Stony Brook University

Topics in Statistical Physics: Protein Stability, Non-Equilibrium Thermodynamics and Bibliometrics

A Dissertation presented

by

Michael Hazoglou

to

The Graduate School

in Partial Fulfillment of the

Requirements

for the Degree of

Doctor of Philosophy

in

Physics

Stony Brook University

December 2016

Stony Brook University

The Graduate School

Michael Hazoglou

We, the dissertation committee for the above candidate for the

Doctor of Philosophy degree, hereby recommend

acceptance of this dissertation

Ken A. Dill - Dissertation Advisor
Professor, Department of Chemistry and Department of Physics and Astronomy

Marivi Fernández-Serra - Chairperson of Defense
Associate Professor, Department of Physics and Astronomy

Harold J. Metcalf
Professor, Department of Physics and Astronomy

Gábor Balázsi
Associate Professor, Department of Biomedical Engineering

This dissertation is accepted by the Graduate School

Charles Taber
Dean of the Graduate School

Abstract of the Dissertation

Topics in Statistical Physics: Protein Stability, Non-Equilibrium Thermodynamics and Bibliometrics

by

Michael Hazoglou

Doctor of Philosophy

in

Physics

Stony Brook University

2016

This dissertation will cover three distinct topics of protein stability, non-equilibrium thermodynamics and scientometrics. In senescent organisms aging is correlated with oxidative damage of proteins. The damage done to proteins destabilizes them inhibiting their function. The implications of a simplified model based on side-chain modification of charged residues using Debye-Hückel theory will be presented. Short length and highly charged proteins are susceptible to destabilization from oxidative damage. Among these proteins already studied in aging several proteins fit this description of being short and highly charged. There is a noticeable enrichment of short- highly-charged proteins in categories of proteins known to be important in aging. Maximum Caliber (MaxCal) is a potential theory of non- equilibrium statistical mechanics. It will be shown how MaxCal is used to derive the Onsager reciprocal relations, Green-Kubo relations and Prigogines Principle and extend these relations beyond the near-equilibrium regime. The last topic is the citation and publication trends of papers and authors, respectively. A discussion of how pure-birth processes can be applied to understanding citation trends and how birth-processes can be used in classifying papers into different categories of performance.

Contents

List of Figures

List of Tables

1 Introduction

This dissertation is broken into three main parts each having to do with projects that I have worked on these last few years. Each part covers disjoint topics in protein stability, non-equilibrium thermodynamics and scientometrics. The only thing connecting these topics is that statistics, probability theory and statistical physics are utilized in all three of them and because of this, the name of this thesis was selected.

The first part will be on the effect of oxidative damage to the thermodynamic stability of proteins and how this influences the stability and function of a proteome of an organism. This study was motivated by aging, as oxidative damage of proteins is more prevalent in older senescent organisms and individuals who suffer from premature aging diseases. It is uncertain if this correlation is a direct cause and effect and which direction it goes (does oxidative damage cause aging or the other way around). The main result of this study is that proteins which are highly charged and/or short in length are highly susceptible to destabilization from oxidation, we identified twenty proteins from the human proteome that have been found implicated in aging two fall into this group of short highly charged proteins. These are mostly proteins that carry out important functions in the nucleus or ribosome as proteins that interact with DNA are positively charged due to the negative backbone of DNA. This section is an expanded version of our work in [15].

The second part describes the statistical mechanics of non-equilibrium steady state systems from a possible theory of non-equilibrium statistical mechanics called Maximum Caliber (MaxCal). A statistical theory for non-equilibrium thermodynamics would be a great achievement with profound applications. With a statistical theory of non-equilibrium processes one can characterize fluctuations, dissipation, transport properties, and possibly understand turbulence and chaotic dynamics from the point of critical phenomena and phase transitions. A statistical formulation of non-equilibrium thermodynamics would be an incredibly powerful tool. Many of the celebrated results of non-equilibrium thermodynamics such as the Onsager reciprocal relation, Prigogine's principle, Green-Kubo equations are reproduced and entropy fluctuation theorem are reproduced under the umbrella of MaxCal. Within the formalism of MaxCal the higher order symmetry relations, relationships which are analogous to the Onsager reciprocal relations for higher order terms in the expansion of constitutive equation are shown to not exist for all orders greater than second order. Much of the key results presented

in this section were published in [27].

The third and last part is focused on the study of scientific writings and their publication and citation, known as scientometrics which is a subset of bibliometrics. These fields are studied with the purpose of understanding how scientists on a whole publish, how they get cited, developing metrics of quality, and predicting future success. This part will develop a model to describe how individuals publish papers and get cited, based on the birth process with cumulative advantage, where the rate of events depends on the number of previous events. By utilizing the mathematical formalism of the birth process one can reproduce the results of Wang, Song and Barabási [72] reducing the number of assumptions that need to go into deriving their primary equation but also providing a theory for the stochastic nature of citation. As far as an individual publishing articles previous studies and our own indicate that as long as a scientist is active they publish papers at a relatively constant rate over their career, characterized by the birth process and a negative binomial distribution. These results are born out from observations on data sets obtained from PubMed and the American Physical Society.

2 Proteins and the Effect of Oxidative Damage on their Stability

In this chapter, I develop a first-principles physical model for how normal oxidative damage processes such as are prevalent in aging biological cells can change the physical properties of the proteins in the cell. This may be important for how cells manifest the effects of aging. Our model begins by recognizing the following key properties: (i) protein molecules are linear chains that fold up into compact structures through large numbers of weak non-covalent interactions, (ii) that protein folding stability is critical for their functioning as biology's workhorse units, (iii) that oxidative damage is randomly distributed across biomolecules in the cell, (iv) that protein molecules are among the most important affected targets, and (v) that a key effect of oxidative damage on proteins is the changing of the charges of some highly charged side chains. Using a polymer statistical physics model of proteins as charged chains, we show how those protein molecules that already are highly charged can be substantially destabilized, and caused to unfold, by even single random oxidative damage events.

At the age of 80 about half the proteins in the human body are damaged by oxidization. Oxidative damage occurs because of the natural metabolic processes (the conversion of nutrients to energy in the presence of oxygen) that occur in the body of organisms. Searching protein databases and processing the data found 20 proteins that where already studied in aging experiments and suggests that proteins which play important roles in DNA maintenance, and protein synthesis are most susceptible to damage by oxidation. This is applicable to the discovery of proteins related to aging and age-related diseases for potential aging studies.

2.1 The Model for Charge-based Destabilization of Proteins

Proteins are the work-horses of the cell. Proteins are enzymes that carry out essential functions for life. They are polymers (peptide chains) composed of twenty possible amino acids with different side chains, see Figure 1 for the chemical structure of an amino acid. The sequence of these amino acid residues (residue refers to the remnant of the amino acid monomer in the chain) determines how the chain folds. The side chains on the backbone of the

peptide chain and their interactions between them will cause the protein to fold into a configuration with the minimum free energy. As the environment of a protein is aqueous much of the hydrophobic residues form the core of the protein which is shielded from the water, the exterior of the protein will prefer residues with charge, polarity as they can interact favorably with surrounding water molecules. Any disturbance in these interactions will harm a protein's stability and if it significant enough to cause the protein to change configuration it will hinder its ability to function.

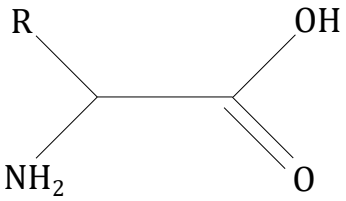


Figure 1: The structure of an amino acid, the is derived from the amine group and carboxyl group (which is acidic) the side chain **R** is anyone of 20 possibilities, which are either negatively charged (acidic), positively charged (basic), or neutral where hydrogen bonds, steric and Van der Waals interactions are more significant for how they interact with other side chains in a peptide chain. Peptide chains are formed by hydrolysis of the amine group with the carboxyl group of another amino acid.

As an organism ages their proteins are damaged by reactive oxygen species, reactive nitrogen species, reactive lipids and glycolytic products, see Figure 2. These reactive species can modify side chains of proteins, cleave the backbone, and covalently bond to lipids, carbohydrates or other proteins [61, 64]. Most of the damage is done by side chain modification, it is an order of magnitude higher than the others. Modification of sides chains in many instances causes a change in the charge [64]. See Table 1 for possible which residues are affected by oxidative damage.

As a first approximation of the effect of oxidative damage on stability of a protein can be modeled by Debye-Hückel theory [17]. Using the linearized Poisson-Boltzmann equation with the assumptions that the net charge is uniformly distributed over the surface of the protein, and the folded (native) and unfolded (denatured) forms are approximately spherical in shape (compact in the native case, compact partially folded molten globules in the denatured case). The change in the folding free energy due to charge effects, ΔG_e, would be [23]

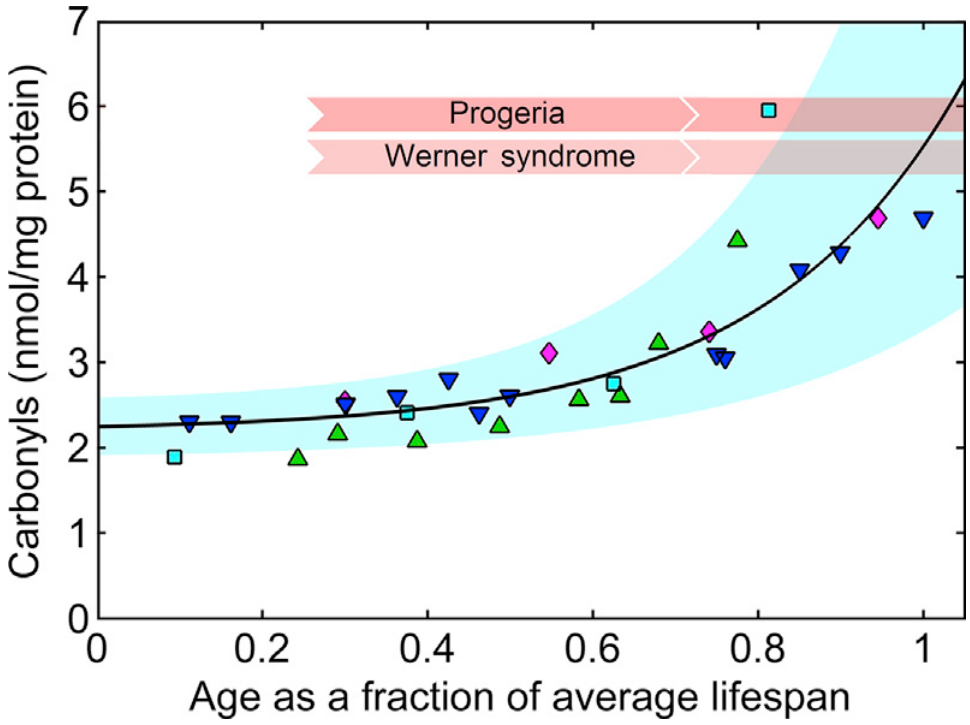


Figure 2: Oxidative damage as measured by carbonyl content versus age normalized to the average lifespan of the species. The purple diamonds are worms [3], the green triangles are flies [63], cyan squares are rats [65], blue upside-down triangles are humans [49]. The black curve is is an exponential fit to all the data with least squares regression, and the cyan band shows the deviation when then parameters are varied by $\pm 15\%$. The pink bands show the levels of carbonylation in accelerated aging diseases [49], which show the same levels of carbonylation as observed in the latest stages of life.

$$\frac{\Delta G_e}{kT} = \frac{Q_d^2 l_b}{2R_d(1+\kappa R_d)} - \frac{Q_n^2 l_b}{2R_n(1+\kappa R_n)} \tag{1}$$

k is the Boltzmann constant, T is the temperature, l_b is the Bjerrum length, $\kappa = \sqrt{2c_s l_b}$ the inverse of the Debye screening length, c_s being the salt concentration, Q_n the net charge of the native protein, Q_d the net charge of the denatured protein, R_n and R_d are the radii of gyration for the protein in the native and denatured form respectively. The definition of ΔG above is taken so that positive values means the native form is stable, this is for

Method of oxidation	Amino Acids affected
Metal-catalyzed oxidation	Arg, Lys, His, Pro, Thr, Tyr, Cys, Met
$^1O_2$	Arg, Lys, His, Pro, Thr, Tyr, Cys, Met
$ONOO^-$	Tyr, Cys, Met
HOCl	Arg, Lys, Pro, Thr, Tyr, Cys, Met
Ozone	Arg, Lys, Pro, Thr, Cys, Met
γ-Ray	Arg, Lys, His, Pro, Thr, Tyr, Trp, Val, Leu, Cys, Met

Table 1: Main methods of oxidative damage to side chains [61].

convenience. Since $R_d > R_n$ and $Q_d \approx Q_n$, $\Delta G < 0$ meaning net charge will only destabilize a protein. Oxidative modification will change the charge of a protein by at most one unit in either direction, $Q \to Q \pm 1$, the change in the folding free energy $\Delta\Delta G$

$$\frac{\Delta\Delta G_e}{kT} = \frac{(\pm 2Q_d + 1)l_b}{2R_d(1+\kappa R_d)} - \frac{(\pm 2Q_n + 1)l_b}{2R_n(1+\kappa R_n)} \tag{2}$$

when the oxidation increases the magnitude of the charge in either direction it destabilizes the protein. For the typical conditions found in a cell calculations are carried out with $l_b = 7.13\text{Å}$, $\kappa = 0.03\text{Å}^{-1}$ which result from $c_s = 0.1\text{M}$ and a relative dielectric constant of 78.5 and a pH of 7. The radii of gyration are functions of chain length, N, these are give by empirical relations $R_n = 2.24N^{0.392}\text{Å}$ and $R_d = 1.927N^{0.598}\text{Å}$ [38]. The equation (1) indicates that a protein with a greater chain length allows the charge to be spread out over the surface of the protein, so longer proteins can accommodate greater charge. For determining the average charge of a residue which is cased on the chance of protonation p and the relevant pK_a

$$p = \frac{1}{1 + 10^{\text{pH}-\text{pK}_a}} \tag{3}$$

for a basic residue this would mean the charge is given by p and for an acidic residue the charge is given by $-(1-p)$ (minus the chance of being deprotonated) the net charge can is the sum of these term for each amino acid in a sequence

$$Q = \sum_{i=\text{base}} \frac{1}{1 + 10^{\text{pH}-\text{pK}_{a,i}}} - \sum_{j=\text{acid}} \frac{1}{1 + 10^{\text{pK}_{a,j}-\text{pH}}} \tag{4}$$

the formula can be used to calculate the net charge of either the native or denatured form of the protein depending on the choice of pK_as. The pK_as used for the denatured form are from [43], and the values used for the native form are estimated from experimental data from [73], see Figure 3. The only significant change in protonation occurs for histidine which in the native form of proteins is protonated approximately 40% of the time compared to the 10% predicted from the $pK_a \approx 6$. Using the average of the pK_a is incorrect as the p is a non-linear function of pK_a. Cysteine ($pK_a \approx 8$) is difficult to model because of it's tendency to form disulfide bridges, there is no simple way to model cysteine's without knowing the structure of protein. So we use the standard value for cysteine.

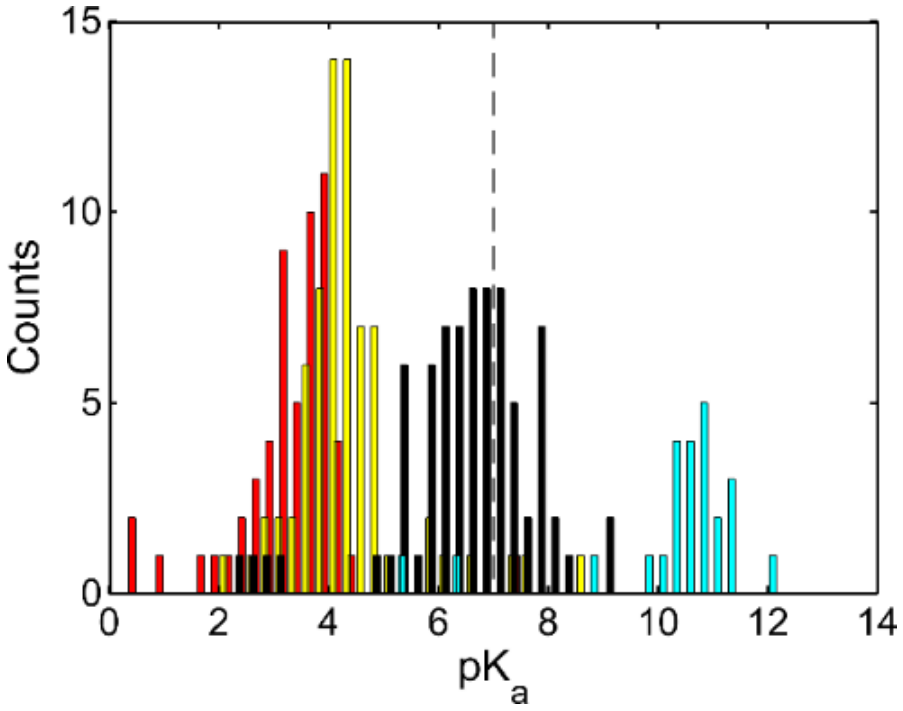


Figure 3: Experimentally measured pK_a values of side chains in the folded state of proteins. From left to right aspartic acid: red, glutamic acid: yellow, histidine: black, and lysine: cyan.

2.2 The Length Distribution of proteins in the Proteome of an Organism is Gamma Distributed

In order to calculate the stability distribution of a proteome (set of all proteins in a organism) we need to know the length distribution $p(N)$ of proteins. From the database UniProt [13], which has protein sequence data, the proteomes of 4 organisms were E. coli, S. cerevisiae, M. musculus and H. sapiens (using only proteins verified experimentally at the protein level). These four organisms were chosen because they are commonly studied organisms in aging. By binning the length of the sequences one sees that the probability distribution of protein sequence length is approximately Gamma distributed [34]

$$P(N) = \frac{N^{k-1} \exp\left(-\frac{N}{\theta}\right)}{\theta^k \Gamma(k)} \tag{5}$$

with scale parameter θ, and shape parameter k, with mean $k\theta$ and variance $k^2\theta^2$. The peak for E.coli (a prokayotic organism which is distinguished from eukaryotes by not having a nucleus) occurs around 200 amino acid residues and around 250 for the other three eukaryotic species. The values of the fitting parameters are shown in Table 2 and a plot showing the fit in Figure 4F. The greater fraction of shorter proteins in the E. coli proteome may allow for a greater portion of the proteome to fold without the help of chaperones, this is likely to result in lower average protein stability. This lower stability imposed evolutionary pressure for E. coli proteins to lower net charge per residue.

Species	k	θ	mean $= k\theta$
E. coli	2.82	105	295
S. cerevisiae	2.49	182	453
M. musculus	2.57	183	472
H. sapiens	2.59	163	421

Table 2: Species-specific parameters for protein length distributions. These parameters were determined by fitting to a Gamma distributions as shown in equation (5). Only proteins experimentally verified at the proteins level are included.

2.3 The Average Net Charge and Variance of Net Charge Depend Linearly on Protein Length

As the model predicts that the oxidative damage to charged side chains changes the stability of a protein, we need to known how many proteins are highly charge for their respective length. From the proteomes of the organisms previously mentioned in the last section 2.2. The charge distribution for all these organism is rather broad and bell shaped. The spread in charge is greater the longer the proteins are. The average and variance of net charge depend approximately linearly on the length. The fitted values for the mean folded net charge are given as

$$\begin{aligned} \mu_{Ecoli} &= 1.1 - 0.0054N && \text{(6a)} \\ \mu_{Scer} &= 4.0 - 0.0068N && \text{(6b)} \\ \mu_{Mmus} &= 1.5 - 0.0005N && \text{(6c)} \\ \mu_{Hsap} &= 2.3 - 0.0003N. && \text{(6d)} \end{aligned}$$

The fitted mean net charge for unfolded proteins as

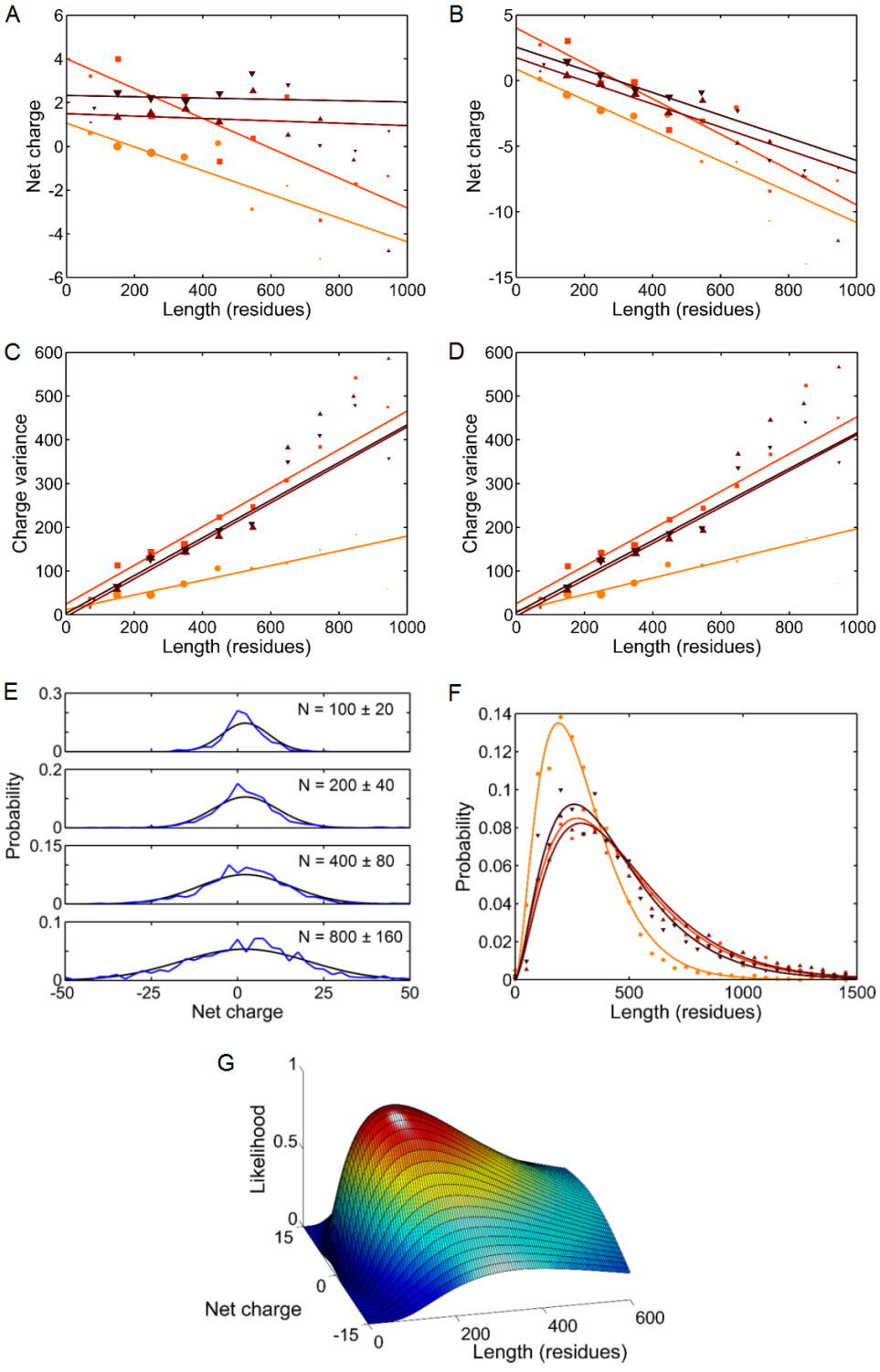
$$\begin{aligned} \mu_{Ecoli} &= 0.9 - 0.0117N && \text{(7a)} \\ \mu_{Scer} &= 4.0 - 0.0135N && \text{(7b)} \\ \mu_{Mmus} &= 1.7 - 00088N && \text{(7c)} \\ \mu_{Hsap} &= 2.6 - 0.0086N. && \text{(7d)} \end{aligned}$$

The variance in net charge for folded proteins as

$$\begin{aligned} \sigma^2_{Ecoli} &= 11.7 + 0.17N && \text{(8a)} \\ \sigma^2_{Scer} &= 24.8 + 0.44N && \text{(8b)} \\ \sigma^2_{Mmus} &= -4.5 + 0.43N && \text{(8c)} \\ \sigma^2_{Hsap} &= 3.4 + 0.43N. && \text{(8d)} \end{aligned}$$

The variance in net charge for the unfolded proteins as

$$\begin{aligned} \sigma^2_{Ecoli} &= 8.8 + 0.19N && \text{(9a)} \\ \sigma^2_{Scer} &= 25.8 + 0.43N && \text{(9b)} \\ \sigma^2_{Mmus} &= -3.0 + 0.41N && \text{(9c)} \\ \sigma^2_{Hsap} &= 5.3 + 0.41N. && \text{(9d)} \end{aligned}$$


A
Net charge
Length (residues)
B
Net charge
Length (residues)
C
Charge variance
Length (residues)
D
Charge variance
Length (residues)
E
Probability
N = 100 ± 20
N = 200 ± 40
N = 400 ± 80
N = 800 ± 160
Net charge
F
Probability
Length (residues)
G
Likelihood
Net charge
Length (residues)

These fits are shown in Figure 4A-D. Again we see a difference between the E. coli and the other three eukaryotic organisms. As was mentioned in the previous section (Section 2.2) E. coli with it's generally shorter proteome, in theory is more sensitive the higher charge and a result of evolutionary pressure has much less variance in charge. The variance in charge in E. coli is about 50% of the other proteomes meaning that the standard deviation of net charge is about 25% in E. coli for proteins of the same length. With all else being equal this would predict E. coli being less susceptible to oxidative damage compared to the other three organism.

