M.S. THESIS

Asymptote of Entropy Production Distribution in the Shortest Path Process on Networks

그물 얼개 위의 최단 경로 과정의 무질서도 발생량 분포의 점근 분포 분석

February 2019

서울대학교 대학원
물리·천문학부
정 재 우

M.S. THESIS

Asymptote of Entropy Production Distribution in the Shortest Path Process on Networks

그물 얼개 위의 최단 경로 과정의
무질서도 발생량 분포의 점근 분포 분석

February 2019

서울대학교 대학원
물리·천문학부
정 재 우

Asymptote of Entropy Production Distribution in the Shortest Path Process on Networks

그물 얼개 위의 최단 경로 과정의 무질서도 발생량 분포의 점근 분포 분석

지도교수 강 병 남

이 논문을 이학석사 학위논문으로 제출함

2019년 2월

서울대학교 대학원

물리·천문학부

정 재 우

정재우의 석사 학위논문을 인준함

2019년 2월

위 원 장: ____________ (인)
부위원장: ____________ (인)
위 원: ____________ (인)

Abstract

Asymptote of Entropy Production Distribution in the Shortest Path Process on Networks

Jaewoo Jung
Department of Physics and Astronomy
The Graduate School
Seoul National University

Entropy production (EP) is a key quantity to measure the irreversibility in both thermal equilibrium states and non-equilibrium phenomena. Assuming complex networks as state spaces, as conformation networks in protein folding problem, one considers data-packet-transport problem on networks. In this case, one can define EP produced during process following one of the shortest paths between two nodes. EP, which is affected by the complexity of the paths, is defined as the logarithmic ratio of the probabilities of the forward and the backward path. This EP satisfies both integral fluctuation theorem (IFT), and detailed fluctuation theorem (DFT). Using extreme value analysis, that the asymptote of cumulative distribution of EP of BA model network is the Gumbel distribution is also confirmed.

keywords: Shortest Path Dynamics, Entropy Production, Network, Extreme Value Analysis, Gumbel distribution
student number: 2016-20319

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Entropy

Generally and historically entropy is the measure of irreversibility. In its early stage, it was defined to understand thermo-equilibrium states. In 1850s and 1860s, R. Clausius firstly developed the concept and stated the second law of thermodynamics explicitly: entropy can not decrease [1].

After ingenious studies of Ludwig Boltzmann (1844-1906), entropy has taken an important role in making the bridge between thermodynamics and statistical mechanics [2]. Gibbs expanded Boltzmann's entropy concept by adopting the concept of ensemble. In later 1940s Shannon expanded thermodynamic entropy into the information realm [3].

Entropy was, at first, one of the quantities which describe systems in equilibrium states, however, in late 20th century, since studies on nonequilibrium phenomena became highlighted, physicists have been trying to extend the definition of entropy from representing equilibrium states to describing nonequilibrium phenomena [4–7].

In the stochastic thermodynamics, entropy production (EP) can be defined in the following way. First, let us consider a specific process from initial and final states of a nonequilibrium process. Given the probability of the initial state, and a transition probability from the initial state to the final state following a specific process, then multiplication of two probability is called a forward probability of the process. Similarly, multiplication of probability of the final state and a transition probability from the final state to the initial state following the process in reverse direction is called a backward probability of the process. Now one can define the EP as the logarithmic ratio of the forward probability to the backward probability. This quantity shows how much flow of the process is irreversible, and is consistently related to heat [4, 8].

By definition, the EP satisfies Integral Fluctuation Theorem (IFT), and in special cases, satisfies Detailed Fluctuation Theorem (DFT). IFT consequently indicates that the second law of thermodynamics is satisfied. Crooks [9], Jarzynski [10], etc. developed the theorem with dissipated work, free energy, etc. It is proved that FT is satisfied in Langevin system [11], and the different proof was also suggested [12]. Since FT was firstly suggested, ceaseless studies for more general FT have been done [13]. In more microscopic scope, EP produced from a single trajectory was studied, and FT was confirmed once again [14]. There is also a Hamiltonian derivation of DFT [15].

1.2 Network Theory

Network is system composed of nodes and links. After Erdős Pál (1913-1996) studied the graph theory and suggested the Erdős-Rényi model (ER) in 1950s, network science rebooted with enhanced computational power in the first decade of new millennium [16]. Physicists tried to design various models which mimic the characteristics of networks in the real world. First, Duncan J. Watts and Steven Strogatz suggested a small world network [17]. Serially, Barabási-Albert model (BA) [18], Huberman-Adimic model [19], the Scale-free Chung-Lu model (SFCL) [20], the Adaptation model [21], etc. were suggested. Nowadays many variations served as bases of researches not only in physics, but also in the other fields of natural science, social sciences, and engineering, from microscopic to macroscopic scales because networks are useful to study affects of system structures. Some kinds of emergence are observed when the system structure is considered.

Network theory also led random walk studies to the new region. Noh successfully embedded network to generic random walk studies, and figure out the probability with which a random walker at the node in steady state, first passage time, etc [22]. After this work, biased random walk on networks, random walk with long range transition on networks are studied. Random walks on networks are applied to search engines, studies on protein folding dynamics, etc.

1.3 Extreme Value Analysis

Extreme Value Theory, also known as Extreme Value Analysis (EVA), is a branch of statistics, which studies extreme values. Given ordered sample from given random variables, this field mainly focuses the probability of occurrence for events which are more extreme than those have already occurred. EVA is practically used in various applied fields, such as structural engineering, finance, earth science, traffic prediction, and geological engineering, etc. Actually, EVA is a special case of order statistics because it deals with the first or the last in the ordered sample.

1.4 the Goal of this Dissertation

In this dissertation, one studies the distribution of EP produced in shortest path processes. One tried to examine properties of the EP which is produced in a kind of biased random walk process along the shortest paths between every pair on the network.

One specially focuses on how one can estimate the limit distributions for EP distributions. One confirmed that the Gumbel distribution is the limit distribution for the all of the distributions which one got from simulation or data. It is generally hard to estimate an analytic functional form of a probability density function from real data, but one was able to estimate it with EVA. Though the entropy production had generally targeted the nonequilibrium processes, EP which is defined in this dissertation is mainly affected by topological characteristics of networks, not by the dynamical interactions. Nevertheless, there are some shortest path dynamics, such as data-packet-transport dynamics, so even though the EP in this paper has the topological origin, it can be helpful to analyze this kind of dynamics.

In chapter 2, one defines EP produced from the process on networks. One also shows that the EP is satisfied Integral Fluctuation Theorem and Detailed Fluctuation Theorem. Next, It is introduced that the core part of the Extreme Value Analysis to determine the functional form of the limiting distribution of the EP distribution. Mainly, the back bone theorem of the Extreme Value Analysis, Fisher-Tippett-Gnedenko theorem is introduced without proof. One also explains how to indirectly determine the functional form of the extreme value function using Fisher-Tippett-Gnedenko theorem in theoretical way.

In chapter 3, one shows the results, mainly the EP distribution on various network models, such as Barabási-Albert (BA) model network, Erdős-Rényi (ER) model network, Scale Free Chung-Lu (SFCL) model network, and coauthorship network. The EP distributions have bell shape with high peak at zero. One focuses on the tail behavior of the distributions, and shows the trial functional form.

In chapter 4, one shows how one can determine the functional form of asymptote of EP distribution. Fisher-Tippett-Gnedenko theorem is used to determine the functional form. There is an issue on determination of the power value which is appearing in determination of the parameters of the Extreme Value Distributions, that for given range of the EP, one has to choose the power values not to make the cumulative distribution functions twisted too much. It is confirmed that EP distribution for BA model network whose system size is 20480 and mean degree is 8 has the asymptote which is the Gumbel distribution.

In chapter 5, one derives the analytic form of EP distribution on 2D lattice. It suggest a clue which might be helpful to understand what kinds of pair or process produce large entropy productions. Because of the symmetry of the 2D lattice, only on the links posed the boundary of the lattice, the nonzero EP is produced.

In chapter 6, one concludes this dissertation.

Chapter 2

Background Theories

2.1 EP from Processes on Network

The more studies on interactions between agents on the networks, the more necessity of studies on noneqilibrium phenomena, because many interactions have to be treated in nonequilibrium state to understand them better. In this context, there were several studies about some EPs on networks: one is by Haye Hinrichsen [23]. This study defined an EP to analyze dynamic message transmission between the nearest neighbors on complex networks, and concluded this EP may help researchers analyze not only the global network structure but also hidden interactions among the nodes.

2.2 EP in Nonequilibrium Statistical Mechanics

Generally the EP is defined with a protocol which is a function of time. It measures the irreversibility of processes following this protocol. It is defined as a $\log$ of the ratio of the probability which is from the specific j-th process following the protocol, from an initial state α to a final state β, from $t = 0$ to $t = \tau$ (written in $\Pi^j_{\alpha\to\beta}$, called the forward probability), to the probability which is from the exactly reverse process, from $t = \tau$ to $t = 2\tau$ following the same protocol, from the new initial state β at $t = \tau$ to the new final state α at $t = 2\tau$ (written in $\tilde{\Pi}^j_{\alpha\to\beta}$, called the backward probability):

$$\Delta S^j_{\alpha\to\beta} = \log\left(\frac{\rho_{t=0}(\alpha)\Pi^j_{\alpha\to\beta}}{\rho_{t=\tau}(\beta)\tilde{\Pi}^j_{\alpha\to\beta}}\right). \tag{2.1}$$

One can split the logarithm into two terms. The former part is Shannon entropy, and the latter part is heat of Schnakenberg formula [24], [25]:

$$\Delta S^j_{\alpha\to\beta} = \log\left(\frac{\rho_{t=0}(\alpha)}{\rho_{t=\tau}(\beta)}\right) + \log\left(\frac{\Pi^j_{\alpha\to\beta}}{\tilde{\Pi}^j_{\alpha\to\beta}}\right). \tag{2.2}$$

IFT is automatically derived from equation 2.1 if ρ and Π are well-defined:

$$\begin{aligned}
\left\langle e^{-\Delta S^j_{\alpha\to\beta}}\right\rangle_{\alpha,\beta,j} &= \sum_{\alpha,\beta}\sum_j \rho_{t=0}(\alpha)\Pi^j_{\alpha\to\beta} e^{-\Delta S^j_{\alpha\to\beta}} \\
&= \sum_{\alpha,\beta}\sum_j \rho_{t=0}(\alpha)\Pi^j_{\alpha\to\beta} \frac{\rho_{t=\tau}(\beta)\tilde{\Pi}^j_{\alpha\to\beta}}{\rho_{t=0}(\alpha)\Pi^j_{\alpha\to\beta}} \\
&= \sum_{\alpha,\beta}\sum_j \rho_{t=\tau}(\beta)\tilde{\Pi}^j_{\alpha\to\beta} = 1.
\end{aligned} \tag{2.3}$$

However, DFT needs more symmetry. If a system is in the steady state, then the EP of the system satisfies DFT [7]. In this paper, DFT is satisfied because of a different reason. The EP in this paper satisfies DFT because it has the following symmetry:

$$\tilde{\Pi}^j_{\alpha\to\beta} = \Pi^j_{\beta\to\alpha}. \tag{2.4}$$

In the following sections, one shows derivations of DFT for given node pair case and whole network case using this symmetry in detail.

