AFRL-RH-WP-TR-2013-0063

CYBERSPACE MATH MODELS

Leslie M. Blaha, Daniel W. Repperger, Katheryn A. Farris, Fairul Mohd-Zaid, Paul R. Havig, Xiaoping Shen, Russell Francis, John P. McIntire, Lyndsey McIntire, David J. Rieksts

Air Force Research Laboratory

JUNE 2013
Final Report

AIR FORCE RESEARCH LABORATORY
711TH HUMAN PERFORMANCE WING,
HUMAN EFFECTIVENESS DIRECTORATE,
WRIGHT-PATTERSON AIR FORCE BASE, OH 45433
AIR FORCE MATERIEL COMMAND
UNITED STATES AIR FORCE

NOTICE AND SIGNATURE PAGE

Using Government drawings, specifications, or other data included in this document for any purpose other than Government procurement does not in any way obligate the U.S. Government. The fact that the Government formulated or supplied the drawings, specifications, or other data does not license the holder or any other person or corporation; or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

Qualified requestors may obtain copies of this report from the Defense Technical Information Center (DTIC).

AFRL-RH-WP-TR-2013-0063 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

\\signed\\
LESLIE M. BLAHA
Battlespace Visualization Branch

\\signed\\
JEFFREY L. CRAIG
Chief, Battlespace Visualization Branch

\\signed\\
WILLIAM E. RUSSELL
Acting Chief, Warfighter Interface Division

This report is published in the interest of scientific and technical information exchange, and its publication does not constitute the Government's approval or disapproval of its ideas or findings.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
11-06-13	Final	07 May 2008 – 31 December 2012

4. TITLE AND SUBTITLE
CYBERSPACE MATH MODELS

5a. CONTRACT NUMBER
FA8650-08-D-6801

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
61102F

6. AUTHOR(S)
Leslie M. Blaha, Daniel W. Repperger, Katheryn A. Farris, Fairul Mohd-Zaid, Paul R. Havig, Xiaoping Shen, Russell Francis, John P. McIntire, Lyndsey McIntire, David J. Rieksts

5d. PROJECT NUMBER
2313

5e. TASK NUMBER
HC

5f. WORK UNIT NUMBER
2313HC57

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
711 HPW/RHCV
Battlespace Visualization Branch
Wright-Patterson Air Force Base, OH 45433

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Air Force Materiel Command
Air Force Research Laboratory
711th Human Performance Wing
Human Effectiveness Directorate
Warfighter Interface Division
Battlespace Visualization Branch
Wright-Patterson Air Force Base, OH 45433

10. SPONSORING/MONITORING AGENCY ACRONYM(S)
AFRL/RHCV

11. SPONSORING/MONITORING AGENCY REPORT NUMBER(S)
AFRL-RH-WP-TR-2013-0063

12. DISTRIBUTION/AVAILABILITY STATEMENT
Distribution A: Approved for public release; distribution unlimited.

13. SUPPLEMENTARY NOTES
88ABW/PA Cleared 08/20/2013; 88ABW-2013-3715

14. ABSTRACT
The goal of this effort was to explore ways of characterizing the complexity, performance, vulnerability, and dynamic properties of networks and complex systems. Techniques investigated included information theory measures and stochastic resonance, together with graphical and statistical assessments. Network persistance was characterized by the Hurst parameter in a network model based on fractional brownian motion, generalized by fast Fourier transform and studied with wavelet analysis. Network uncertainty was cahracterized with approximate entropy. General applicability to other complex systems were were studied in the areas of sunspot cycles, chatbot detection, genetic data, and image processing.

15. SUBJECT TERMS
Complex Networks, Visualization, Image Fusion, Cyberspace, Information Theory

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT:	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON (Monitor)
a. REPORT	b. ABSTRACT	c. THIS PAGE			Leslie M. Blaha
Unclassified	Unclassified	Unclassified	SAR	113	19b. TELEPHONE NUMBER *(Include Area Code)*

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39-18

THIS PAGE INTENTIONALLY LEFT BLANK

Table of Contents

Executive Summary
Final Report to Dr. Robert J. Bonneau, AFOSR Program Officer

Laboratory Research Initiation Request Project: Quantifying Cyberspace Situational Awareness, Performance, Vulnerability, and the Design of Optimal Cyber Attacks in Complex Networks

Contract/Grant FA8650-08-D-6801, September 2008 to August 2011

Final accomplishments: This final progress report describes work conducted under the AFOSR Laboratory Request to Initiate Research (LRIR). Our goal was to quantify cyberspace vulnerability through a probability model with multiple parameters. We primarily focused on two key areas: 1.) Quantifying the level of network persistence (ie. long range memory), and, 2.) Quantifying the level of network uncertainty (ie. predictability).

1.) **Quantifying the level of network persistence (long range memory).** Fractional Brownian motion (fBm) is chosen as a basic design model for complex networks. It has been used as a theoretical framework to study non-stationarity and long- range dependence. Fractional Brownian motion is characterized by the Hurst parameter (H) which requires sophisticated techniques that often yield ambiguous results. We utilized a Fast Fourier Transform (FFT) based method to generalize fBm. These synthetic fBm data are then analyzed by using wavelet multi- resolution analysis and confirmed with extensive numerical simulations. Further details are reported in the list of archival publications and conference talks.