The mean and variance are of net charge for fixed length are sufficient as the net charge is normally distributed, see Figure 4E. Previous studies on D. Melanogaster and S. cerevisiae have also observed Gaussian distributions in net charge for fixed length [36, 37]. The distribution in net charge is given as a conditional distribution

$$P(Q|N) = \frac{1}{\sigma(N)\sqrt{2\pi}} \exp\left(-\frac{(Q-\mu(N))^2}{2\sigma^2(N)}\right) \tag{10}$$

the joint distribution of charge and length is given by the product of Eq. (5) and Eq. (10)

$$P(Q,N) = P(Q|N)P(N) = \frac{1}{\sigma(N)\sqrt{2\pi}} \exp\left(-\frac{(Q-\mu(N))^2}{2\sigma^2(N)}\right) \frac{N^{k-1}\exp\left(-\frac{N}{\theta}\right)}{\theta^k \Gamma(k)} \tag{11}$$

which is plotted in Figure 4G for the parameters of H. sapiens. The narrower

Figure 4 *(preceding page)*: Species specific charge and length distributions. The each species is labeled. E.coli in light orange circles, S. cerevisiae in orange squares, M. musculus in brown upward triangles, and H. sapiens in black downward triangles. The size of each symbol indicates the size of each bin. A) average folded protein charge versus length B) average unfolded protein charge as function of length C) variance in net charge as a function of folded protein length D) variance in net charge a function of unfolded protein length E) net charge distribution for fixed protein lengths F) the Length distribution of proteins G) joint probability distribution of charge and length

distribution in charge at shorter lengths causes the peak to occur at ≈ 130 (instead of ≈ 250 as given from the marginal Gamma distribution) amino acids and a charge $+2$.

2.4 Stability distribution of Proteins

The fraction of time the protein is folded, f, can be determined by

$$f = \frac{1}{1 + \exp(-\Delta G)} \tag{12}$$

ΔG is the folding free energy. Adding $\Delta\Delta G_e$ to the free energy of folding would give the appropriate estimate after oxidative damage. The phenomenological studies of proteins suggest that the average free energy of a protein should depend linearly on the length of the chain [23, 74] but the scatter in the data has to be properly accounted. The a proteins should be stable enough to fold ($\Delta G > 0$) but not so stable that it cannot change configuration to perform its function, so there must some peak in the stability distribution reflecting this fact. The analytical calculation of Zeldovich et al. [74] suggests a skewed stability distribution for this purpose a Gamma distribution is used to estimate the empirical distribution from a high quality data on mesophiles [59]. The Gamma distribution is fit by the method of maximum likelihood using a linear dependence on length in the scale parameter. The average stability $\overline{\Delta G}(N)$

$$\frac{\overline{\Delta G}(N)}{kT} = 1.06 + 0.0369N \tag{13}$$

the fit is shown in Figure 5. The data is collapsed into one Gamma distribution (inset of Figure 5) by subtracting off $\Delta G_{min} = 1.06kT$ and then dividing by the length dependent term. Protein stabilities for an organism can be modeled by a conditional distribution $P(\Delta G|N)$ that is a gamma distribution for given length. Even though big proteins are more stable on average there are proteins of all that are barely stable and susceptible to oxidative destabilization. This is the reason why we avoid complicated combinations of $\overline{\Delta G}$ and $\Delta\Delta G$ and focus instead on the effects of oxidative damage protein based solely on $\Delta\Delta G$.

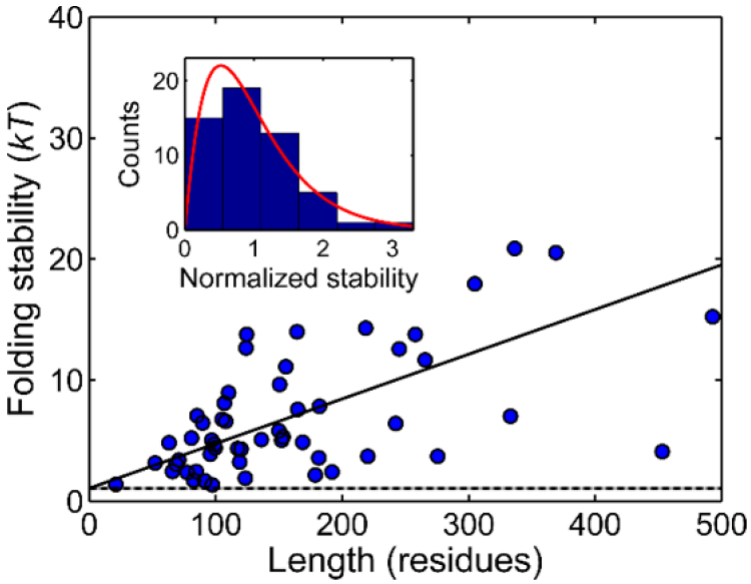


Figure 5: Experimentally measured stability of proteins for mesophilic organisms (circles), with the length dependent average stability $\overline{\Delta G}(N)$ (solid line) and length independent minimum ΔG_{min} (dashed line). The inset is the normalized histogram of experimental stabilities normalized by $\overline{\Delta G}(N)$ compared to a Gamma distribution with mean equal to one.

2.5 Estimating the Stability Distribution of a proteome from its length distribution

Although we have stability data that is not specifically from one species [59] we will use this assuming the stability distribution is somewhat universal. The joint probability $P(\Delta G, N) = P(\Delta G|N)P(N)$ can be integrated over length N for the particular organism of interest, in this case H. sapiens

$$P_{H.sap.}(\Delta G) = \int_0^\infty P(\Delta G|N)P_{H.sap.}(N)dN \tag{14}$$

where $P_{H.sap.}(N)$ is given in Eq. (5) with the parameters listed in Table 2. The resulting distribution is shown in Figure 6 where we can see a substantial difference due to the variability than what would be determined from just the in mean in Eq. (13). This suggests that the a substantial portion of human proteins are near the boundary of stability. The peak stability is about $4kT$ and a mean around $10kT$, the effects of oxidative damage are on the same

order of magnitude, as seen in Figure 7. These marginally stable proteins are also sensitive to other age-related phenomena such as DNA mutation and protein mistranslation resulting in amino acid substitution.

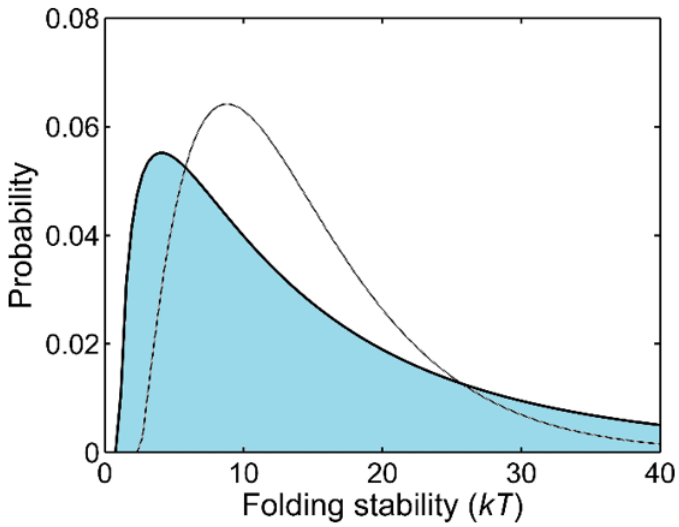


Figure 6: The estimated human stability distribution based off of equation (14) (solid line with blue shading underneath) and the (incorrect) estimate based on substituting Eq. (13) into Eq. (5) (Jacobian included)

2.6 Even if the Net Change in Charge Caused by Oxidative Damage is Zero it Will Destabilize a Proteome

As mentioned earlier oxidative damage can increase or decrease the net charge of a protein. When an oxidation event increases the magnitude of the net charge, the protein is destabilized, but the opposite can happen as well where the proteins net charge is brought closer to zero. Even if the average ΔG is unchanged due to are equal amounts of stabilizing and destabilizing oxidative damage the proteome as a whole will have a higher fraction of unfolded proteins. The nonlinear dependence of f in Eq. (12) means that for proteins with $\Delta G \leq 5kT$ would suffer a substantial change in f from a destabilizing event than a stabilizing one as f is near saturated (nearly one). The significance of this nonlinear effect can be seen by taking the stability distribution

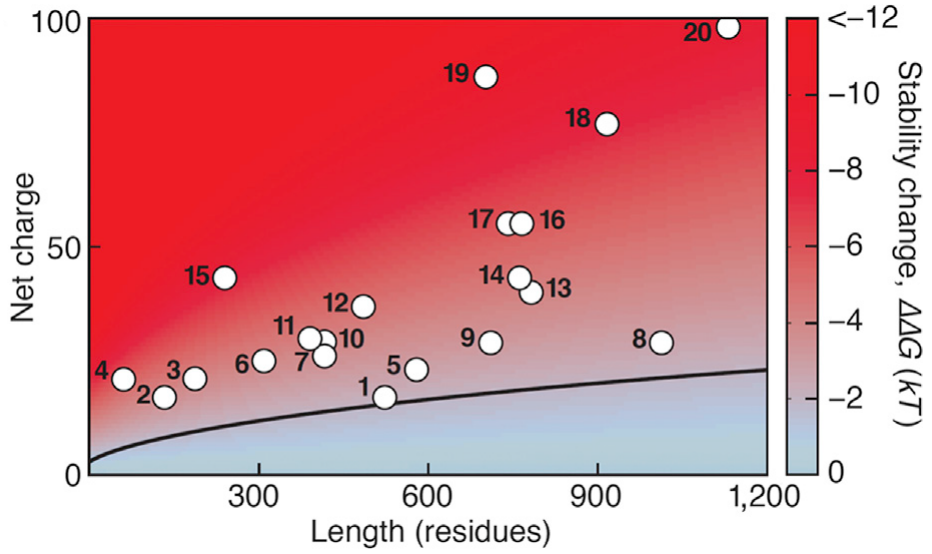


Figure 7: A colormap showing the magnitude of oxidative destabilization from Eq. (2) for the human proteome. The solid black line is one standard deviation from neutrality. The circles are proteins that have been identified as important to aging and are listed in table 3.

of the undamaged and singly-oxidized ($Q \to Q \pm 1$ with equal probability) human proteome and changing variables from ΔG to f, see Figure 8 and it's inset. Despite the number of marginally stable proteins being small they contribute substantially to the fraction of proteins that are unfolded. As a result of this nonlinearity more focus will be placed on destabilization, as it would apply greater stress on the chaperones and proteasomes responsible for folding and degrading damaged protein.

Some caveats about the calculation of destabilization from Equation (2) and the expressions for the radius of gyration, particularly of the denatured state $R_d = 1.927 N^{0.598} \mathring{A}$, should be mentioned for clarity. The radius of gyration for the native state is relatively easy to predict from empirical relations as it is a compact form. The expression for the radius of gyration of the denatured state were obtained by chemical denaturant experiments [38]. Under typical physiological conditions these would be overestimates of the denatured state radius, as it would most likely be a compact molten globular structure. Overestimating the radius of gyration of the denatured state underestimates the denatured state's contribution to the destabilization.

	Gene Name	Function	Charge	Length (aa)
1	HSF1	transcription regulator	-17	529
2	H2AFX	histone	+17	143
3	IGF1	hormone	+20	195
4	SHFM1	proteasome	-21	70
5	HSP90AA1	protein folding	-23	585
6	NFKBIA	transcription factor binding	-25	317
7	RBBP7	histone binding	-26	425
8	PARP1	poly ADP ribosylation	+29	1014
9	MTA1	histone deacetylase	+29	714
10	RBBP4	histone acetylase	-29	425
11	TERF2IP	telomere	-30	399
12	MDM2	E3 ubiquitin ligase	-37	291
13	ELN	structure	+40	786
14	TOP1	transcription regulator	+43	765
15	RPS6	ribosome	+43	249
16	APP	receptor binding	-55	770
17	SIRT1	histone deacetylase	-55	747
18	BCLAF1	transcription regulator	+77	920
19	PJA2	E3 ubiquitin ligase	-87	708
20	TERT	telomerase	+98	1132

Table 3: A set of twenty human proteins that have been implicated in aging or aging-related processes [69] that fall into the set of highly charged proteins.

The denatured state of a protein with charge and hydrophobicity profiles with a stable, folded native state are captured by a swollen molten globule-like state. The radius of this molten globule state have been observed to posses a similar length dependence as the native state but with a radius roughly 50% greater [71]. Using this relation for R_d in Equation (2) decreases $\Delta\Delta G_e$ by roughly 40%, without much of an affect of the length and charge dependence of $\Delta\Delta G_e$ effectively leave the predictions unchanged expect for a scaling.

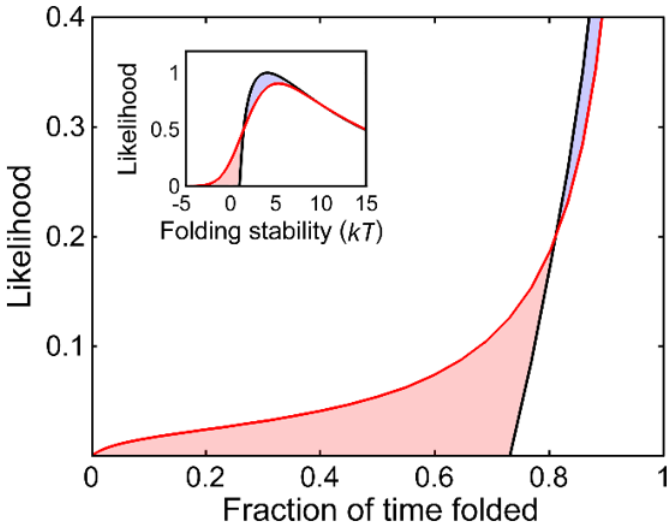


Figure 8: The inset is a comparison of the stability distribution of the human proteome when all proteins are unoxidized in black, and single oxidated in red. The main plot is the same data as in the inset but plotted against the fraction of time each protein is folded given by Eq. (12).

2.7 Proteins Sensitive to Oxidative Destabilization are Prone to Disorder Before Oxidation

The model of oxidative damage predicts that proteins which are highly charged per unit length are the most susceptible to oxidative-induced destabilization. The model also predicts that even without damage proteins with high charge density should have more difficulty folding, see Eq. (1). One can test this prediction by determining where highly charged proteins fall on an Uversky plot [71]. For our purposes highly charged proteins are define as more than two standard deviations away from neutral. An Uversky plot discribes a protein in terms of its average hydrophobicity per amino acid versus its net charge per amino acid, it has been shown to predict whether proteins have a stable folded form or are innately disordered. Only arginine, lysine, aspartic acid and glutamic acid are counted as charged amino acids in the Uversky plots. The hydrophobicity of each amino acid is defined by the Kyte-Doolittle scale in such a way that is shifted and rescaled to lie between zero and one, least to most hydrophobic respectively.

Of the 14,079 proteins in the human proteome available on UniProt, that

Function	GO term	Outliers	Total	Enrichment	p-Value
Nucleosome	0000786	51	60	12.2	$< 2.2 \times 10^{-16}$
Ribosome	0005840	85	195	6.3	$< 2.2 \times 10^{-16}$
Nucleolus	0005730	133	781	2.4	$< 2.2 \times 10^{-16}$
Telomere organization	0032200	8	62	1.9	0.064
Histone modification	0016570	31	293	1.5	1.3×10^{-2}
Transcription	0006351	188	1828	1.5	7.8×10^{-9}
DNA replication	0006260	18	198	1.3	0.15
Signaling	0023052	206	4181	0.7	1.6×10^{-10}
Lipid metabolism	0006629	21	944	0.3	1.3×10^{-11}
Precursor metabolism & energy	0006091	7	358	0.3	1.3×10^{-5}
Nucleotide synthesis	0009165	3	186	0.2	7.9×10^{-4}
Amino acid synthesis	0008652	0	94	0	1.1×10^{-3}

Table 4: The proportion of proteins that are highly charged outliers by GO categories. The highly charged outliers are defined to be more than two standard deviations away from neutrality. The total number of highly-charged proteins in the human proteome are 979 out of the 14,079 verified experimentally at the protein level. The p-value is determined by a one-tailed Fisher's exact test.

have been experimentally verified at the protein level, 80% are in the stable region of the Uversky plot, see Figure 9. This fraction decreases to 30% for the set of highly charged proteins, the ones most susceptible to oxidative damage. This supports the prediction that proteins with high charge which would be destabilized by oxidation, are less stable when unoxidized. This result does not mean that 70% of our predicted proteins lack a stable folded state, because the Uversky plot treat proteins as monomers (lacking

binding partners). Many proteins in the human proteome do not function as monomers they are part of multi-protein and RNA-protein complexes. The protein SHFM1 is an example of this with a net charge per amino acid of 0.31 and average hydrophobicity per amino acid of 0.35 puts it in the intrinsically disordered region of the Uversky plot, but it is known to have a stable form when it is a member of a larger complex.

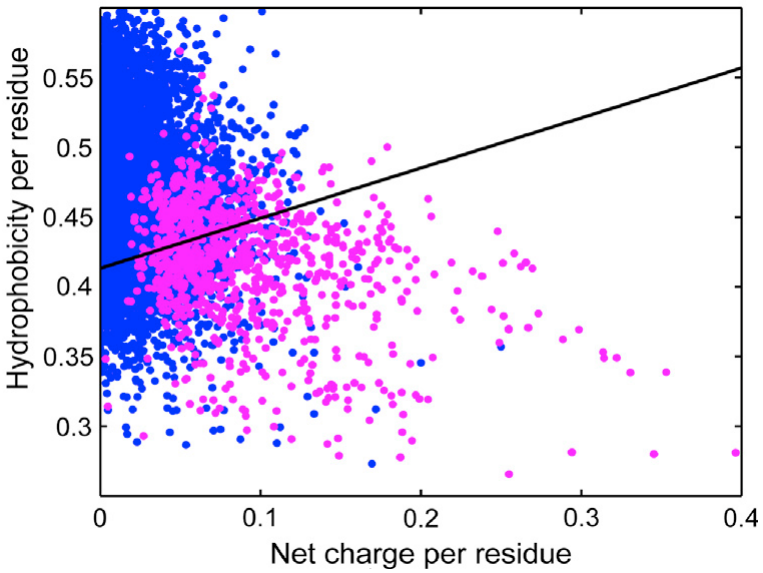


Figure 9: The black line marks the boundary between folded (above the line) and those with unstructured tendencies (below the line). The purple circles are human proteins predicted to be at higher risk (more than two standard deviations from neutrality) of which a higher portion are in unstructured region than the rest of the H. sapiens proteome, shown as blue circles.

2.8 Many Proteins in the Human Proteome are Destabilized by Oxidative Damage

Figure 7 shows key predictions from the model. The color map shows the oxidative destabilization predicted from the equation (2). The proteins that are suspected to destabilize are shorter and highly charged (in the red region). Most proteins will only be affected by less than $2kT$ by a single oxidative event, as about two-thirds of the human proteome lie in the low risk blue

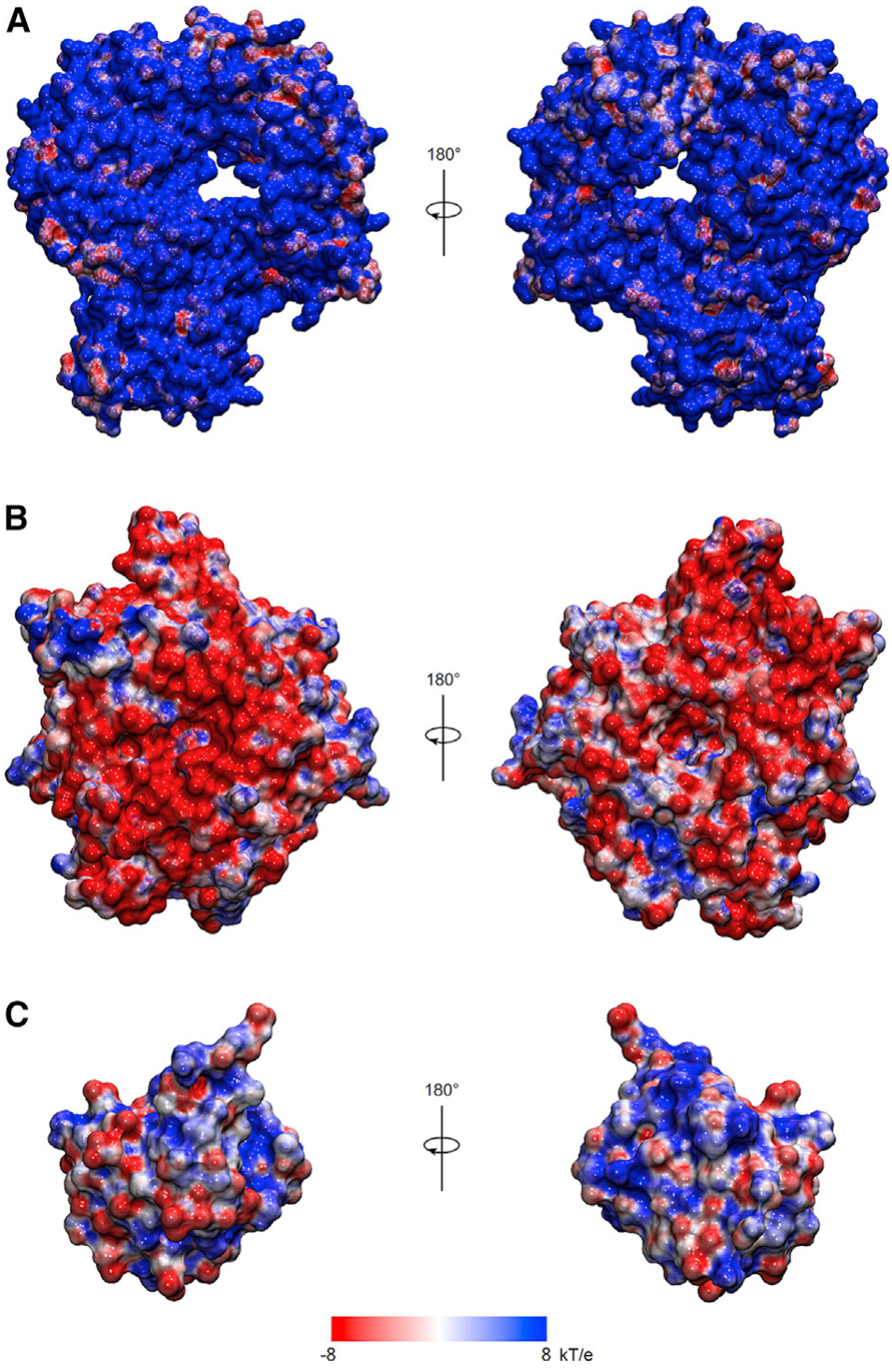
region. The twenty points in figure 7 are human proteins implicated in aging. These are listed in table 3 with their net charge and length.