2.3 Given Pair Case: a Simple Example

The shortest path between two given nodes are the path whose number of the links to travel from the one to the another are minimum. To construct the forward (backward) probabilities, one multiplies all of the reciprocals of the number of the links from the nodes towards the target (the source), which are on the shortest paths. As one did this process with the computation, the mutually exclusive and collectively exhaustive

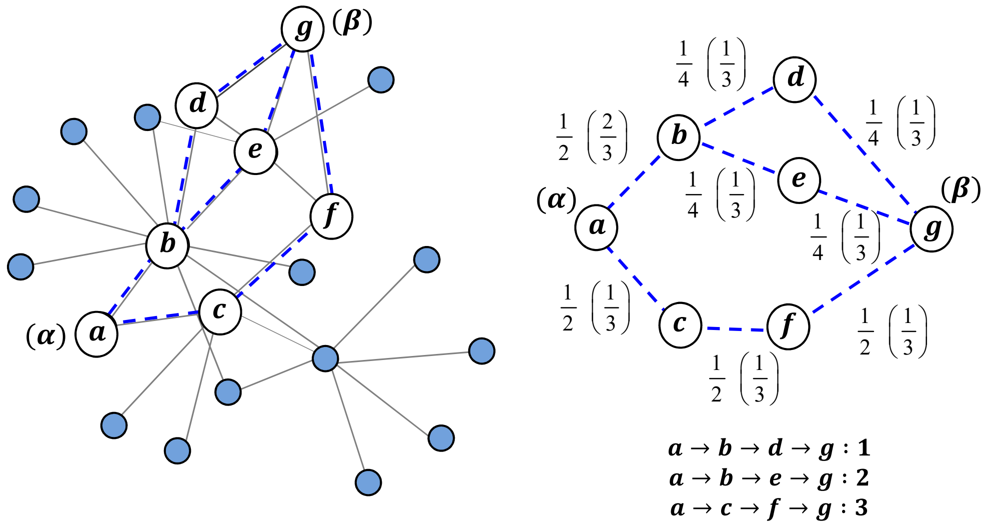

Figure 2.1: (Left) One can pick two nodes, say α and β, on the network. (Right) One can find the nodes which consist of the shortest paths between α and β. There are three shortest paths between α and β in this case. Numbers with or without parentheses mean the probabilities for each link (see Figure 2.2).

way one use to search the shortest paths between all of the possible pairs of nodes on the network is an important part. There is well-known Dijkstra algorithm for this kind of search [26]. The algorithm is similar to breadth-first search (BFS). Using the algorithm, one can find the whole processes between the given two nodes, one node as the source and the other as the target, going along the shortest paths between them.

Let us consider an example to explain how the forward and the backward probabilities are assigned. Figure 2.2 shows the subgraph of the network in the right panel of Figure 2.1, consisting of given two nodes and links on the shortest paths between them. Two given nodes α and β take roles as the source and the target, respectively. When one considers the first shortest path, one can figure out that the forward probability does not have to be the same to the backward probability, so there are nonzero EPs.

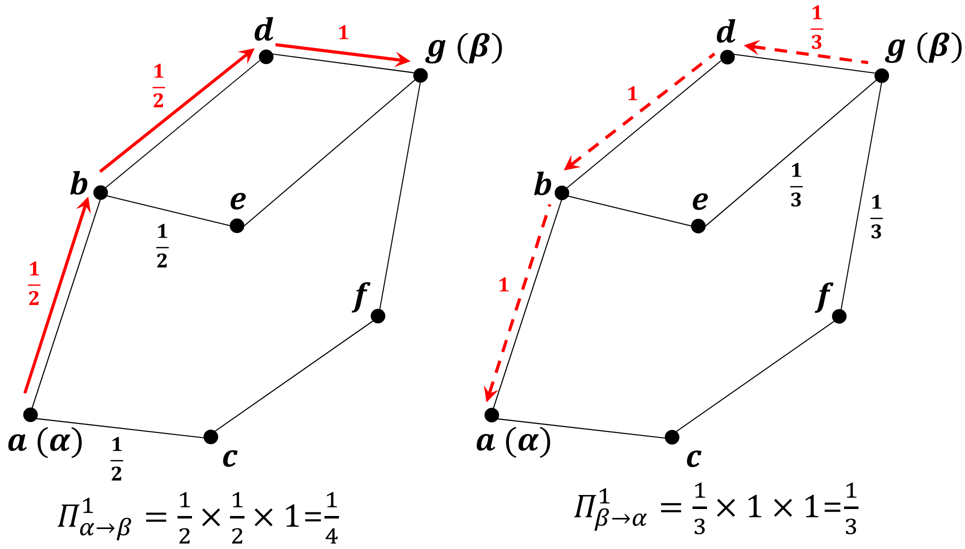

Figure 2.2: The EP of a given pair. One starts from the source α (node a). There are two different options to go along the shortest paths heading to the target, so each link $a \to b$ and $a \to c$ has probability $\frac{1}{2}$. If one chooses to go to the node b, then there are 2 different options to go along the shortest paths heading to the target, so each link $b \to d$ and $b \to e$ has probability $\frac{1}{2}$. Similarly, one can choose node d and β (node g), with probability $\frac{1}{2}$ and 1, respectively. The forward probability of the shortest path $a(\alpha) \to b \to d \to g(\beta)$ is, therefore, $\frac{1}{2} \times \frac{1}{2} \times 1 = \frac{1}{4}$. In the same way, one can find the backward probability of the shortest path is $\frac{1}{3} \times 1 \times 1 = \frac{1}{3}$.

The sequence of nodes of the first shortest path is $a(\alpha) \to b \to d \to g(\beta)$. One has two options to follow the shortest paths heading to the target at node a (to node b and to node c). Therefore, each link $a \to b$ and $a \to c$ has probability $\frac{1}{2}$. If one chooses to go to the node b, then there are two options to follow the shortest paths heading to the target, to node d and to node e. Therefore, each link $b \to d$ and $b \to e$ has probability $\frac{1}{2}$. Lastly, if we are at node d, then there is the only one option to go along the shortest paths heading to the target, to node $g(\beta)$, so the link $d \to g$ has probability 1. Now, the forward probability of the shortest path $a(\alpha) \to b \to d \to g(\beta)$ is $\frac{1}{2} \times \frac{1}{2} \times 1 = \frac{1}{4}$.

pathway (j)	forward pathway	$\Pi^j_{\alpha\to\beta}$	backward pathway	$\Pi^j_{\beta\to\alpha}$
1	$a(\alpha) \to b \to d \to g(\beta)$	$\frac{1}{4}$	$g(\beta) \to d \to b \to a(\alpha)$	$\frac{1}{3}$
2	$a(\alpha) \to b \to e \to g(\beta)$	$\frac{1}{4}$	$g(\beta) \to e \to b \to a(\alpha)$	$\frac{1}{3}$
3	$a(\alpha) \to c \to f \to g(\beta)$	$\frac{1}{2}$	$g(\beta) \to f \to c \to a(\alpha)$	$\frac{1}{3}$

Table 2.1: Probabilities comparison: forward and backward pathways

One can similarly calculate the backward probability of the path. Table 2.1 shows all of the shortest paths and their forward and backward probabilities.

Now, one can use these probabilities to get the EP explicitly. First, starting from the equation 2.1, plug 1 for every ρ because α and β are given. Since there is no time dependence of all of the processes going from the source to the target, one removes time dependence in each Π. It is also very easy to show the following relation is satisfied:

$$\tilde{\Pi}^j_{\alpha\to\beta} = \Pi^j_{\beta\to\alpha}. \tag{2.5}$$

Finally, the EP of the j-th path from α to β becomes:

$$\Delta S^j_{\alpha\to\beta} = \log\left(\frac{\Pi^j_{\alpha\to\beta}}{\tilde{\Pi}^j_{\alpha\to\beta}}\right) = \log\left(\frac{\Pi^j_{\alpha\to\beta}}{\Pi^j_{\beta\to\alpha}}\right) = -\Delta S^j_{\beta\to\alpha}. \tag{2.6}$$

On the other hand, these processes, from α to β following the shortest paths, may be similar to data-packet-transport through the router network. When one sends a file through the Internet, the file is divided into small fractions. Each of them are called data packet after they are attached information about the start and the target router. Data packets travel from the initial router to the final router along the most efficient ways. There could be several equivalently efficient ways, so each data packets may travel through different ways.

