} \neq 1} p_c^{(h)}([A_1, B_1 \to A_{\mathrm{j}}, A_1 + B_1 - A_{\mathrm{j}}], E_1, E_2, t),$$

Having added heterogeneity to our system, we must now show that the transition probabilities still sum to 1 and similar to Section 5.1, the argument for entropy maximization still holds.

5.3.6 Probability mass function of transition probability in a heterogeneous setting

Once again we have to show that the sum of transition probabilities is 1 , this time in a heterogeneous setting. Notice that even with the addition of heterogeneity, the same sequence of arguments as in Section 5.3.4 may be used.

5.3.7 Trajectory of entropy growth in a heterogeneous setting

This is an extension of Section 5.1 to a heterogeneous setting. Once again, we assume the system is isolated. We denote a few probability notations first:

- $p([k, L], t)$: probability of finding an entity with user base A_k and feature vector s_L.
- $p_c([[\mathrm{i}, K], [k, L] \rightarrow [\mathrm{j}, K], [m, L]], t)$: the probability that the collision leads to website with user base level A_i and feature vector s_K moving to user-base level A_j and A_k user-base website moving to level A_m, with feature vector s_L at time t.

The probability of an entity at user base level A_i with feature vector s_K moving to user base level A_j at time t is expressed as

$$w([\mathrm{i}, K] \rightarrow [\mathrm{j}, K], t) = \sum_L \sum_k \sum_m p([k, L], t)\, p_c([[\mathrm{i}, K], [k, L] \rightarrow [\mathrm{j}, K], [m, L]], t)$$

The Equation (5.1), is then rewritten as follows

$$\begin{aligned}
\frac{dp([\mathrm{i},K],t)}{dt} &= \sum_{\mathrm{j}} p([\mathrm{j},K],t) \sum_{L}\sum_{m}\sum_{k} p([m,L],t)\, p_c([[\mathrm{j},K],[m,L] \to [\mathrm{i},K],[k,L]],t) - \\
&\qquad p([\mathrm{i},K],t) \sum_{L}\sum_{k}\sum_{m} p([k,L],t)\, p_c([[\mathrm{i},K],[k,L] \to [\mathrm{j},K],[m,L]],t) \\
\frac{dp([\mathrm{i},K],t)}{dt} &= \sum_{L}\sum_{\mathrm{j}}\sum_{m}\sum_{k} [p([\mathrm{j},K],t)p([m,L],t)\, p_c([[\mathrm{j},K],[m,L] \to [\mathrm{i},K],[k,L]],t) - \\
&\qquad p([\mathrm{i},K],t)p([k,L],t)\, p_c([[\mathrm{i},K],[k,L] \to [\mathrm{j},K],[m,L]],t)]
\end{aligned}$$

Using this in the expression $\frac{dS}{dt}$, we have

$$\begin{aligned}
\frac{dS}{dt} &= -\sum_{\mathrm{i}}\sum_{K}\sum_{\mathrm{j}}\sum_{L}\sum_{m}\sum_{k} [p([\mathrm{j},K],t)p([m,L],t)\, p_c([[\mathrm{j},K],[m,L] \to [\mathrm{i},K],[k,L]],t) - \\
&\qquad p([\mathrm{i},K],t)p([k,L],t)\, p_c([[\mathrm{i},K],[k,L] \to [\mathrm{j},K],[m,L]],t)] \ln p([\mathrm{i},K],t) \\
\frac{dS}{dt} &= -\sum_{\mathrm{i}}\sum_{K}\sum_{\mathrm{j}}\sum_{L}\sum_{m}\sum_{k} [p([\mathrm{j},K],t)p([m,L],t)\, p_c([[\mathrm{j},K],[m,L] \to [\mathrm{i},K],[k,L]],t) - \\
&\qquad p([\mathrm{i},K],t)p([k,L],t)\, p_c([[\mathrm{i},K],[k,L] \to [\mathrm{j},K],[m,L]],t)] \ln p([k,L],t)
\end{aligned}$$

Similarly , we may also write

$$\begin{aligned}
\frac{dS}{dt} &= \sum_{\mathrm{i}}\sum_{K}\sum_{\mathrm{j}}\sum_{L}\sum_{m}\sum_{k} [p([\mathrm{j},K],t)p([m,L],t)\, p_c([[\mathrm{j},K],[m,L] \to [\mathrm{i},K],[k,L]],t) - \\
&\qquad p(\mathrm{i},t)p(k,t)\, p_c([[\mathrm{i},K],[k,L] \to [\mathrm{j},K],[m,L]],t)] \ln p([\mathrm{j},K],t) \\
\frac{dS}{dt} &= \sum_{\mathrm{i}}\sum_{K}\sum_{\mathrm{j}}\sum_{L}\sum_{m}\sum_{k} [p([\mathrm{j},K],t)p([m,L],t)\, p_c([[\mathrm{j},K],[m,L] \to [\mathrm{i},K],[k,L]],t) - \\
&\qquad p(\mathrm{i},t)p(k,t)\, p_c([[\mathrm{i},K],[k,L] \to [\mathrm{j},K],[m,L]],t)] \ln p([m,L],t)
\end{aligned}$$

so that adding them we have,

$$\frac{dS}{dt} = -\sum_{\mathrm{i}}\sum_{K}\sum_{\mathrm{j}}\sum_{L}\sum_{m}\sum_{k} [p([\mathrm{j},K],t)p([m,L],t)\, p_c([[\mathrm{j},K],[m,L] \to [\mathrm{i},K],[k,L]],t) - \\ p([\mathrm{i},K],t)p([k,L],t)\, p_c([[\mathrm{i},K],[k,L] \to [\mathrm{j},K],[m,L]],t)] \ln \frac{p([\mathrm{i},K],t)p([k,L],t)}{p([m,L],t)p([\mathrm{j},K],t)}$$

It can now be seen that the same argument used in Section 5.1 may be used here to show that $\frac{dS}{dt} > 0$ until equilibrium is reached.

6. SIMULATION

The question we seek to answer is: *with the given rules of interaction, can we predict the evolution and emergence of hotspots in large systems?*. Let's clarify what we mean by this: Take the case of the World Wide Web. We already have access to a list of significant perturbations affecting the system over the period of 2014-2018. Can we predict the time-dependence of the DAUs of all the entities when subjected to these perturbations? Understanding the time-evolution of DAUs will enable us to detect the emergence of hotspots in the system.

We also consider the ongoing pandemic. For the spread of COVID-19, we are interested in capturing the heterogeneity of the number of observed infections across the globe. Given variations in the (1) proximity to the origin of the virus, (2) population density of the cities, (3) time of introduction of non-pharmaceutical intervention measures and (4) number of undetected cases initially entering the local population, estimating the number of infections is a non-trivial undertaking. To get precise estimates of the infections we need robust and high-resolution geo-spatial information. Since we're interested in only capturing the heterogeneity, we work with a coarse description of the system using the above mentioned descriptor variables.

6.1 Datasets

Reddit

The data we use to evaluate the model pertains to active users of online games. It is estimated that around 1 billion people game online [51]. This includes a mixture of console, smartphones and PCs. The projected year-on-year growth is 6.4% [52]. However, the user activity measures from online gaming sites themselves are not publicly available. Hence, we use the activity on the social media site Reddit as an approximation. Historical reddit data is publicly available and accessible through a free API. Almost all of the online gaming companies have a dedicated discussion forum where users discuss about latest game releases, strategies to play etc. We record the Daily Active Users (DAU) on these forums as the measure of popularity of online gaming companies. We evaluate the model in the time period from 2009 to 2018 (data from 2009 to 2014 is used to arrive at a probability distribution for the magnitude of perturbations). The perturbations impacting this system are (1) updates to existing games (2) introduction of new games. A time stamped record of these perturbations are available on [50]. Their impact is estimated by measuring the change in DAU of the respective gaming company forums. The two independent and measurable parameters for predicting the magnitude and direction of the perturbation are the count and average rating of pre-release responses. These are copies of the games handed out to popular gamers before the release date, who then post their review on sites like [50].