2.) **Quantifying the level of network uncertainty (predictability).** To quantify network uncertainty, we used a powerful tool based on information theory called approximate entropy (ApEn). Our study began with a literature survey on information entropy and was initiated in the summer of 2010. Initial numerical experiments were designed to study the relationship between the persistence parameter H and the uncertainty parameter ApEn. Further details are reported in the list of archival publications and conference talks.

Changes in research objectives, if any: Yes. The original proposal emphasized specific applications, such as modeling cyberspace with an actual cyber network. However, due to

the untimely passing of Dr. Daniel Repperger, we made some modifications such as taking a more general approach to assessing non-stationary time-series data with long- range memory through the analysis of sunspot cycles and stock market prices. The underlying structure of a complex network is the same whether you are analyzing the stock market or the internet. Hence, our studies still analyzes characteristics of a complex network and the data we used was more from a motivation of practicality because it was publicly available, parsed, and "cleaned up".

Change in AFOSR program manager, if any: Yes. The project has been managed by Dr. Paul R. Havig and Ms. Katheryn Farris since the original laboratory PI, Dr. Daniel Repperger of the 711th HPW, deceased on January 03, 2010 and Co-PI, Dr. Jeffrey McDonnell of AFIT, was deployed. We engaged in strong collaboration with Dr. Xiaoping Annie Shen, Mathematics Professor of Ohio University. She was sponsored by Summer Faculty Fellowship Program (SFFP) and worked with us full-time during the summers of 2009, 2010 and 2011. She also contracted out part of her time from October, 2010 to May, 2011 to continue research efforts under the complex networks domain.

Extensions granted or milestones slipped, **if any**: None

Archival Publications, Conference Presentations and Other:

Archival publications

1. X. A. Shen, K. A. Farris and P. R. Havig, "Quantifying Effects of Long-Range Memory on Predictability of Long-Range Systems," Proceedings of the 2011 National Aerospace and Electronics Conference (NAECON), July 20-22, 2011, Dayton, OH.

2. D. W. Repperger and K. A. Farris, Stochastic resonance – interpretations from a nonlinear control theory perspective, International Journal of Systems Science, July, 2010.

3. J. P. McIntire, P. R. Havig, K. A. Farris and L. K. McIntire, "Graphical and Statistical Communication Patterns of Automated Conversational Agents in Collaborative Computer-Mediated Communication Systems," Proceedings of the 2010 National Aersospace and Electronics Conference (NAECON), July 14-16, 2010, Dayton, OH.

Conference presentations

4. X. Shen, Sparse statistical data analysis based on the L1-norm - case study, Special session on Sparse Data Representations and Applications, AMS Spring Southeastern

Section Meeting, Statesboro, GA, March 18-20, 2011.

5. X. Shen, Digital signal representation with Slepian series, Special session on Control Systems and Signal Processing, AMS Spring Southeastern Section Meeting, Statesboro, GA, March 18-20, 2011.

6. X. Shen, K. A. Farris, D. Riesksts, and P. R. Havig, Shannon entropy and relatives: A brief review and new development, Special Session on Computational and Applied Mathematics, AMS Section Meeting, Richmond, VA, November 6-7, 2010.

7. D. W. Repperger, J. S. Shattuck, and K.A. Farris, Complex Network Analysis in Power Grids, 35th Annual Dayton-Cincinnati Aerospace Sciences Symposium, March 9, 2010, Dayton, Ohio. (*presented posthumously)

8. D. W. Repperger, K. A. Farris and R. Bradford, "Fractional Calculus – A Paradigm for Understanding Complex Systems," 34th Annual Dayton-Cincinnati Aerospace Sciences Symposium, March 3, 2009, Dayton, Ohio.

9. D. W. Repperger, K. A. Farris and R. Bradford, "Computational Studies of Complex Networks Using Information Flow," 34th Annual Dayton-Cincinnati Aerospace Sciences Symposium, March 3, 2009, Dayton, Ohio.

Other presentations

10. X. A. Shen, D. J. Rieksts, K. A. Farris and P. R. Havig, (Poster) Visualizing Complexity of Sunspot Cycles Via Wavelet Transform, the Summer Intern Research Day, August 3, WPAFB 2010, Dayton, OH.

11. K. A. Farris and P. R. Havig, Complexity Analysis and Information Visualization, Ohio University, March 5, 2010, Athens, Ohio.

12. D. W. Repperger, Studies on Information Theoretic Variables and Decision Making, seminar talk at Quantitative Psychology Program, the Ohio State University, October 19, 2009, Columbus, Ohio.

New discoveries, inventions, or patent disclosures during this reporting period:

Invention: Fractal Filter Apparatus Based on Wavelets, D. W. Repperger, X. A. Shen, A. R. Pinkus and K. A. Farris. US Air Force Invention Number: AF# 1092, approved 11/16/2009. Patent application pending.