2.4 Fluctuation Theorems on a Given Pair

By definition, the EP explicitly satisfies two fluctuation theorems, Integral Fluctuation Theorem (IFT) and Detailed Fluctuation Theorem (DFT). One may confirm IFT as follows:

$$\begin{aligned}
\left\langle e^{-\Delta S^j_{\alpha\to\beta}} \right\rangle &= \sum_j \Pi^j_{\alpha\to\beta} e^{-\Delta S^j_{\alpha\to\beta}} \\
&= \sum_j \Pi^j_{\alpha\to\beta} \frac{\Pi^j_{\beta\to\alpha}}{\Pi^j_{\alpha\to\beta}} \\
&= \sum_j \Pi^j_{\beta\to\alpha} = 1.
\end{aligned} \tag{2.7}$$

Let us consider all of the pathways between the given two nodes α and β to confirm DFT. One can label them and choose the j-th path. Then the EPs of the cases which are starting from α to β and starting from β to α going along the j-th path have opposite signs (see Equation 2.6). Now one considers the probability density function for the EP (denoted by ΔS) of the forward and backward paths, $P_f(\Delta S)$ and $P_b(\Delta S)$ respectively. One can start with the mathematical definition of the $P_f(\Delta S)$:

$$\begin{aligned}
P_f(\Delta S) &= \sum_j \{\delta(\Delta S - \Delta S^j_{\alpha\to\beta})\Pi^j_{\alpha\to\beta}\} \\
&= \sum_j \{\delta(\Delta S - \Delta S^j_{\alpha\to\beta})\Pi^j_{\beta\to\alpha} e^{\Delta S^j_{\alpha\to\beta}}\} \\
&= \sum_j \{\delta(\Delta S + \Delta S^j_{\beta\to\alpha})\Pi^j_{\beta\to\alpha} e^{\Delta S}\} \\
&= P_b(-\Delta S) e^{\Delta S}
\end{aligned} \tag{2.8}$$

$$\therefore \frac{P_f(\Delta S)}{P_b(-\Delta S)} = e^{\Delta S}.$$

Thus, EP in the dissertation satisfies DFT.

2.5 Fluctuation Theorems on the Whole Network

One can extend this logic to the whole network. IFT is automatically satisfied if DFT is satisfied, so here it is shown only that the EP satisfies DFT.

EP for the j-th path starting from α to β and starting from β to α satisfied following relation as in the above case.

$$\Delta S^{j}_{\alpha\to\beta} = \log\left(\frac{\rho(\alpha)\rho(\beta|\alpha)\Pi^{j}_{\alpha\to\beta}}{\rho(\beta)\rho(\alpha|\beta)\Pi^{j}_{\beta\to\alpha}}\right) = -\Delta S^{j}_{\beta\to\alpha} \tag{2.9}$$

Where $\rho(n)$ is an arbitrary weight function to choose a node n among nodes of the given network. One can define

$$q(\alpha,\beta) = \frac{\rho(\alpha)\rho(\beta|\alpha)}{\sum_{\gamma,\delta}\rho(\gamma)\rho(\delta|\gamma)}, \tag{2.10}$$

and q is the normalized probability. Then one may prove DFT using the definition of the probability density function for forward paths:

$$\begin{aligned}
P_f(\Delta S) &= \sum_{\alpha,\beta}\sum_{j}\{\delta(\Delta S - \Delta S^{j}_{\alpha\to\beta}) \times q(\alpha,\beta)\Pi^{j}_{\alpha\to\beta}\} \\
&= \sum_{\alpha,\beta}\sum_{j}\{\delta(\Delta S - \Delta S^{j}_{\alpha\to\beta}) \times q(\beta,\alpha)\Pi^{j}_{\beta\to\alpha}e^{\Delta S^{j}_{\alpha\to\beta}}\} \\
&= \sum_{\alpha,\beta}\sum_{j}\delta(\Delta S + \Delta S^{j}_{\beta\to\alpha}) \times q(\beta,\alpha)\Pi^{j}_{\beta\to\alpha}e^{\Delta S} \qquad (2.11) \\
&= P_b(-\Delta S)e^{\Delta S}
\end{aligned}$$

$$\therefore \frac{P_f(\Delta S)}{P_b(-\Delta S)} = e^{\Delta S}.$$

Because IFT is derived from DFT, the EP is satisfied IFT and the second law of thermodynamics. Moreover, P_f and P_b are same because one treats the whole network and it means that every path takes all of possible roles as forward and backward paths.

The above proof is for an arbitrary normalized probability distribution $q(\alpha, \beta)$. Taking this advantage, there are two straightforward trial for $q(\alpha, \beta)$. First, one could assign the uniform weight to the nodes. In this case, the EP of the j-th shortest path from the source α to the target β becomes

$$\begin{aligned} \Delta S^j_{\alpha\to\beta} &= \log\left(\frac{\frac{1}{N}\frac{1}{N-1}\Pi^j_{\alpha\to\beta}}{\frac{1}{N}\frac{1}{N-1}\Pi^j_{\beta\to\alpha}}\right) \\ &= \log\left(\frac{\Pi^j_{\alpha\to\beta}}{\Pi^j_{\beta\to\alpha}}\right). \end{aligned} \tag{2.12}$$

Second, one could assign the nodes the weight which linearly depends on the degrees of each node. The EP of the j-th shortest path from the source α to the target β now becomes

$$\begin{aligned} \Delta S^j_{\alpha\to\beta,dd} &= \log\left(\frac{\frac{k_\alpha}{2L}\frac{k_\beta}{2L-k_\alpha}\Pi^j_{\alpha\to\beta}}{\frac{k_\beta}{2L}\frac{k_\alpha}{2L-k_\beta}\Pi^j_{\beta\to\alpha}}\right) \\ &= \log\left(\frac{\Pi^j_{\alpha\to\beta}}{\Pi^j_{\beta\to\alpha}}\right) + \log\left(\frac{2L-k_\beta}{2L-k_\alpha}\right). \end{aligned} \tag{2.13}$$

These two examples above satisfied DFT explicitly.

It is also easy to confirm that DFT is satisfied by computation. Figure 2.3 shows the result. On the horizontal axis, there are the ratios of the probability distribution function values of the ΔS and $-\Delta S$, and on the vertical axis, there are corresponding exponential values of ΔS. The vertical axis is $\log$ scale. The red square points are data, and the blue dashed line is $y = x$ line. The data points are perfectly on the line. One confirmed DFT by not only visually with the figure but also value by value, despite not presenting the latter here.

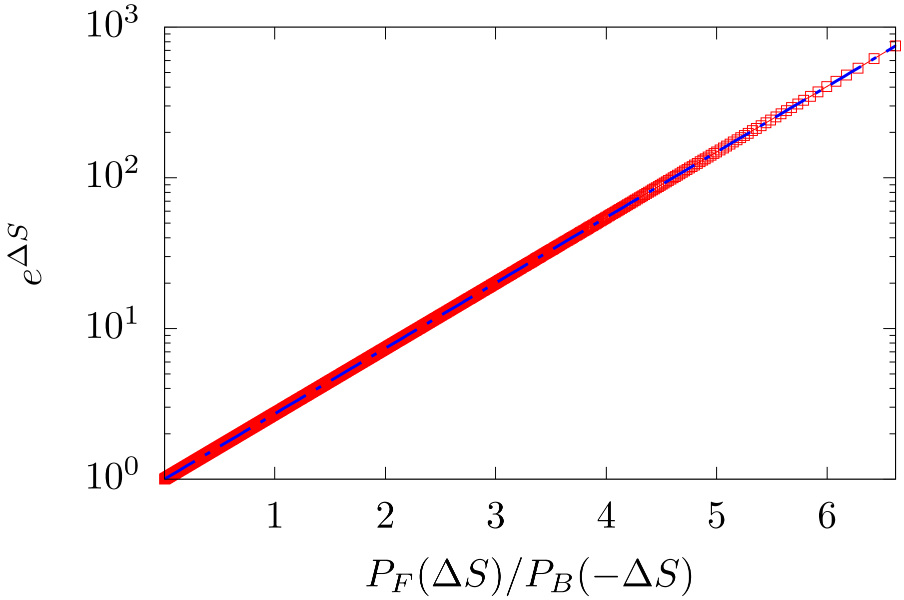

Figure 2.3: Numerical confirmation of Detailed Fluctuation Theorem. Perpendicular axis is $\log$ scale. The red square points are data, and the blue dashed line is $y = x$ line. The data points are perfectly on the line.

2.6 Extreme Value Analysis

In these days, EVA has been applied to analyze extreme events, as its name tells. For instance, EVA is used to estimate the probabilities of occurrence of centennially severe flood, extreme tornado, mountain fire, or probability of mutation, maximum biomass, abnormal traffic occurrence in data transportation, etc.

The well-known theorem, Fisher-Tippett-Gnedenko theorem says that there are only three kinds of limiting distributions:

Theorem 1. *(Fisher-Tippett-Gnedenko) Let $X_1, ..., X_n$ be random variables following an independent, identical distribution, and M_n be the maximum of them. If there are ordered pairs of real sequences, (a_n, b_n), such that $a_n > 0$ and*

$$\lim_{n\to\infty} Pr\left(\frac{M_n - b_n}{a_n} \leq x\right) = F(x),$$

then $F(x)$ must be one of Gumbel, Fréchet, and the Inverse Weibull distribution.

If one wants to see the proof of the theorem, then see [27]. Here are cumulative distribution functions of them respectively:

$$F(x; \mu, \beta) = e^{-e^{-\frac{x-\mu}{\beta}}} \tag{Gumbel}$$

$$F(x; \mu, \alpha > 0) = e^{-x^{-\alpha}} \quad (x > 0) \tag{Fréchet}$$

$$F(x; \mu, \alpha > 0) = \begin{cases} e^{-(-x)^{\alpha}} & (x < 0) \\ 1 & (x \geq 0) \end{cases} \tag{Inverse Weibull}$$

This theorem describes the general result for asymptote of extreme value distribution in order statistics. This theorem does not help one discern if the renormalized distribution has an asymptote, but only useful to determine the asymptote if it exists. In this dissertation, one uses this fact to figure out the functional form of the asymptote of EP.

2.7 Generalized Extreme Value Distribution

Generalized Extreme Value (GEV) distribution is the generalized version for three extrmema distributions, namely, the Gumbel, the Fréchet, and the Inverse Weibull distribution. GEV distribution has three parameters which determines the distribution. These are location (μ), scale ($\beta > 0$), shape ($-\infty < \xi < \infty$). General form is

$$F(x; \mu, \beta) = e^{-(1+\xi\frac{x-\mu}{\beta})^{-\frac{1}{\xi}}} \tag{general form}$$

One should think $\xi = 0$ case is that ξ goes to zero limit. If ξ is zero, then GEV distribution turns out the Gumbel distribution. If ξ is positive or negative, then GEV distribution turns out the Fréchet distribution or the Inverse Weibull distribution, respectively.