COVID-19

To simulate the spread of infections from COVID-19 through human to human interactions we need the population densities and an estimate of the initial number of undetected cases entering the local population. For the density, we source available data from Wikipedia articles on individual cities/counties. The initial number of cases to a city is estimated as being proportional to the incoming traffic from the source of the virus. Ideally we need granular data to estimate the initial number of undetected COVID-19 cases. The absence of this data leads us to approximate this quantity as being proportional to the traffic of tourists to a city from the source of the virus.

The daily number of infections is obtained from USAFACTS.org [53]. We use this to qualitatively assess our model's capability to predict the disease spread.

6.2 Estimating model parameters

Reddit

To estimate the model parameters we use the data from 2009 - 2014. We identify all the instances of perturbations affecting the system. The reddit forums are scraped for keywords referencing the perturbation, over the next 400 days. We have already described how we've estimated the magnitude of the perturbations. We classify the entities based on the following features:

- First Person Shooter
- Real-time Strategy
- Massively Mulitplayer Online Game
- Player vs Environment
- Player vs Player

Using these, we design a feature vector for each gaming company. For example, if a company designs First Person Shooter games alone, its feature vector would be $[1, 0, 0, 0, 0]$.

COVID-19

Epidemiology studies are generally concerned with how changing parameters can modify the spread of diseases. To show the validity of our model we try to capture the heterogeneity in the peak number of infections across cities with different population densities. The observed number of COVID-19 cases are available on the World Wide Web [53]. The population density and annual tourist traffic from countries across the globe and annual traffic in large airports are obtained from Wikipedia. We approximate the initial number of COVID cases entering a local population as proportional to an estimate of the number of tourists from the origin of the virus entering the local population. We evaluate the infection numbers in the following 15 cities:

- Chicago, Milan, Tokyo, Kuala Lumpur, Seattle, New York, Indianapolis, Singapore, Rome, Paris, London, Seoul, Ho Chi Minh City, Taipei, Los Angeles, Rome.

External Perturbations

- **Demography based infection probability**: The age group spread of the population. The age groups are $0-30, 30-60, 60-100$. Depending on the age group, the probability of hospitalisation-after-infection varies [54].

- **Threshold for infection spread**: The minimum amount of virus quantity required in a host's body so that the host may then become infected.

- **Social Distancing**: As a Bernoulli random variable to decide if an interaction will take place. Here p is the probability of a success. The value of p should proportionally increase with social distancing strictness.

- **Infection Testing Rate**: The publicly available testing rates are scaled to match our population count. Let the number of tests in the simulation be estimated as n. Then in every iteration, n people are randomly selected for testing, and if infected, they are quarantined.

- **Population density in a city**: Let the initial virus load be A_1 and possible virus load after interaction with an infected be A_2. Let β_2 denote the density equivalent. Then the probability of this event occurring is $\propto \mathrm{e}^{-\frac{(A_1-A_2)^2}{\beta_2}}$. We set the density $\beta_2 = \alpha \times$ population density.

- **Initial number of infections (Boundary Condition)**: Let the historic tourist fraction to a city be x. Let the fraction of air-traffic to nearby airports relative to air-traffic to all concerned cities in the associated country be y. Let the relative air traffic to the concerned country from the origin of the virus be z. Let the number of initial cases in the city of origin be e. Let the relative air traffic to this city be

k. Then we approximate the initial number of infected cases to the said city as $x \cdot y \cdot z \cdot \mathrm{e} \cdot k$.

We do not assume the presence of any sort of restrictions on the daily movement of the local populace (excluding social distance). We allow for each entity to interact with some other entity in every iteration (i.e. every day). Hence we don't expect the trajectory of disease spread to be inline with the actual numbers. Each iteration of the simulation is one day. We randomly divide the population into pairs of two and let them interact. For any pair of entities, the virus load after interaction is estimated through the *collision* model taking into consideration the demography based hospitalisation after infection probability and population density. Following this, the threshold for infection spread and social distancing are introduced to ascertain if the interaction is truly successful.

6.3 Algorithms

In Section 6.2, we have explained how we measure the impact of a perturbation on the reddit forums. In Section 5.3.1 we have detailed how we predict the magnitude of impact of such perturbations. We describe the mechanism of interaction in Algorithm 1. The entities are made to undergo random binary collisions obeying the interaction rule (in Section 5.3.2).

Algorithm 1: N entities competing for user-base through ”collisions”

```
Result: User-base positions of N entities
initialise the user-base of N entities to their respective DAU, $d$ to 0, a list of 4-variable tuples for each entity;
while $d <=$ prediction time-horizon do
    initialise i to 0;
    while i <= N-1 do
        if list for entity is non-empty then
            initialise $k$ to first element of list;
            while $k$ ! = end of list do
                I = first element of tuple, $\tau$ = second element of tuple;
                K(t) = third element of tuple, $t_*$ = fourth element of tuple;
                update K(t) = $I * e^{-(d-t_*)/\tau}$;
                point $k$ to next element of list;
            end
        end
        if perturbation introduced for entity then
            create and add tuple of initial value, time constant, current value (same as initial value), time of origin;
        end
        increment i by 1;
    end
    initialise epoch to 0;
    while epoch <= $N-1$ do
        randomly divide N entities into 2 groups;
        initialise j to 0, set $M$ to size of smaller of two groups;
        while j <= $M$ do
            with uniform prob. remove two entities, one from each group;
            follow the collision rule to let the two entities interact;
            update their respective user-base to values following collision;
            increment j by 1;
        end
        increment epoch by 1;
    end
end
increment $d$ by 1;
```

Algorithm 2: N entities competing for user-base through preferential attachment

```
Result: User-base positions of N entities
initialise the user-base of N entities to their respective DAU, $d$ to 0, a list of 4-variable tuples for each entity;
while $d$ <= prediction time-horizon do
    initialise i to 0;
    while i <= N-1 do
        if list for entity is non-empty then
            initialise $k$ to first element of list;
            while $k$ ! = end of list do
                I = first element of tuple, $\tau$ = second element of tuple;
                K(t) = third element of tuple, $t_*$ = fourth element of tuple;
                update K(t) = $I * e^{-(d-t_*)/\tau}$;
                point $k$ to next element of list;
            end
        end
        if perturbation introduced for entity then
            create and add tuple of initial value, time constant, current value (same as initial value), time of
              origin;
        end
        increment i by 1;
    end
    for each entity update new user-base amount by sampling from total user-base amount with sampling
      weight = existing user-base + K(t)
end
increment $d$ by 1;
```

Algorithm 3: N entities competing for user-base through cascade effect

```
Result: User-base positions of N entities
initialise the user-base of N entities to their respective DAU, $d$ to 0, a list of 4-variable tuples for each entity;
while $d <=$ prediction time-horizon do
    initialise i to 0;
    while i <= N-1 do
        if list for entity is non-empty then
            initialise $k$ to first element of list;
            while $k$ != end of list do
                I = first element of tuple, $\tau$ = second element of tuple;
                K(t) = third element of tuple, $t_*$ = fourth element of tuple;
                update K(t) = $I * \mathrm{e}^{-(d-t_*)/\tau}$;
                point $k$ to next element of list;
            end
        end
        if perturbation introduced for entity then
            create and add tuple of initial value, time constant, current value (same as initial value), time of
              origin;
        end
        increment i by 1;
    end
    Readjust number of users adopting a game by sampling from total user-base proportional to perturbations;
    Randomly select a neighbour of an entity and with non-zero probability let the neighbour adopt the user's
      preference;
    update total user-base levels for each product;
end
increment $d$ by 1;
```

6.4 Validation

Reddit

Previous literature has extensively treated the spread of popularity of competing entities from a user perspective (i.e. the spread of innovations/products amongst a (homogeneous) network of users). Broadly speaking, there are two models that have been used to study the competition for user-base amongst entities on the World Wide Web.