International Journal of Systems Science
Vol. 41, No. 7, July 2010, 897–907

Taylor & Francis
Taylor & Francis Group

Stochastic resonance – a nonlinear control theory interpretation

D.W. Repperger† and K.A. Farris*

711 Human Performance Wing, Air Force Research Laboratory, AFRL/RHCV, Wright-Patterson Air Force Base, Dayton, OH 45433-7022, USA

(*Received 12 June 2009; final version received 3 November 2009*)

Stochastic resonance (SR) is an effect that has been known (Benzi, R., Sutera, A., and Vulpiani, A. (1981), 'The Mechanism of Stochastic Resonance', *Journal of Physics*, A14, L453–L457) for almost three decades and has been extensively studied in biology, statistics, signal processing and in numerous other eclectic areas (Wiesenfeld, K., and Moss, F. (1995), 'Stochastic Resonance and the Benefits of Noise: From Ice Ages to Crayfish and Squids', *Nature*, 373, 33–36). Herein, a nonlinear control theory analysis is conducted on how to better understand the class of systems that may exhibit the SR effect. Using nonlinear control theory methods, equilibrium points are manipulated to create the SR response (similar to shaping dynamical response in a phase plane). From this approach, a means of synthesising and designing the appropriate class of nonlinear systems is introduced. New types of nonlinear dynamics that demonstrate the SR effects are discovered, which may have utility in control theory as well as in many diverse applications. A numerical simulation illustrates some powerful attributes of these systems.

Keywords: nonlinear dynamics; stochastic resonance; biological systems

1. Introduction

Stochastic resonance (SR) provides an enabling property to a certain class of nonlinear systems, which may be viewed as an optimisation of system design, and commonly occurs in nature. In this article, a method of synthesising such systems is conducted by manipulation of equilibrium points, which is a common practice in nonlinear control theory. The term 'resonance' refers to a unimodal peak of the SR curve displayed, for example, in Figure 1. The term 'stochastic' indicates noise injection into a system. The vertical axis in Figure 1 portrays an output variable or property of a system, which is desired to be optimised. In the signal processing literature, for example, the y-axis could represent the signal-to-noise ratio (SNR) gain of a system (Loerincz, Gingl, and Kiss 1996; Chapeau-Blondeau 1997). An SR is a property of the input–output characteristics of a nonlinear system, which is shown in the sequel. The analysis and synthesis of nonlinear systems can be approached a number of ways, e.g. by sliding mode control (Diong 2004) and passivity methods (Son, Yang, Jo, Shim, and Seo 2004). Manipulating the response of a nonlinear system is a topic of interest, e.g. mitigating a limit cycle using specified feedback (Repperger, Buckholtz, and Daniels 2004). Some related problems in the control theory area involve chaotic systems (Serletis and Andreadis 2000; Apostolou and King 2002; Liao and Chen 2003; Wei and Billing 2004; Wang and Ip 2005). Another property of nonlinear systems, which will be shown later, involves the bifurcation characteristics of the locus of equilibrium points. Bifurcation studies are investigated in singular systems (Li and Liu 1999) within biological systems (Hritonenko and Yatsenko 2007). Equilibrium points can be managed (Shin and Chung 2002), a topic to be discussed in the ensuing analysis. Finally, the properties of systems that exhibit SR appear in the most basic forms of biological response (Repperger, Phillips, and Neidhard 2001) and can even be linked to the manner in which image processing is performed in nature (Xu, Jiang, Wu, and Repperger 2009). It should be mentioned that the SR effect can also occur in leaky integrate-and-fire neuron models, which have been investigated in Repperger et al. (2001) and are outside the scope of the procedures discussed herein.

To continue the discussion of the manner how SR works, the horizontal axis usually represents the power of the noise intensity in the input signal as indicated in Figure 1. Figure 2 presents a block diagram description of a commonly studied SR system. The input signal is $S(t)$ and is typically deterministic; the measured output signal is $y(t)$. The noise term $\eta(t)$ is usually assumed to be zero mean and Gaussian, but this is not a

*Corresponding author. Email: katheryn.farris@wpafb.af.mil
†This paper is dedicated in memory of Dr. Daniel Repperger, who passed away on January 3, 2010.

ISSN 0020–7721 print/ISSN 1464–5319 online

DOI: 10.1080/00207720903494692
http://www.informaworld.com

XML Template (2010) [1.6.2010-4:40pm] [897–907]
{TANDF}TSYS/TSYS_I_41_07/TSYS_A_449887.3d (TSYS) [PRINTER STAGE]

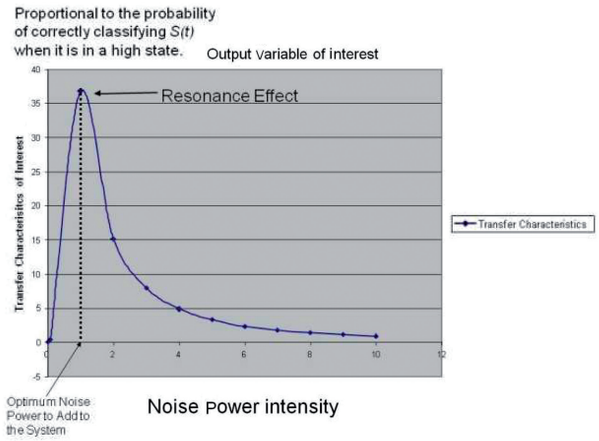


Figure 1. SR resonance effect.

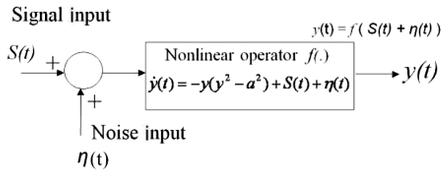


Figure 2. Block diagram of SR process.