2.7.1 the Gumbel Distribution

If the shape parameter ξ is zero, then the GEV distribution becomes the Gumbel distribution. The general functional form of cumulative distribution function of the Gumbel distribution is

$$F(x;\mu,\beta) = e^{-e^{-\frac{x-\mu}{\beta}}}. \qquad \text{(general form)}$$

The basic statistics are below (where γ is the Euler-Mascheroni constant):

$$\mu \qquad \text{(mode)}$$

$$\mu - \beta\log(\log 2) \qquad \text{(median)}$$

$$E(X) = \mu + \beta\gamma \qquad \text{(mean)}$$

$$\sigma^2(X) = \frac{(\beta\pi)^2}{6}. \qquad \text{(variance)}$$

The standard form of cumulative distribution function is when $\mu = 0$ and $\beta = 1$:

$$F(x;\mu=0,\beta=1) = e^{-e^{-x}}. \qquad \text{(standard form)}$$

From this standard form, one can easily find out the probability density function of the Gumbel distribution:

$$P(x) = e^{-x-e^{-x}}. \qquad \text{(p.d.f.)}$$

The basic statistics turn out

$$\mu = 0 \qquad \text{(mode)}$$

$$-\log(\log 2) \approx 0.37 \qquad \text{(median)}$$

$$E(X) = \gamma \qquad \text{(mean)}$$

$$\sigma^2(X) = \frac{\pi^2}{6} \approx 1.64. \qquad \text{(variance)}$$

Linearization of the cumulative distribution function F is

$$-\log[-\log F(x)] = \frac{x-\mu}{\beta}. \qquad \text{(linearization)}$$

2.7.2 the Fréchet Distribution

If the shape parameter ξ is positive, then the GEV distribution becomes the Fréchet distribution. The general functional form of cumulative distribution function of the Fréchet distribution (if $x > 0$) is

$$F(x > \mu; \mu, \alpha > 0, \beta > 0) = e^{-(\frac{x-\mu}{\beta})^{-\alpha}} \qquad \text{(general form)}$$

where $\alpha = \frac{1}{\xi}$.

The basic statistics are below:

$$\mu + \beta\left(\frac{\alpha}{1+\alpha}\right)^{\frac{1}{\alpha}} \qquad \text{(mode)}$$

$$\mu + \frac{\beta}{\sqrt[\alpha]{\log 2}} \qquad \text{(median)}$$

$$E(X) = \begin{cases} \mu + \beta\Gamma(1-\frac{1}{\alpha}) & (\alpha > 1) \\ \infty & (0 < \alpha < 1) \end{cases} \qquad \text{(mean)}$$

$$\sigma^2(X) = \begin{cases} \beta^2\left(\Gamma\left(1-\frac{2}{\alpha}\right) - \Gamma\left(1-\frac{1}{\alpha}\right)^2\right) & (\alpha > 2) \\ \infty & (0 < \alpha < 2). \end{cases} \qquad \text{(variance)}$$

The standard form of cumulative distribution function is when $\mu = 0$ and $\beta = 1$:

$$F(x > 0; \mu = 0, \beta = 1) = e^{-x^{-\alpha}}. \qquad \text{(standard form)}$$

From this standard form, one can easily find out the probability density function of the Gumbel distribution:

$$P(x) = \alpha x^{-1-\alpha} e^{-x^{-\alpha}}. \qquad \text{(p.d.f.)}$$

The basic statistics turn out

$$\left(\frac{\alpha}{1+\alpha}\right)^{\frac{1}{\alpha}} \qquad \text{(mode)}$$

$$\frac{1}{\sqrt[\alpha]{\log 2}} \qquad \text{(median)}$$

$$E(X) = \begin{cases} \Gamma\left(1-\frac{1}{\alpha}\right) & (\alpha > 1) \\ \infty & (0 < \alpha < 1) \end{cases} \qquad \text{(mean)}$$

$$\sigma^2(X) = \begin{cases} \left(\Gamma\left(1-\frac{2}{\alpha}\right) - \Gamma\left(1-\frac{1}{\alpha}\right)^2\right) & (\alpha > 2) \\ \infty & (0 < \alpha < 2). \end{cases} \qquad \text{(variance)}$$

2.7.3 the Inverse Weibull Distribution

If the shape parameter is negative, then the GEV distribution becomes the Inverse Weibull distribution. The general functional form of cumulative distribution function of the Inverse Weibull distribution (if $x < 0$) is

$$F(x < \mu; \mu \leq 0, \alpha > 0, \beta > 0) = e^{-(\frac{\beta}{-x+\mu})^{\alpha}} \qquad \text{(general form)}$$

where $\alpha = -\frac{1}{\xi}$.

The basic statistics are below:

$$-\mu - \beta \left(\frac{\alpha}{1+\alpha} \right)^{\frac{1}{\alpha}} \qquad \text{(mode)}$$

$$-\mu - \frac{(-\beta^{\alpha})^{-\frac{1}{\alpha}}}{\sqrt[\alpha]{\log 2}} \qquad \text{(median)}$$

$$E(X) = \begin{cases} -\mu - \beta\Gamma(1 - \frac{1}{\alpha}) & (\alpha > 1) \\ \infty & (0 < \alpha < 1) \end{cases} \qquad \text{(mean)}$$

$$\sigma^2(X) = \begin{cases} \beta^2 \left(\Gamma\left(1 - \frac{2}{\alpha}\right) - \Gamma\left(1 - \frac{1}{\alpha}\right)^2 \right) & (\alpha > 2) \\ \infty & (0 < \alpha < 2). \end{cases} \qquad \text{(variance)}$$

The standard form of cumulative distribution function is when $\mu = 0$ and $\beta = 1$:

$$F(x < 0; \mu = 0, \beta = 1) = e^{-(-x)^{-\alpha}}. \qquad \text{(standard form)}$$

From this standard form, one can easily find out the probability density function of the Gumbel distribution:

$$P(x) = \alpha(-x)^{-1-\alpha} e^{-(-x)^{-\alpha}}. \tag{p.d.f.}$$

The basic statistics turn out

$$-\left(\frac{\alpha}{1+\alpha}\right)^{\frac{1}{\alpha}} \tag{mode}$$

$$-\frac{1}{\sqrt[\alpha]{\log 2}} \tag{median}$$

$$E(X) = \begin{cases} -\Gamma\left(1-\frac{1}{\alpha}\right) & (\alpha > 1) \\ \infty & (0 < \alpha < 1) \end{cases} \tag{mean}$$

$$\sigma^2(X) = \begin{cases} \left(\Gamma\left(1-\frac{2}{\alpha}\right) - \Gamma\left(1-\frac{1}{\alpha}\right)^2\right) & (\alpha > 2) \\ \infty & (0 < \alpha < 2). \end{cases} \tag{variance}$$

2.8 Determination of Parameters in the Gumbel Distribution

Fisher-Tippett-Gnedenko Theorem can be used to determine the functional form of the asymptote of the cumulative distribution. In this subsection, one is going to narrow down into the Gumbel distribution case. For the other two distribution cases or more detailed explanation, see [28].

Before one get started, let us consider N samples, consisting of ℓ realizations of random variables. That is, each sample has length ℓ. The probability that the largest value among those ℓ realizations is less than x is given as $F^{\ell}(x)$. Now, there are in total $N\ell$ realizations of random variables, and N maximum value of each sample. Similarly to the above, the probability that the largest value among those $N\ell$ realizations is less than x is given as $F^{N\ell}(x)$. If there is an asymptote for $\ell \to \infty$ limit, maximum distributions of N and $N\ell$ realizations must be the same for large N and ℓ.

One can define an asymptote $G(z)$ as fallows:

$$\mathrm{Prob}\left\{\frac{M_l - \beta_\ell}{\alpha_\ell} \leq z\right\} \to G(z), \tag{2.14}$$

where M_l is a random variable representing the maximum value from a set of size ℓ. α_ℓ and β_ℓ are appropriate sequences of ℓ satisfying Equation 2.14.

$G(z)$ also satisfies following, namely, the stability postulate;

$$G^n(z) = G(a_n z + b_n). \tag{2.15}$$

Now one will show why if $a_n = 1$, then one can conclude that $G(x)$ is the Gumbel distribution. If $a_n = 1$, the maximum distributions can be overlapped by shifts along x axis without change of scale. The amount of translation is b_n. Therefore, Equation 2.14 becomes

$$G^n(x) = G(x + b_n). \tag{2.16}$$

Equation 2.16 has two unknown functions. One is the parameter b_n which is considered as a sequence of n above, and the other is $G(x)$ itself.

To determine b_n, one should take $-\log$ first and take $\log$ once more:

$$\log n + \log(-\log G(x)) = \log(-\log(G(x + b_n)). \tag{2.17}$$

when we move the double $\log$ term of the $G(x)$, Equation 2.17 becomes

$$\log(-\log(G(x+b_n)) - \log(-\log G(x)) = \log n. \tag{2.18}$$

It means that the increment of b_n for the given x makes the increase of $\log n$ for the double $\log$ of the $G(x)$. Hence,

$$\log(-\log G(x)) - \frac{x \log n}{b_n} = k, \tag{2.19}$$

where the constant k is independent of x.