The electrostatic potential of the surface of high-risk proteins differs from the lower-risk proteins as shown in figure 10. Three examples are shown in figure 10 of A) a highly positively charged telomerase reverse transcriptase (TERT) B) a negatively charged protein nucleosome-remodeling factor subunit (RBBP4) C) and the relatively neutral ubiquitin. Compared to ubiquitin, TERT and RBBP4 have almost uniformly positive and negative potentials on their surfaces, respectively. Any change net increase in the magnitude of charge, by oxidative damage or otherwise will destabilize the TERT/RBBP4 proteins locally or possibly globally due to the increase in electrostatic repulsion.

Using GO terms provided by Gene Ontology Consortium as categories [6] we investigate how these different categories compare in their proportion of high-risk proteins to the proteome as a whole. The greatest enrichment occurs for proteins mainly involved in DNA binding, see Table 4. Due to the negative charge of the backbone of DNA, the enrichment of high-risk proteins in DNA binding categories makes sense as these proteins would need a strong positive charge to bind DNA. Aging-implicated proteins are involved is the following dysfunctions altered packing of DNA around histones, abnormal histone modification, telomere destabilization, decreased transcriptional response to stress, and decreased translation and degradation of proteins [45]. The protein categories conducting these processes implicated in aging are consistent with the predictions from the model. Many of these functions are carried out in the nucleus, suggesting that keeping nucleus devoid of damaging species protects high-risk proteins from destabilization.

2.9 Experimental Point Mutation Studies and Similarity to Stability Changes from Oxidation

Debye-Hückel theory is observed to capture the change in stability from pH-induced unfolding [23] and charge-ladder experiments [25], another test of this models predictions, equation (2), is from the experimental studies on the destabilization effect of single point mutations. Ideally is would be best to compare the model with actual oxidation data composed of known oxidation sites and stability change, these experiments have yet to be conducted (most likely due to their difficulty). One can however use single point mu-


A
180°
B
180°
C
180°
-8
8 kT/e

tation data as proxy, since certain amino acids acts as good approximations to common oxidation products due to their similar hydrophobicity [54, 55]. From a mutation curated dataset [70], keeping only those mutations that: mutate charged amino acids to amino acids that are good proxies to their oxidation products, and involved solvent-exposed residues (residues accessible to oxidation). The distribution of stability changes can be compared with the distribution predicted with equations (2) and (11) for the human proteome, shown in Figure 11. If the compact denatured state were used, the predicted distribution would be narrower with a standard deviation roughly half of that shown in Fig. 11.

Not all mutation should be used as proxies for oxidation. For the purpose of predicting the effects of oxidation on stability only the mutations to methione, cysteine, alanine and theorine have been used as they are known to be similar to amino-adipic and glutamic semialdehyde due to their similar hydrophobicity. Amino-adipic semialdehyde and glutamic semialdehyde are the most abundant carbonylated products of side chain oxidation [54]. Picking only mutations of charged residues (arginine, lysine, aspartic acid and glutamic acid) to the four previously mentioned proxy residues, resulted in combined distribution of stability changes shown in Fig. 11. The average stability change of arginine, lysine, aspartic acid and glutamic acid are $-1.1kT$, $-0.3kT$, $-0.5kT$ and $-0.2kT$ respectively, with an cumulative average of $-0.5kT$. The model predicts an equal number of stabilizing and destabilizing mutations due to the assumption that increasing and decreasing charge are equally likely, but the small bias of $-0.5kT$ does not change the conclusion that the overall effect of oxidative damage is destabilizing, it in fact enhances destabilization. The width of the predicted distribution is similar to the experimental one, suggest fluctuations and the magnitude of the perturbation predicted by the model is realistic.

The comparison of the average and variance of stability change to experimental data is only appropriate if the experimental set of proteins has similar

---

Figure 10 *(preceding page)*: The surface potential of A) the positively charged telomerase reverse transcriptase (point 20 in Figure 7 and Table 3; PDB: 3KYL) and B) the negatively charged nucleosome-remodeling factor subunit RbAp48 (point 10 in Figure 7 and Table 3; PDB: 2XU7) differ greatly from the weak potential at the surface of C) the neutral protein ubiquitin (PDB: 1UBQ).

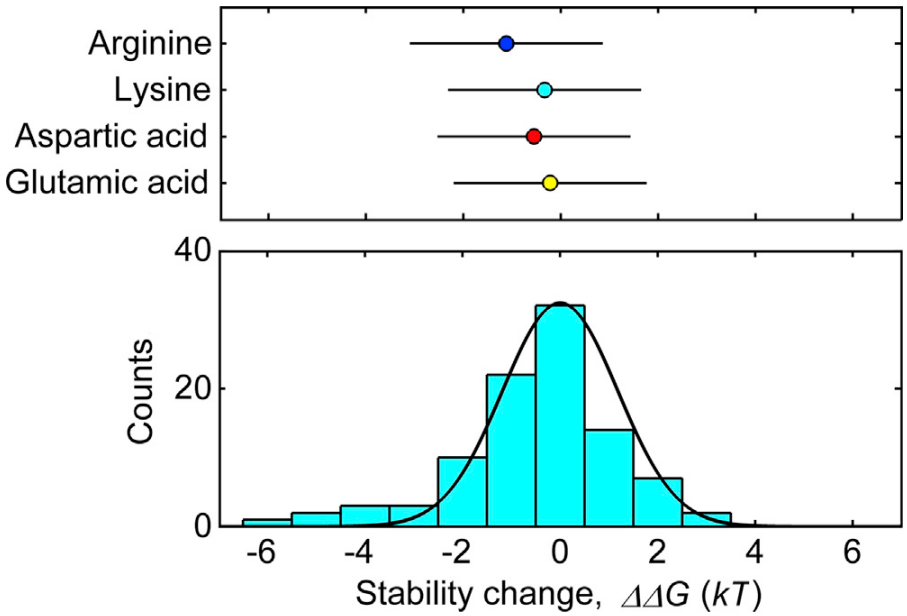


Figure 11: Experimental effect of the charge modification $Q \to Q \pm 1$ on protein stability from mutation data (histogram) compared to model prediction for the human proteome (black curve). The top is only from mutations that turned charged residues to uncharged oxidation-product analogs, used in producing the histogram. The bars are one standard deviation.

charge and length characteristics of the human proteome. Luckily, the experimental set has an average net charge per unit length close to that of the human proteome, and the set contains a similar number of mutations that bring the net charge closer to neutrality as those that bring it farther away. So it is fair to compare experimental dataset for the model's predictions.

3 Non-equilibrium Thermodynamics and the Theory of Maximum Caliber

A longstanding challenge in non-equilibrium statistical physics is to establish its root foundations. The foundation of equilibrium thermodynamics is the Second Law, which is a variational principle. It says that states of equilibrium can be predicted as those states that have maximum entropy. Microscopically, the Second Law is manifested as the Boltzmann Distribution Law. No such variational principle is yet accepted as the basis for predicting the microscopic processes in dynamical forces and flows, such as in Ficks Law diffusion, Fouriers Law of heat flow and others. Our lab has explored a principle called Maximum Caliber as the possible variational basis for non-equilibrium statistical physics.

In this chapter, we take some key steps along that road. We show that some of the major known near-equilibrium results of non-equilibrium statistical mechanics can be derived from Max Cal, and we go beyond in showing what MaxCal predicts far from equilibrium. It will start with a brief review of the method of Maximum Entropy (MaxEnt) going over several important features of constraints and their relation to the Lagrange multipliers that enforce them. Then introduce Maximum Caliber (MaxCal) as potential theory for describing the non-equilibrium statistical mechanics. The Onsager reciprocal relations and Prigogine's principle will be discussed and derived from the newer and simpler perspective of MaxCal. These arguments are then generalized to determine if similar reciprocal relations or principles hold further away from equilibrium. More general results can be obtained for far-from equilibrium systems, offering expressions in terms of fluctuations near-equilibrium and an alternative derivation of the entropy fluctuation theorem. Finally, this part ends with a discussion of the assumed known distribution $q[\Gamma]$ of dynamical fluctuations at equilibrium and difficulties of naive attempts to use MaxCal for determining it. Much of the key results from this section are summarized in my paper and the supplementary material [27].

3.1 Maximum Entropy Methods in Statistical Mechanics

This section will introduce the method of Maximum Entropy (MaxEnt) as it was originally used by Jaynes to derive the ensembles of statistical mechanics

[32, 33]. The key quantity here is the (Shannon) entropy of the probability distribution. This quantity is maximized as it is a measure of the spread of the distribution. This created the least bias distribution consistent with certain information which is known to be relevant to the system. The Shannon entropy satisfies four key properties as given by Shore and Johnson [56, 62],

- unique
- invariant under coordinate transformation
- subset independent, this means that the relative probability of two subsets of results are independent of other subsets
- system independent, if the probability distributions of two systems are independent of one another they factorize and the entropy is additive

The mathematical form of the entropy consistent with these four axioms is,

$$S = -\sum_i p_i \ln \frac{p_i}{q_i} \tag{15}$$

for a discrete probability distribution p_i with prior q_i. The Shannon entropy will be identified with the entropy of thermodynamic system. Boltzmann constant will be equal to 1 in these units, in other words temperature and energy will be measured in the same units. If we identify p_i with the probability of the state i in a physical system with energy E_i. If one knows the average energy of the system $\langle E \rangle$ it can be imposed as a constraint on the system by the typical method of a Lagrange multipliers. One would also impose the typical requirement that the probabilities p_i are normalized. With uniform prior the Lagrange function is

$$-\sum_i p_i \ln p_i - \beta \left(\sum_i p_i E_i - \langle E \rangle \right) - \alpha \left(\sum_i p_i - 1 \right) \tag{16}$$

This function is maximized in the usual way by taking derivatives with respect to the set of p_i, α and β and equating them to zero. This gives the following set of equations

$$
\begin{aligned}
-\ln p_i - 1 - \beta E_i - \alpha &= 0 && (17)\\
\sum_i p_i E_i - \langle E \rangle &= 0 && (18)\\
\sum_i p_i - 1 &= 0 && (19)
\end{aligned}
$$

which the solution for p_i is

$$p_i = \exp(-\beta E_i - \alpha - 1) = \frac{\exp(-\beta E_i)}{Z} \tag{20}$$

where $Z = \sum_i \exp(-\beta E_i)$ is the partition function. This is the Boltzmann distribution which finds its use in the canonical ensemble. These results are very easy to interpret in hindsight. With $\beta = \frac{\partial S}{\partial \langle E \rangle}$ (a property of Lagrange multipliers appearing in constrained optimization [16]) because we identified the Shannon entropy with the entropy of the system we immediately see $\beta = 1/T$ where T is the temperature of the system. Averaging equation (17) gives

$$-\alpha - 1 = \beta \langle E \rangle - S = \beta F \tag{21}$$

which when using equation (20) is equivalent to

$$F = -T \ln Z. \tag{22}$$

The cumulants of energy can be calculated by taking derivatives of $\ln Z$ with respect to $-\beta$ just as one would typically do in the canonical ensemble. Much of this example can be generalized. The microcanonical ensemble would only constrain the normalization and the states are restricted to fixed energy or other conserved quantities if they exist. For example by adding the average number of particles as a constraint to the example above much of the same arguments can be repeated yielding the grand canonical ensemble. Adding the average magnetization as a constraint instead gives the magnetic Gibbs ensemble, which is typically used in the study of the Ising and other magnetic models (rarely being recognized as a different ensemble).

3.2 MaxCal as a Generalization of MaxEnt Methods for Non-Equilibrium Statistical Mechanics

Much of the mathematics in section 3.1 still will apply in this section. The main difference between Maximum Caliber (MaxCal) and what is normally done in statistical mechanics, is that in MaxCal the probability distribution is over a space of trajectories instead of states of the system. By trajectory one means a time-ordered sequence of states. Trajectories will be labeled with Γ and the probability distribution over the trajectories $p[\Gamma]$ and a prior distribution $q[\Gamma]$. With any constraints on the average of $A_i^\Gamma(t)$, $\langle A_i(t)\rangle$, which is dependent on the trajectory Γ and the time t. It is important to note that the constraints are imposed over the time range of the trajectories so the Lagrange multipliers $\lambda_i(t)$ are functions of time and they are integrated over time. The Lagrange function for MaxCal is

$$-\int d\Gamma\, p[\Gamma]\ln\left(\frac{p[\Gamma]}{q[\Gamma]}\right)-\sum_i\left(\int\lambda_i(t)\int d\Gamma\, p[\Gamma]A_i^\Gamma(t)-\langle A_i(t)\rangle dt\right) - \alpha\left(\int d\Gamma\, p[\Gamma]-1\right) \quad (23)$$

where the integration over Γ is some properly defined sum over paths. The probability distribution consistent with these constraints using the same methods from above

$$p[\Gamma]=\frac{q[\Gamma]}{Z}\exp\left(-\sum_i\int\lambda_i(t)A_i^\Gamma(t)\,dt\right) \quad (24)$$

where $Z=\int d\Gamma\, q[\Gamma]\exp\left(-\sum_i\int\lambda_i(t)A_i^\Gamma(t)\,dt\right)$ is the dynamical partition function. The logarithm of the dynamical partition function is the cumulant generating function

$$-\frac{\delta\ln Z}{\delta\lambda_i(t)}=\langle A_i(t)\rangle \quad (25a)$$

$$\frac{\delta^2\ln Z}{\delta\lambda_i(t)\delta\lambda_j(\tau)}=\langle A_i(t)A_j(\tau)\rangle-\langle A_i(t)\rangle\langle A_j(\tau)\rangle \quad (25b)$$

the derivatives here are variational derivatives which pick out a particular time in the trajectory. As an inference method MaxCal determines the

probability of the trajectory to depends on the $A_i^\Gamma(t)$ at each time along the trajectory, so when one determines the path average $\langle A_i(t')\rangle$ for a trajectory defined along the time interval $t_a \leq t \leq t_b$, the average will weight part of the trajectory that has occurred after the time of interest (when $t > t'$) this is causality violating. So one must be careful when evaluating averages and correlations to including only the causal parts of the trajectories. Since we will be looking at systems in steady state (time-translation invariant) and evaluating expectation values when the system is microscopically time reversible the distinction between past and future is non-existent allowing one to ignore the issue.

3.3 The Onsager Reciprocal Relations

The Onsager reciprocal relations are one of the earliest and most celebrated results in Non-equilibrium thermodynamics particularly transport phenomena. The reciprocal relations are a relationship between transport coefficients in the near equilibrium regime. If there is a system with two flowing quantities J_1 and J_2, driven by thermodynamics gradients X_1 and X_2 in the near equilibrium regime they are related by

$$J_1 = L_{11}X_1 + L_{12}X_2 \tag{26a}$$
$$J_2 = L_{21}X_1 + L_{22}X_2 \tag{26b}$$

an example of this is a thermoelectric effect, heat and electricity flow in a metal is coupled, temperature and electrical potential gradients induce both heat transfer and electrical currents. The statement of the Onsager reciprocal relations is

$$L_{12} = L_{21} \tag{27}$$

more generally if one has a system with n fluxes J_i and n driving forces X_i with $i = 1, ..., n$ in the linear regime the fluxes are

$$J_i = \sum_{j=1}^{n} L_{ij}X_j \tag{28}$$

the reciprocal relations state that the matrix of transport coefficients L_{ij} is symmetric ($L_{ij} = L_{ji}$).

Microscopic reversibility is the condition that a fluctuation in a at time t followed by another fluctuation b at $t + \Delta t$ is just as likely as a fluctuation in

b at time t and a fluctuation in a at time $t + \Delta t$. Equivalently this can also be stated that every forward trajectory is just as like as its time reversed at equilibrium. Onsager's original derivation required microscopic reversibility and assumes near-equilibrium [50, 51]. This derivation is quite involved. I will show in a later section that it can be obtained with those assumption in a few lines of simple calculation from MaxCal.

3.4 Prigogine's Principle of Minimum Entropy Production

Prigogine's principle of minimum entropy production states that for systems with coupled flows near equilibrium, and some imposed driving forces, the remaining driving forces adopt values to minimize the entropy production [39]. For the example of a system with two coupled fluxes J_1 and J_2, with a fixed driving force X_1 the entropy production per unit volume σ is

$$\sigma = J_1 X_1 + J_2 X_2 \geq 0 \tag{29}$$

this can be derived by applying balance equations and local thermodynamic equilibrium as shown in chapter 15 of reference [39]. The entropy production in steady state must be positive according to the second law of thermodynamics. This would require that the matrix of linear response coefficients is positive definite. Substituting in equations (26) and the reciprocal relations gives

$$\sigma = L_{11} X_1^2 + 2 L_{12} X_1 X_2 + L_{22} X_2^2 \tag{30}$$

with X_1 fixed differentiating with respect to X_2 and setting that expression equal to zero gives

$$\frac{\partial \sigma}{\partial X_2} = 2(L_{12} X_1 + L_{22} X_2) = 2 J_2 = 0 \tag{31}$$

so the unconstrained fluxes go to zero according to Prigogine's principle. It is guaranteed to be a minimum since matrix of transport coefficients are positive definite. There is a subtlety with Prigogine's principle. The entropy production per volume σ is truncated to second order in driving forces as the calculation is done near equilibrium. The derivative of entropy production in equation (31) is a first order quantity in driving forces effectively ignoring terms of the order of σ. It is a coincidence that these mathematically inappropriate truncations lead to experimentally observed absence of flux

for unspecified driving forces, this still means that the entropy production is not minimized. Detailed criticisms of this truncation is given in writings by John Ross [29, 30, 5]. Prigogine's principle will be shown in later sections to be a specialized case of MaxCal and a more general far from equilibrium expression.

3.5 MaxCal for Steady State Systems Near and Far From Equilibrium

In this section I will re-derive results from previous sections 3.3 3.4 and show how to generalize the MaxCal results to systems further away from equilibrium. The treatment will be similar to what is found in [27]. For a spatially homogeneous system in steady state which satisfies microscopic reversibility with two coupled fluxes of J_1 and J_2 can be imposed as constraints for the averages of $j_1^\Gamma(t)$ and $j_2^\Gamma(t)$ and a prior $q[\Gamma]$ for the probability distribution of trajectories at equilibrium as having macroscopic fluxes drive the system away from equilibrium and when they become irrelevant constraints (the Lagrange multipliers equal zero and the average flux is zero as a result) the system is characterized by the prior $q[\Gamma]$. Using equation (24) with these constraints gives

$$p[\Gamma] = \frac{q[\Gamma]}{Z} \exp\left(\int \lambda_1(t) j_1^\Gamma(t) + \lambda_2(t) j_2^\Gamma(t)\, dt \right) \tag{32}$$

here we use the opposite sign convention for the Lagrange multipliers $\lambda_1(t)$ and $\lambda_2(t)$ as it will be more convenient for expanding. Equations (25) differ by this sign convention there are no negative signs appearing in the equations. The fluxes J_1 and J_2 at time $t = 0$ as we are in steady state do not depend on time, these fluxes will be expanded around $\lambda_1(t) = \lambda_2(t) = 0$

$$J_1 = \frac{\delta \ln Z}{\delta \lambda_1(0)} \approx \int \lambda_1(t) \langle j_1^\Gamma(0) j_1^\Gamma(t) \rangle_{\lambda=0} + \lambda_2(t) \langle j_1^\Gamma(0) j_2^\Gamma(t) \rangle_{\lambda=0}\, dt + \mathcal{O}(\lambda^2) \tag{33a}$$

$$J_2 = \frac{\delta \ln Z}{\delta \lambda_2(0)} \approx \int \lambda_1(t) \langle j_2^\Gamma(0) j_1^\Gamma(t) \rangle_{\lambda=0} + \lambda_2(t) \langle j_2^\Gamma(0) j_2^\Gamma(t) \rangle_{\lambda=0}\, dt + \mathcal{O}(\lambda^2) \tag{33b}$$

where the angle brackets $\langle\, \rangle_{\lambda=0}$ are averages over the equilibrium distribution $q[\Gamma]$ (equivalently the average when both λs equal zero for all time). There

are no zero order terms as $\langle j_1^\Gamma(t)\rangle_{\lambda=0}$ and $\langle j_2^\Gamma(t)\rangle_{\lambda=0}$ are zero at equilibrium. The integration over time t is taken from $-\infty$ to ∞ this will be implicit for convenient notation, this long time limit is analogous to the thermodynamic limit it will give the steady state properties of the system just as the thermodynamic limit gives the bulk properties of a system.