- In [3], the classic preferential attachment mechanism is used, with random shifts in popularity to represent external factors. The authors were able to verify critical features observed in empirical analysis of the networks considered. We'll refer to this model as b1. The datasets used here were the entire Wikipedia and the Chilean Web. The popularity of a document on Wikipedia is determined by considering the number of hyperlinks pointing to it and the traffic to it expressed by the number of clicks on it. The fractional change in these measurements is used to represent the variation of popularity with time. Upon plotting the distribution of this measure, the authors were able to find heavy tailed distribution. The authors also observed pages receiving intermittent spikes in traffic owing to exogenous factors. Given these two factors, the model tries to accomodate both the scale free distribution observed and the spikes in attention. To model the powerlaw distribution, a preferential attachment mechanism is adopted. Every page is assigned a rank such that the probability that a page receives a unit of attention is proportional to its existing attention/popularity rank. The authors observe that this preferential attachment model does not sufficiently capture the long-tailed distribution of relative change in traffic. They argue that the long-tailed nature is because the model fails to account for external factors affecting the current popularity levels. A simple and straightforward mechanism to do this is to randomly alter the popularity rank such that with a non-zero probability every item in the ranked list is moved to the front of the list, chosen uniformly between the current rank and 1. This is defined as the rank-shift model. This new model is able to capture the ditribution of traffic and the long-tailed nature of the distribution of fractional change in attention.

- In [55], the cascade effect model is used to study the competition amongst memes for user limited user-base. The agent based simulation was able explain the wide heterogeneity in the popularity of memes. We'll refer to this model as b2. The model has agents with limited attention. The memes are essentially competing with one another for this finite resource. The dataset used here is a temporal snapshot of Twitter. A notion of heterogeneity is introduced in terms of the interests of the users sharing/accepting memes from their neighbours. Based on the set of memes tweeted by a user, her interests are defined by the set of the 10 most recent memes. The Maximum Information Path is used to measure similarity between two users. Using empirical data, the authors observe the heterogeneous behaviour in meme popularity. The model has a fixed number of users (agents). Each agent maintains a time-ordered record of memes. Each agent is assigned a screen. At every instant,

> each user receives memes shared by her connections. With a probability p_s, the agent chooses to post a new meme which is also stored in memory. Otherwise with a probability $1-p_s$ the user checks her screen and selects a meme on the screen with probability p_m. Then with a probability p_k either a meme from memory is chosen or this previously selected meme is posted with probability $1-p_m$. To capture the notion of limited user attention, posts on the screen and in memory remain only for a finite time. The parameters are chosen from fitting empirical data. The model is able to capture the long-tailed distributions of meme popularity, time to death. The authors try to argue here that external factors aren't required to capture the essentially features of propagation of meme popularity.

We need to measure how well the DAU time series can be predicted by our model as compared to the benchmarks. We utilise statistical hypothesis testing to arrive at a conclusion on this. We use the RMSE (root mean squared error) of our prediction vs the actual DAU of the entities, sampled every 50-days over the test period (2014-2018), as our experiment statistic. We perform a one-sided t-test [56] to study if our model shows a statistically significant improvement in prediction of the DAU time series when compared to the RMSE from the benchmark models specified above. We perform 1000 simulations of each of the three models (the entropic force, b1 and b2). The convergnce plots are in Appendix B. Let the mean RMSE of the entropy model be μ and the mean RMSE of one the benchmark models be μ_0. Then our hypothesis test is as follows:

$$H_o : \mu_0 = \mu$$
$$H_a : \mu_0 > \mu$$

We use a 95% confidence interval. Below are the results of the experiments.

The results show that at the very least, our model performs better than existing models in the literature in predicting the evolution of user-base amongst a system of entities on the World Wide Web. An interesting phenomenon observed in real-world

Table 6.1. Average RMSE Values

model type	**RMSE (averaged over 1000 runs)**
Entropy	2191
Preferential attachment (b1)	4469
Cascade effect (b2)	8939

Table 6.2. t-test Results

vs. model	t_0	$t_{0.05,999}$	**p-value**
b1	55.88	1.6464	< 0.00001
b2	82.77	1.6464	< 0.00001

networks like the World Wide Web is the existence of a long-tailed distribution of traffic across the web entities. In Figure 6.1, we plot the distribution of average DAU from 1000 simulations. As can be seen it is well fit by a power-law curve with scale 2.7. Hence, collision based model is also able to capture this peculiar feature of the World Wide Web.

A key consequence of being able to predict the evolution of DAU across the entities is the possible early detection of hotspots. The simulation should be able to detect such abnormal events. We shall classify a hotspot as a 15% or more deviation in DAU over a 100-day period. We compare the actual % change against that predicted by the model. We choose 15% because the dataset shows 95 percentile of perturbations impact the user-base with less than this amount. Hence it is a good approximation of what a hotspot means in the context of this dataset. We use the 100-day period as we have estimated it to be the average time a perturbation impacts an entity. Interestingly, we observe that on an average (over the 1000 simulation runs), hotspots are predicted by the model to within 2% of the actual magnitude of user-base. The results are posted in Appendix C.

Once we isolate this system, as per the interaction rules we expect the system to tend towards a state of maximum entropy after a sufficiently long time. We plot the distribution of the average state of the system over the 1000 simulation runs in Figure 6.2. We can see that it is well fit by a gamma distribution with shape $a = 2.7$ and scale $= 1$ ($R^2_{adj} = 0.89$).