Now one is interested in b_n. Starting from 2.16,

$$G^{(mn)}(x) = G(x + b_{(mn)}). \tag{2.20}$$

The other way to express the $G^{mn}(x)$ is the taking m-th power of Equation 2.16:

$$(G^n(x))^m = (G(x+b_n))^m = G((x+b_n)+b_m) = G^{mn}(x) \tag{2.21}$$

Now one can get a recurrence relation for b_n regardless of the G, that is

$$b_n + b_m = b_{mn}. \tag{2.22}$$

The solution of the Equation 2.22 is

$$b_n = c \log n + d \tag{2.23}$$

where c and d are constants. If one substitutes $\frac{b_n - d}{c}$ for the $\log n$ in Equation 2.19, one can realize that $d = 0$ because k is independent of x. Now, b_n is a function of n:

$$b_n = c \log n \tag{2.24}$$

One should solve the other part of the functional equations in Equation 2.22. Instead of $\log n$, one can use $\frac{b_n}{c}$ in Equation 2.22:

$$\log(-\log G(x)) = \frac{x}{c} + k. \tag{2.25}$$

As x increases, $G(x)$ increases to 1, and $-\log G(x)$ decreases to zero. So, c must be a negative number. One can introduce positive number β and μ as

$$\beta = -c, \qquad \mu = -ck = \beta k. \tag{2.26}$$

Now, one can rewrite Equation 2.19 and get

$$b_n = -\beta \log n, \tag{2.27}$$

i.e., b_n is a decreasing function of n. One can also remove the two logarithms in the Equation 2.25:

$$G(x) = e^{-e^{-\frac{x-\mu}{\beta}}}. \tag{2.28}$$

Let us go back to the comparison between $F^{\ell}(x)$ and $F^{N\ell}(x)$. One tried to find the finite but large ℓ_0 which satisfies the following relation that is similar to the stability postulate because the stability postulate is just the ideal case as $\ell \to \infty$ limit;

$$\{F^{\ell_0}(x)\}^N = F^{\ell_0}(a_N x + b_N). \tag{2.29}$$

For the appropriate ℓ_0 and N, if one check whether $a_N = 1$ and $b_N \sim c \log N$, then these indirectly determine α_ℓ as 1 and $\beta_\ell \sim \log \ell$. The relation between F and G in Equation 2.14 and these implications tell us that

$$F^{\ell}(x) = G(x - \beta_\ell). \tag{2.30}$$

This means that one can use Equation 2.30 to find out the parameters needed to be determined in the Gumbel distribution. Chapter 4 shows the process in detail.

Chapter 3

Simulations

3.1 the EP on Model Networks

One used 4 kinds of network models to make networks to investigate the properties of the EP: BA, ER, SFCL (exponent of the degree distribution $\gamma = 2.2$), SFCL ($\gamma = 2.5$). All of them had mean degree of 8. System sizes were various from $4096 (= 2 \times 2^{11})$ to $20480 (= 10 \times 2^{11})$ with increment of 2048. One used 300 ensembles for each model. One used giant clusters of ER and SFCL.

Each panel in Figure 3.1 shows the EP distribution of each network for different system sizes. Except ER case, the other 3 cases show that the distributions seem to converge to asymptotic distributions of each case. For ER case, the distributions for each size seems like the same. All of the distributions show very high peak at zero. This is partly because there are enormous paths whose distances are not bigger than 2, which produce the zero EP.

One can compare the EP distributions of different networks. Figure 3.2 shows the comparison of the EP distributions of the networks at the same size. BA has the broadest range of the EP and the others follow in order of SFCL ($\gamma = 2.2$), SFCL ($\gamma = 2.5$), and ER. The binning size is 0.25 and the perpendicular axis is $\log$ scale. The figure also contains the EP distribution of the co-authorship network (December 2010 data).

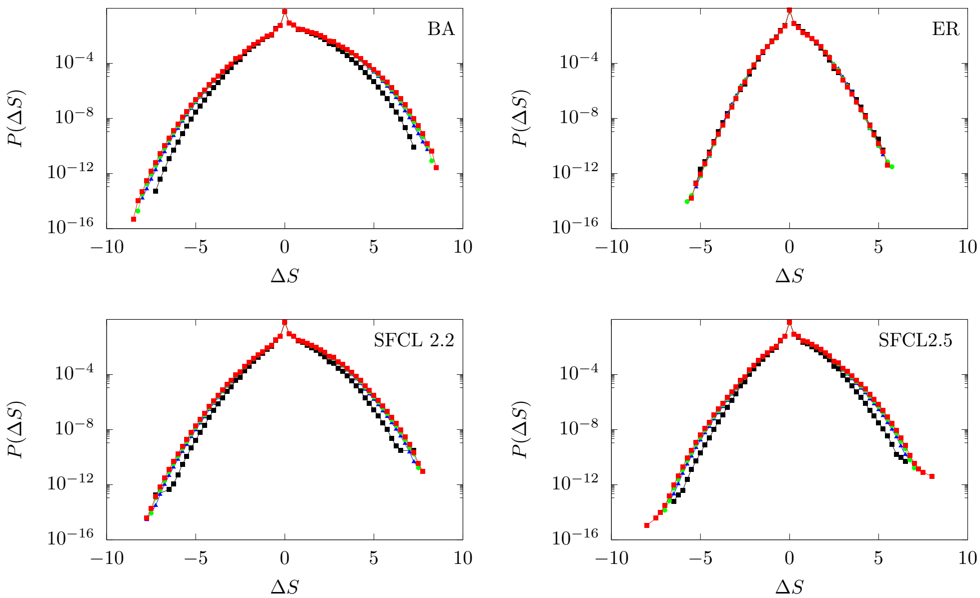

Figure 3.1: Each panel shows the EP distribution of each model with binning size 0.25 for different system sizes. One used 300 ensembles for each size. Except ER case, the other 3 cases show that the distributions seem to converge to asymptotic distributions. For ER case, the distributions for each size seem like the same.

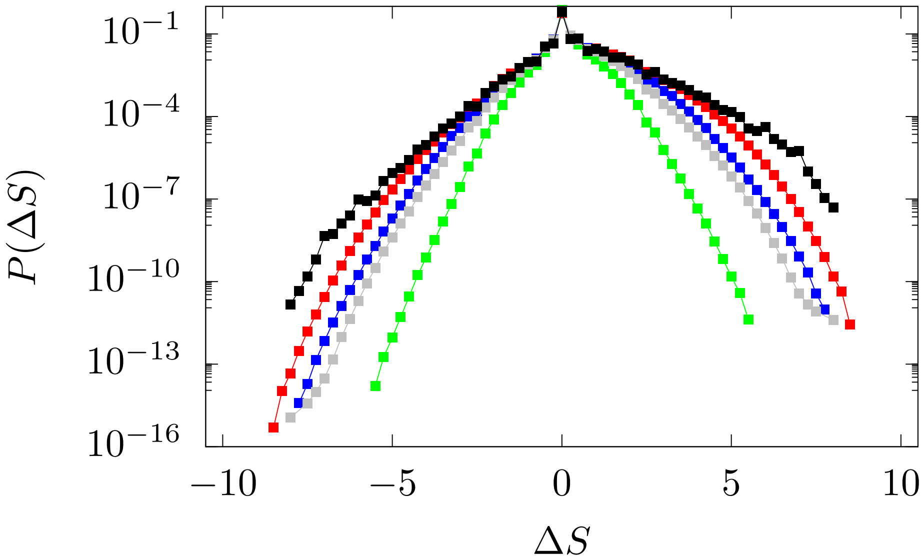

Figure 3.2: The EP distributions of 4 model networks and co-authorship network. Black, green, grey, blue, and red points are respectively co-authorship network, ER network, SFCL ($\gamma = 2.2$) network, SFCL ($\gamma = 2.5$) network, BA network. One used 300 ensembles with system size 20480 and mean degree 8 for each model network. co-authorship network is December 2010 data. The binning size is 0.25 and the perpendicular axis is $\log$ scale. All of the distributions show very high peak at zero. and the positive EP region has higher values of distribution than the negative region.

3.2 the EP on Empirical Networks

One studied the EP distributions of co-authorship networks. Figure 3.3 shows the EP distributions of giant clusters of co-authorship networks at different times. In these cases, the bell shape and the system size tendency is observed too. Detailed information about these networks are given in Table 3.1. The binning size is 0.25. Though it is hard to directly compare these 4 cases because they have different network characteristics, the bell shape distribution with high peak around zero, and that the larger system size, the broader range of the EPs are in common with the model network cases. On the other hand, there are jiggling patterns with the distributions because these distributions are for each network at each time, not averaged over ensembles like in the model network cases.

Time	System size	Number of links	k_{min}	$< k >$	k_{max}
Dec 2003	1075 (195)	1639 (441)	1 (1)	3.05 (4.52)	32 (32)
Oct 2006	3851 (1269)	7624 (3355)	1 (1)	3.96 (5.29)	63 (63)
Dec 2008	6729 (3079)	14729 (9034)	1 (1)	4.38 (5.87)	94 (94)
Dec 2010	9530 (4991)	22116 (15156)	1 (1)	4.64 (6.07)	120 (120)

Table 3.1: Network characteristics of co-authorship networks. Numbers in brackets are that of giant clusters.

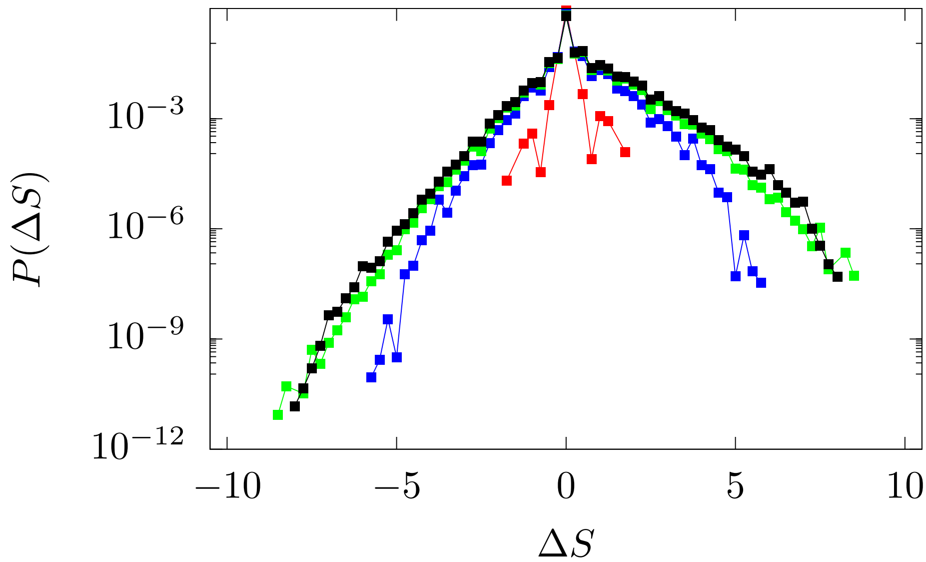

Figure 3.3: The EP distributions for co-authorship networks at 4 different times (red: Dec 2003, blue: Oct 2006, green: Dec 2008, black: Dec 2010). Detailed information about these networks are given in Table 3.1. The binning size is 0.25. The perpendicular axis is $\log$ scale. One can see bell shape distribution with high peak around zero, the range of the EP is getting broader as the system size increases.