One can show that $\lambda_1(t)$ and $\lambda_2(t)$ are time-independent if J_1 and J_2 are time independent. The Caliber

$$\mathcal{C} = \int d\Gamma\, p[\Gamma] \ln \frac{p[\Gamma]}{q[\Gamma]} = \ln Z - \int \left[\lambda_1(t) J_1 + \lambda_2(t) J_2\right]\, dt \tag{34}$$

is a convex function of $\lambda_1(t)$ and $\lambda_2(t)$ (since $\delta \ln Z/\delta\lambda_i(t)\delta\lambda_j(\tau)$ with $i, j = 1, 2$ is positive semi-definite being a Hessian matrix). It is guaranteed a unique solution for the quantities $\lambda_1(t)$ and $\lambda_2(t)$. The time independence of this solution follows from uniqueness when applying the trial solution $\lambda_1(t) = \lambda_1$ and $\lambda_2(t) = \lambda_2$ to the condition $\partial\mathcal{C}/\partial\lambda_i = 0$ (being a maximum) for $i = 1, 2$ yielding

$$J_i = \frac{1}{\tau}\frac{\partial \ln Z}{\partial \lambda_i} \tag{35}$$

where τ is the duration of a trajectory. Using the constancy of the Lagrange multipliers λ_1 and λ_2 gives the linear response of fluxes as

$$J_1 \approx \lambda_1 \int \langle j_1^\Gamma(0) j_1^\Gamma(t)\rangle_{\lambda=0}\, dt + \lambda_2 \int \langle j_1^\Gamma(0) j_2^\Gamma(t)\rangle_{\lambda=0}\, dt + \mathcal{O}(\lambda^2) \tag{36a}$$

$$J_2 \approx \lambda_1 \int \langle j_2^\Gamma(0) j_1^\Gamma(t)\rangle_{\lambda=0}\, dt + \lambda_2 \int \langle j_2^\Gamma(0) j_2^\Gamma(t)\rangle_{\lambda=0}\, dt + \mathcal{O}(\lambda^2) \tag{36b}$$

if we identify the Lagrange multipliers (up to some constant factor) as driving forces as the fluxes have no other functional dependence the coefficients $L_{ik} = \int \langle j_i^\Gamma(0) j_k^\Gamma(t)\rangle_{\lambda=0}\, dt$ are transport coefficients. This equality is Green-Kubo formula which uses equilibrium fluctuations to calculate the response of the system to driving forces near-equilibrium. Applying microscopic reversibility gives

$$\langle j_1^\Gamma(0) j_2^\Gamma(t)\rangle_{\lambda=0} = \epsilon_1\epsilon_2 \langle j_1^\Gamma(t) j_2^\Gamma(0)\rangle_{\lambda=0} \tag{37}$$

where ϵ_1, ϵ_2 are the parities of the fluxes J_1 and J_2 respectively. For example fluxes of heat, mass and electric charge have $\epsilon = -1$ since they are flows of quantities which are unaffected by time reversal. An example of a flux with $\epsilon = +1$ would be a flux of momentum (e.g. shear stress). Integrating eq. (37) over time results in

$$L_{12} = \epsilon_1\epsilon_2 L_{21} \tag{38}$$

this is a more general result than Onsager's reciprocal relations called the Onsager-Casimir relations [7, 8].

The Caliber in equation (34) for a system with a fixed driving force λ_1 and allowing λ_2 to vary the maximum of the caliber occurs when

$$\frac{\delta \mathcal{C}}{\delta \lambda_2(\tau)} = -\int \lambda_1(t) \frac{\delta J_1(t)}{\delta \lambda_2(\tau)} + \lambda_2(t) \frac{\delta J_2(t)}{\delta \lambda_2(\tau)} \, dt = 0 \tag{39}$$

when truncated to lowest order at steady state and using the standard Onsager reciprocal relation

$$-\lambda_1 L_{12} - \lambda_2 L_{22} + \mathcal{O}(\lambda^2) \approx -J_2 = 0 \tag{40}$$

gives the same result as the Prigogine's principle as was shown in section 3.4. Equation (39) can be thought of a generalized version of Prigogine's principle which applies even when far from equilibrium.

3.6 Lack of Symmetry Relations in Higher-Order Expansions of Flux

The expansion of fluxes J_1, J_2 in terms of driving forces λ_1, λ_2 can be carried out to any order. The terms of second order in the expansion of J_1

$$\frac{\lambda_1^2}{2} \int dt d\tau \langle j_1^\Gamma(0) j_1^\Gamma(t) j_1^\Gamma(\tau) \rangle_{\lambda=0} + \lambda_a \lambda_b \int dt d\tau \langle j_1^\Gamma(0) j_1^\Gamma(t) j_2^\Gamma(\tau) \rangle_{\lambda=0} + \frac{\lambda_b^2}{2} \int dt d\tau \langle j_1^\Gamma(0) j_2^\Gamma(t) j_2^\Gamma(\tau) \rangle_{\lambda=0} \tag{41}$$

and the equivalent formula for J_2 will be shown to be zero when $\epsilon_1 = \epsilon_2 = -1$. For the time ordering $0 \leq t \leq \tau$ applying time-reversal to the third order moments of fluxes gives

$$\langle j_l^\Gamma(0) j_m^\Gamma(t) j_n^\Gamma(\tau) \rangle_{\lambda=0} = -\langle j_n^\Gamma(0) j_m^\Gamma(\tau - t) j_l^\Gamma(\tau) \rangle_{\lambda=0} \tag{42}$$

where $l, m, n = 1, 2$ applying time translation to the above expression

$$\langle j_l^\Gamma(0) j_m^\Gamma(t) j_n^\Gamma(\tau) \rangle_{\lambda=0} = -\langle j_n^\Gamma(-\tau) j_m^\Gamma(-t) j_l^\Gamma(0) \rangle_{\lambda=0}, \tag{43}$$

and microscopic reversibility for the alternative ordering $t \leq 0 \leq \tau$

$$\langle j_l^\Gamma(t) j_m^\Gamma(0) j_n^\Gamma(\tau) \rangle_{\lambda=0} = -\langle j_n^\Gamma(-\tau) j_m^\Gamma(0) j_l^\Gamma(-t) \rangle_{\lambda=0} \tag{44}$$

these show that for the integrations done in Eq (41) there will be regions of the function that have the same magnitude but opposite signs therefore canceling out. This is expected as $J_i(-\lambda_1, -\lambda_2) = -J_i(\lambda_1, \lambda_2)$, reversing the driving forces should reverse the fluxes in this case. So the next leading order non-trivial relations occur at third order, for which the terms for J_1 are integrals over fourth order cumulants of the fluxes

$$\begin{aligned}&\langle j_1^\Gamma(0) j_l^\Gamma(t) j_m^\Gamma(\tau) j_n^\Gamma(s)\rangle_{\lambda=0} - \langle j_1^\Gamma(0) j_l^\Gamma(t)\rangle_{\lambda=0}\langle j_m^\Gamma(\tau) j_n^\Gamma(s)\rangle_{\lambda=0} \\ &- \langle j_1^\Gamma(0) j_m^\Gamma(\tau)\rangle_{\lambda=0}\langle j_l^\Gamma(t) j_n^\Gamma(s)\rangle_{\lambda=0} - \langle j_1^\Gamma(0) j_n^\Gamma(s)\rangle_{\lambda=0}\langle j_l^\Gamma(t) j_m^\Gamma(\tau)\rangle_{\lambda=0}. \quad (45)\end{aligned}$$

the first order coefficients make a reappearance at the this order but application of microscopic reversibility and other symmetries of the system do not yield any useful relations between terms at third order. It is known through other methods that there are no simple relations among higher order terms in the expansion of fluxes [7, 8, 9].

3.7 Entropy Production Fluctuation Theorem from MaxCal

Let us start with Eq. (32) and divide it by its time reversed trajectory Γ_R assuming both $\epsilon_1 = \epsilon_2 = -1$

$$\frac{p[\Gamma]}{p[\Gamma_R]} = \exp\left(2\int \lambda_1(t) j_1^\Gamma(t) + \lambda_2(t) j_2^\Gamma(t)\, dt\right) \quad (46)$$

as $\lambda_i \propto X_i$ the relative probability of the forward process to the reverse is proportional to the exponent of entropy produced along the trajectory as can be seen from Eq. (29). This result is a type of entropy production fluctuation theorem (see [14] and the references therein). Rearranging Eq. (46) and integrating over paths Γ

$$\left\langle \exp\left(-2\int \lambda_1(t) j_1^\Gamma(t) + \lambda_2(t) j_2^\Gamma(t)\, dt\right)\right\rangle = \langle \exp\left(-\tau\sigma\right)\rangle = 1 \quad (47)$$

here σ is the time averaged entropy production and τ is the time of the trajectory and the angle brackets here are averages over trajectories. Using Jensen's inequality $\langle \exp(x)\rangle \geq \exp\langle x\rangle$ and Eq. (47) gives $1 \geq \exp(-\tau\langle\sigma\rangle)$ taking the logarithm and negating both sides gives

$$0 \leq \langle\sigma\rangle \quad (48)$$

which is a statement of the second law of thermodynamics. This is an interesting result in that a well known inequality can be obtained directly from fluctuation theorems like Eq. (46). Fluctuation theorems are important in that they allow for bounds to be calculated on physical processes recently this has been applied to self-replicating systems [19].

3.8 Expanding Flux Around A Point Away From Equilibrium

So far the expansions of fluxes in terms of driving force have only been carried out around equilibrium ($\lambda = 0$), but there is nothing in the formalism that prevents us from choosing another arbitrary point for expansion. We will assume here for the rest of this section that the fluxes have odd time reversal parity, $\epsilon_1 = \epsilon_2 = -1$. Taking the example from previous sections and expanding the flux around the point $\lambda^* = \lambda_1^*, \lambda_2^*$ gives

$$J_1 \approx J_1^* + (\lambda_1 - \lambda_1^*) \int \langle j_1^\Gamma(0) j_1^\Gamma(t) \rangle_{\lambda=\lambda^*} \, dt + (\lambda_2 - \lambda_2^*) \int \langle j_1^\Gamma(0) j_2^\Gamma(t) \rangle_{\lambda=\lambda^*} \, dt \tag{49a}$$

$$J_2 \approx J_2^* + (\lambda_1 - \lambda_1^*) \int \langle j_2^\Gamma(0) j_1^\Gamma(t) \rangle_{\lambda=\lambda^*} \, dt + (\lambda_2 - \lambda_2^*) \int \langle j_2^\Gamma(0) j_2^\Gamma(t) \rangle_{\lambda=\lambda^*} \, dt \tag{49b}$$

since the expansion is no longer around equilibrium there is a non-vanishing zeroth order term. With the expression in Equation (32) we can re-write the brackets as

$$\langle j_a^\Gamma(0) j_b^\Gamma(t) \rangle_{\lambda=\lambda^*} = \left\langle j_a^\Gamma(0) j_b^\Gamma(t) \exp\left(\int \lambda_1^* j_1^\Gamma(\tau) + \lambda_2^* j_2^\Gamma(\tau) \, d\tau \right) \right\rangle_{\lambda=0} \tag{50}$$

where $a, b = 1, 2$ in such an expression it would be tempting to apply microscopic reversibility since the right hand side is evaluated at equilibrium but since the exponential has an argument that is integrated over time it is not a simple reordering of factors, the result is

$$\left\langle j_a^\Gamma(0) j_b^\Gamma(t) \exp\left(\int \lambda_1^* j_1^\Gamma(\tau) + \lambda_2^* j_2^\Gamma(\tau) \, d\tau \right) \right\rangle_{\lambda=0} = \left\langle j_b^\Gamma(0) j_a^\Gamma(t) \exp\left(- \int \lambda_1^* j_1^\Gamma(\tau) + \lambda_2^* j_2^\Gamma(\tau) \, d\tau \right) \right\rangle_{\lambda=0} \tag{51}$$

this is equivalent to

$$\langle j_a^\Gamma(0) j_b(t)\rangle_{\lambda=\lambda^*} = \langle j_b^\Gamma(0) j_a(t)\rangle_{\lambda=-\lambda^*} \tag{52}$$

so it is not a simple equality between coefficients at the same values of λ_1, λ_2.

Another way of deriving the previous result is by utilizing equation (46) to determine the appropriate time reversed form. By multiplying both sides by $j_a^\Gamma(0) j_b^\Gamma(t) p[\Gamma_R]$ and averaging over trajectories

$$\langle j_a^\Gamma(0) j_b^\Gamma(t)\rangle_{\lambda=\lambda^*} = \left\langle j_b^\Gamma(0) j_a^\Gamma(t) \exp\left(-2\int \lambda_1^* j_1^\Gamma(\tau) + \lambda_2^* j_2^\Gamma(\tau)\, d\tau\right)\right\rangle_{\lambda=\lambda^*} \tag{53}$$

one must time-reverse the integrand in the right hand side. Using equation (32) one can write

$$\left\langle j_b^\Gamma(0) j_a^\Gamma(t) \exp\left(-2\int \lambda_1^* j_1^\Gamma(\tau) + \lambda_2^* j_2^\Gamma(\tau)\, d\tau\right)\right\rangle_{\lambda=\lambda^*} = \left\langle j_b^\Gamma(0) j_a^\Gamma(t) \exp\left(-\int \lambda_1^* j_1^\Gamma(\tau) + \lambda_2^* j_2^\Gamma(\tau)\, d\tau\right)\right\rangle_{\lambda=0} \tag{54}$$

the right hand side of this equality is the same as the right hand side of equation 51, establishing the result in Eq. (52).

3.9 The Dynamical Fluctuations of Equilibrium and the Distribution $q[\Gamma]$

So far nothing has been said as to what the distribution of trajectories for a system in equilibrium, $q[\Gamma]$, takes as a mathematical form. Having a mathematical form for $q[\Gamma]$ very important as the actual calculation of expectation values rely on this distribution and much of making new and useful predictions falters if this unavailable. Currently the author does not know what general form $q[\Gamma]$ for each commonly used ensemble.

For the case of the microcanonical ensemble the situation is the simplest and $q[\Gamma]$ can be determined explicitly. The microcanonical ensemble describes an isolated system with known energy E. For a Hamiltonian system its dynamics are deterministic and conserve energy E. Because of the deterministic Hamiltonian evolution there is mapping between any initial condition to a trajectory. This mapping allows the equality $q[\Gamma] = q(\Gamma(0))$ where $q(\Gamma(0))$ is

equilibrium distribution at the time $t = 0$. This basically makes the second law of thermodynamics equivalent to MaxCal. For this special case we can only have trajectories that are consistent with those provided by Hamilton's equations and the multiplicity of these trajectories is the same as that of the multiplicity of microcanonical ensemble $\Omega(E)$. The distribution $q[\Gamma]$ can be written in terms of a product of Dirac functionals (one for each coordinate represented by the shorthand of Δ)

$$q[\Gamma] = \frac{\Delta[\Gamma - z(t, \Gamma(0))]}{\Omega(E)} \tag{55}$$

where $z(t, \Gamma(0))$ is the particular trajectory in phase space starting with initial conditions $\Gamma(0)$. This means that one can integrate the equations of motion and obtain any dynamical correlations or expectation values. The most practical way of doing this would be with molecular dynamics simulations so this does not offer any new insights. This is also stifled by the nature of the microcanonical ensemble, being an isolated system it does not easily lend itself to describing a nonequilibrium steady state system, which would be exchanging matter, energy, charge, etc. with its surroundings.

The natural question to ask is: how does a bath influence the distribution of trajectories of a system at equilibrium? This can be built out a large isolated system where the majority of the degrees of freedom are assigned to a bath or a set of reservoirs and the energy can exchange between the subsystem and its baths. Equation (55) still applies to the collection of reservoirs and the system but in order to determine the distribution $q[\Gamma_s]$ for just the degrees of freedom associated with the system, Γ_s, one would need to integrate over the degrees of freedom of the bath after solving the set of equation of motion for the reservoir and system. Solving equations of motion is not the business of statistical mechanics and should be avoided like the plague. We know from the canonical ensemble that the only information we need to specify is the temperature of the bath (or equivalently the energy per degree of freedom) to obtain the Boltzmann distribution. The question now becomes do the dynamics of the reservoirs play a role in $q[\Gamma]$ or does only the temperature matter?

From the point of view of MaxCal several R.M.L. Evans has suggested a method for a system exchanging energy with a bath. R.M.L Evans considered imposing the time averaged energy as a constraint and the corresponding Lagrange multiplier being the inverse temperature as in section 3.1, claiming that for long trajectories this choice would result in the Boltzmann distri-

bution [20]. The mapping of a general trajectory to a configuration, as the length of time of the trajectory goes to infinity does not make much sense as there should be no restriction on trajectories (every trajectory is possible although some are so rare they are practically impossible). If use a Feynman like path integral with this weight of the energy the resulting two point probability distribution would similar the density matrix of the equivalent quantum mechanical problem in thermodynamic equilibrium [22]. For a single degree of freedom this would be

$$p(x_b, t_b; x_a, t_a) \propto \int_{x_a}^{x_b} Dx(t) \exp\left(-\frac{\beta}{t_b - t_a} \int_{t_a}^{t_b} \frac{m\dot{x}^2(t)}{2} + V(x(t))\, dt\right) \quad (56)$$

this can be written in terms of the solutions of the equivalent Schrödinger equation which is (see [22] for details)

$$-\frac{t_b - t_a}{\beta}\frac{\partial \psi}{\partial t} = \frac{(t_b - t_a)^2}{2m\beta^2}\frac{\partial^2 \psi}{\partial x^2} + V(x)\psi(x,t) \quad (57)$$

which is a Wick rotated Schrödinger equation ($t \to -it$) with $\hbar$ replaced by $\frac{t_b - t_a}{\beta}$. The two point probability can be written in terms of eigensolutions of the above equation $\psi_n(x) \exp\left(-\frac{\beta}{t_b - t_a}\epsilon_n t\right)$

$$p(x_b, t_b; x_a, t_a) \propto \sum_{n=0}^{\infty} \psi_n^*(x_b)\psi_n(x_a) \exp\left(-\beta\epsilon_n \frac{t_b + t_a}{t_b - t_a}\right) \quad (58)$$

it is important to keep in mind ϵ_n are not energy levels, simply eigenvalues, there are no energy levels as we are still dealing with a classical thermodynamics. In general $\psi_n(x,t)$ is complex though for one dimension we can always find a combination eigensoltions that are real. It is simple to see this does not result in the Boltzmann distribution as $\psi_0 \neq \exp(-\beta V(x))$ where ϵ_0 is the smallest eigenvalue of all the ϵ_n. So despite what would appear intuitively simple as applying the same constraints as in section 3.1 for canonical ensemble does not yield sensible results assuming all trajectories are possible no matter how improbable.

If we started with the idea of keeping the dynamics to some extent, allowing fluctuations around the classical equations of motion by constraining the action then results would be similar because the Lagrangian would replace the energy in the equations above. Constraining the Lagrangian does limit

to equations of motion as the Lagrange multiplier goes to infinity and for finite values it gives non-dissipative stochastic motion [44], but it is not a viable option to the question of "what is $q[\Gamma]$?".

Given the previous constraints and allowing all possible paths may not be viable solution to the currently posed but what if physically not all paths are possible? We will now examine what the consequences of the Kolmogorov-Arnold-Moser (KAM) theorem will have on the allowed trajectories of an integrable Hamiltonian system in contact with a thermal reservoir. Taking a classical integrable Hamiltonian system as an example with Hamiltonian $H(p_k, q_k)$, with $k = 1, 2, \ldots N$ since this system is integrable it has as many conserved quantities as there are degrees of freedom [68]. The $2N$ dimensional phase space is foliated by all the invariant N-tori which are define by the N action variables $I_j = \oint_{c_j} p_k dq_k$ (summation over repeated indices is implied, the integration is done over c_j, the unique cycles of the invariant torus). The name invariant torus comes from the fact that all trajectories starting on the torus remain on it. Now adding a reservoir which is integrable with a Hamiltonian $H_r(P_j, Q_j)$ and introducing a weak interaction term between the reservoir and system $h_{int}(p_k, q_k, P_j, Q_j)$ such that it breaks integrability, the total Hamiltonian is $\mathcal{H}(p_k, q_k, P_j, Q_j) = H(p_k, q_k) + H_r(P_j, Q_j) + h_{int}(p_k, q_k, P_j, Q_j)$. By the KAM theorem, of the original set of tori to the integrable part of the system-reservoir Hamiltonian $H(p_k, q_k) + H_r(P_j, Q_j)$ (which are the products of the invariant tori of the two independent systems) the ones with sufficiently irrational frequencies (corresponding to trajectories which are not quasi-periodic) are preserved under the perturbation $h_{int}(p_k, q_k, P_j, Q_j)$. Since trajectories cannot cross the invariant tori act as barriers closing off annular sections of the phase space. So even the assumption made before that all transitions are possible may not be generically true and may depend on how strongly coupled the system is to the bath.

Jarzynski in [31] has suggested that according to Hamiltonian mechanics where the system is in contact with several reservoirs at given temperatures, near or far from equilibrium, it satisfies the fluctuation theorem for some process sending the system between state a, represented by phase space point z_a and state b with z_b generating entropy ΔS in its surrounding reservoirs is

$$\frac{P(z_b, \Delta S | z_a)}{P(z_a^*, -\Delta S | z_b^*)} = \exp(\Delta S / k) \tag{59}$$

where z^* is the time reverse of z which only negates the momenta, and k is

the Boltzmann constant. Because $P(z_b, \Delta S|z_a)$ is a conditional probability the product of it and another transition probability like it, $P(z_c, \Delta s, |z_b)$ for example give

$$\frac{P(z_c, z_b, \Delta S + \Delta s|z_a)}{P(z_a^*, z_b^* - \Delta S - \Delta s|z_c^*)} = \exp\left(\frac{\Delta S + \Delta s}{k}\right) \tag{60}$$

this argument can be continued indefinitely for an arbitrary number of time points which are arbitrarily close together making $p(z_n, z_{n-1}, \ldots, z_1, \Delta S|z_0)$ limit to a probability functional (denoted with square brackets)

$$\frac{P[z(t), \Delta S|z(0)]}{P[z_R(t), \Delta S|z_R(0)]} = \exp(\Delta S/k) \tag{61}$$

where $z_R(t) = z^*(\tau - t)$ is the time-reversed trajectory of $z(t)$ for $0 \leq t \leq \tau$. This is similar to the fluctuation theorem in Eq. (46), but the difference is in the choice made in the defining of entropy production ΔS. Jarzynski says each reservoir is prepared in a state with known temperature and the entropy produced ΔS is calculated by

$$\Delta S = -\sum_i \frac{Q_i}{T_i} \tag{62}$$

where the sum runs over the index of all baths and $Q_i = H_i^b(y_i(\tau)) - H_i^b(y(0))$ is determined by the change in energy of each of reservoir. This definition of entropy production assumes the temperature of the reservoirs does not change as energy is added to them, which is a reasonable assumption if the bath is so large that the added/removed energy is not changing the temperature appreciably. The expression in Equation (61) is appealing as any preferred direction of time for the trajectories can be chalked up to the production of entropy in the surroundings, and it was derived from completely reversible Hamiltonian dynamics.

3.10 MaxCal Describes Dissipative Systems of Few Degrees of Freedom Experiencing Thermal Fluctuations

So far only systems with many degrees of freedom have been considered. These are the types of systems that can be given a macroscopic description, but dissipation can occur at multiple length scales, and Langevin dynamics

is typical used to model dissipative processes with few degrees of freedom by implicitly modeling a solvent as thermal noise plus a drag effect. It is important to understand what quantities are key for a mathematical description of this type dynamics. It will be shown that the relevant details are the fluctuations of the force, velocity and the work done by the random force. The results are exactly those of a path integral formulation of Langevin dynamics, a review of those techniques is given here [10].