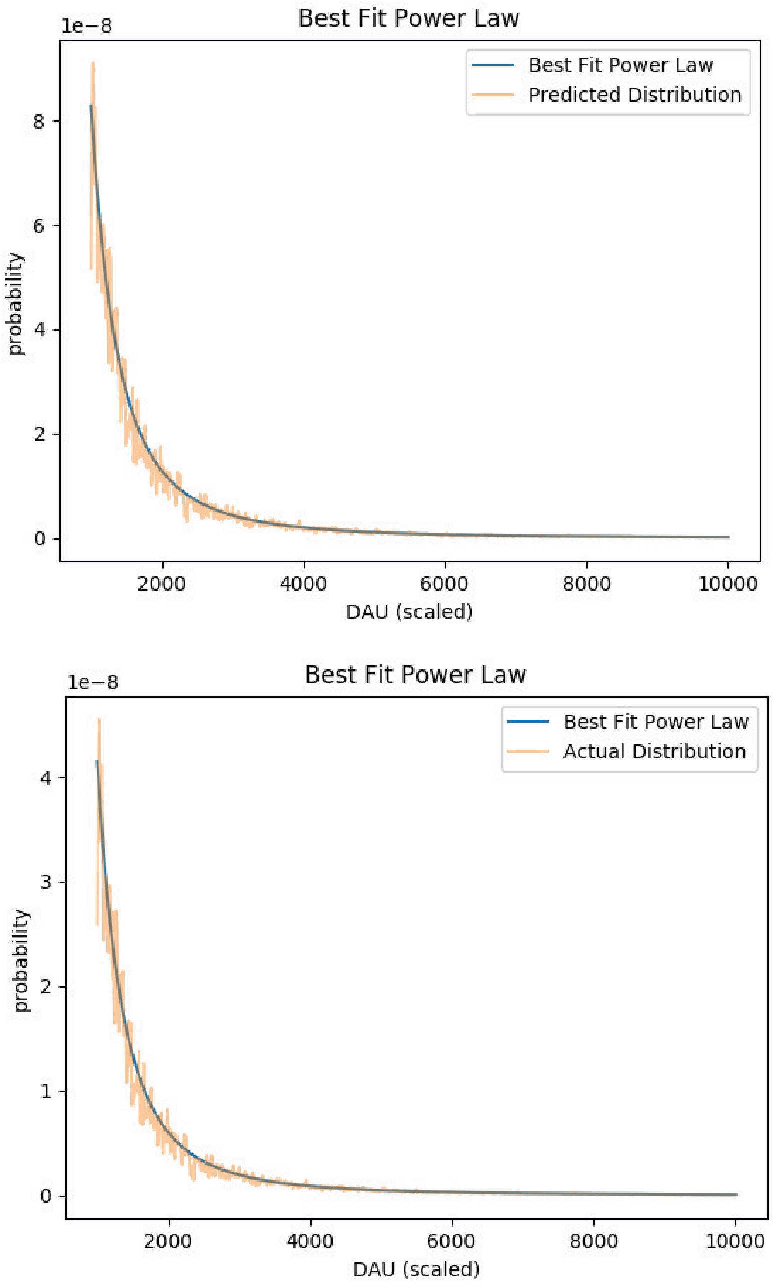

Figure 6.1. a) Power Law fit for predicted DAU ($R^2_{adj} = 0.87$) b) Power Law for actual DAU ($R^2_{adj} = 0.9$)

COVID-19 Study

We want to determine if the entropy based model can capture the heterogeneity of the number of infections across the globe. Since we don't have any historical data to

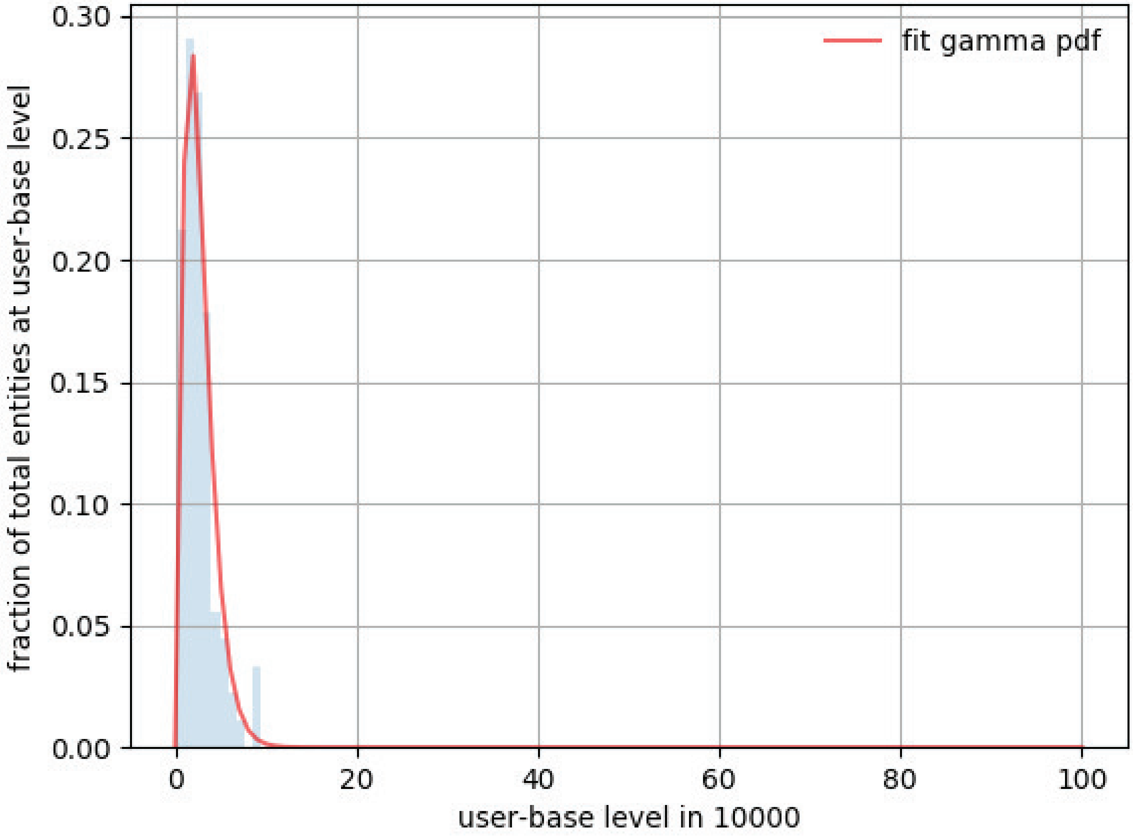

Figure 6.2. a) Distribution of average user-base levels after 5000 iterations once the system is isolated

calibrate the model parameters (i.e threshold for infection spread, social distancing, α), we need to evaluate the model on a range of values of these parameters to ensure its robustness. The infection testing rate and initial number of infections are purely derived from available data and hence are not subject to design. In [54], the authors record one the earliest hospitalization-after-infection rates across China (using data until February 25, 2020). We use this infection rate across the age groups to estimate unknown model parameters. We shall evaluate the model by simulating the spread of COVID-19 from March to December of 2020.

The threshold for infection and the sensitivity to deviations in post interaction level of virus are related to each other in the following sense: Let the threshold be β. Then the sensitivity to deviations should be such that, in a sufficiently large sample of Monte Carlo Simulations of interaction between two persons, the fraction of instances where a successful transmission takes place is inline with the hospitalisation-after-infection rates from [54]. We use a grid-search approach to arrive at suitable values of α corresponding to β. The estimated α are in Appendix E.

Additionally, we have to set discrete levels for the social distancing measure (Appendix E). We perform a t-test with $n = 50$ simulations for 15 cities excluding Tokyo, for each combination of the variables. Let $\mu_0 = 1$. Let μ be the mean accuracy of the simulations. The test is

$$H_o : \mu_0 = \mu$$

$$H_a : \mu_0 > \mu$$

We use a 99% confidence interval. In order to show that there isn't a significant statistical deviation in ranking accuracy, we need to show a lack of evidence to reject the null hypothesis. In Appendix E, we've tabulated the results. The experiment's results suggest our entropy model (with the included perturbations) is able to satisfactorily capture the heterogeneity in infection numbers around the globe. In Figure E.1, we show the heterogeneous nature of infections across two cities in the US and the corresponding predicted infection numbers.

A Curious Observation

Experimentally, we notice that Tokyo is expected to have the highest peak number of infections going against the data on actual number of COVID-19 cases. On further inspection, we've found that this anomaly has been independently reported [57]. Despite the high population density and minimal social distancing measures, the number of infected cases is rather low as of July, suggesting that existing social practices or a form of immunity to corona viruses already exists in the local population. The entropy model we've designed doesn't accommodate for this variability.

7. SUMMARY-CONCLUSION

The overarching theme of this thesis is to arrive at a general *ab initio* model for predicting the evolution of large systems. Existing phenomenological models work well for systems that are at or near equilibrium but fall short when modelling real-world large complex systems like the World Wide Web (*i.e.* systems far away from equilibrium).

We commence with a deterministic model of user-base flow in the World Wide Web, in Section 3. Here we treat the entities on the Web as a network interacting with each other. We were able to show the existence of a set of sufficient conditions which if met by an entity would signal its imminent parabolic growth of its accrued user-base. This result although powerful, is restricted to a purely deterministic setting.