3.3 Tail Behavior of the EP Distributions

Network	a	b	c
BA	0.27	0.96	1
ER	0.72	0.92	0.7
SFCL2.2	0.50	0.92	0.75
SFCL2.5	0.49	0.99	0.77

Table 3.2: The parameters in functional form (Equation 3.1) for each model network.

One looked for an analytic form of the tail of the EP distributions. Firstly one could try a functional form,

$$P(x) \sim e^{-e^{ax^c+b}} \tag{3.1}$$

Then one could find suitable parameters a, b, and c. Figure 3.4 shows tail behaviors of the EP distributions. Both axes are $\log$ scales. The red points are data, and the blue lines are fitting lines. One used 300 ensembles whose system sizes are 20480 and mean degrees are 8, for each case. The parameters for each network are shown in the Table 3.2.

Because there are too many parameters and there is no solid mathematical reason justifying the functional form, Equation 3.1, one should need to use the methodology based on more rigorous theory. Chapter 4 shows how the theory which is introduced in Chapter 3 is used practical analysis.

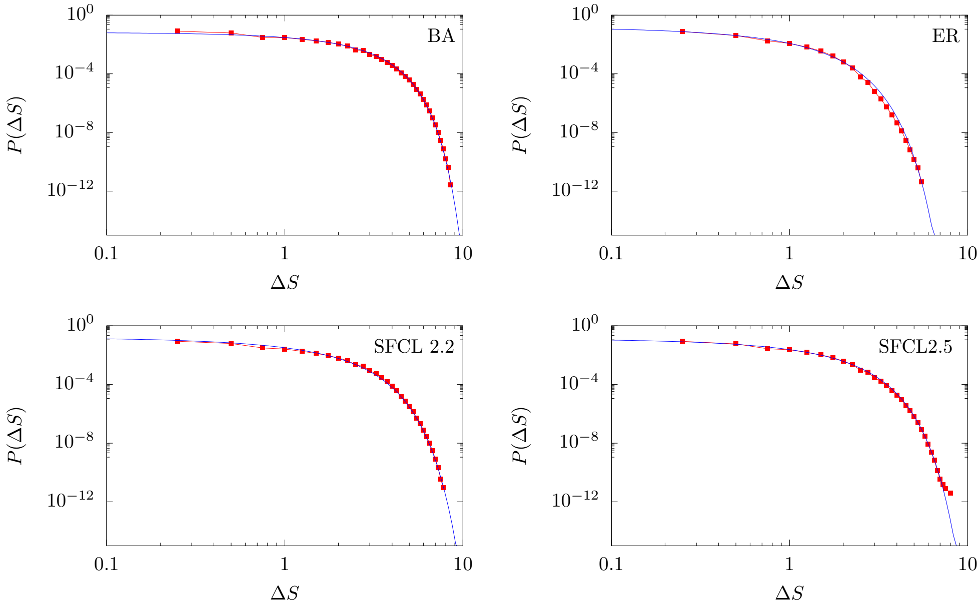

Figure 3.4: Tail behaviors of the EP distributions. Both axes are $\log$ scales. Red points are data, and blue lines are fitting lines (see Equation 3.1). One used 300 ensembles whose system sizes are 20480 and mean degrees are 8, for each case. The parameters in functional form (Equation 3.1) for each network are shown in Table 3.2.

Chapter 4

Asymptote of EP Distribution

In this chapter, one investigates how to determine the functional form of asymptote of the EP distributions, using Fisher-Tippett-Gnedenko theorem.

4.1 Determination of the Asymptote with Exact Data

In Chapter 3, one discussed how one can determine the functional form of the maxima distribution. For one assembled data from all of the possible ways between all of the possible node pairs on the ensembles (though Figure 3.2 is binned), one can calculate the exact cumulative distribution functions of EP of given ensembles. After getting the cumulative distributions, it is straightforward how to determine if the distribution is the Gumbel distribution. For this, one should calculate a cumulative distribution function to the power of several $\ell(\equiv \ell_0 n$, see Equation 2.29) values, and translate it along the x axis to confirm if the ℓ-th power functions overlap each other. Roughly speaking, if they overlap without change in scale, the functional form is the Gumbel distribution.

Figure 4.1 shows the halfway through the process of determination of the functional form using the exact data of the BA network case (system size is 20480, mean degree is 8). On the left panel, there are cumulative functions to the ℓ powers. The red, blue, and black lines are $\ell = 3000$, 6000, and 9000 case, respectively ($\ell_0 = 1000$ and $n = 3$, 6, and 9). On the right panel, one can see the three lines moved horizontally with appropriate amounts and overlapped well. It indicates that the asymptote of the cumulative distribution is the Gumbel distribution. The extent how much one should

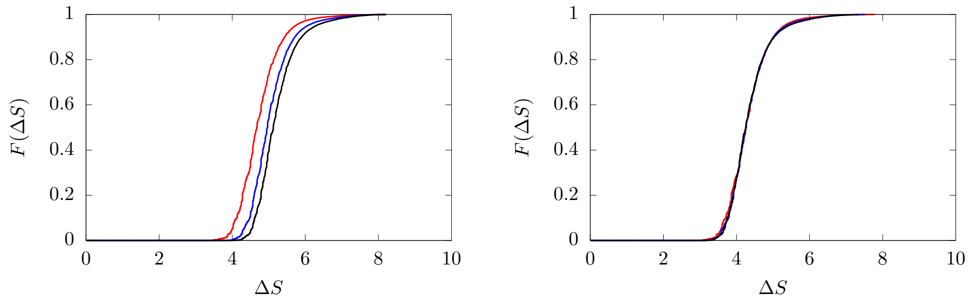

Figure 4.1: (Left) Each line is ℓ-powered functions of the cumulative distribution function of the BA network case (system size is 20480). The red, blue, and black lines are $\ell = 3000, 6000$, and 9000 case, respectively. (Right) One can make the three lines collapsed just with translating along x axis, without change of scale. This fact indicates that the asymptote of the cumulative distribution is the Gumbel distribution. The extent how much one should translate each line for different ℓ value is offered on the right panel in Figure 4.2.

translate each line is offered on the right panel in Figure 4.2.

On the other hand, though one uses the exact result from the whole data, determination might not be perfect because there is the limitation to the range of the EP value. This is radically because the system size is not infinite. Theoretically, one can assume that the complete probability distribution ranged from $-\infty$ to ∞ is given, but in reality, making a network as big as one can run simulation and the finite size effect will be hidden is just the best way to avoid the finite size effect. That is, though the sufficiently big size is used here and the lines on the right panel are well overlapped, one should notice that the overlap might not appear for the smaller system size case.

Assuming $\ell_0 = 1000$, Figure 4.2 shows the information to determine the parameters of the Gumbel distribution. Recall the Equation 2.15, 2.29, 2.30:

$$G^n(z) = G(a_n z + b_n) \tag{4.1}$$

$$\{F^{\ell_0}(x)\}^N = F^{\ell_0}(a_N x + b_N) \tag{4.2}$$

$$F^{\ell}(x) = G(x - \beta_\ell) \tag{4.3}$$

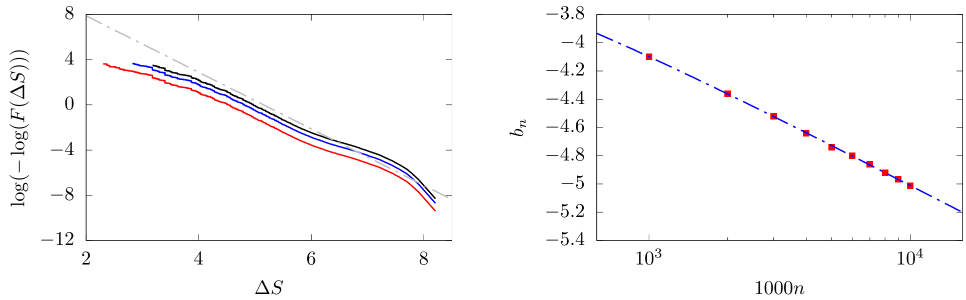

Figure 4.2: (Left) One can realize that $a_n = 1$ easily from the fact that the lines are overlapped without change in scale. To determine β, one can take the double $\log$ of the cumulative distribution function, and draw it versus ΔS. One can find that the slopes of the three lines are the same in quite long interval of the EP. The slope is $\beta = -2.5$. (Right) To determine n-dependence of b_n, one can plot and fit the data of the extent how much each line, which are the $n-$powered functions of the F^{ℓ_0}, should be translated to be collapsed. x axis is $1000n(= \ell_0 n)$ and log scale, which means that b_n is the $\log$ function of the system size, as it is expected in Equation 2.27.

Now a_n and a_N is simply 1, because the lines are overlapped without change in scale. For β (not β_ℓ, but β in Equation 2.26), one took the $\log$ twice (c.f. Equation 2.25) and found out the common slope of the lines. Though there are some intervals in which the slope of the lines is changed, one can determine that $\beta = -2.5$. For b_n, one investigated the b_n values which is the same to the amount of translation to make the lines overlap each other (on the right panel in Figure 4.1). The tendency of b_n is shown on the right panel in Figure 4.2. One can confirm that

$$b_n \sim -0.42 \log n. \tag{4.4}$$

4.2 Determination of the Asymptote with Sampling Data

One can use the probability distribution functions in Figure 3.2 because all of the possible cases are counted. But in reality, one can not count the whole possible cases. So, generally, one should run a sampling process with the probability distribution from data. Now, a question naturally arises: is the result from the sampling data consistent to that from the exact data?

The first step of the sampling process with the probability distribution from the whole data is picking ℓ-numbers from probability distribution, in total N-times. Each time, one writes down the maximum value among the ℓ-numbers. After N iterations, there are N maximum values (See subsection 2.8).

Using Equation 2.29, one can try if

$$\{F^{\ell_0}(x)\}^n = F^{\ell_0}(a_n x + b_n) \tag{4.5}$$

is satisfied with an appropriate ℓ_0. For that there is the limit process that $\ell(\equiv \ell_0 n)$ goes to ∞ in Fisher-Tippett-Gnedenko Theorem, ℓ_0 (and also ℓ) has to be sufficiently large. On the other hand, because there must be a restriction on the range of x because the distribution is from data, not mathematically given, one should stay in the region of ℓ sufficiently far from the end of the range of x that the finite range effect does not affect on the distribution quite much. To go through this dilemma, there is the only way: gather as many data for constructing the probability distribution as possible, so many that the probability distribution covers sufficiently large range.