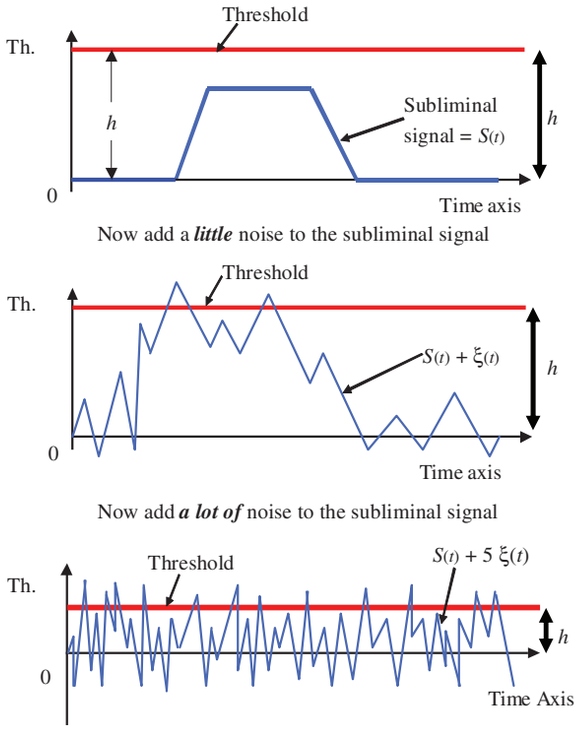


Figure 3. Application 1 – subthreshold identification of $S(t)$.

strict requirement. The dynamical equation in Figure 2 is classically described by

$$\dot{y}(t) = -y(y^2 - a^2) + S(t) + \eta(t) \tag{1}$$

By measurement of the scalar variable $y(t)$, subjected to an injected noise $\eta(t)$, the goal is to correctly determine if the deterministic signal $S(t)$ may be in a high or low state. This model is very akin to how biological systems work and respond in nature. Two physical viewpoints on applications are presented on how the SR effect is traditionally envisioned in the literature.

2. A biological and statistical viewpoint on why the SR effect occurs

At first it appears counter intuitive that the system in Equation (1) may gain some advantage in its output capabilities related to $y(t)$ when adding the noise $\eta(t)$ to the input signal $S(t)$. From the extant literature, however, there are two important physical applications that demonstrate that the shape of the curve in Figure 1 can be realised and used to optimise system performance. The first perspective is in terms of correct statistical detection of the signal $S(t)$ being either in a high or low state. Since this article has the focus on the nonlinear control theory aspects of the design of SR systems, the prior applications will be briefly described here with more specific material presented in Appendix 1.

2.1. *Application 1: correct detection of S(t) when it is subthreshold*

In Figure 3, the input signal $S(t)$ is a trapezoid and is below a threshold h. The resonance curve in Figure 1 for this application has the vertical axis being proportional to the correct detection of $S(t)$ when it is truly in a high state. Appendix 1 describes these details (cf Repperger, Phillips, Berlin, Neidhard-Doll, and Haas 2005 for a haptics study related to Figure 3, which represents a more practical application of this effect involving human subjects and subliminal force signals that helped improve tracking performance and situational awareness).

2.2. *Application 2: the bipotential well problem*

The second application occurs in the physics literature. In Figure 4, the signal $S(t)$ may represent the state of the ball in a bipotential well. The ball can be in either a high or low state (right or left well) and is subthreshold by having insufficient kinetic energy of motion as compared to the potential energy necessary to switch states. Similar to Application 1, the resonance curve in Figure 1 has the vertical axis proportional to the

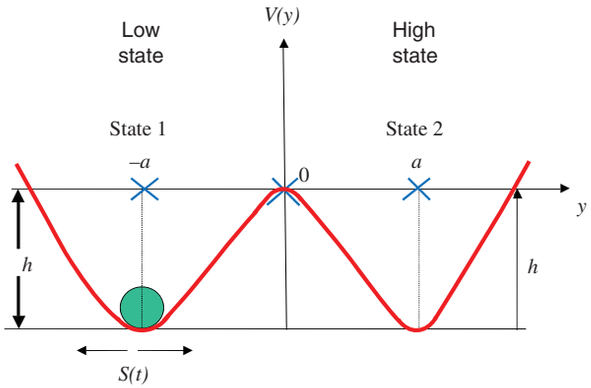


Figure 4. Application 2 – the bipotential well problem.

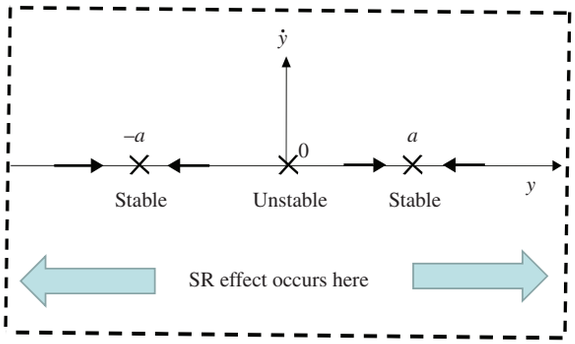


Figure 5. The equilibrium points from Figure 4 that induce the SR effect to occur.

correct placement of $S(t)$ in the high state (the right most potential well) when $S(t)$ is high. Again, these details are described further in Appendix 1.

3. A physics perspective of the SR system and equilibrium points

Figure 4 presents a physical platform to examine how the dynamics of SR arises and its relationship to Equation (1). Examining the homogeneous portion of Equation (1), the equilibrium points can be determined by letting $y(t) \to \infty$ or letting $\dot{y}(t) \to 0$.