To capture the stochastic nature of real world systems we have employed the tools of statistical mechanics (Section 4). In particular, we utilize the *H-Theorem* and hypothesize that isolated large systems are driven by an innate tendency to reach a state of maximum entropy. We identify the sufficient conditions that an interaction rule must satisfy in order for it to drive a closed system towards a configuration of maximum entropy. Subsequently, we design an interaction rule that satisfies the sufficient condition, and extend the rule to incorporate stochastic perturbations of the system as well as heterogeneity among the interacting constituent entities that make up the system.

For the World Wide Web, we compare the performance of our model against existing models in the literature (Section 6.4). We use the RMSE of actual DAU vs predicted DAU sampled every 50 days over the period from 2014-2018. We use a statistical hypothesis test over 1000 simulations to evaluate the performance. We are able to show a statistically significant improvement in performance through the collision-based model. Importantly, our model is able to capture the emergence of hotspots to within 2% of the actual magnitude of user-base. Additionally, our model captures the power-law distribution of user-base across the entities of the Web.

For the COVID-19 infection study, we perform a t-test to measure the accuracy. We performed an experiment to ensure that the ranking of cities by the peak number of infections based on our simulations does not have a statistically significant deviation from the actual ranking. We record that our model does reproduce the actual ranking, except in the case of Tokyo.

Our results suggest that the principle of entropy maximization might indeed be the guiding principle for understanding the evolution of large systems. Our work, thus, introduces a new paradigm in understanding the behaviour of large complex systems.

REFERENCES

[1] L. Boltzmann, "Further studies on the thermal equilibrium of gas molecules," in *The kinetic theory of gases: an anthology of classic papers with historical commentary*, World Scientific, 2003, pp. 262–349.

[2] J. G. Webster, *The marketplace of attention: How audiences take shape in a digital age.* Mit Press, 2014.

[3] J. Ratkiewicz, S. Fortunato, A. Flammini, F. Menczer, and A. Vespignani, "Characterizing and modeling the dynamics of online popularity," *Physical review letters*, vol. 105, no. 15, p. 158 701, 2010.

[4] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *science*, vol. 286, no. 5439, pp. 509–512, 1999.

[5] S. Fortunato, A. Flammini, and F. Menczer, "Scale-free network growth by ranking," *Physical review letters*, vol. 96, no. 21, p. 218 701, 2006.

[6] J.-P. Onnela and F. Reed-Tsochas, "Spontaneous emergence of social influence in online systems," *Proceedings of the National Academy of Sciences*, vol. 107, no. 43, pp. 18 375–18 380, 2010.

[7] D. M. Boyd and N. B. Ellison, "Social network sites: Definition, history, and scholarship," *Journal of computer-mediated Communication*, vol. 13, no. 1, pp. 210–230, 2007.

[8] B. Ribeiro and C. Faloutsos, "Modeling website popularity competition in the attention-activity marketplace," in *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, ACM, 2015, pp. 389–398.

[9] B. Hood, V. Hwang, and J. King, "Inferring future business attention," *Yelp Challenge, Carnegie Mellon University*, 2013.

[10] D. Chakrabarti, Y. Wang, C. Wang, J. Leskovec, and C. Faloutsos, "Epidemic thresholds in real networks," *ACM Transactions on Information and System Security (TISSEC)*, vol. 10, no. 4, p. 1, 2008.

[11] A. Ganesh, L. Massoulié, and D. Towsley, "The effect of network topology on the spread of epidemics," in *INFOCOM 2005. 24th Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings IEEE*, IEEE, vol. 2, 2005, pp. 1455–1466.

[12] H. W. Hethcote, “The mathematics of infectious diseases,” *SIAM review*, vol. 42, no. 4, pp. 599–653, 2000.

[13] B. A. Prakash, A. Beutel, R. Rosenfeld, and C. Faloutsos, “Winner takes all: Competing viruses or ideas on fair-play networks,” in *Proceedings of the 21st international conference on World Wide Web*, ACM, 2012, pp. 1037–1046.

[14] K. S. McCurley, “Income inequality in the attention economy,” *Google Research*, 2007.

[15] M. H. Goldhaber, “The attention economy and the net.,” *First Monday*, vol. 2, no. 4, 1997.

[16] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan, “Group formation in large social networks: Membership, growth, and evolution,” in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, 2006, pp. 44–54.

[17] E. M. Jin, M. Girvan, and M. E. Newman, “Structure of growing social networks,” *Physical review E*, vol. 64, no. 4, p. 046 132, 2001.

[18] S. R. Kairam, D. J. Wang, and J. Leskovec, “The life and death of online groups: Predicting group growth and longevity,” in *Proceedings of the fifth ACM international conference on Web search and data mining*, ACM, 2012, pp. 673–682.

[19] E. M. Rogers, *Diffusion of innovations*. Simon and Schuster, 2010.

[20] M. Granovetter, “Threshold models of collective behavior,” *American journal of sociology*, vol. 83, no. 6, pp. 1420–1443, 1978.

[21] D. Acemoglu, A. Ozdaglar, and E. Yildiz, “Diffusion of innovations in social networks,” in *Decision and control and European control conference (CDC-ECC), 2011 50th IEEE conference on*, IEEE, 2011, pp. 2329–2334.

[22] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins, “Information diffusion through blogspace,” in *Proceedings of the 13th international conference on World Wide Web*, ACM, 2004, pp. 491–501.

[23] N. T. Bailey *et al.*, *The mathematical theory of infectious diseases and its applications*. Charles Griffin & Company Ltd, 5a Crendon Street, High Wycombe, Bucks HP13 6LE., 1975.

[24] R. M. Anderson and R. M. May, *Infectious diseases of humans: dynamics and control.* Oxford university press, 1992.

[25] P. S. Dodds and D. J. Watts, "Universal behavior in a generalized model of contagion," *Physical review letters*, vol. 92, no. 21, p. 218 701, 2004.

[26] J. Leskovec, L. A. Adamic, and B. A. Huberman, "The dynamics of viral marketing," *ACM Transactions on the Web (TWEB)*, vol. 1, no. 1, p. 5, 2007.

[27] R. Albert and A.-L. Barabási, "Statistical mechanics of complex networks," *Reviews of modern physics*, vol. 74, no. 1, p. 47, 2002.

[28] M. E. Newman, "The structure and function of complex networks," *SIAM review*, vol. 45, no. 2, pp. 167–256, 2003.

[29] M. Molloy and B. Reed, "A critical point for random graphs with a given degree sequence," *Random structures & algorithms*, vol. 6, no. 2-3, pp. 161–180, 1995.

[30] R. Cohen, A. F. Rozenfeld, N. Schwartz, D. Ben-Avraham, and S. Havlin, "Directed and non-directed scale-free networks," in *Statistical Mechanics of Complex Networks*, Springer, 2003, pp. 23–45.

[31] A.-L. Barabâsi, H. Jeong, Z. Néda, E. Ravasz, A. Schubert, and T. Vicsek, "Evolution of the social network of scientific collaborations," *Physica A: Statistical mechanics and its applications*, vol. 311, no. 3, pp. 590–614, 2002.

[32] L. Zhu and K. Lerman, "Attention inequality in social media," *arXiv preprint arXiv:1601.07200*, 2016.

[33] L. Hamill and N. Gilbert, "Social circles: A simple structure for agent-based social network models," *Journal of Artificial Societies and Social Simulation*, vol. 12, no. 2, p. 3, 2009.

[34] S. Mossa, M. Barthelemy, H. Stanley, and L. A. N. Amaral, "Truncation of power law behaviour in "scale-free" network models due to information filtering," *Physical Review Letters*, vol. 88, no. 13, p. 138 701, 2002.