After Equation 4.5 is checked, the sequential process is the same to the exact process: find β from taking double $\log$ of the both sides, find b_n by gathering the amount of translation for overlapping.

One can compare the results from the sampling data and the exact data in Figure 4.3. The panels on the first row are the same to Figure 4.1. The panels on the second row are the results from sampling data. The red, blue, and black lines are cases that ℓ_0 is 1000 and $n = 3, 6, 9$, respectively. One can notice that it is also true that the three lines are collapsed just by translating along x axis with the amount which one has found in the exact data case, without change of scale. This fact implies that one can use the sampling process to determine the functional form of the asymptote, though the distribution constructed by sampling process can never be the exact distribution.

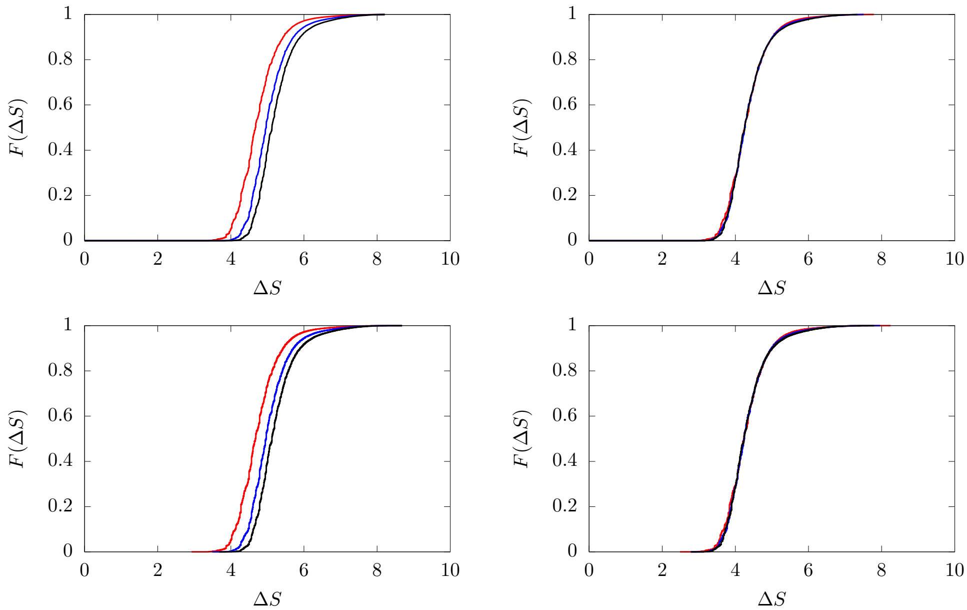

Figure 4.3: (Up) The upper two panels are the same to Figure 4.1. (Down) Results of the sampling data cases. The red, blue, and black lines are cases of $\ell_0 = 1000$ and $n = 3, 6$, and 9, respectively. One can notice that it is also true that the three lines are collapsed just by translating along x axis with the amount which one has found in the exact data cases, without change of scale. This fact implies that one can use the sampling data to determine the functional form of the asymptote.

Figure 4.4 shows the data and the fitting lines whose parameters are determined by the exact data. One can see that the choice of range of the power n (or, the sample length ℓ for sampling way) is appropriate to avoid finite range problem and is sufficiently large to be used in the determination of parameters.

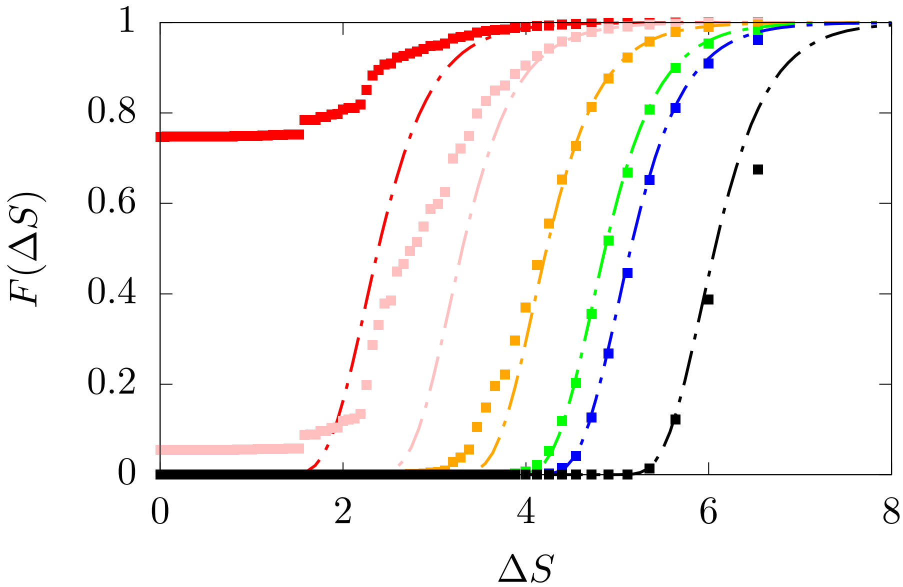

Figure 4.4: For BA network (system size 20480) case, one found out fitting lines for various powers. the square solid dots represent data, and the dashed lines represent the fitting lines which is determined in the method which is introduced in this chapter. The red, pink, orange, green, blue, and black represent the power $n = 0.01, 0.1, 1, 5, 10, 100$ cases. One can see that the fitting lines do not fit well if ΔS is bigger around 6. This is because of the limitation of the number of the data.

Chapter 5

EP Distribution on 2D Lattice

One can not find out the analytic form of the EP distribution for networks because of their complexity, but for 2D lattice, one can find the analytic form of the EP distribution. In this chapter, one shows the derivation of the EP distribution on 2D lattice.

5.1 Derivation of EP Distribution on 2D Lattice

Let us consider $m{\times}n$ lattice. One may set the left-bottom-most vertex at the origin, and then the total number of the shortest paths from the origin to (m, n) is

$$\binom{m+n}{n} = \binom{m+n}{m}. \tag{5.1}$$

Now one can simply express a path with a sequence like:

$$(\rightarrow, \uparrow, \rightarrow, \cdots, \rightarrow, \uparrow). \tag{5.2}$$

One can figure out that the first series of the same kind of arrows contributes with the EP (the next section explains the reason in detail) of $(-\log 2)\times$(number of the arrows). Similarly, the last series of the same kind of arrows contributes with the EP of $\log 2\times$(number of the arrows). The middle part of the sequence between the two series contributes zero EP. Therefore, the length of the two series are important. Let ℓ_h and ℓ_t respectively denote the length of the head and tail series, then the EP can be written by $(\ell_t - \ell_h) \times \log 2$.

To calculate the weight of a specific EP value, say, $\Delta \log 2$ (where $\Delta \equiv \ell_t - \ell_h$), let us cut the sequence of a path into three parts and two dividers:

$$(head, divider, body, divider, tail).$$

The length of the body is $m + n - (\ell_t + \ell_h) - 2$.

If there are only one kind of arrow in both the head and the tail, for example, all of them are right arrows, then the two dividers are up arrows, and there are $m - (\ell_t + \ell_h)$ right arrows and $n - 2$ up arrows in the body. The probability for a path of this kind is $(\frac{1}{2})^{d-\ell_t}$, and the number of combinations of this kind is

$$\binom{m + n - (\ell_h + \ell_t) - 2}{n - 2} = \frac{(m + n - (\ell_h + \ell_t) - 2)!}{(m - (\ell_h + \ell_t))!(n - 2)!}. \tag{5.3}$$

If the head and the tail have different kinds of arrows, for example, right arrows are in the head and up arrows are in the tail, then the former divider is up arrow, the latter divider is right arrow, and there are $m - \ell_h - 1$ right arrows and $n - \ell_t - 1$ up arrows in the body. The probability for a path of this kind is $(\frac{1}{2})^{d-\ell_t}$, and the number of combinations of this kind is

$$\binom{m + n - (\ell_h + \ell_t) - 2}{n - \ell_t - 1} = \frac{(m + n - (\ell_h + \ell_t) - 2)!}{(m - \ell_h - 1))!(n - \ell_t - 1)!}. \tag{5.4}$$

Finally,

$$\begin{aligned} P(\Delta \log 2) = \sum_{\ell_h, \ell_t, \Delta = \ell_t - \ell_h} \left(\frac{1}{2}\right)^{d-\ell_t} & (m + n - (\ell_h + \ell_t) - 2)! \\ \times \Bigg\{ & \frac{1}{(m - (\ell_h + \ell_t))!(n - 2)!} + \frac{1}{(n - (\ell_h + \ell_t))!(m - 2)!} \\ & + \frac{1}{(m - \ell_h - 1))!(n - \ell_t - 1)!} + \frac{1}{(n - \ell_h - 1))!(m - \ell_t - 1)!} \Bigg\}. \end{aligned} \tag{5.5}$$

5.2 Qualitative Analysis

Let us consider 2D lattices to analyze simply how the EP distribution related to symmetry of the shortest paths between the two given nodes.

Assume that the source and the target are located as far as possible. Now one can imagine two extreme examples of networks: 10×10 lattice, and 2×18 lattice. The distances between the two nodes are 20 for each lattice case. In the former case (Figure 5.1), it is more possible that one shortest path is consisted of the symmetric links (Figure 5.3). Because the symmetric links in the shortest paths produce zero EP, most of the shortest paths produces the EP around zero, and stretches in narrower range than the 2×18 case.

However, in the latter case (Figure 5.2), most links are asymmetric links. Because asymmetric links produce the nonzero EP, the EP distribution for this case stretches in broader range than the 10×10 case.

Reciprocal of average of reciprocal distance, which is called diameter of network, may be affect the distribution. It is because there are only a fewer way between pairs whose distances are longer than diameter, rather than between pairs whose distances are around diameter. This may be because if the distance is so long that there are only few shortest paths, then it is more probable that the EP is zero. When the distance between a pair is less than the typical distance, which is the diameter of the network, the longer distance, however, contributes the variety of the shortest paths. Thus there are more chances to produce the finite EP between this pair.