Thus

$$\dot{y}(t) = -y(y^2 - a^2) \to 0 \Rightarrow y = 0, \quad y = \pm a \text{ as equilibrium points} \tag{2}$$

and Figure 5 portrays a plot of these equilibrium points in a phase plane. The velocity flow arrows on the y-(horizontal) axis indicate that the equilibrium point $y = 0$ is unstable, and the two equilibrium points $y = \pm a$ are both stable. The equilibrium diagram in Figure 5 concurs with Figure 4 where, physically, the unstable equilibrium (the origin) is coincident with the top of the small hill between the two potential wells. The bottoms of the two potential wells represent the location of the two stable equilibrium points at the locations $y = \pm a$.

From a physics perspective, one can also view the dynamics of the homogeneous portion of Equation (1) as a force equation associated with a potential energy function $V(y)$. This paradigm describes the potential well in Figure 4 as the spatial integral of the negative force gradient. For example, let $V(y)$ represent the potential function, then the homogeneous portion of Equation (1) is given by

$$\frac{\partial V(y)}{\partial y} = -[-y(y^2 - a^2)] = -\text{force spatial gradient}, \tag{3}$$

which integrates to

$$V(y) = -\frac{a^2}{2}y^2 + \frac{1}{4}y^4 + c, \tag{4}$$

where c is constant. Thus $V(y)$ of Equation (4) represents the quartic potential energy function curve in Figure 4, derived from systems that show the SR effect. This concept can now be generalised to other examples.

4. Generalising SR systems using nonlinear dynamics principles

From the discussions in Appendix 1, the following definitions will apply to classes of nonlinear systems that would exhibit the SR effect:

Definition: A nonlinear system $\dot{y}(t) = f(y, S(t), \eta(t))$ will exhibit an SR effect if the function $f(\cdot)$ has three or more equilibrium points. The equilibrium points must have an unstable equilibrium point interlaced between two stable equilibrium points, similar to Figure 5.

Discussion: It is clear that $f(\cdot)$ must be nonlinear. In order to have three or more equilibrium points requires $\dot{y}(t) = f(y, S(t), \eta(t)) = \mathrm{O}(y^3)$, which must be of order 3 or higher, excluding linear systems. As in Figure 5, the domain of the state space, where this SR effect will occur is locally near the three equilibrium points. This definition can now be expanded to a much wider class of systems that occur in nonlinear dynamics.

5. Relating the SR effect to a super critical pitchfork bifurcation

By relating the class of systems that may exhibit the SR effect to properties of their equilibrium points as described in the prior section, generalisations now can be obtained to different systems in the areas of

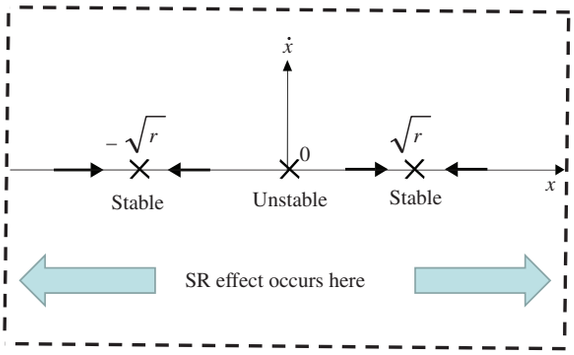


Figure 6. The equilibrium points from Equation (5).

nonlinear dynamics (Strogatz 1995). A familiar example of systems that have the same property, as in the definition given previously, include nonlinear dynamical systems that fall in a class of equilibrium points determined by a critical pitchfork bifurcation. For example, consider the nonlinear system with dynamics (a well-studied problem):

$$\dot{x} = rx - x^3 = x(r - x^2) \tag{5}$$

with Figure 6 describing the set of all equilibrium points of Equation (5) if $r > 0$. It is clear that Equation (5) is akin to the prior discussion on the classical equation that produces the SR effect. The relationship of Equation (5) and systems that produce the SR effect can now be described in terms of the concepts of a vector flow field.

5.1. *A vector flow field representation of Equation (5)*

For a system of the form $\dot{x} = f(x)$ to determine the equilibrium points when $f(x)$ may be a nonlinear function, it can be obtained graphically by plotting $f(x)$ in the phase plane of $\dot{x}$ *versus* x. The intersections of $f(x)$ with the x-axis ($\dot{x} = 0$) represent equilibrium point solutions, which can be determined graphically. For Equation (5), the three possible cases $r < 0$, $r = 0$, and $r > 0$ are displayed in Figure 7. The case $r > 0$ corresponds to systems that show the SR effect. If one now considers r as the independent variable and a graph is now made of the class of all equilibrium points on the vertical axis, as in Figure 8, the region where the SR effect occurs appears in the right half of that diagram.

Thus in Figure 8, which includes all three cases ($r < 0$, $r = 0$, and $r > 0$), one can design the two stable equilibrium points by adjusting the $r > 0$ value to cover any area of the phase plane desired. From Figure 6, this extends the domain and range of the SR effect to

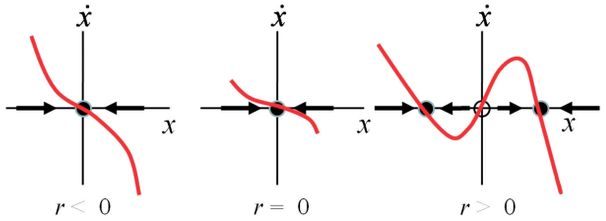


Figure 7. Vector flow fields for different values of r in Equation (5).