[35] M. Kochen, *The small world.* Ablex Pub., 1989.

[36] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world'networks," *nature*, vol. 393, no. 6684, pp. 440–442, 1998.

[37] J. Leskovec, J. Kleinberg, and C. Faloutsos, "Graphs over time: Densification laws, shrinking diameters and possible explanations," in *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, ACM, 2005, pp. 177–187.

[38] P. L. Krapivsky, S. Redner, and F. Leyvraz, "Connectivity of growing random networks," *Physical review letters*, vol. 85, no. 21, p. 4629, 2000.

[39] S. N. Dorogovtsev, J. F. F. Mendes, and A. N. Samukhin, "Structure of growing networks with preferential linking," *Physical review letters*, vol. 85, no. 21, p. 4633, 2000.

[40] H. A. Simon, "On a class of skew distribution functions," *Biometrika*, vol. 42, no. 3/4, pp. 425–440, 1955.

[41] N. Gilbert, *Agent-based models*, 153. Sage, 2008.

[42] E. Kiesling, M. Günther, C. Stummer, and L. M. Wakolbinger, "Agent-based simulation of innovation diffusion: A review," *Central European Journal of Operations Research*, vol. 20, no. 2, pp. 183–230, 2012.

[43] J. W. Gibbs, *Elementary principles in statistical mechanics.* Courier Corporation, 2014.

[44] G. Bianconi and A.-L. Barabási, "Bose-einstein condensation in complex networks," *Physical review letters*, vol. 86, no. 24, p. 5632, 2001.

[45] C. N. Yang, "Journey through statistical mechanics," *International Journal of Modern Physics B*, vol. 2, no. 06, pp. 1325–1329, 1988.

[46] S. Bornholdt and H. Ebel, "World wide web scaling exponent from simon's 1955 model," *Physical Review E*, vol. 64, no. 3, p. 035 104, 2001.

[47] H. Reiss, *Methods of thermodynamics.* Courier Corporation, 2012.

[48] H. B. Callen, "The application of onsager's reciprocal relations to thermoelectric, thermomagnetic, and galvanomagnetic effects," *Physical Review*, vol. 73, no. 11, p. 1349, 1948.

[49] J. M. Vilar and J. Rubi, "Thermodynamics "beyond" local equilibrium," *Proceedings of the National Academy of Sciences*, vol. 98, no. 20, pp. 11 081–11 084, 2001.

[50] Meta Critic, *Game reviews*, [Online; accessed 19-December-2019], 2019. [Online]. Available: https://www.metacritic.com.

[51] Mordor Intelligence, *Online gaming industry*, [Online; accessed 17-December-2019], 2019. [Online]. Available: https://www.mordorintelligence.com/industry-reports/global-games-market.

[52] Statista, *Online games*, [Online; accessed 17-December-2019], 2019. [Online]. Available: https://www.statista.com/outlook/212/100/online-games/worldwide.

[53] Facts,USA, *Coronavirus outbreaks stats and data*, [Online; accessed 8-May-2020], 2020. [Online]. Available: https://usafacts.org/issues/coronavirus/.

[54] R. Verity, L. C. Okell, I. Dorigatti, P. Winskill, C. Whittaker, N. Imai, G. Cuomo-Dannenburg, H. Thompson, P. G. Walker, H. Fu, *et al.*, "Estimates of the severity of coronavirus disease 2019: A model-based analysis," *The Lancet infectious diseases*, 2020.

[55] L. Weng, A. Flammini, A. Vespignani, and F. Menczer, "Competition among memes in a world with limited attention," *Scientific reports*, vol. 2, p. 335, 2012.

[56] D. C. Montgomery, *Introduction to statistical quality control.* John Wiley & Sons, 2007.

[57] A. Iwasaki and N. D. Grubaugh, "Why does japan have so few cases of covid-19?" *EMBO Molecular Medicine*, vol. 12, no. 5, e12481, 2020.

[58] B. G. Kyle, "Entropy: Reflections of a classical thermodynamicist, chemical and process thermodynamics," 1984.

A. EMPIRICAL ESTIMATIONS

In Figure A.1a, we have plotted the time over which a perturbation remains in the system. This is measured as the interval from the time of introduction of the perturbation to the time when mention of the perturbation in the subreddit attenuates to ≈ 1.

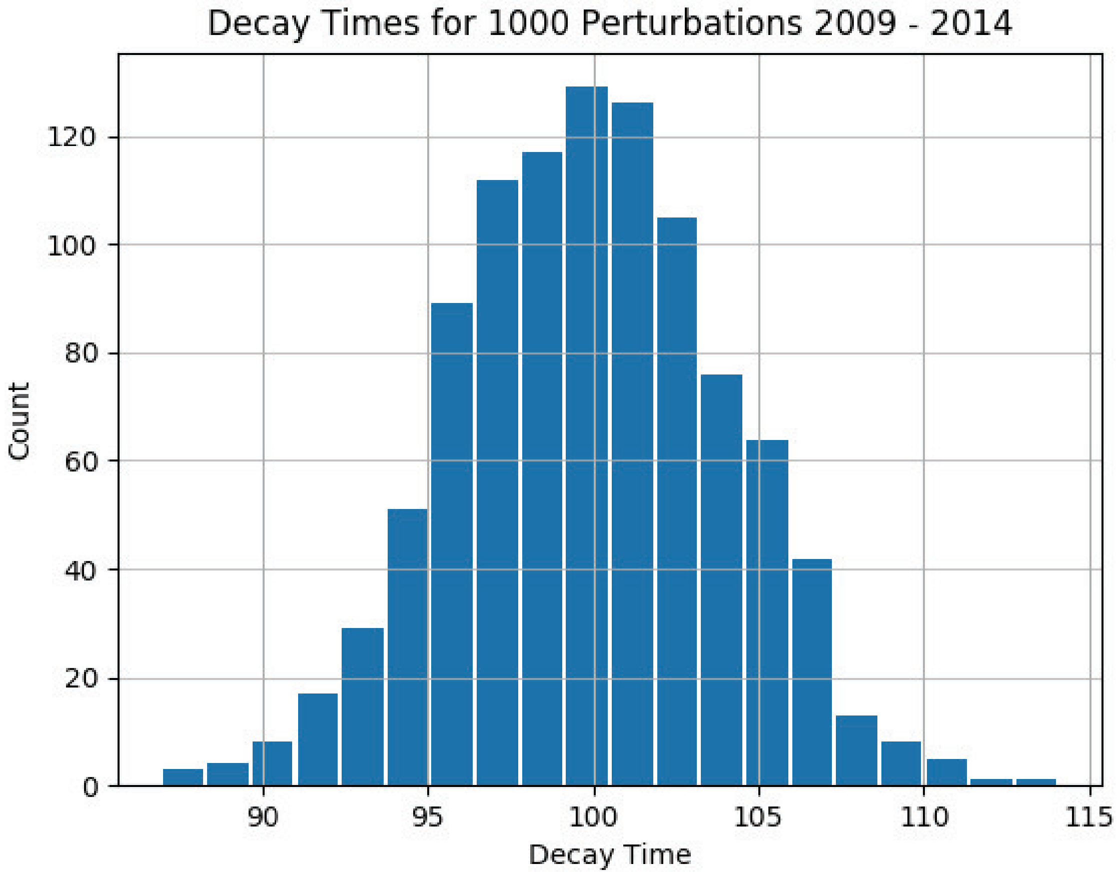

Figure A.1. a)Decay Times

B. CONVERGENCE PLOTS

Here we plot the lagging mean of RMSE across simulation runs for each of the 3 models.