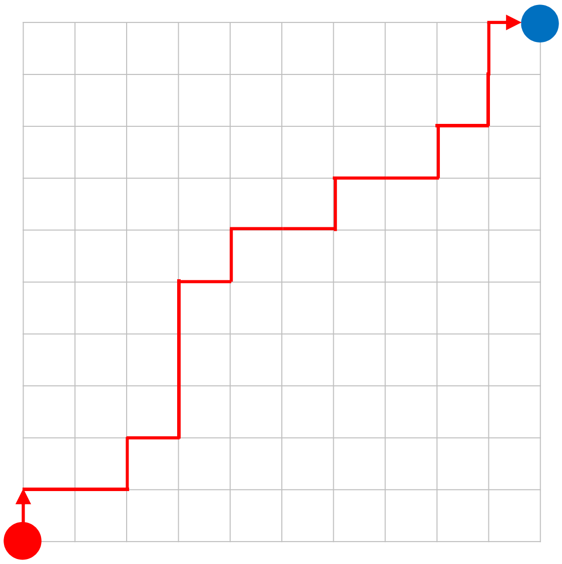

Figure 5.1: 2D lattice example: 10×10 lattice case. The red and blue nodes are the source and the target, respectively. The red path is the one of the shortest paths. One can see that there are 18 symmetric links (see Figure 5.3) on the red path, not producing the EP. Merely the 2 asymmetric links (see Figure 5.4) contribute the EP produced. Because the first link produces $-\log 2$ and the last link produces $\log 2$, the EP that is produced from the red path happens to be zero in total.

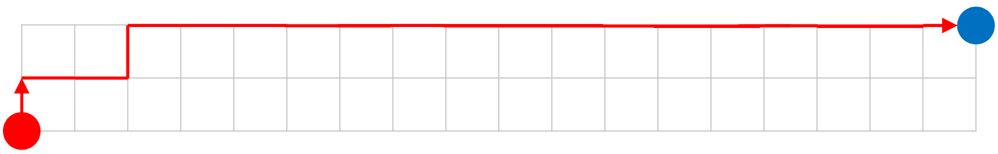

Figure 5.2: 2D lattice example: 2×18 lattice case. The red node is the source, and the blue node is the target. The red line is the one of the shortest paths from the source to the target. Different from Figure 5.1, there are asymmetric links more than symmetric links. These asymmetric links contribute to the EP.

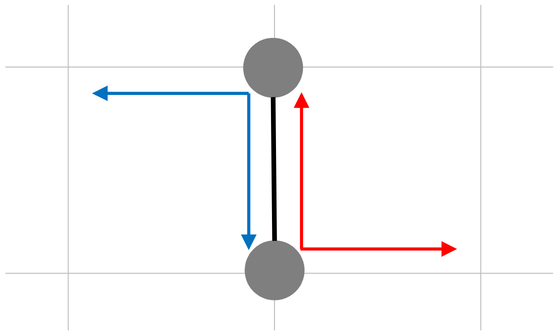

Figure 5.3: A symmetric link. There are two options to go forward (red arrows) from the lower node (lower gray circle), and two options to come backward (blue arrows) from the upper node (upper gray circle). Therefore, the bold link carries a weight from the lower node to the upper node as much as from the upper node to the lower node. That is, this link carries flow symmetrically, so it does not produce the EP.

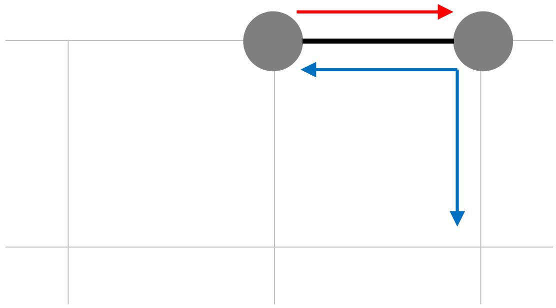

Figure 5.4: An asymmetric link. There is the only one option to go forward (red arrows) from the left node (left gray circle). However, there are two options to come backward (blue arrows) from the right node (right gray circle). Therefore, the bold link carries a weight from the left node to the right node as twice as from the right node to the left node. That is, this link carries flow asymmetrically, so it produces the nonzero EP, $\log 2$.

Chapter 6

Conclusion

Entropy is the quantity to understand the equilibrium states and nonequilibrium phenomena. Including Hinrichsen, there are some attempts to embed the EP to the processes on the complex networks. This may be partly because networks can serve as state spaces, for example, the conformation network in the protein folding problem.

In this dissertation, one studied the entropy production (EP) that is defined with the topological characteristics, especially the shortest paths of the networks. The asymmetric structure of the subgraph consisting of the shortest paths affects the EP and the shape of the EP distribution. On complex networks, system size also matters. It may be because the distance may be related on the asymmetric structure of the shortest paths between pairs on complex network. If the distance is longer than the diameter of the network, then it may have an affect to decrease the EP.

One found out the functional form of the asymptote of EP distributions using extreme value analysis, specifically Fisher-Tippett-Gnedenko Theorem. The asymptote is the Gumbel distribution. The fitting lines are fit well to the data in the range in which data is not affected by finite range effect.

One was able to derive the analytic form of the EP distribution for a special case, on 2D lattice. There are two kinds of links, symmetric and asymmetric links, and only the latter can contribute nonzero EP. Therefore the more asymmetric links are in a shortest path, the more EP produces from the shortest path.

Various fields such as data packet transport, logistics, traffic transportation engineering, etc. need to consider processes going along the shortest paths on the given network. One expects that the EP can be helpful to study this subjects because the EP is based on the shortest paths between the source and the target, and represents the asymmetry of the flow on the shortest paths.

Bibliography

[1] R. Clausius, *The Mechanical Theory of Heat – with its Applications to the Steam Engine and to Physical Properties of Bodies*, J. Van Voorst, London, (1867).

[2] L. Boltzmann, *Wiener Berichte*, **53**, 195 (1866).

[3] C. Shannon, *The Bell Syst. Tech. Jounal.*, **27**(3), 379 (1948).

[4] U. Seifert, *Rep. Prog. Phys.* **75**, 126001 (2012).

[5] D. J. Evans, E.G.D.Cohen, and G. P. Morriss, *Phys. Rev. Lett.* **71**, 2401 (1993).

[6] D. J. Evans and D. J. Searles, *Phys. Rev. E* **50**, 1645 (1994).

[7] G. Gallavotti and E. G. D. Cohen, *Phys.Rev.Lett.* **74**, 2694 (1995).

[8] J. D. Noh, C. Kwon, and H. Park, *Phys. Rev. Lett.* **111**, 130601 (2013).

[9] G. E. Crooks, *J. Stat. Phys.* **90**, 1481 (1998).

[10] C. Jarzynski, *Phys. Rev. Lett.* **78**, 2690 (1996).

[11] J. Kurchan, *J. Phys. A* **31**, 3719 (1998).

[12] C. Maes, *J. Stat. Phys.* **95**, 367 (1998).

[13] J. L. Lebowitz and H. Spohn, *J. Stat. Phys.* **95**, 333 (1999).

[14] U. Seifert, *Phys. Rev. Lett.* **95**, 040602 (2005).

[15] C. Jarzynski, *J. Stat. Phys.* **98**, 77 (2000).

[16] P. Erdős and A. Rényi, *Publ. Math. Debrecen* **6**, 290 (1959).

[17] D. J. Watts, and S. H. Strogatz, *Nature* **393**(6684), 440–442 (1998).

[18] A.-L. Barabási, and R. Albert, *Science* **286**, 509 (1999).

[19] B. A. Huberman, and L. A. Adamic, *Nature* **401**, 131 (1999).

[20] K.-I. Goh, B. Kahng, and D. Kim, *Phys. Rev. Lett.* **87**, 278701 (2001).

[21] K.-I. Goh, B. Kahng, and D. Kim, *Phys. Rev. Lett.* **88**, 108701 (2002).

[22] J.D. Noh, and H. Rieger, *Phys. Rev. Lett.* **92**, 118701 (2004).

[23] H. Hinrichsen, T. Hoßfeld, M. Hirth, and P. Tran-Gia, G. Ghoshal et al. (Eds.): *Complex Networks IV*, SCI **476**, 47 (2013).

[24] R. Ziener et al., *J. Stat. Mech.* **2015**, P08014 (2015).

[25] J. Schnakenberg, *Rev. Mod. Phys.* **48**, 4 (1976).

[26] E. W. Dijkstra, *Numerische Mathematik* **1**, 269 (1959).

[27] L. De Haan, and A. Ferreira, *Extreme Value Theory: An Introduction*, Springer, New York, (2006).

[28] E. J. Gumbel, *Statistics of Extremes*, Dover, New York, (1958).

초 록

열 및 통계물리에서 무질서도 발생량 (Entropy Production) 은 평형 상태와 비평형 현상을 이해하는 데에 중요한 양이다. 여기서는 그물 얼개 (Network) 가 상태 공간 (State Space) 으로 사용될 수 있다는 것에서 착안하여, 그물 얼개 위에서의 데이터 패킷 전송 문제를 다루었다. 이 과정에서 그물 얼개 위에서의 무질서도 발생량을 정의했다. 이 양은 출발점에서 끝점을 향해 갈 때 어느 한 최단경로를 따라갈 확률과, 그 최단경로를 따라 돌아올 확률의 로그 비율로 정의된다. 이 무질서도 발생량은 통합적인 요동 정리 (Integral Fluctuation Theorem, IFT) 와 자세한 요동 정리 (Detailed Fluctuation Theorem, DFT) 를 만족한다. 무질서도 발생량 분포의 특성을 알아보기 위해 극단값 통계학 (Extreme Value Statistics) 을 사용해 그 점근 분포를 조사했다. 이에 따라 바라바시-알버트 모형 그물 얼개에서 얻은 무질서도 발생량 분포 함수의 점근 분포가 굼벨 분포 (Gumbel distribution) 라는 것을 밝혔다. 마지막으로 2차원 격자 위에서 무질서도 발생량의 분포 함수를 수학적으로 유도했다.

주요어: 최단 경로 과정, 무질서도 발생량, 그물 얼개, 극단값 통계학, 굼벨 분포
학번: 2016-20319