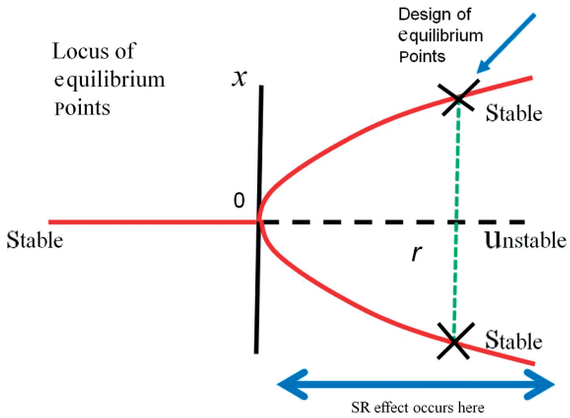


Figure 8. The pitchfork bifurcation representation of the class of all equilibrium points.

any portion of the phase plane required for a particular application.

6. Alternative properties of SR systems

It is not just a coincidence that the $V(y)$ function in Equation (4) is an even function (symmetric about the vertical axis in Figure 4). This means replacing y by $-y$ yields the same equations and similar equilibrium points. This is a property of systems that give rise to super critical pitchfork bifurcations. Figure 8 also shows this requisite symmetry in the pitchfork shape. The term 'super' in super critical refers to the $-x^3$ term in the homogeneous term of $\dot{x} = -x^3 + a^2x$, which provides global stability as $x \to \pm\infty$. This is a consequence of the two stable equilibrium points $x = \pm a$, which are attracting for trajectories that have $|x| > a$, thus ensuring global stability.

7. Other nonlinear systems that exhibit the SR effect

Once the methodology for designing nonlinear systems that exhibit the SR effect is obtained, the results can now be easily generalised to other nonlinear systems. Several other applications that also are derived from nature are discussed. Some of these applications are

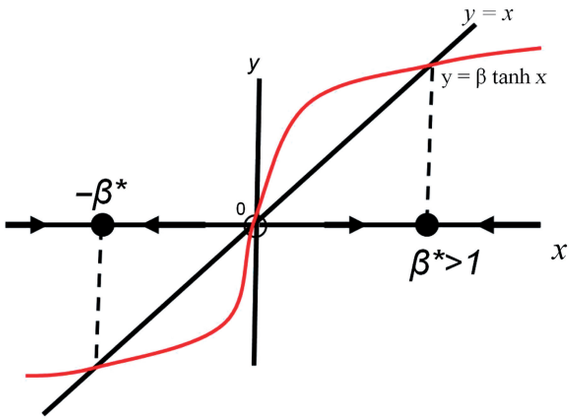


Figure 9. Example 1 equilibrium points.

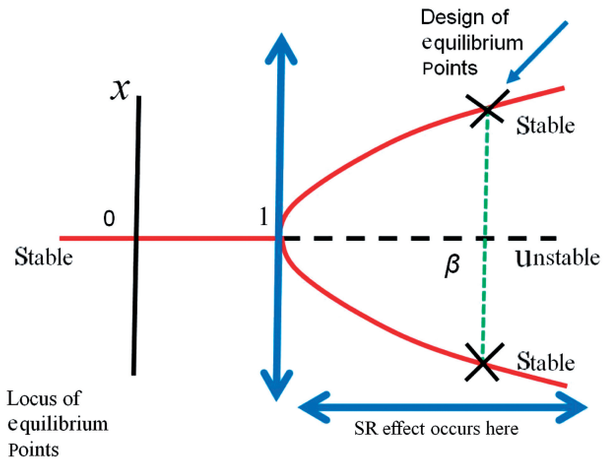


Figure 10. The pitchfork bifurcation for Example 1.

well-known in the nonlinear dynamics area (Strogatz 1995), but have not been previously discussed within the context of the SR effect, which is a type of optimisation. This demonstrates that the subsequent applications also have some optimisation properties in their outputs of their physical structures, which may not be known at this time. Several other applications have already demonstrated that they exhibit pitchfork bifurcations involving the locus of their equilibrium points. The first application involves models of magnets and neural networks (Palmer 1989).

7.1. *Example 1 – mechanical models of magnets and neural networks*

From Palmer 1989, the following equation is known to exhibit a pitchfork bifurcation as the parameter β is varied with $\beta > 1$:

$$\dot{x} = -x + \beta \tanh x. \tag{6}$$

To show this result and to determine the equilibrium points, a plot *versus* x is made of both sides of the equation $x = \beta \tanh x$ in Figure 9 for common intersection points as shown (setting $\dot{x} = 0$). The pair of equilibrium points $\pm\beta^*$ have the requisite symmetry. Figure 10 shows the resulting pitchfork bifurcation diagram. Again, the design of the two stable equilibrium points can be placed anywhere in the phase plane desired through the choice of the $\beta > 1$ parameter in Equation (6).

The next application is a biological example and includes a bifurcation and a catastrophe cusp curve. The model is for a sudden outbreak of insects and shows the interaction of insects with their food supply and the effect of predators (Ludwig, Jones, and Holling 1978).