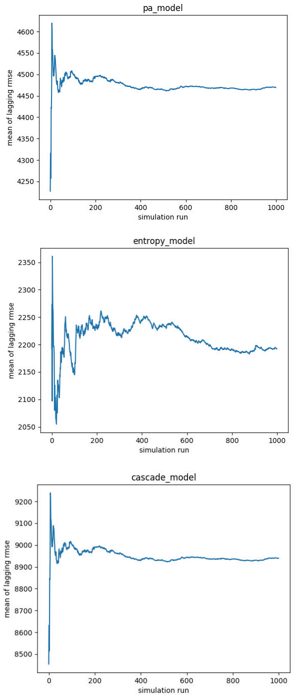

Figure B.1. Convergence plots for a) preferential attachment model, b) entropy-based model, c) cascade effect based model

C. PREDICTION EMERGENCE OF HOTSPOTS

Table C.1. Predicted vs. Actual change in Daily Active Users (a)

date	actual_change	predicted_change
2015-08-18	0.388288	0.350742
2014-08-26	0.831094	0.550722
2018-12-04	0.753645	0.695370
2015-02-17	0.180983	0.211810
2016-11-22	0.207015	0.193559
2017-08-21	0.326308	0.326640
2017-07-26	0.362151	0.313675
2015-09-01	0.210019	0.277959
2015-09-15	0.443388	0.270366
2016-11-08	0.316609	0.285963
2015-08-25	0.158862	0.153144
2013-11-12	0.194014	0.146860
2018-09-25	0.310303	0.306083
2017-09-12	0.300710	0.282940
2014-09-09	0.176157	0.187725
2018-09-14	0.186543	0.137408
2016-10-25	0.322832	0.291494
2018-07-13	0.189420	0.155817
2016-07-26	0.184936	0.163917
2015-06-12	0.181120	0.161066
2016-06-17	0.289179	0.332455
2018-12-04	0.195213	0.254286
2017-05-30	0.277095	0.254341
2015-02-17	0.852333	0.781008

In Tables (C.1,C.2), we have displayed the predictive performance of the entropy based model. The 1^{st} column marks the date of the perturbation. The 2^{nd} column represents the actual change in DAU due to the perturbation and the 3^{rd} column is the predicted change in DAU.

Table C.2. Predicted vs. Actual change in Daily Active Users (b)

date	actual_change	predicted_change
2016-06-28	0.320954	0.216146
2017-05-02	0.381113	0.336923
2016-07-19	0.185952	0.121904
2014-11-18	0.190014	0.183388
2017-08-01	0.162102	0.135741
2016-09-13	0.166855	0.133941
2018-01-23	2.019831	2.255991
2014-12-16	0.211077	0.313293
2017-03-27	0.198320	0.278170
2015-07-21	0.166824	0.139513
2015-11-30	0.184524	0.166459
2017-03-30	1.194280	1.383813
2017-09-26	0.340274	0.338192
2018-05-15	0.652558	0.744999
2016-08-02	0.264101	0.327900
2017-11-14	0.174526	0.227170
2015-10-13	0.270061	0.351184
2016-10-13	0.299536	0.342183
2015-09-15	0.211452	0.300023
2016-12-01	0.179757	0.193537
2016-08-23	0.213049	0.291780
2017-08-29	1.784665	1.875254
2018-05-29	0.265843	0.375416
2015-08-21	0.653490	0.674729
2014-06-24	0.546812	0.452296
2016-12-06	0.297427	0.325384
2016-04-21	0.485130	0.529604
2017-04-21	0.269286	0.243246
2018-10-19	0.193738	0.221009
2016-02-09	0.182510	0.206909
2016-01-19	0.307617	0.299922
2018-12-13	0.218207	0.252423
2017-06-13	0.206621	0.130548
2016-04-19	0.647346	0.741916
2017-08-15	0.656456	0.772220

D. BOLTZMANN'S H-THEOREM

We should take a moment to appreciate the fact that with this theorem and its ensuing derivation, Boltzmann kick-started the field of statistical mechanics and made it a useful tool in physics.

We closely follow the derivation in [58] to explain in relatively simplistic terms, the arguments weaved by Boltzmann to arrive at the H-theorem. Consider f as a distribution function of the number of particles n_{i} in the spatial region $\delta x_{\mathrm{i}}\delta y_{\mathrm{i}}\delta z_{\mathrm{i}}$ and having momentum $\delta p_{x_{\mathrm{i}}}\delta p_{y_{\mathrm{i}}}\delta p_{z_{\mathrm{i}}}$. We relate the number of particles to the density function as follows:

$$n_{\mathrm{i}} = f(.)\delta x_{\mathrm{i}}\delta y_{\mathrm{i}}\delta z_{\mathrm{i}}\delta p_{x_{\mathrm{i}}}\delta p_{y_{\mathrm{i}}}\delta p_{z_{\mathrm{i}}} \tag{D.1}$$

The term $\delta x_{\mathrm{i}}\delta y_{\mathrm{i}}\delta z_{\mathrm{i}}\delta p_{x_{\mathrm{i}}}\delta p_{y_{\mathrm{i}}}\delta p_{z_{\mathrm{i}}}$ can be thought of at the volume of a cell in the 6 dimensional phase space of the 3 position coordinates and 3 momentum coordinates. We can denote this 'volume' of a unit in the 6-dimensional space as dv_{i}. Utilizing this representation, Boltzmann proceeded to make the following definition of a quantity H:

$$H = \sum_{\mathrm{i}} f_{\mathrm{i}} \ln(f_{\mathrm{i}}) dv_{\mathrm{i}}$$

It may be trivially shown that we can rewrite this function H as

$$H = N \sum_{\mathrm{i}} \frac{n_{\mathrm{i}}}{N} \ln\left(\frac{n_{\mathrm{i}}}{N}\right) + constant$$

If we can take $\frac{n_{\mathrm{i}}}{N}$ as the probability P_{i} of a particle being found in the i^{th} cell of the phase space, we can rewrite H as $\sum_{\mathrm{i}} p_{\mathrm{i}} \ln(p_{\mathrm{i}})$ + constant. From the definition of mechanical entropy (S) [45], we can see that H may be related to S as $H = -\frac{S}{k}+$ constant, but only as the system approaches equilibrium since the definition of mechanical entropy assumes the relation $p_{\mathrm{i}} = \frac{n\mathrm{i}}{N}$ when system has reached a state of equilibrium. Consequently, in regions close to equilibrium, the time derivative of S can extracted from the time derivative of H which inturn we have shown can be obtained from the derivative of the particle distribution function $f(.)$. We now proceed to show that $\frac{dH}{dt} \leq 0$. Since we are summing over the differential volume in the definition of H, we can regard f as being continuous in the phase space so that we may redefine H as $\int ... \int f \ln f dv$. Since we are integrating over the six coordinates of the phase space (3 position and 3 momentum coordinates), $\frac{dH}{dt}$,

$$\frac{dH}{dt} = \int ... \int (\frac{df}{dt} \ln f + \frac{df}{dt}) dv$$

Using the Leibniz rule, we see that the second term in the above integro differential equation is 0. Further, making the assumption that f is independent of position (in order to make the derivation easier), we can write

$$\frac{dH}{dt} = V \int \int \int \frac{df}{dt} \ln f d\omega$$

where, $d\omega$ is the momenta differential $dp_x dp_y dp_z$. Using the assumption that f() is independent of the position, we can write

$$N_{\mathrm{i}} = V f_{\mathrm{i}} \delta\omega_{\mathrm{i}} \tag{D.2}$$

i.e the number of particles with the momentum in the range $d\omega_{\mathrm{i}}$, with volume of container being V. Now lets observe the collision between two particles with initial momenta in ranges $d\omega_{\mathrm{i}}, d\omega_{\mathrm{j}}$ such that the collide to produce particles with momenta ranges $d\omega_l, d\omega_m$. By using the assumption of molecular chaos, Boltzmann argues that the collision will proportionally increase N_l, N_m and decrease $N_{\mathrm{i}}, N_{\mathrm{j}}$ such that the frequency is $C f_{\mathrm{i}} f_{\mathrm{j}}$. Here C is a collision constant. We can determine the rate of change of the number of particles in the ranges (momentum) as

$$C_{\mathrm{ij}} f_{\mathrm{i}} f_{\mathrm{j}} = \frac{-dN_{\mathrm{i}}}{dt} = \frac{dN_m}{dt} = ...$$