Consider a mechanical system of a particle with a single degree of freedom which is in contact with a bath with a given temperature. Let's say that we want to know the probability, $p(x_b, v_b, t_b|x_a, v_a, t_a)$, that given a particular initial position and velocity, x_a and v_a respectively, the chance that it will have position x_b and velocity v_b at time t_b. By mathematical necessity requiring both initial and final position and velocity to be specified means that path integration must be done on a function of x, $\dot{x}$ and $\ddot{x}$ or equivalently $x(t)$, $\dot{x}$, $\dot{x}$ with the condition $v(t) = \dot{x}(t)$.

The conditional probability, $p(x_b, v_b, t_b|x_a, v_a, t_a)$, for the process considered here will depend on the equation of motion of the system and the temperature of the bath (related to fluctuations in velocity of the system). We will consider only constraints that are generally quadratic functions of the equation of motion and the velocity. There are a total of 5 constraints of this nature

- $\langle (m\ddot{x} - F)^2 \rangle$ imposed with Lagrange multipliers c_0
- $\langle v^2 \rangle$ imposed with c_1
- $\langle v(F - m\ddot{x}) \rangle$ imposed with c_2
- $\langle m\ddot{x} - F \rangle$ imposed with c_3
- $\langle v \rangle$ imposed with c_4

and two more if we include normalization of $p(x_b, v_b, t_b|x_a, v_a, t_a)$ and $p(x_a, v_a, t_a)$. $F = -\partial V(x)/\partial x = -V'(x)$ is a conservative force field on the particle and we require that the definition of velocity holds for all times $\dot{x}(t) = v(t)$, this will be imposed with a Dirac functional. By having this Dirac functional we can free switch $\dot{x}$ and v under any path integration of either $x(t)$ or $v(t)$.

The maximizing the Caliber subject to these constraints gives the path weights

$$\exp\left(-\int_{t_a}^{t_b} c_0 \left(\dot{v} + \frac{V'(x)}{m} \right)^2 + c_1 v^2 \right.$$
$$\left. + c_2 v \left(\dot{v} + \frac{V'(x)}{m} \right) + c_3 \left(\dot{v} + \frac{V'(x)}{m} \right) + c_4 v \, dt \right) \quad (63)$$

the (normalized) path integration of these over position and velocity subject to the definition $v(t) = \dot{x}(t)$ fixed by the Dirac functional gives $p(x_b, v_b, t_b | x_a, v_a, t_a)$ as

$$p(x_b, v_b, t_b | x_a, v_a, t_a) =$$
$$\int Dk(t) \int_{x_a}^{x_b} Dx(t) \int_{v_a}^{v_b} Dv(t) \, \exp\left(-\int_{t_a}^{t_b} c_0 \left(\dot{v} + \frac{V'(x)}{m} \right)^2 + c_1 v^2 + \right.$$
$$\left. c_2 v \left(\dot{v} + \frac{V'(x)}{m} \right) + c_3 \left(\dot{v} + \frac{V'(x)}{m} \right) + c_4 v - ik(t)[\dot{x}(t) - v(t)] \, dt \right) \quad (64)$$

with the corresponding Fokker-Planck type equation (derivation is shown in the following section 3.11)

$$\frac{\partial p(x, v, t)}{\partial t} =$$
$$- v \frac{\partial p(x, v, t)}{\partial x} + \frac{\partial}{\partial v} \left[\left(\frac{c_2 v}{2c_0} + \frac{V'(x)}{m} \right) p(x, v, t) \right] + \frac{1}{2c_0} \frac{\partial^2 p(x, v, t)}{\partial v^2}. \quad (65)$$

the corresponding Langevin equations

$$\frac{dx}{dt} = v \quad (66a)$$

$$\frac{dv}{dt} = -\frac{V'(x)}{m} - \frac{\zeta v}{m} + \frac{\eta(t)}{m} \quad (66b)$$

$\eta(t)$ is a Gaussian white noise driving term, which gives the last conditions

$$c_0 = \frac{\beta m^2}{2\zeta} \quad (67a)$$

$$c_1 = \frac{\zeta\beta}{2} \tag{67b}$$

$$c_2 = m\beta \tag{67c}$$

3.11 Derivation of Corresponding Fokker-Plank Equation

Completing the squares

$$p(x_b, v_b, t_b|x_a, v_a, t_a) =$$
$$\int Dk(t) \int_{x_a}^{x_b} Dx(t) \int_{v_a}^{v_b} Dv(t) \exp\left(- \int_{t_a}^{t_b} c_0 \left(\dot{v} + \frac{V'(x)}{m} + \frac{c_2}{2c_0} v + \frac{c_3}{2c_0} \right)^2 \right.$$
$$\left. + \left(c_1 - \frac{c_2^2}{4c_0} \right) \left(v + \frac{c_4}{2c_1 - \frac{c_2^2}{2c_0}} \right)^2 - \frac{c_3^2}{4c_0} - \frac{c_4^2}{4\left(c_1 - \frac{c_2^2}{4c_0} \right)} - ik(t)[\dot{x}(t) - v(t)]\, dt \right) \tag{68}$$

We can take the short time step approximation $\epsilon \to 0$ to derive the Fokker-Plank equation and because of the delta function replace $\dot{x}(t)$ with $v(t)$

$$p(x_b, v_b, t+\epsilon) = \int_{-\infty}^{\infty} dk \int_{-\infty}^{\infty} dv_a \int_{-\infty}^{\infty} dx_a \sqrt{\frac{c_0}{\pi\epsilon}} \times$$
$$\exp\left(-\epsilon \left[c_0 \left(\frac{v_b - v_a}{\epsilon} + \frac{V'(x_a)}{m} + \frac{c_2}{2c_0} v_a + \frac{c_3}{2c_0} \right)^2 + \left(c_1 - \frac{c_2^2}{4c_0} \right) \left(v_a + \frac{c_4}{2c_1 - \frac{c_2^2}{2c_0}} \right)^2 \right.\right.$$
$$\left.\left. - \frac{c_3^2}{4c_0} - \frac{c_4^2}{4\left(c_1 - \frac{c_2^2}{4c_0} \right)} \right] + ik[x_b - x_a - \epsilon v_a] \right) p(x_a, v_a, t) \tag{69}$$

the integral over k gives a Dirac delta function and the integration over x_a is easily done yielding

$$p(x_b, v_b, t+\epsilon) = \int_{-\infty}^{\infty} dv_a \sqrt{\frac{c_0}{\pi\epsilon}} \exp\left(-\epsilon \left[c_0 \left(\frac{v_b - v_a}{\epsilon} + \frac{V'(x_b - \epsilon v_a)}{m} + \frac{c_2}{2c_0} v_a + \frac{c_3}{2c_0} \right)^2 \right.\right.$$
$$\left.\left. + \left(c_1 - \frac{c_2^2}{4c_0} \right) \left(v_a + \frac{c_4}{2c_1 - \frac{c_2^2}{2c_0}} \right)^2 - \frac{c_3^2}{4c_0} - \frac{c_4^2}{4\left(c_1 - \frac{c_2^2}{4c_0} \right)} \right] \right) p(x_b - \epsilon v_b, v_a, t) \tag{70}$$

substitute $\eta = (1 - \epsilon\frac{c_2}{2c_0})v_a - v_b - \epsilon V'(x_b)/m - \frac{c_3}{2c_0}$ and expanding keeping terms of order ϵ.

$$\exp\left(-\epsilon\left(c_1 - \frac{c_2^2}{4c_0}\right)\left(v_a + \frac{c_4}{2c_1 - \frac{c_2^2}{2c_0}}\right)^2\right) =$$

$$\exp\left(-\epsilon\left(c_1 - \frac{c_2^2}{4c_0}\right)\left(\eta + v_b + \epsilon\frac{V'(x_b)}{m} + \frac{c_3}{2c_0} + \frac{c_4(1 - \epsilon c_2/2c_0)}{2c_1 - \frac{c_2^2}{2c_0}}\right)^2 \Bigg/ \left(1 - \epsilon\frac{c_2}{2c_0}\right)^2\right) \approx$$

$$\exp\left(-\epsilon\left(c_1 - \frac{c_2^2}{4c_0}\right)\left(\eta + v_b + \frac{c_3}{2c_0} + \frac{c_4}{2c_1 - \frac{c_2^2}{2c_0}}\right)^2\right)$$

$$\approx 1 - \epsilon\left(c_1 - \frac{c_2^2}{4c_0}\right)\left(\eta + v_b + \frac{c_3}{2c_0} + \frac{c_4}{2c_1 - \frac{c_2^2}{2c_0}}\right)^2 \quad (71)$$

$$\exp\left(-\epsilon c_0\left(\frac{v_b - v_a}{\epsilon} + \frac{V'(x_b - \epsilon v_a)}{m} + \frac{c_2}{2c_0}v_a + \frac{c_3}{2c_0}\right)^2\right) \approx$$

$$\exp\left(-\epsilon c_0\left(\frac{\eta}{\epsilon} - \frac{\epsilon v_a V''(x_b)}{m}\right)^2\right) \approx \exp\left(-c_0\frac{\eta^2}{\epsilon}\right) \quad (72)$$

as it will not contribute to leading order in ϵ

$$p\left(x_b - \epsilon v_b, \frac{v_b + \eta + \epsilon V'(x_b)/m + c_3/(2c_0)}{1 - \epsilon\frac{c_2}{2c_0}}, t\right) \approx p(x_b, v_b, t) - \epsilon v_b\frac{\partial p(x_b, v_b, t)}{\partial x_b}$$
$$+ \left(\epsilon\frac{c_2}{2c_0}v_b + \eta + \epsilon\frac{V'(x_b)}{m} + \frac{c_3}{2c_0}\right)\frac{\partial p(x_b, v_b, t)}{\partial v_b} + \frac{\eta^2}{2}\frac{\partial^2 p(x_b, v_b, t)}{\partial v_b^2} \quad (73)$$

$$p(x_b, v_b, t + \epsilon) \approx p(x_b, v_b, t) + \epsilon\frac{\partial p(x_b, v_b, t)}{\partial t} \quad (74)$$

at zeroth order we obtain an identity (the probability equals itself) at first order one obtains

$$p(x_b,v_b,t)+\epsilon\frac{\partial p(x_b,v_b,t)}{\partial t}\approx\int_{-\infty}^{\infty}\frac{d\eta}{1-\epsilon\frac{c_2}{2c_0}}\sqrt{\frac{c_0}{\pi\epsilon}}\exp\left(-c_0\frac{\eta^2}{\epsilon}\right)\times$$
$$\left(1-\epsilon\left(c_1-\frac{c_2^2}{4c_0}\right)\left(\eta+v_b+\frac{c_3}{2c_0}+\frac{c_4}{2c_1-\frac{c_2^2}{2c_0}}\right)^2+\epsilon\frac{c_3^2}{4c_0}+\epsilon\frac{c_4^2}{4\left(c_1-\frac{c_2^2}{4c_0}\right)}\right)\times$$
$$\left[p(x_b,v_b,t)-\epsilon v_b\frac{\partial p(x_b,v_b,t)}{\partial x_b}+\left(\epsilon\frac{c_2}{2c_0}v_b+\eta+\epsilon\frac{V'(x_b)}{m}+\frac{c_3}{2c_0}\right)\frac{\partial p(x_b,v_b,t)}{\partial v_b}\right.$$
$$\left.+\frac{\eta^2}{2}\frac{\partial^2 p(x_b,v_b,t)}{\partial v_b^2}\right] \quad (75)$$

Note that the substitution introduces a term of order ϵ from the Jacobian $(1-\frac{\epsilon c_2}{2c_0})^{-1}\approx 1+\frac{\epsilon c_2}{2c_0}$. Keeping only terms of order ϵ and removing the b subscripts

$$\frac{\partial p(x,v,t)}{\partial t}=\left(\frac{c_3^2}{4c_0}\left(1-c_1+\frac{c_2^2}{4c_0}\right)-\left(c_1-\frac{c_2^2}{4c_0}\right)\frac{c_3}{2c_0}\frac{c_4}{2c_1-\frac{c_2^2}{2c_0}}+\frac{c_2}{2c_0}\right)p(x,v,t)$$
$$-\left(c_1-\frac{c_2^2}{4c_0}\right)v^2p(x,v,t)$$
$$-v\frac{\partial p(x,v,t)}{\partial x}+\left(\frac{c_2v}{2c_0}+\frac{V'(x)}{m}+\frac{c_3}{2c_0}\right)\frac{\partial p(x,v,t)}{\partial v}+\frac{1}{2c_0}\frac{\partial^2 p(x,v,t)}{\partial v^2} \quad (76)$$

which can be grouped as

$$\frac{\partial p(x,v,t)}{\partial t}=$$
$$\left(\frac{c_3^2}{4c_0}\left(1-c_1+\frac{c_2^2}{4c_0}\right)-\left(c_1-\frac{c_2^2}{4c_0}\right)\frac{c_3}{2c_0}\frac{c_4}{2c_1-\frac{c_2^2}{2c_0}}-\left(c_1-\frac{c_2^2}{4c_0}\right)v^2\right)p(x,v,t)$$
$$-v\frac{\partial p(x,v,t)}{\partial x}+\frac{\partial}{\partial v}\left[\left(\frac{c_2v}{2c_0}+\frac{V'(x)}{m}+\frac{c_3}{2c_0}\right)p(x,v,t)\right]+\frac{1}{2c_0}\frac{\partial^2 p(x,v,t)}{\partial v^2} \quad (77)$$

requiring normalization or the conservation of probability gives the condition that $4c_0c_1=c_2^2$ and $c_3=0$.

4 Bibliometrics and the Dynamics of Publication and Citation

In this chapter, we make a statistical-physical model of the scientific citation process. It has interesting dynamics. After a research paper is initially published, that paper begins to attract citations from other subsequent publications in the same area. However, in contrast to simple physical processes which have rate coefficients that are independent of time, the citation process is one in which the rate changes with time, in multiple ways. First, in addition to having an initial citation velocity, highly regarded papers also receive accelerating citations over time. This is an example of 'the rich get richer, where the 'rich here refers to papers that are already being highly cited. Second, over longer timescales, the citation rates of even the best papers begin to fall, as newer papers become more relevant. Here, by studying a large database of the papers published by PubMed and the American Physical Society, we are able to classify papers into categories based on their citation fingerprints.

Positions in fields of science are particularly competitive, so much so that "publish or perish" has become a mantra. As a result of this pressure many scientist have turned their attention to scientometrics, and bibliometrics to study general trends in the career of scientists. The advantages of such studies could be significant, as considerable amounts of money and time are spent by organizations reviewing grant proposals and evaluating individuals for positions. Expert opinion is struggling in some places to be consistent. For example NIH percentile scores of proposals do not correlate well with their productivity unless they are highly ranked in terms of scoring [21]. A similar lacking of consistency in expert opinion has been observed with two different groups of experts reviewing the same NSF proposals. The proposals were ranked by the two groups of experts and there was poor correlation between the two rankings [12]. Any models that can provide reliable and useful projections could reduce the burden on financial resources. The vast majority of these studies focus on easily obtainable data such as citations, citation networks and publication counts, though more studies have been focused on citations rather than publication.

Various schemes and metric have been developed in order to improve our understanding of scientometrics, we will briefly mention some important results. Impact factor has been used a measure of journal quality and has

been used to quantify the quality of a scientists publications, though it has some serious flaws [4, 1, 60]. Individualized metrics such as h-index and g-index have been developed as measures of productivity and impact of a scientist's body of work [28, 18]. At the level of individual papers the work of Stringer et al. [67, 66] have developed a way to determine the eventual accumulated total of citations by a paper. Wang et al. [72] have developed a cumulative advantage theory for the citation trajectory of a paper. The distribution of citation count per paper have been shown to follow a universal log-normal distribution when normalized by a field specific average number of citations per article [57]. The very early work of Lotka has shown that the frequency of individuals with a given number of total publication goes as a power law [46, 11]. The cross-sectional studies of publication and age have shown that despite what is some might expect old age does not result in a decline in productivity for active scientists [24, 40, 41, 58].

In the following section I will give a brief overview of the datasets used for the analysis of publication and citation trends. We will see that scientists publish papers at a steady pace much like previous research has suggested but with new detail that the number of papers published each year is given by a negative binomial distribution. The reason to focus on the probability of publishing a certain number of papers a year is that it's a necessary ingredient for projecting how cited a scientist will be, as citations would be accumulated by published papers both new and old. Then I will discuss the use of birth processes to for describing (the stochastic process) of citation in scientific papers, whose evolution depends on the number of previous citations, I will show how the main result of Wang, Song and Barabási follows trivially from these process and introduce a simple model based on the two-mechanism model, and that these two models are rather similar in performance. Then I will show that we can classify papers into three different types of clusters which have different behaviors. This type of clustering is useful in making predictions as it is easier to classify papers into these clusters than it is to make very accurate predictions of citations using Bayesian methods of projection which are normally limited by the model and the trends observed in the training data.

4.1 The Data from PubMed Database and the American Physical Society

The studies performed here where done on two separate datasets. One dataset is from the American Physical Society (APS) which has both pairwise citation information for all papers published in APS journals and relevant meta-data, which can be downloaded upon request. Each paper has its own digital object identifier (DOI) along with other data such as authors names and publication data. The data from the PubMed databased was scraped. Each paper has a unique identifier and information on the authors and publication year. The author names in PudMed do not have have unique identifiers, so name clashes must be accounted for and removed.

From the APS data we can determine how citation trajectories for individual papers with a resolution of days. This allows one to get better estimates of parameters as the time between each citation, and the time to get the first citation as pieces of information, compared to taking year-wise counts. In the PubMed database many papers do not have their exact date of publication, there is merely the year of publication, so in the PubMed data we are stuck to using year-wise counts.

4.2 The Pure Birth Process and the Mathematics of a Process with Cumulative Advantage

In this section we will review a special case of birth-death processes, the pure birth process. In a birth-death process which involves an increase in the state variable by one, called birth and a decrease in the state variable by one called death. The pure birth process is useful in bibliometrics citations and papers don't disappear the is no death process (retraction of a paper is an extremely rare event). This formalism allows the user a lot of flexibility in describing bibliometric data, as long as the four defining properties of the birth process are true the rate can be refined with different functional forms to model the data. The birth process which is described in terms of a conditional probability distribution $p(N(t+\tau) - N(t) = n | N(t) = m) \equiv p_{m,n}(t, \tau)$ for the number of events $N(t)$ at the time t, has the following properties:

- There can be no events if time does not pass $N(0) = 0$
- The process is Markovian depending only on the current state $N(t), t \geq 0$

- $p_{m,1}(t, \delta t) = \lambda_m(t)\delta t + o(\delta t)$ for $m = 0, 1, 2, 3, \ldots$
- $p_{m,n}(t, \delta t) = o(\delta t)$ for $n > 1$, $m = 0, 1, 2, 3, \ldots$

here $\lambda_{m+n}(t)$ is the rate which can depend on the total number of events and the time. These conditions can be used to obtain a set of differential equations to solve for $p_{m,n}(t, \tau)$ [52], but a general solution can be obtained in terms of a recursive integral equation which is simpler and straight forward to understand. The probability of no events occurring in a time $\tau + d\tau$ is $p_{m,0}(t, \tau + d\tau)$, this is equivalent to the product $p_{m,0}(t, \tau)(1 - \lambda_m(t + \tau)d\tau)$ where $p_{m,0}(t, \tau)$ is the probability that no events occur from t to $t + \tau$ and $1 - \lambda_m(t + \tau)d\tau$ is the probability that no events occur in the interval of length $d\tau$ after the time $t + \tau$. This gives

$$p_{m,0}(t, \tau + d\tau) = p_{m,0}(t, \tau)[1 - \lambda_m(t + \tau)d\tau + o(d\tau)] \tag{78}$$

expanding $p_{m,0}(t, \tau + d\tau)$ and canceling out the first term in the expansion with the first term of the left hand side

$$\frac{dp_{m,0}(t, \tau)}{d\tau} d\tau = -\lambda_m(t + \tau)p_{m,0}(t, \tau)d\tau \tag{79}$$

which can be rearranged and integrated over τ on the interval of interest

$$p_{m,0}(t, \tau) = \exp\left(-\int_0^\tau \lambda_m(t + \tau')d\tau'\right) \tag{80}$$

with this expression the general solution for $p_{m,n}(t, \tau)$, for $n \geq 1$, can be built through recursion of the following integral equation

$$p_{m,n}(t, \tau) = \int_t^{t+\tau} p_{m,n-1}(t, u - t)\lambda_{m+n-1}(u)p_{m+n,0}(u, t + \tau - u)\, du \tag{81}$$

which can be understood as breaking the process down into three components. The first is the chance of getting $n - 1$ events in the time interval $u - t$, the next is probability of a single event between u and $u + du$ which is $\lambda_{m+n-1}(u)du$, and for the remaining time the probability that no events occur is $p_{m+n,0}(u, t + \tau - u)$ and we integrate over u to account for all possible realizations of the same process. The differential equations for which equations (80) (81) are solution for

$$\frac{\partial p_{m,0}(t, \tau)}{\partial \tau} = -\lambda_m(t + \tau)p_{m,0}(t, \tau) \tag{82a}$$

$$\frac{\partial p_{m,n}(t,\tau)}{\partial \tau} = -\lambda_{m+n}(t+\tau)p_{m,n}(t,\tau) + \lambda_{m+n-1}(t+\tau)p_{m+n-1}(t,\tau) \quad (82b)$$

for all $n \geq 1$, which are derived in reference [52] in detail (alternatively, take the derivative of equations (80) (81) with respect to τ).

For the the purposes of this thesis we will consider the case $\lambda_n(t) = \lambda(t)(an+b)$, with $a, b > 0$ and $\lambda(t) \geq 0$ for all t. In this case the equations (82) can be written compactly in terms of the probability generating function

$$P_m(z,t,\tau) = \sum_{n=0}^{\infty} p_{m,n}(t,\tau)z^n \quad (83)$$

by differentiating the generating function with respect to τ and replacing the right hand side of the above equation with equations (82) yields

$$\frac{\partial P_m(z,t,\tau)}{\partial \tau} = \lambda(t+\tau)\times \left(-\sum_{n=0}^{\infty}[a(m+n)+b]p_{m,n}(t,\tau)z^n + \sum_{n=1}^{\infty}[a(m+n-1)+b]p_{m,n-1}(t,\tau)z^n \right) \quad (84)$$

which is straight forwardly simplified to

$$\frac{\partial P_m(z,t,\tau)}{\partial \tau} = \lambda(t+\tau)\left((z-1)[am+b]P_m + az(z-1)\frac{\partial P_m}{\partial z}\right) \quad (85)$$

this partial differential equation can be solved by the method of characteristics.

$$\frac{d\tau}{d\gamma} = \frac{1}{\lambda(t+\tau)} \quad (86a)$$

$$\frac{dz}{d\gamma} = az(1-z) \quad (86b)$$

here γ is a dummy variable that parameterizes the flow which makes the above partial differential equation into an ordinary differential equation,

$$\int_0^{\tau} \lambda(t+\tau')\,d\tau' = \int_t^{t+\tau} \lambda(u)\,du = \gamma(t,\tau) \quad (87a)$$

$$z = \frac{1}{1 - k\exp(-a\gamma)} \tag{87b}$$

are the characteristics of the flow, we can solve equation (85) by writing it as an ordinary differential equation

$$\frac{dP_m}{d\gamma} = -\frac{(am+b)k\exp(-a\gamma)}{1-k\exp(-a\gamma)}P_m \tag{88}$$

which is integrated as

$$P_m(z,t,\tau) = P_m\left(t, \gamma(t,\tau) = 0\right)\left(\frac{1-k\exp(-a\gamma)}{1-k}\right)^{\frac{b}{a}+m} \tag{89}$$

replacing k in terms of z and γ from equation (87)b

$$P_m(z,t,\tau) = \left(\frac{\exp(-a\gamma(t,\tau))}{1-z[1-\exp(-a\gamma(t,\tau))]}\right)^{\frac{b}{a}+m} \tag{90}$$

where this is the probability generating function for the negative binomial distribution

$$p_{m,n}(t,\tau) = \frac{\Gamma(b/a+m+n)}{n!\Gamma(b/a+m)}[e^{-a\gamma(t,\tau)}]^{\frac{b}{a}+m}[1-e^{-a\gamma(t,\tau)}]^n \tag{91}$$

the mean number of events in the time interval τ given m events in the time t

$$\langle n(\tau)|m,t\rangle = \left(\frac{b}{a}+m\right)\left[e^{a\gamma(t,\tau)}-1\right] \tag{92}$$

this formula will be important for comparison in later sections. The negative binomial distribution is characteristic of a process with cumulative advantage where the frequency of events increases as there are more events. In the context of citations this is would model the discovery of a paper through the references of another more recent paper. In the limit that the parameter $a \to 0^+$ removes the cumulative advantage effect and the distribution reduces to a Poisson distribution.