7.2. *Example 2 – the interaction of insects with their food supply and predators*

Let $N(t)$ represent the insect population, which is affected by its growth rate determined by food supply and death rate determined by predators:

$$\dot{N}(t) = \text{growth rate death rate (predator attacks).} \tag{7}$$

The growth rate includes the typical logistics model that reaches saturation.

$$\text{Growth rate} = RN\left(1 - \frac{N}{K}\right). \tag{8}$$

Thus, the growth is initially exponential ($\dot{N}(t) = RN(t)$) and after K individuals appear, the growth rate then turns negative and the population saturates. To model the death rate, the predators are birds that start to reduce $N(t)$ after they reach the threshold $N > A$ and is modelled via:

$$\text{Death rate} = \frac{BN^2}{A^2 + N^2} \tag{9}$$

Thus, Equation (7) now has the complete form:

$$\dot{N} = RN\left(1 - \frac{N}{K}\right) - \frac{BN^2}{A^2 + N^2}. \tag{10}$$

A normalisation is conducted by the choice of variables (Strogatz 1995) of $x = \frac{N}{A}, \tau = \frac{Bt}{A}, r = \frac{RA}{B}$, $k = \frac{K}{A}$ and Equation (10) simplifies to the dimensionless form:

$$\frac{\mathrm{d}x}{\mathrm{d}\tau} = rx\left(1 - \frac{x}{k}\right) - \frac{x^2}{1 + x^2} \tag{11}$$

To find the equilibrium points of Equation (11) requires the simultaneous solution of $(\mathrm{d}x/\mathrm{d}\tau = 0)$

$$xy_2(x) = xr\left(1 - \frac{x}{k}\right) = \frac{x}{1 + x^2}x = xy_1(x). \tag{12}$$

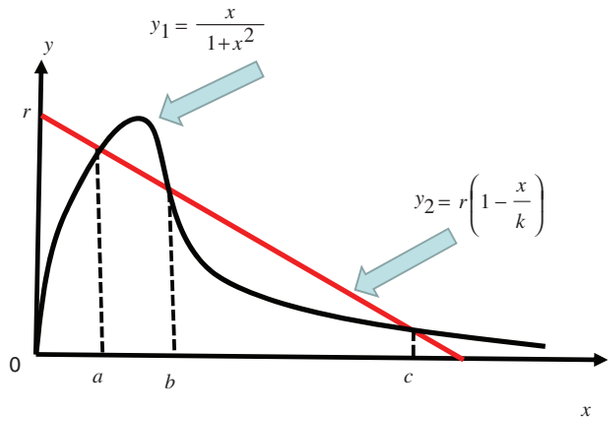


Figure 11. Simultaneous plots of $y_1(x)$ and $y_2(x)$.

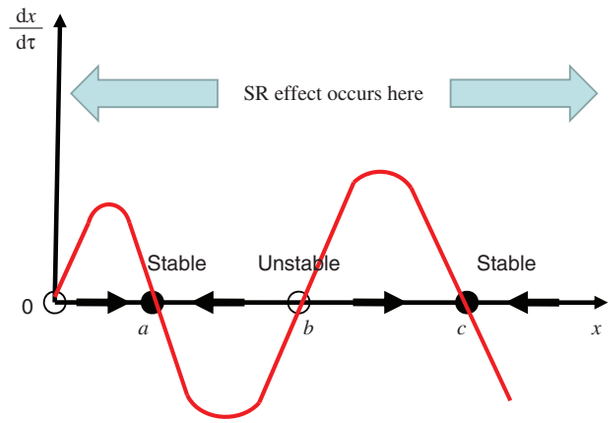


Figure 12. Plots of $(y_2(x) - y_1(x))$ from Figure 11.

The equilibrium point at $x = 0$ will be considered separately from the case $x > 0$, which is now explored by the functions $y_1(x)$ and $y_2(x)$ for their common intersections. If k is large, Figure 11 shows the simultaneous plots of the left- and right-hand sides of Equation (12) demonstrating three equilibrium point intersections. They are shown in Figure 12 and they satisfy the properties, where an SR effect will occur (two stable equilibrium points separated by an unstable equilibrium point). The bifurcation curve is very interesting because it is a slice of a cusp of a catastrophe curve. Figure 13 portrays the bifurcation curve, where the SR effect will occur and Figure 14 shows this curve as a top view projection of a cusp catastrophe surface.

Finally, it is noted in Figure 12 that the symmetry property now disappears about the origin for the set of all equilibrium points (satisfying $x = -x$) are not always true. The next example also loses this symmetry when small variations are made in the assumptions of the underlying dynamics using the classical SR model of Equations (1) and (5).

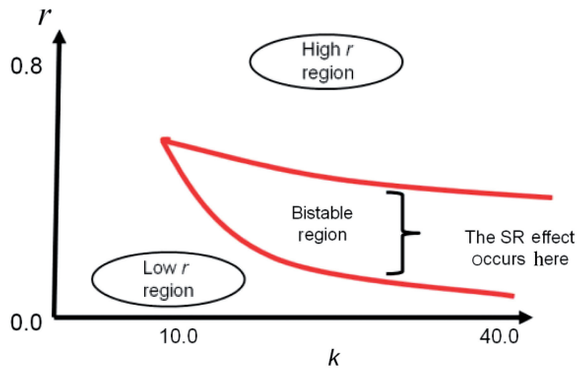


Figure 13. SR region for Example 2.