Using this in Equation D.2, we can see that

$$\frac{dN_l}{dt} = V\frac{df_l}{dt}\delta\omega_l$$

This similarly hold for particles in the other involved momenta differentials. Notice that we can use the above relation in the summation expression for $\frac{dH}{dt}$ so that the contribution from the collision of two particles in momenta differentials $d\omega_{\mathrm{i}}, d\omega_{\mathrm{j}}$ to the summation is

$$-Cf_{\mathrm{i}}f_{\mathrm{j}}\ln f_{\mathrm{i}} - Cf_{\mathrm{i}}f_{\mathrm{j}}\ln f_{\mathrm{j}} + Cf_{\mathrm{i}}f_{\mathrm{j}}\ln f_l + Cf_{\mathrm{i}}f_{\mathrm{j}}\ln f_m$$

Now considering the reverse collision process i.e. the collision of particles in momenta differentials $d\omega_l, d\omega_m$ to produce particles in momenta differentials $d\omega_{\mathrm{i}}, d\omega_{\mathrm{j}}$. We can write a similar expression as before for the frequency of collision so that

$$C_{ml}f_{\mathrm{i}}f_{\mathrm{j}} = \frac{dN_{\mathrm{i}}}{dt} = \frac{-dN_m}{dt} = \dots$$

By the application of Liouville's theorem, Boltzmann shows that the collision constant in the forward collision C_{ij} is the same as C_{ml}. Hence for this pair of forward and reverse collisions, the constituent terms in the summation representation of $\frac{dH}{dt}$ may be written as

$$C(f_{\mathrm{i}}f_{\mathrm{j}} - f_l f_m)\ln\left(\frac{f_m f_l}{f_{\mathrm{i}}f_{\mathrm{j}}}\right)$$

This expression is always negative or zero for positive values of $f_{\mathrm{i}}, f_{\mathrm{j}}, f_l, f_m$. Given that for spherical particles binary collisions occur in forward and reverse pairs, we can state that $\frac{dH}{dt} \leq 0$ or $\frac{dS}{dt} \geq 0$. Since by definition H is bounded from below, $\frac{dH}{dt}$ has tend to a limit where $\frac{dH}{dt} = 0$.It can be shown that in the case that $\frac{dH}{dt} = 0$, the distribution functions $f()$ take the form of a Maxwell Boltzmann distribution which inturn corresponds to a state of maximum entropy [58].

E. COVID-19 SIMULATION PARAMETERS

Range of infection thresholds and corresponding α values selected using grid search.

Infection Threshold	Age Group	α
5	0-30	0.000001
	30-60	0.00002
	60-90	0.00008
10	0-30	0.000005
	30-60	0.00006
	60-90	0.00009
20	0-30	0.000007
	30-60	0.00009
	60-90	0.0001

$1-p$: for the Bernoulli random variable representing social distancing. Here p is the probability of a successful interaction. The social distancing measure is stratified into 3 levels: Low, Medium, High. We define the Low, Medium and High levels as the following:

Sl.No	Low	Medium	High
1	0.2	0.4	0.7
2	0.2	0.5	0.8
3	0.3	0.6	0.9

- Low: mild lockdown where advisories are sent regarding restriction of movement, workplaces are made online and social gatherings are avoided.
- Medium: moderate lockdown where workplaces are made online, movement of people is moderately restricted and social gatherings are avoided.
- High: severe lockdown where movement of people is completely restricted eg. Wuhan.

Results of t-tests for all variable combinations

Infection Threshold	Low	Medium	High	**t-score (p value)**
5	0.2	0.4	0.7	$-2.3(p > 0.01)$
10	0.2	0.4	0.7	$-2.2(p > 0.01)$
20	0.2	0.4	0.7	$-2.4(p > 0.01)$
5	0.2	0.5	0.8	$-2.1(p > 0.01)$
10	0.2	0.5	0.8	$-2.2(p > 0.01)$
20	0.2	0.5	0.8	$-2.25(p > 0.01)$
5	0.3	0.6	0.9	$-2.21(p > 0.01)$
10	0.3	0.6	0.9	$-2.21(p > 0.01)$
20	0.3	0.6	0.9	$-2.22(p > 0.01)$

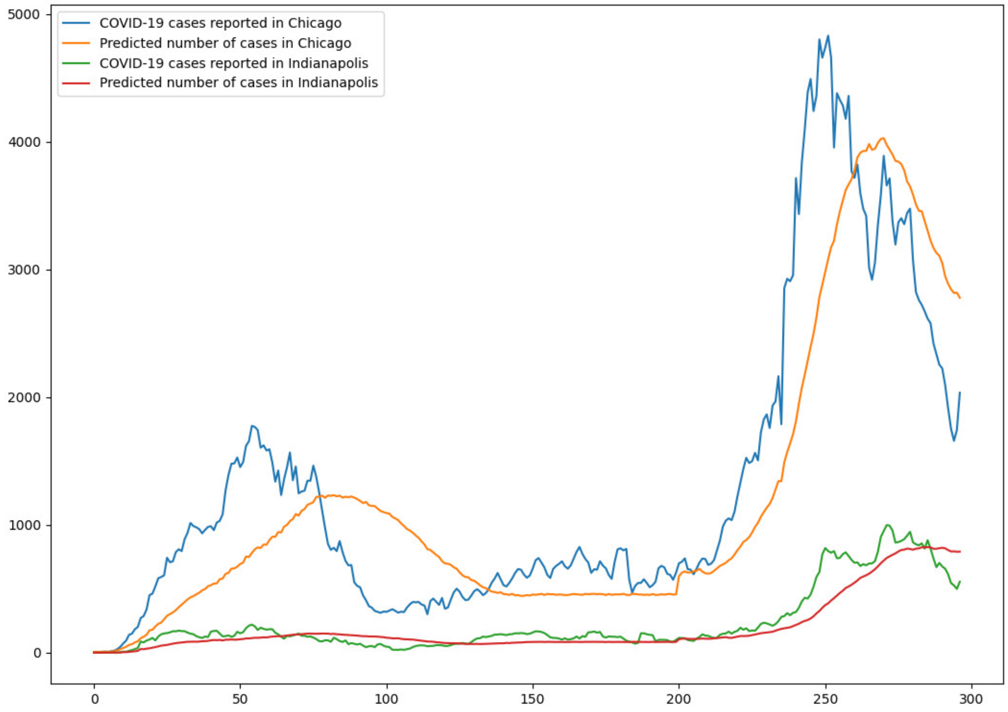

Figure E.1. Actual vs Predicted count of daily infections in Chicago and Indianapolis, from March 10, 2020 to December 31, 2020. (0^{th} index on the x-axis represents March 10, 2020.)

VITA

Nikhil was born in India and has had most of his formative years in Saudi Arabia. He has completed a Bachelor's Degree from Amrita Vishwa Vidyapeetham followed by his doctoral studies under Dr. Nagabhushana Prabhu.