Given data containing the times t_1, t_2, t_3, ... , t_n for each of the n events that occurs in a time interval from 0 to T such that $0 < t_1 < t_2 < t_3 < \cdots < t_n < T$ the model parameters can be extracted by maximizing the likelihood of observing the data given the model, without needing to repeat

multiple trials. This is convenient since we will be using the birth process to describe the citations of a paper and we cannot repeat history to get multiple trials, to maximize the probability $p_{m,n}(t,\tau)$. Knowing the probability of no events in a time interval τ given m events occurred in a time t, shown in equation (80) allows us to determine the probability of any number of events occurring in the time interval τ given m events occurred in a time t, $p_{m,n>0}(t,\tau) = 1 - p_{m,0}(t,\tau)$. The probability $p_{m,n>0}(t,\tau)$ is the cumulative distribution function for the waiting time distribution, since it is the chance of observing at least one event if one waits for a time τ to pass. Differentiating $p_{m,n>0}(t,\tau)$ with respect to τ will give the waiting time distribution or the probability that an event occurs after a time τ passes between τ and $\tau + d\tau$

$$f(\tau|m,t)d\tau = \lambda_m(t+\tau)\exp\left(-\int_0^\tau \lambda_m(t+u)\,du\right)\,d\tau \tag{93}$$

so the probability density of n events occurring at the times t_1, t_2, t_3, $\ldots$, t_n with the condition that $0 < t_1 < t_2 < t_3 < \cdots < t_n < T$ is simply the products of the above formula which give the likelihood function

$$\mathcal{L} = f(t_1,t_2,t_3,\ldots,t_n) = \exp\left(-\int_0^T \lambda_{N(t)}(t)\,dt\right)\prod_{i=1}^{n}\lambda_{i-1}(t_i) \tag{94}$$

note that in the integral the rate $\lambda_{N(t)}(t)$ has an implicit time dependence on the number of events that happened up to the time t. Explicitly $N(t)$ can be written as

$$N(t) = \sum_{i=1}^{n}\theta(t-t_i) \tag{95}$$

where $\theta(x)$ is the Heaviside step function which is unity for $x > 0$ and zero $x < 0$. The formula in equation (94) is general as we made no assumptions about the mathematical form of the rate. It is quite easy to interpret that maximum likelihood estimate (MLE) maximizes the rates at the points in time when events occur, but minimize the rate between the times of events.

Alternatively, if one just has count data for fixed time intervals (the units are rescaled here for convenience to be equal to 1 with no loss of generality), the likelihood function for T time intervals would be given by

$$\mathcal{L} = p(n_1, n_2, n_3, \ldots, n_T) = \prod_{i=1}^{T-1} p_{n_i, n_{i+1}}(i, 1) \tag{96}$$

where n_i is the number of events in the ith interval. Again, MLE methods can be used to obtain the most likely parameters for the process

4.3 Two-Mechanism model of Citation

Due to the work of Peterson et. al [53] in this section we propose a model that is inspired by their two mechanisms model for citation. The direct mechanism is when one person finds an article out of a group of many and decides to cite it, and the indirect mechanism is when an author finds an article through another articles references and decides to cite the referenced article. Mathematically we motivate the rate for our birth process as

$$\lambda_n(t) = \lambda_{\text{indirect}}(t) + \lambda_{\text{direct}}(t) \tag{97}$$

if we say a paper is found selected almost randomly from a body of relevant papers

$$\lambda_{\text{direct}}(t) \propto \frac{1}{N(t)} \tag{98}$$

$N(t)$ is the number of relevant papers at time t, and the indirect rate is proportional to the number of citations a paper has received at a given time $n(t)$

$$\lambda_{\text{indirect}}(t) \propto \frac{n(t)}{N(t)} \tag{99}$$

and since the number of papers increases exponentially with time $N(t) \propto \exp(rt)$ [42, 72].

$$\lambda_n(t) = (an + b)r \exp(-rt). \tag{100}$$

This choice of rate gives $\gamma(t, \tau) = \exp(-rt) - \exp(-r(t + \tau))$, the mean number of citations from Eq. (92), with $t = 0$ and $m = 0$

$$\langle n(\tau)|0, 0\rangle = \frac{b}{a}\left[\exp\left(a(1 - e^{-r\tau})\right) - 1\right] \tag{101}$$

this result will be relevant when it comes to rescaling and collapsing the citation data from the Physical Review Corpus.

For long times we can use this to determine the expected number of new citations a paper will receive given some m and t

$$\lim_{\tau\to\infty} \langle n(\tau)|m,t\rangle = \left(\frac{b}{a}+m\right)\left[\exp\left(ae^{-rt}\right)-1\right] \tag{102}$$

setting $m=0$ and $t=0$ gives the expected number of total citations

$$\lim_{\tau\to\infty} \langle n(\tau)|0,0\rangle = \frac{b}{a}\left[\exp\left(a\right)-1\right] \tag{103}$$

and the relevant time scale for this process is $1/r$ with our original motivation this will tell us the doubling time from

$$t_{2\times} = \frac{\ln(2)}{r} \tag{104}$$

a typical doubling time of 13 years would result in a $r \approx 0.05\,\text{year}^{-1}$

4.4 Log-Normally Distributed Rate Gives the Model of Wang-Song-Barabási

We also consider the rate given by

$$\lambda_n(t) = \frac{(an+b)}{t\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(\ln t-\mu)^2}{2\sigma^2}\right) \tag{105}$$

there are several reasons for this, as the probability distribution between citations for papers with a fixed total number of citations and age is log-normally distributed [72]. Additionally the age of cited literature in references of papers published in the same year are log-normally distributed [26].

The log-normal rate gives the mean number of citations from Eq. (92), with $t=0$ and $m=0$

$$\langle n(\tau)|0,0\rangle = \frac{b}{a}\left[\exp\left(a\Phi\left(\frac{\ln\tau-\mu}{\sigma}\right)\right)-1\right] \tag{106}$$

where $\Phi(x)$ is the cumulative distribution function for the normal distribution given by

$$\Phi(x) \equiv \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp(-y^2/2) dy = \frac{1}{2} \left[1 + \mathrm{erf}\left(\frac{x}{\sqrt{2}} \right) \right] \tag{107}$$

so $\gamma(t, \tau) = \Phi(t + \tau) - \Phi(t)$ for this model.

Equation (106) is a formula similar in form to a model proposed in [72] which will be referred to as the Wang-Song-Barabási (WSB) model. The formulation here differs from original in that it does not require one of their assumptions, that the number of papers increases exponentially with time. In their paper the factor b/a is set fixed to 30 as it is interpreted as the typical number of references in a paper.

The expected number of new citations given m in time t are

$$\lim_{\tau \to \infty} \langle n(\tau) | m, t \rangle = \left(\frac{b}{a} + m \right) \left[\exp\left(a \left[1 - \Phi\left(\frac{\ln t - \mu}{\sigma} \right) \right] \right) - 1 \right] \tag{108}$$

and a characteristic time scale for impact

$$T^* = \exp(\mu - \sigma^2) \tag{109}$$

this time is where the peak occurs in the log-normal in equation (105).

4.5 The Negative Binomial Distribution from the Polya Process and Other Motivations, for describing the Mechanism of Near Constant Publication

Cross-sectional studies of publication of aging scientists have shown that the publication rate of scientists is relatively constant [24, 40, 41, 58]. To describe this process from some sort of mechanistic point we need to be consistent with this observation of constant publication rate observed in previous studies. Two possible models can produce constant rate processes with over-dispersion (variance that is greater than the mean) for large values of mean publication rate, as seen in figure 12.

In the previous sections it was shown that a process with a rate that increases linearly with the number of events results in a negative binomial distribution. The Polya Process is described by a rate

$$\lambda_n(t) = \frac{n + \alpha}{t + \beta} \tag{110}$$

which gives the stationary distribution

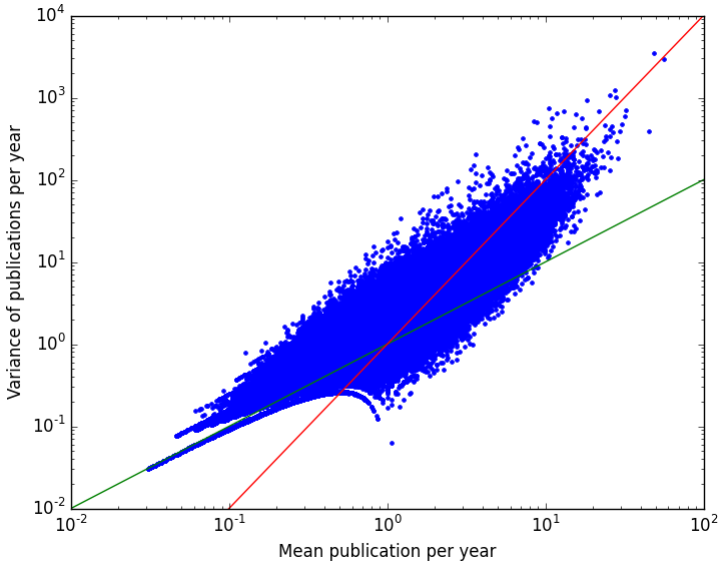


Figure 12: Yearly variance of publication versus the mean publication rate for each unique author in PubMed dataset (in blue). The green line is the curve $y = x$ and the red line is the curve $y = x^2$.

$$p_{m,n}(t,\tau) = \frac{\Gamma(\alpha+n+m)}{n!\Gamma(\alpha+m)}\left(\frac{\beta+t}{\beta+t+\tau}\right)^{\alpha+m}\left(\frac{\tau}{\beta+t+\tau}\right)^{n} \tag{111}$$

with the expected value of new events being

$$\langle n(\tau)|m,t\rangle = \tau\frac{\alpha+m}{\beta+t} \tag{112}$$

this is easy to interpret as the number of events m occurring in time t update our knowledge of the process, for long enough times t the rate of the process will be given by what was observed as it will approximate to m/t. If we where tracking the process from when it began the mean would be $\alpha\tau/\beta$ so the process can thought of as a constant rate process, such that eventually for large enough t the approximation of the negative binomial distribution to a Poisson distribution becomes exact.

Alternatively, the negative binomial distribution can be obtained if we take a Poisson process and take the rate to be distributed by a gamma

distribution

$$p(n|\alpha,p) = \int_0^\infty \frac{\lambda^n}{n!}\exp(-\lambda)\,\lambda^{\alpha-1}\frac{\exp(-\lambda(1-p)/p)}{\left(\frac{p}{1-p}\right)^\alpha\Gamma(\alpha)}\,d\lambda$$
$$= \frac{\Gamma(n+\alpha)}{n!\Gamma(\alpha)}(1-p)^\alpha p^n. \quad (113)$$

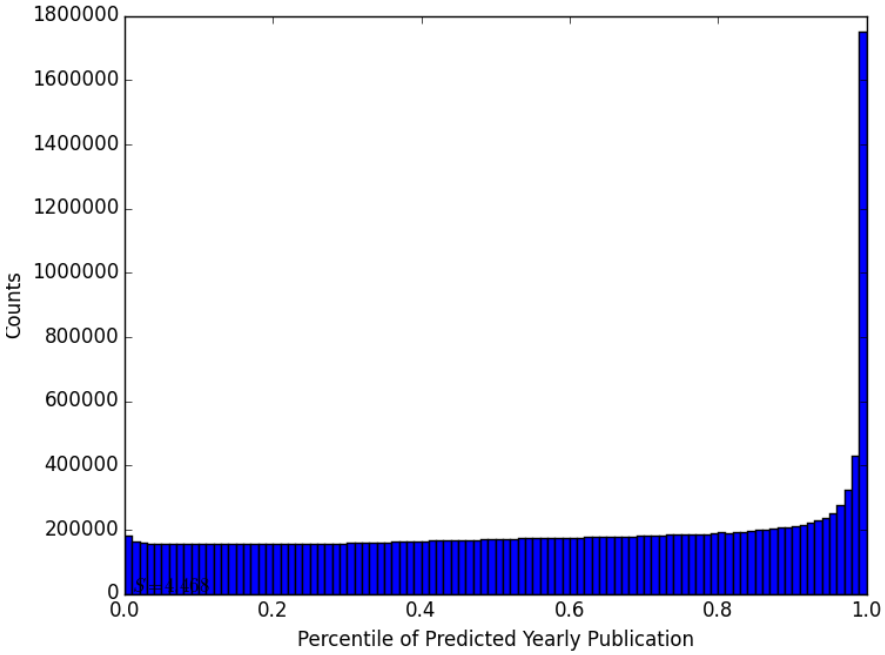
where the scale parameter of the gamma distribution $\beta = \frac{1-p}{p}$.

These two models are similar in that with no prior information besides the model parameters the expected number of publications are

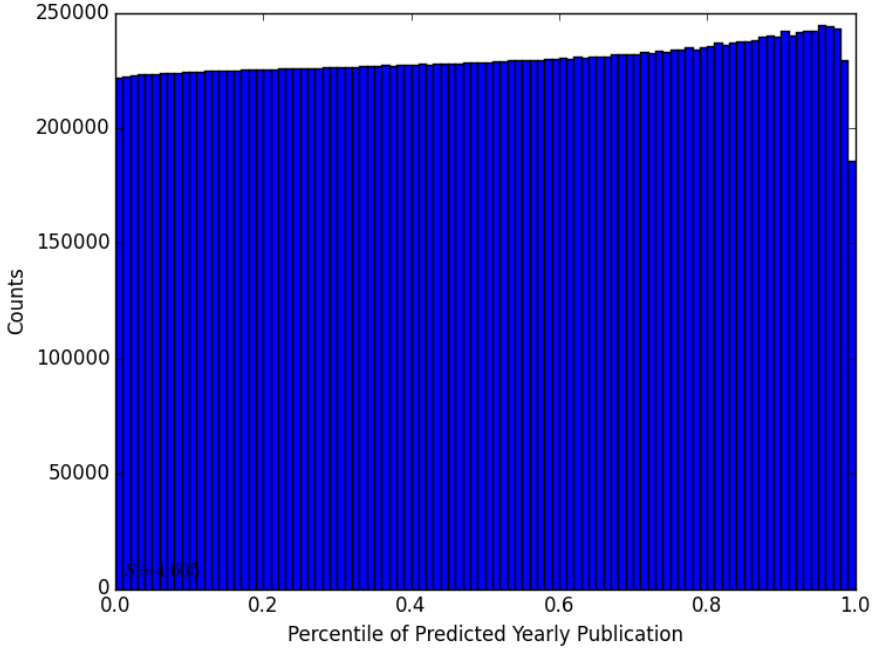
$$\langle n(\tau)\rangle = \frac{\alpha\tau}{\beta} \quad (114)$$

which is the same for both the birth process with $m = 0$ and $t = 0$ and the Poisson process with a gamma distributed rate (with $\beta = \frac{1-p}{p}$), but the two distributions are not the same.

If we use each individual's yearly paper counts to fit a negative binomial distribution given by the Polya process or the one in equation (113) we can compare each individual's publication records by the percentiles predicted by these models. First we extract the model parameters for each individual with 15 or more years of data using maximum likelihood estimation. As the negative binomial probability distribution is discrete the percentiles will be discrete too. In order to get a uniform distribution (assuming the model fits the data properly) when plotting the histogram of the percentiles, one must evenly distribute the count across multiple bins. Take for example the year with n publications, the percentiles of n and $n-1$ are calculated from the cumulative distribution function, and this year is even distributed across all bins that fall in this range. The results of these plots are shown in figure 13. Most of the deviation from these models is at the highest percentiles where in the Polya process about 8% of the data is underestimated and in the time independent negative binomial process (Eq. (113)) where less than 1% of the data is overestimated. Ideally the entropy would be $\ln(100) \approx 4.6051$ for the Polya process it is 4.468 and the time independent process is it about 4.605. From these results we can see that the yearly publication of scientists is well approximated by a negative binomial distribution with time independent parameters. This is agrees with previous studies that indicate that the productivity of scientists is relatively constant with time [40, 41, 58].



(a) Polya Process



(b) Negative binomial with time independent parameters.

It has been observed that the total number of articles an author produces is power law distributed, indicative of a cumulative advantage this is very much the case in the PubMed database, see figure 14. Of the two models used to fit the data the Polya process was lacking despite it have the better motivation based off of cumulative advantage. This would either mean that there is a fluctuating rate in publication or that the cumulative advantage effect only lasts for a short amount of time (about a year or significantly less) and then resets. Both of these would approximate to the time independent negative binomial distribution. As the PubMed data is limited to yearly counts there is no way to extract more information to distinguish between these two models.

Using the data from APS after accounting for clashing author names similar features are observed as in the PubMed dataset. The same over

---

Figure 13 *(preceding page)*: The distribution of percentiles for publications per year for each author in the PubMed database with more than 15 years of data, with the percentiles calculated from the a) Polya Process and b) equation (113). The more uniform the percentiles are distributed the better the model describes the data.

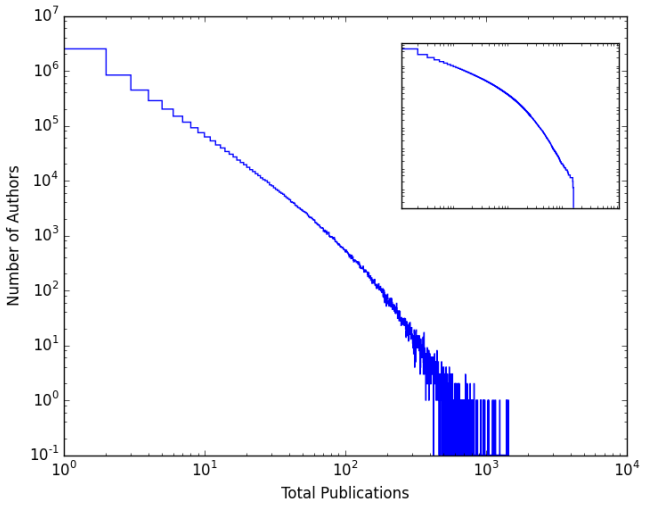


Figure 14: The distribution of total publications per author from PubMed. The inset is the survival function.

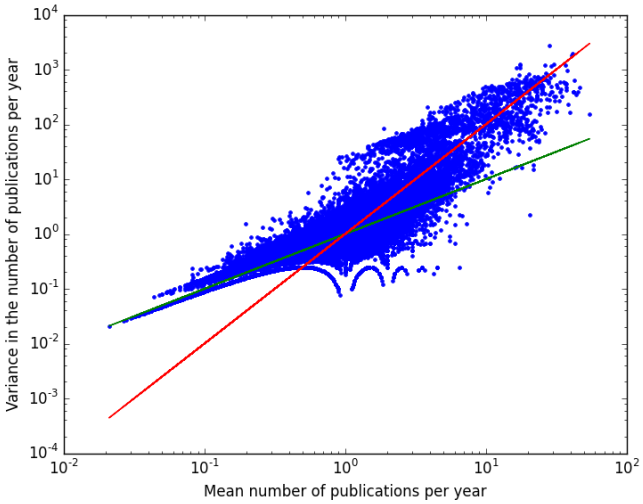


Figure 15: Yearly variance of publication versus the mean publication rate for each unique author in APS dataset (in blue). The green line is the curve $y = x$ and the red line is the curve $y = x^2$.

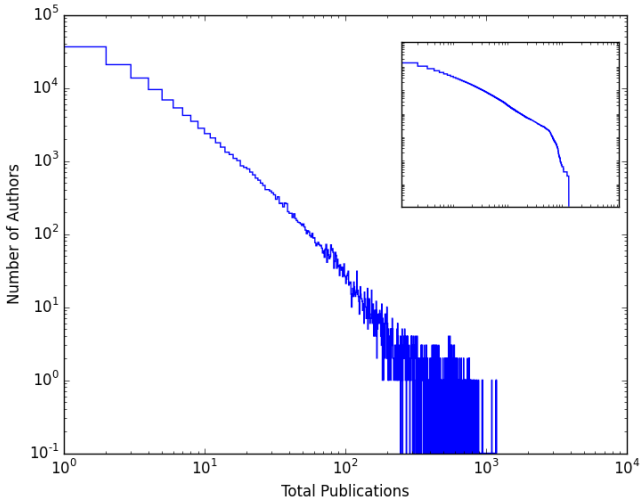


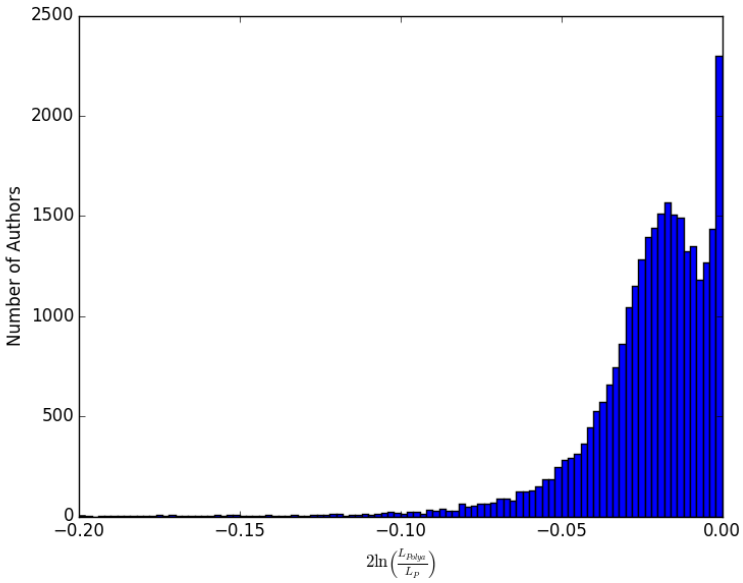
Figure 16: The distribution of total publications per author from APS. The inset is the survival function.

dispersion observed in figure 12 occurs in APS set (see figure 15) and power law tails are observed in both the set (see figure 16). Since we can resolve the citations down to the scale of days we can improve our estimation of the parameters for the models. So we compared how these models perform compared to a Poisson process with constant rate in figure 17. The Polya process performs the same or worse than the Poisson as the G-score (twice the logarithm of the likelihood ratio) is negative so there is no reason to even consider the Polya process above a Poisson process. The model in equation 113 performs better than to about the same as a Poisson process. The question of whether or not to accept the negative binomial distribution is not quite as easy as doing a likelihood ratio test since the number of observations are not enough to take the limit of a χ^2-distribution.

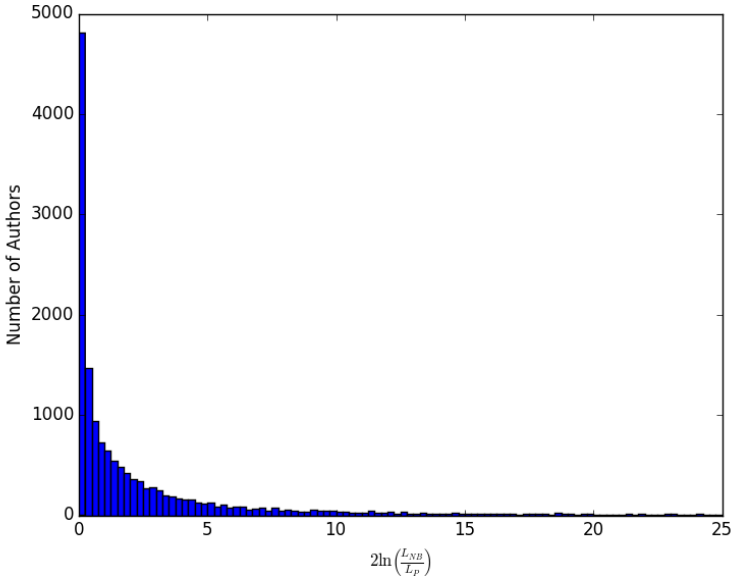
4.6 The Vast Majority of Citation Histories can be Reduced to a Few Parameters

The citation information provided by the APS dataset is easily fit two both models in sections 4.3 and 4.4 by maximum likelihood estimation (MLE) using the likelihood given in equation (94). From the parameters obtained from MLE one can rescale the data using by manipulating the expected number of citations from equations (101) and (106), such that the expected trajectories are straight lines with slope of unity and intercept of zero. Plots of these rescaled trajectories are seen in figure 19 for all papers with 30 or more citations. The a lower limit was placed on the parameter a in both models of 10^{-8} to avoid dividing by zero. There is a particularly strong power law like trend in the parameters a and b in WSB model as determined by MLE which goes over several orders of magnitude. The MLE values of a and b are compared to $b = 30a$ which reflects that in the work of D. Wang, et al. [72] that $m = 30$.

Comparing the two models it seems that the WSB model performs better than the direct-indirect mechanism though the WSB model based on the birth process has one more parameter so improvement in the fit is to be expected. A better way of comparing the two models would be comparing the likelihoods. The distribution of G-score (twice the logarithm of the likelihood ratio with the WSB model in the numerator) is shown in figure 20. We can see that the distribution falls very closely to zero despite the WSB model having one additional degree of freedom it fails to outdo the direct-indirect



(a) Polya Process



(b) Negative binomial with time independent parameters.

Figure 17: Histogram of G-score (twice the log-likelihood ratio) for each author in the APS dataset from the a) Polya Process and b) equation (113).

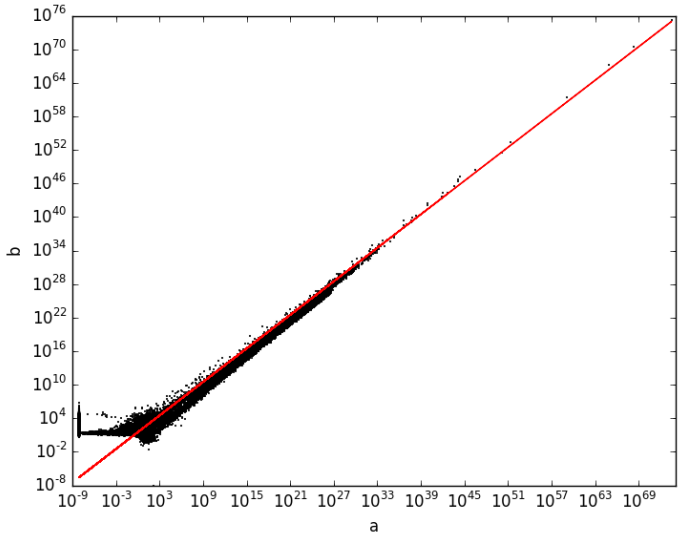


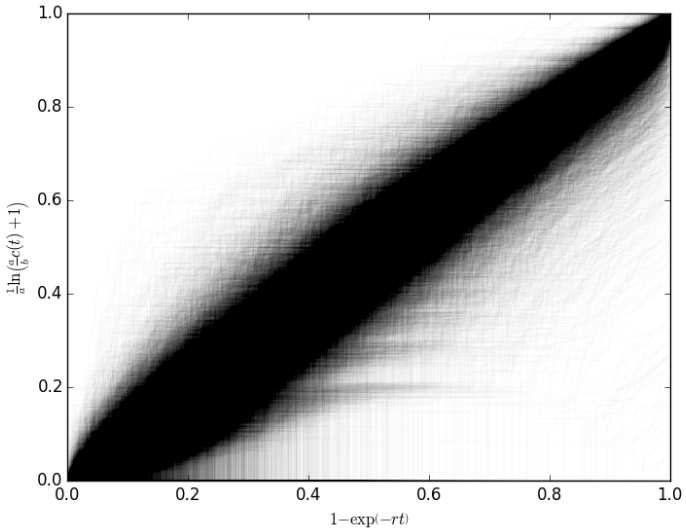
Figure 18: The scatter of b versus a as determined by MLE for the WSB model. The red line is the line $b = 30a$ which reflects the assumption in the work of D. Wang, et al. [72] that $m = 30$ in all their calculations.

model in terms of likelihood in about 28% of cases and the vast majority of G-score fall so close to zero that the addition of a parameter does not yield much improvement. As the direct-indirect mechanism is much simpler and faster to obtain parameters we will focus on it primarily.

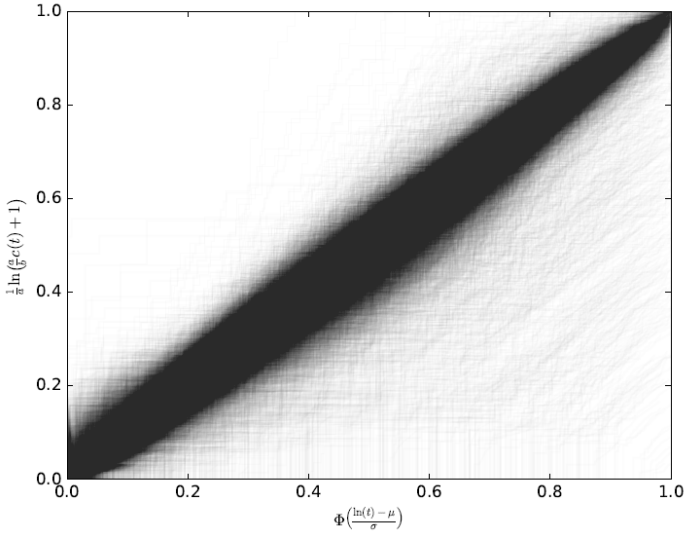
4.7 Clustering Papers Based on Model Parameters May Give Insight into Future Performance

Determining the parameters for a paper with a few citations is a difficult problem as estimates can have rather wide distributions, this can be a problem for making accurate statistical predictions. A better method is to classify papers based on a rough idea of their behavior. This can be achieved if we look at the order of magnitude of the parameters (the logarithm of their values) and cluster papers by their similarity to other papers. We can then use relatively short training periods to classify papers with a good level of accuracy.

Having fit the papers with at least ten citations with the direct-indirect



(a) Direct-indirect mechanism



(b) Wang-Song-Barabási model

Figure 19: Plots of rescaled citation trajectories with a minimum of 30 citations, each trajectory is rescaled based off of the expected number of citations from (a) the direct-indirect mechanism and (b) Wang-Song-Barabási model using the maximum likelihood estimates for the parameters of each model. With these rescaling the data should fall mostly along the line with slope of unity and intercept of zero (y=x).

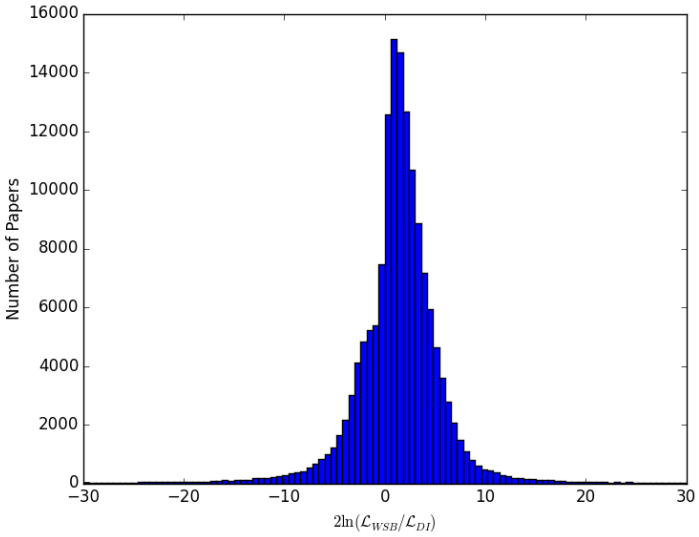
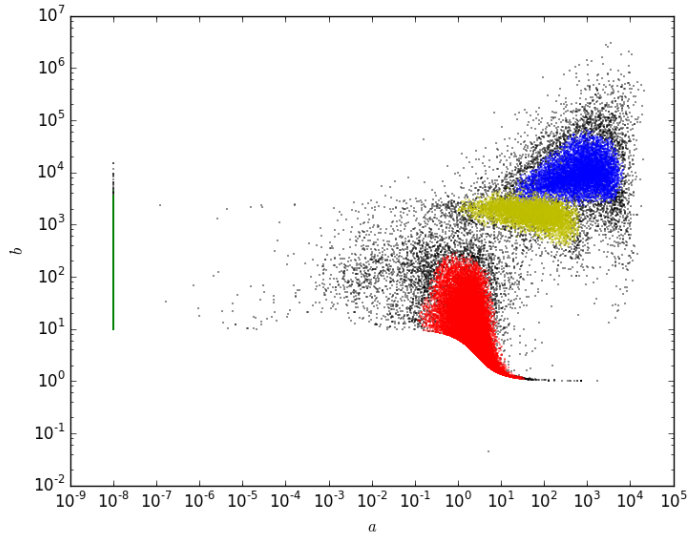


Figure 20: Histogram of the G-scores of the WSB model and the direct-indirect mechanism.

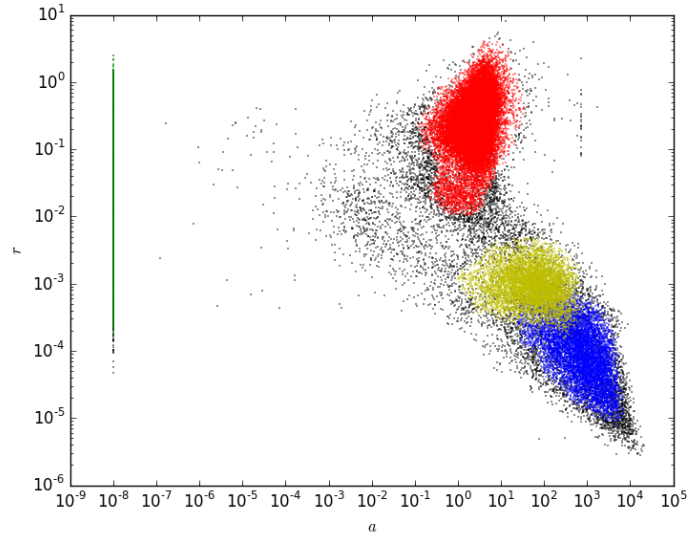
(DI) model we can make an attempt to cluster the papers into groups characterizing them by some similarity in behavior. I propose that you can cluster the papers by Euclidean distance between the logarithm of the parameter values (linear distance on a log-scale). This compares papers with parameters that are of similar order of magnitude. We utilized a density based method for clustering the papers called DBSCAN.

The papers mostly fall into one of 4 clusters shown in figure 21. These 4 clusters have different typical behavior which is shown in figure 22 for the ten most cited papers in each cluster. One can see that the red cluster (cluster 0) has the most citations and they all occur relatively quickly. The green cluster (cluster 1) reaches their respective maxima relatively quickly as they as cited mostly in their first ten years. The yellow and blue clusters (clusters 2 and 3 respectively) have rather similar long term growth, the blue cluster having a slower response.

The composition of the journals in each cluster was examined compared to the whole, p-values where calculated to test the significance using a Fisher exact test at a significance level of $0.01/(5 \times 12) \approx 1.67 \times 10^{-4}$. Dividing by 60 accounts for the multiple times the hypothesis is tested reducing the



(a) b vs. a



(b) r vs. a

Figure 21: Scatter plot showing the clustering of papers based on the DBSCAN algorithm with $\epsilon = 0.406$ and a minimum threshold of 50 papers projected onto (a) the b and a axes and (b) the r and a axes. The different clusters are color coded as red, green, blue and yellow. The black points are papers categorized as noise.

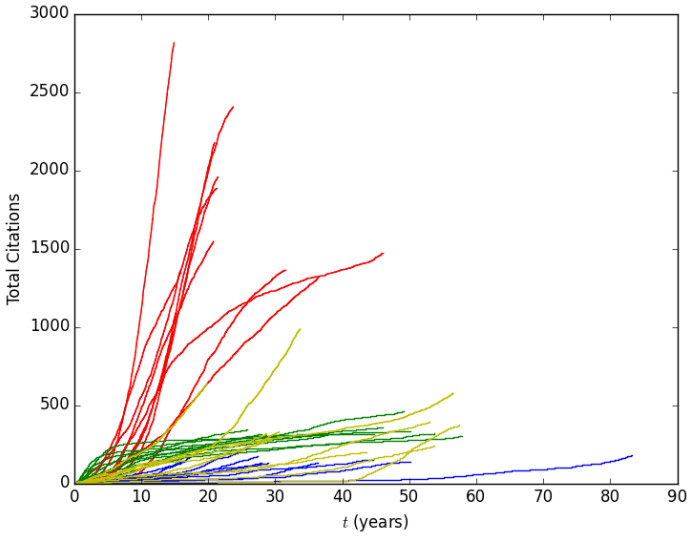


Figure 22: Citation trajectories of the ten most cited papers in each cluster. The colors indicate the clusters the papers belong to and are the same as figure 21

odds of accepting the alternative by chance. The raw counts of papers by their journal and cluster are shown in table 6 as well as the whole. The enrichment of certain journals the clusters compared to the whole is shown in table 5. The p-values appear in table 7. Looking at cluster 0 (the one in red) we notice a significant enrichment in papers from Phys. Rev. B, Phys. Rev. Lett., and Rev. Mod. Phys. which are well regarded journals in the Physical Review Corpus. Where as these same journals are lacking papers to a significant degree in cluster 1 (green), and it is the only cluster enriched in papers from Phys. Rev. which existed before the split in 1970. Clusters 2 and 3 (yellow and blue) have the fewest papers from before the 1970 split, comparatively, and these clusters have far more papers from Phys. Rev. A, C, D and E. If we are to interpret cluster 0 as the one with the fastest and highest impact (in terms of having the most citations), cluster 1 having the lowest impact with rapid saturation (reaches obsolescence quickly) this comes along with the negligible values of $a = 1 \times 10^{-8}$ (this was the lower bound used to avoid a divide by zero error) which indicate weak cumulative advantage effect, and clusters 2 and 3 as having slower long term impact, then the

Journal	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Noise
PhysRev	0.786275	1.185251	0.212699	0.191920	1.661312
PhysRevA	0.944248	1.009865	1.156630	1.233410	0.867695
PhysRevB	1.098570	0.977874	0.962043	0.907453	0.766124
PhysRevC	0.881826	0.997911	1.485603	1.421313	0.897456
PhysRevD	0.809020	1.028589	1.363962	1.231208	1.279976
PhysRevE	0.991213	0.952417	1.575415	1.588778	0.513525
PhysRevLett	1.081608	0.970159	0.920078	0.994517	0.928858
PhysRevSTAB	0.812700	0.735513	3.726188	2.893501	0.842557
PhysRevSTPER	0.335239	0.532798	2.049403	4.774277	6.255983
PhysRevSeriesI	0.670477	0.710397	0.000000	0.000000	8.341311
PhysRevX	1.005716	0.355199	0.000000	0.000000	10.426639
RevModPhys	1.281528	0.635657	0.679437	1.098278	3.767135

Table 5: Relative ratios of journal compositions of each cluster compared to the whole. Values greater than one are enriched while values less than unity are unenriched. The cells highlighted in yellow are statistically significant at a significance level of 1.67×10^{-4}.

enrichments observed also tell us a similar story. The fastest growing subfield in physics is condensed matter, which is published primarily in Phys. Rev. B since 1970 which is enriched in cluster 0. The other subfields of atomic and molecular, nuclear, particle, and statistical physics (Phys. Rev. A, C, D, and E respectively) appear more enriched in clusters 2 and 3. If the reader is not convinced that condensed matter is the fastest growing subfield of physics the number of Ph.D.s awarded in last few years is nearly double of its closest competitor particle physics [47, 48, 2].

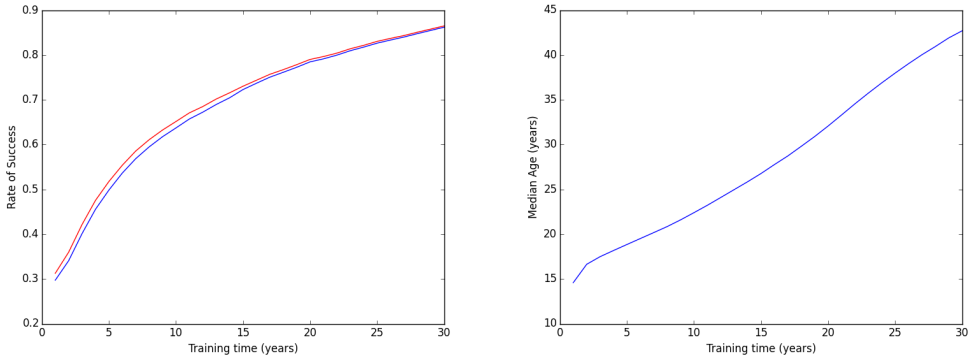
Using a training period limited to the first few years of each paper whose age exceeds that limit had its parameters calculated and then clustered. Then these were compared to the clusters determined by their full history and a success rate was determined based on how many where correctly classified based on the training period. We used training periods of 1 to 30 years. With only five years of training we can successfully place the papers fit in their correct clusters about half the time (49.85%) and the success monotonically increases with the amount of training time see figure 23. The relatively high rate of success is not due to papers with ages close to the training period. For example, the set with a training period of 30 years, over half the papers

Journal	whole	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Noise
PhysRev	13682	3209	9131	142	110	1090
PhysRevA	15309	4312	8705	864	791	637
PhysRevB	39985	13103	22016	1877	1520	1469
PhysRevC	7877	2072	4426	571	469	339
PhysRevD	18361	4431	10634	1222	947	1127
PhysRevE	6010	1777	3223	462	400	148
PhysRevLett	48246	15566	26355	2166	2010	2149
PhysRevSTAB	99	24	41	18	12	4
PhysRevSTPER	10	1	3	1	2	3
PhysRevSeriesI	5	1	2	0	0	2
PhysRevX	20	6	4	0	0	10
RevModPhys	1478	565	529	49	68	267

Table 6: Raw Counts of Papers with at least ten citations each within each cluster compared to the whole.

Journal	p-value 0	p-value 1	p-value 2	p-value 3	p-value Noise
PhysRev	4.60e-40	2.30e-40	4.48e-139	1.11-128	1.18e-57
PhysRevA	1.87e-05	6.58e-03	1.45e-06	4.28e-10	1.42e-05
PhysRevB	1.20e-28	4.86e-05	1.82e-03	8.71e-07	5.56e-34
PhysRevC	1.08e-08	1.19e-02	1.13e-19	8.25e-14	2.98e-03
PhysRevD	1.12e-43	3.90e-04	5.69e-28	1.07e-11	1.17e-17
PhysRevE	1.35e-02	1.01e-03	3.55e-20	2.01e-18	3.84e-20
PhysRevLett	4.13e-26	6.09e-08	2.47e-07	1.10e-02	2.67e-06
PhysRevSTAB	7.27e-02	2.51e-02	7.24e-06	1.35e-03	1.95e-01
PhysRevSTPER	2.03e-01	1.95e-01	3.47e-01	8.36e-02	2.18e-02
PhysRevSeriesI	4.36e-01	3.89e-01	7.88e-01	8.14e-01	4.64e-02
PhysRevX	2.35e-01	2.81e-02	3.86e-01	4.40e-01	6.96e-07
RevModPhys	1.33e-07	5.60e-21	1.05e-03	3.92e-02	8.71e-67

Table 7: The p-values corresponding with tables 5 and 6 included for completeness.



(a) (b)

Figure 23: (a) Success Rate vs. Training time in years. The red curve treats clusters 2 and 3 as one cluster (due to there similarities confusing them can be considered "forgivable") and the blue curve keeps them separate. (b) Median Age vs. Training time in years for the samples in figure 23a. The relatively high rate of success is not due to papers with ages close to the training period. For example, the set with a training period of 30 years, over half the papers are more than 40 years old.

are more than 40 years old as seen in figure 23b. So even with a very narrow window of time such as 5 or 6 years we can get a rough idea of what half of the papers we are looking at will be doing for many years to come.

4.8 Bayesian Methods Allow for Purely Evidence Based Prediction of Future Outcomes

The advantage of a probabilistic model is that it lends itself easily to all the inference tools of Bayesian analysis. This allows someone to make evidence based predictions, using only the model on previously observed data. One can utilize Monte Carlo methods such as Markov Chain Monte Carlo (MCMC) to sample from the following distribution without having to evaluate the (implied) integration in the denominator if it is analytically intractable.

$$p(a, b, \{\alpha\}|t_1, t_2, \ldots, t_m; t) = \frac{p(t_1, t_2, \ldots, t_m; t|a, b, \{\alpha\})p(a, b, \{\alpha\})}{p(t_1, t_2, \ldots, t_m; t)} \quad (115)$$

where $p(a, b, \{\alpha\}|t_1, t_2, \ldots, t_m; t)$ is the posterior distribution, $p(a, b, \{\alpha\})$ is the prior distribution which represents the users prior knowledge, $p(t_1, t_2, \ldots, t_m; t|a, b, \{\alpha\})$ is the likelihood, and $p(t_1, t_2, \ldots, t_m; t) = \int p(t_1, t_2, \ldots, t_m; t|a, b, \{\alpha\})p(a, b, \{\alpha\})\, da\, db\, d\{\alpha\}$. The set $\{\alpha\}$ is the set of all parameters belonging to the intrinsic time dependence of the process $\lambda(t)$.

The posterior distribution $p(a, b, \{\alpha\}|t_1, t_2, \ldots, t_m; t)$ calculated from the above equation either analytically or by sampling can be used in making predictions by integrating with the appropriate probability distribution over the parameters. For example the density of certain events occurring at particular times is given by

$$p(t'_1, t'_2, \ldots, t'_n; \tau|t_1, t_2, \ldots, t_m; t) = \int p(t'_1, t'_2, \ldots, t'_n; \tau|a, b, \{\alpha\})p(a, b, \{\alpha\}|t_1, t_2, \ldots, t_m; t)\, da\, db\, d\{\alpha\} \quad (116)$$

alternatively since it is particularly awkward to talk about a probability density such as the one above we can answer questions like "what is the probability of a paper getting n citations in the next t years?". Such a results can be obtained from integrating the posterior with equation (91) over the parameters

$$p_{m.n}(\tau|t_1, t_2, \ldots, t_m; t) = \int p_{m,n}(t, \tau|a, b, \{\alpha\})p(a, b, \{\alpha\}|t_1, t_2, \ldots, t_m; t)\, da\, db\, d\{\alpha\} \quad (117)$$

this can address practical questions such as what is the probability that a paper never gets cited again by setting $n = 0$ and taking the limit as $\tau \to \infty$. Another option is to use moments of the distribution such as equation (101) and (106) instead of the probability distribution and integrate over the model parameters. Lets takes for example the WSB model for a papers with a total of 250 citations at the end of 2013 its digital object identifier (DOI) is PhysRevB.55.3015 we will calculate the expected (based off of Eq. (106)) number of citations based off of the first 7 years of the papers history and compare it to what is actually observed. If we look at the projection in figure 24 based off the first 7 years the projection seems to do pretty well over the next 9 years, but such projections should be taken with a grain of salt.

Looking at how the posterior is distributed in a corner plot (figure 25) we can see even though we had over a hundred citations in the training period,

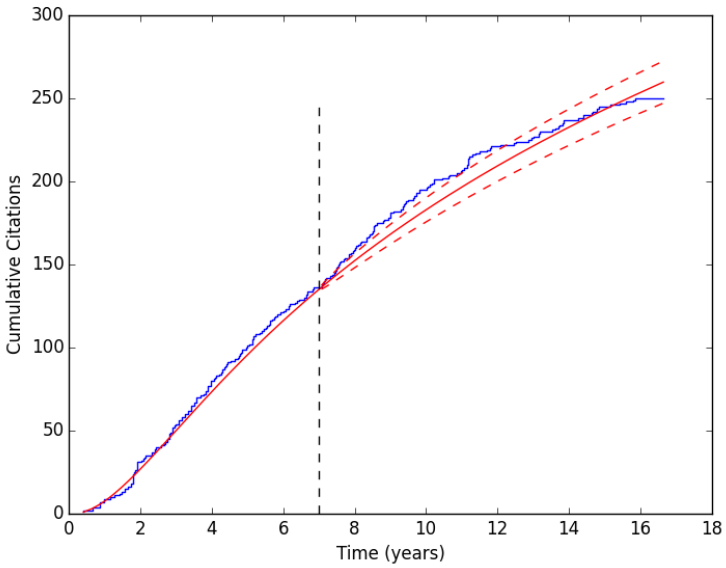


Figure 24: Using the first 7 years of data (this is marked by the black dashed line) the expected number of citations (in red) is inferred using Bayesian analysis is compared to the cumulative citation count (in blue). The red dashed lines represent plus and minus one standard deviation.

the parameters are not very sharply defined so any projection is much fuzzier than the red curve plotted in figure 24 or a standard deviation would indicate, and different runs can give somewhat different results. This is problem with any type of extrapolation with the any of these models projecting forward requires a good estimate of the parameters of a paper (sharp posterior) which requires a lot of data and there is no way to obtain multiple histories for the same paper. Anyone should be skeptical of people claiming any stochastic model can predict the future of a large body of papers with high accuracy even for long training periods. Examples of papers that would be difficult to characterize would be ones called "sleeping beauties" [35], these are papers that are characterized by a significant increase in citation many years after publication. The original paper on Einstein-Podolsky-Rosen (EPR) paradox with DOI PhysRev.47.777 is an example of such a paper where obtained most of its citations 50 years after it publications. By the clustering done before for 5 to 30 years of training this paper was lumped in cluster 1 as it appeared to stagnate very quickly. Though it was put into the noise group its parameter values of $a = 44.6$, $b = 40.5$ and $r = 2.4 \times 10^{-3}$ it is closest to cluster 3 which has slow long term growth. Sleeping beauty papers would require either expertise in their field or hindsight in order to determine that they are indeed beauties and not members of cluster 1. No simple stochastic model involving a point process can predict if a paper will be a sleeping beauty.

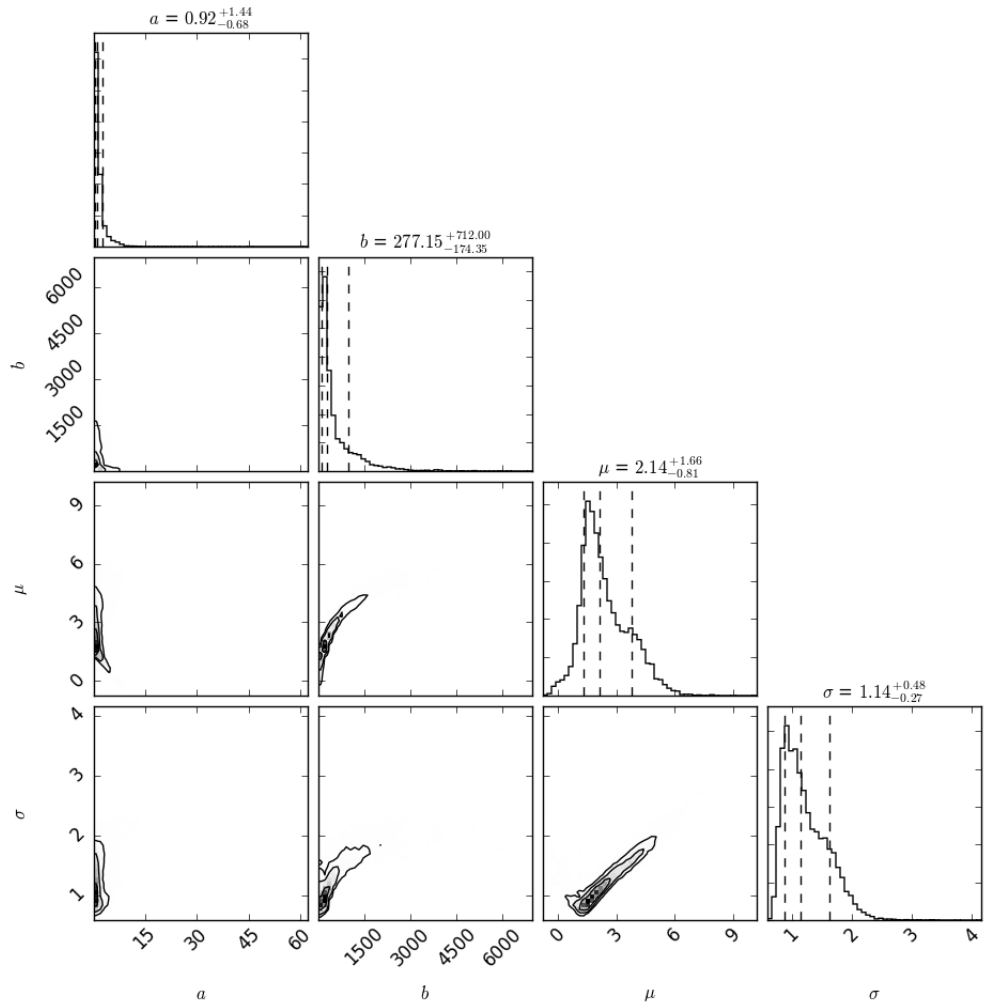


Figure 25: Using the first 7 years of data the posterior distribution of the parameters for the paper with DOI PhysRevB.55.3015.

5 Concluding Remarks

Hopefully upon reading each of the sections of this thesis I have convinced the reader of the main detail of each middle section. The first is that short highly charged proteins are susceptible to oxidative destabilization. The second is that Maximum Caliber (MaxCal) is a theoretical framework which also all known results of non-equilibrium thermodynamics to fall under one umbrella, and has the potential to generalize those results regimes far from equilibrium. Lastly, although precise predictions of citations trends are difficult, papers can be categorized into three groups of poor citation, high citation with quick impact and sleeping beauties that take many year before they obtain most of their citations. By no means are these the only important details or applications. For example, the fact that short and highly charged proteins are susceptible to oxidative destabilization means we can use this to find other proteins to study for aging studies and the study of age-related diseases. The clustering of papers can be combined with data from authors, with a suitable model, to predict roughly how many papers of each class the author will produce in the future. If we can characterize dynamical fluctuations at equilibrium by some statistical theory we could use MaxCal exactly like we would perform calculations in statistical mechanics.

References

[1] Not-so-deep impact. Nature, 435:1003–1004, 2005.

[2] Physics PhDs granted by Subfield: Classes of 2013 & 2014 combined. https://www.aip.org/sites/default/files/statistics/physics-trends/fall16-phdsubfield-p1.pdf, 2016.

[3] Hiroshi Adachi, Yoshisada Fujiwara, and Naoaki Ishii. Effects of oxygen on protein carbonyl and aging in caenorhabditis elegans mutants with long (age-1) and short (mev-1) life spans. The Journals of Gerontology Series A: Biological Sciences and Medical Sciences, 53A(4):B240–B244, 1998.

[4] Bruce Alberts. Impact factor distortions. Science, 340(6134):787–787, 2013.

[5] Bjarne Andresen, E. C. Zimmermann, and John Ross. Objections to a proposal on the rate of entropy production in systems far from equilibrium. The Journal of Chemical Physics, 81(10):4676–4677, 1984.

[6] Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, Midori A. Harris, David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C. Matese, Joel E. Richardson, Martin Ringwald, Gerald M. Rubin, and Gavin Sherlock. Gene ontology: tool for the unification of biology. Nat. Genet., 25:25 – 29, 2000.

[7] G.N. Bochkov and Yu. E. Kuzovlev. General theory of thermal fluctuations in nonlinear systems. Zh. Eksp. Teor. Fiz., 72:238–247, 1977.

[8] G.N. Bochkov and Yu. E. Kuzovlev. Fluctuation-dissipation relations for nonequilibrium processes in open systems. Zh. Eksp. Teor. Fiz., 76:1071–1088, 1979.

[9] G.N. Bochkov and Yu. E. Kuzovlev. Nonlinear fluctuation-dissipation relations and stochastic models in nonequilibrium thermodynamics: I. generalized fluctuation-dissipation theorem. Physica A: Statistical Mechanics and its Applications, 106(3):443–479, 1981.

[10] Carson C. Chow and Michael A. Buice. Path integral methods for stochastic differential equations. The Journal of Mathematical Neuroscience (JMN), 5(1):1–35, 2015.

[11] Kee H. Chung and Raymond A. K. Cox. Patterns of productivity in the finance literature: A study of the bibliometric distributions. The Journal of Finance, 45(1):301–309, 1990.

[12] S Cole, JR Cole, and GA Simon. Chance and consensus in peer review. Science, 214(4523):881–886, 1981.

[13] The UniProt Consortium. Uniprot: a hub for protein information. Nucleic Acids Research, 43(D1):D204–D212, 2015.

[14] Gavin E. Crooks. Entropy production fluctuation theorem and the nonequilibrium work relation for free energy differences. Phys. Rev. E, 60:2721–2726, Sep 1999.

[15] Adam M. R. de Graff, Michael J. Hazoglou, and Ken A. Dill. Highly Charged Proteins: The Achilles' Heel of Aging Proteomes. Structure, 24(2):329 – 336, 2016.

[16] Andrew M. Gleason David Mumford Daniel E. Flath Brad G. Osgood Patti Frazer Lock Douglas Quinney David O. Lomen Karen Rhea David Lovelock Jeff Tecosky-Feldman Thomas W. Tucker Deborah Hughes-Hallett, William G. McCallum. CALCULUS SINGLE AND MULTIVARIABLE, 4TH ED. John Wiley & Sons, Inc., 2005.

[17] K.A. Dill and S. Bromberg. Molecular Driving Forces: Statistical Thermodynamics in Chemistry and Biology. Garland Science, 2003.

[18] Leo Egghe. Theory and practise of the g-index. Scientometrics, 69(1):131–152, 2006.

[19] Jeremy L. England. Statistical physics of self-replication. The Journal of Chemical Physics, 139(12), 2013.

[20] R M L Evans. Detailed balance has a counterpart in non-equilibrium steady states. Journal of Physics A: Mathematical and General, 38(2):293, 2005.

[21] Ferric C Fang, Anthony Bowen, and Arturo Casadevall. Nih peer review percentile scores are poorly predictive of grant productivity. eLife, 5:e13323, feb 2016.

[22] R.P. Feynman, A.R. Hibbs, and D.F. Styer. Quantum Mechanics and Path Integrals. Dover books on physics. Dover Publications, 2005.

[23] Kingshuk Ghosh and Ken A. Dill. Computing protein stabilities from their chain lengths. Proceedings of the National Academy of Sciences, 106(26):10649–10654, 2009.

[24] Y. Gingras, V. Larivière, B. Macaluso, and J.-P. Robitaille. The Effects of Aging on Researchers' Publication and Citation Patterns. PLoS ONE, 3:e4048, December 2008.

[25] Irina Gitlin, Jeffrey D. Carbeck, and George M. Whitesides. Why are proteins charged? networks of chargecharge interactions in proteins measured by charge ladders and capillary electrophoresis. Angewandte Chemie International Edition, 45(19):3022–3060, 2006.

[26] B. M. Gupta. Growth and obsolescence of literature in theoretical population genetics. Scientometrics, 42(3):335–347, 1998.

[27] Michael J. Hazoglou, Valentin Walther, Purushottam D. Dixit, and Ken A. Dill. Communication: Maximum caliber is a general variational principle for nonequilibrium statistical mechanics. The Journal of Chemical Physics, 143(5), 2015.

[28] J. E. Hirsch. An index to quantify an individual's scientific research output. Proceedings of the National Academy of Sciences of the United States of America, 102(46):16569–16572, 2005.

[29] K.L.C. Hunt, P.M. Hunt, and John Ross. Dissipation in steady states of chemical systems and deviations from minimum entropy production. Physica A: Statistical Mechanics and its Applications, 147(12):48 – 60, 1987.

[30] K.L.C. Hunt, P.M. Hunt, and John Ross. Deviations from minimum entropy production at steady states of reacting chemical systems arbitrarily close to equilibrium. Physica A: Statistical Mechanics and its Applications, 154(1):207 – 211, 1988.

[31] C. Jarzynski. Hamiltonian derivation of a detailed fluctuation theorem. Journal of Statistical Physics, 98(1):77–102, 2000.

[32] Edwin T Jaynes. Information theory and statistical mechanics. Physical Review, 106(4):620, 1957.

[33] Edwin T Jaynes. Information theory and statistical mechanics II. Physical Review, 108(2):171, 1957.

[34] Zhang Jianzhi. Protein-length distributions for the three domains of life. Trends in Genetics, 16:107–109, 2000.

[35] Qing Ke, Emilio Ferrara, Filippo Radicchi, and Alessandro Flammini. Defining and identifying sleeping beauties in science. Proceedings of the National Academy of Sciences, 112(24):7426–7431, 2015.

[36] Runcong Ke and Shigeki Mitaku. Gaussian analysis of net electric charge of proteins in the ¡i¿drosophila melanogaster¡/i¿ genome. Chem-Bio Informatics Journal, 4(3):101–109, 2004.

[37] Runcong Ke and Shigeki Mitaku. Local repulsion in protein structures as revealed by a charge distribution analysis of all amino acid sequences from the saccharomyces cerevisiae genome. Journal of Physics: Condensed Matter, 17(31):S2825, 2005.

[38] Jonathan E. Kohn, Ian S. Millett, Jaby Jacob, Bojan Zagrovic, Thomas M. Dillon, Nikolina Cingel, Robin S. Dothager, Soenke Seifert, P. Thiyagarajan, Tobin R. Sosnick, M. Zahid Hasan, Vijay S. Pande, Ingo Ruczinski, Sebastian Doniach, and Kevin W. Plaxco. Random-coil behavior and the dimensions of chemically unfolded proteins. Proceedings of the National Academy of Sciences of the United States of America, 101(34):12491–12496, 2004.

[39] Dilip Kondepudi and Ilya Prigogine. From Heat Engines to Dissipative Structures. John Wiley & Son, 1998.

[40] Svein Kyvik. Age and scientific productivity. differences between fields of learning. Higher Education, 19(1):37–55, 1990.

[41] Svein Kyvik and Terje Bruen Olsen. Does the aging of tenured academic staff affect the research performance of universities? Scientometrics, 76(3):439–455, 2008.

[42] Peder Olesen Larsen and Markus von Ins. The rate of growth in scientific publication and the decline in coverage provided by science citation index. Scientometrics, 84(3):575–603, 2010.

[43] A.L. Lehninger, D.L. Nelson, and M.M. Cox. Lehninger Principles of Biochemistry. W. H. Freeman, 2005.

[44] T.L. Lin, R. Wang, W.P. Bi, A. El Kaabouchi, C. Pujos, F. Calvayrac, and Q.A. Wang. Path probability distribution of stochastic motion of non dissipative systems: a classical analog of feynman factor of path integral. Chaos, Solitons & Fractals, 57:129 – 136, 2013.

[45] Carlos López-Otín, Maria A. Blasco, Linda Partridge, Manuel Serrano, and Guido Kroemer. The hallmarks of aging. Cell, 153(6):1194 – 1217, 2013.

[46] Alfred J. Lotka. The frequency distribution of scientific productivity. Journal of the Washington Academy of Science, 16(12):317–324, 1926.

[47] Patrick. J. Mulvey and Starr Nicholson. Physics Graduate Degrees: Results from the Enrollments and Degrees & the Degree Recipient Follow-up Surveys. https://www.aip.org/sites/default/files/statistics/graduate/graddegrees-p-08.pdf, 2011.

[48] Patrick. J. Mulvey and Starr Nicholson. Trends in Physics PhDs. https://www.aip.org/sites/default/files/statistics/graduate/trendsphds-p-12.2.pdf, 2014.

[49] Cynthia N Oliver, Bong-Whan Ahn, Elena J Moerman, Samuel Goldstein, and Earl R Stadtman. Age-related changes in oxidized proteins. Journal of Biological Chemistry, 262(12):5488–5491, 1987.

[50] Lars Onsager. Reciprocal relations in irreversible processes. I. Physical Review, 37:405, 1931.

[51] Lars Onsager. Reciprocal relations in irreversible processes. II. Physical Review, 38:2265, 1931.

[52] H.H. Panjer and G.E. Willmot. Insurance risk models. Society of Acturaries, 1992.

[53] George J. Peterson, Steve Press, and Ken A. Dill. Nonuniversal power law scaling in the probability distribution of scientific citations. Proceedings of the National Academy of Sciences, 107(37):16023–16027, 2010.

[54] Drazen Petrov and Bojan Zagrovic. Microscopic analysis of protein oxidative damage: Effect of carbonylation on structure, dynamics, and aggregability of villin headpiece. Journal of the American Chemical Society, 133(18):7016–7024, 2011. PMID: 21506564.

[55] Drazen Petrov and Bojan Zagrovic. Are current atomistic force fields accurate enough to study proteins in crowded environments? PLoS Comput Biol, 10(5):1–11, 05 2014.

[56] Steve Pressé, Kingshuk Ghosh, Julian Lee, and Ken A Dill. Principles of maximum entropy and maximum caliber in statistical physics. Reviews of Modern Physics, 85(3):1115, 2013.

[57] Filippo Radicchi, Santo Fortunato, and Claudio Castellano. Universality of citation distributions: Toward an objective measure of scientific impact. Proceedings of the National Academy of Sciences, 105(45):17268–17272, 2008.

[58] Kristoffer Rørstad and Dag W. Aksnes. Publication rate expressed by age, gender and academic position a large-scale analysis of norwegian academic staff. Journal of Informetrics, 9(2):317 – 333, 2015.

[59] Lucas Sawle and Kingshuk Ghosh. How do thermophilic proteins and proteomes withstand high temperature? Biophysical Journal, 101(1):217 – 227, 2011.

[60] Per O Seglen. Why the impact factor of journals should not be used for evaluating research. BMJ, 314(7079):497, 1997.

[61] Emily Shacter. Quantification and significance of protein oxidation in biological samples*. Drug Metabolism Reviews, 32(3-4):307–326, 2000. PMID: 11139131.

[62] J. Shore and R. Johnson. Properties of cross-entropy minimization. IEEE Transactions on Information Theory, 27(4):472–482, Jul 1981.

[63] R S Sohal, S Agarwal, A Dubey, and W C Orr. Protein oxidative damage is associated with life expectancy of houseflies. Proceedings of the National Academy of Sciences, 90(15):7255–7259, 1993.

[64] Earl R. Stadtman. Protein oxidation and aging. Free Radical Research, 40(12):1250–1258, 2006.

[65] Pamela E. Starke-Reed and Cynthia N. Oliver. Protein oxidation and proteolysis during aging and oxidative stress. Archives of Biochemistry and Biophysics, 275(2):559 – 567, 1989.

[66] Michael J. Stringer, Marta Sales-Pardo, and Lus A. Nunes Amaral. Statistical validation of a global model for the distribution of the ultimate number of citations accrued by papers published in a scientific journal. Journal of the American Society for Information Science and Technology, 61(7):1377–1385, 2010.

[67] Michael J. Stringer, Marta Sales-Pardo, and Lus A. Nunes Amaral. Effectiveness of journal ranking schemes as a tool for locating information. PLoS ONE, 3(2):1–8, 02 2008.

[68] M. Tabor. Chaos and integrability in nonlinear dynamics: an introduction. Wiley-Interscience publication. Wiley, 1989.

[69] Robi Tacutu, Thomas Craig, Arie Budovsky, Daniel Wuttke, Gilad Lehmann, Dmitri Taranukha, Joana Costa, Vadim E. Fraifeld, and Joo Pedro de Magalhes. Human ageing genomic resources: Integrated databases and tools for the biology and genetics of ageing. Nucleic Acids Research, 41(D1):D1027–D1033, 2013.

[70] Nobuhiko Tokuriki, Francois Stricher, Joost Schymkowitz, Luis Serrano, and Dan S. Tawfik. The stability effects of protein mutations appear to be universally distributed. Journal of Molecular Biology, 369(5):1318 – 1332, 2007.

[71] Vladimir N Uversky. Natively unfolded proteins: a point where biology waits for physics. Protein science, 11(4):739–756, 2002.

[72] Dashun Wang, Chaoming Song, and Albert-László Barabási. Quantifying long-term scientific impact. Science, 342(6154):127–132, 2013.

[73] Michael S. Wisz and Homme W. Hellinga. An empirical model for electrostatic interactions in proteins incorporating multiple geometry-dependent dielectric constants. Proteins: Structure, Function, and Bioinformatics, 51(3):360–377, 2003.

[74] Konstantin B. Zeldovich, Peiqiu Chen, and Eugene I. Shakhnovich. Protein stability imposes limits on organism complexity and speed of molecular evolution. Proceedings of the National Academy of Sciences, 104(41):16152–16157, 2007.