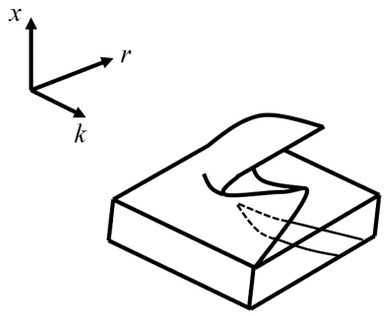


Figure 14. Catastrophe cusp and bifurcation.

7.3. *Example 3 – an approximation to the classical SR method of Equations (1) and (5)*

A small variation of the homogeneous form of Equations (1) and (5) can also produce the SR effect, which is not symmetric, but may have utility for certain other applications. Starting with

$$\dot{x} = h + rx - x^3. \tag{13}$$

If $h = 0$, this is identical to Equation (5). However, if h small, one would expect the SR effect may still occur. To investigate this small variation to Equation (5), plot $y = -h$ and $y = rx - x^3$ for common points of intersection (the equilibrium points of Equation (13)) as shown in Figure 15. Three equilibrium points will occur (two stable interlaced between an unstable) if the term h is smaller than h_c given by

$$h < h_c = \frac{2r}{3}\sqrt{\frac{r}{3}}. \tag{14}$$

Appendix 2 derives this result. Figure 16 portrays the bifurcation curve for the equilibrium points and the region, where the SR effect will occur in $h - r$ space.

Finally, additional nonlinear systems are shown to generate the SR effect from a number of other procedures (Repperger, Alderman, and Djouadi 2006)

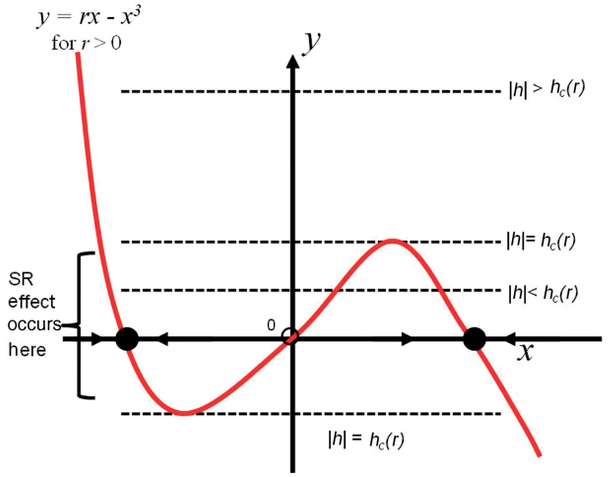


Figure 15. *y versus x* for $r>0$.

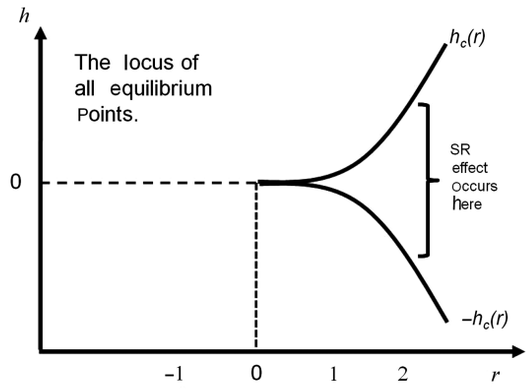


Figure 16. *h versus r* showing bifurcation.

by manipulating the spatial location of the equilibrium points.

7.4. *Other systems that exhibit the SR effect*

As mentioned earlier, there are almost an infinite number of ways to generate nonlinear systems that show the SR effect. In Repperger et al. (2006) the attribute of enhancing SNR gain was leveraged by using the SR effect. Some of the systems with nonlinear dynamics are given below. In all cases, the phase plane was partitioned into rectangular regions, with the pattern of two stable equilibrium points being separated by a single unstable equilibrium point. This technique of the 'management of the physical placement of equilibrium points or manipulation of the shape of the potential function' is not unlike applications in control theory, which typically would be concerned with placing closed loop poles in certain regions of a pole-zero diagram. Additional systems that show the SR effect include ($b_{i+1}>b_i>0$, $a_{i+1}>a_i$):

$$\dot{x} = -(x+a_1)(x+a_2)(x-a_3)(x-a_4)(x-a_5) \quad (15)$$

$$\dot{x} = -(x+a_1)(x+a_2)(x+a_3)(x-a_4)(x-a_5)(x-a_6) \times (x-a_7)+a_8(e^{-b_1x}-e^{-b_2x})+a_9(e^{-b_3x}-e^{-b_4x}) \quad (16)$$

and so on. Hence the types of nonlinear systems that can show the SR effect are practically unlimited because the a_i and b_i terms in Equations (15) and (16) are somewhat arbitrary.

8. A numerical simulation to show the SR effect

To show the potential efficacy of the SR effect, it is well known in the extant literature (Loerincz et al. 1996; Chapeau-Blondeau 1997) that the SR effect can *significantly amplify* the SNR of certain signals, especially under high noise conditions. This provides an explanation in biology on how creatures in nature can detect the presence of mating partners or predators with abnormally high values of sensitivity. The nonlinear sensing systems of these animals may have dynamics, where this SR effect occurs to promote sensitivity and specificity in the identification of objects.

The question of how much signal-to-noise amplification can be achieved presents an intriguing issue. It has been hypothesised that SR may amplify the SNR of an input signal $S(t)$ a factor of 10,000 or more (Loerincz et al. 1996) and the following example was selected to examine this effect through an exhaustive numerical simulation (Repperger et al. 2006).

With reference to Figure 2, the signal $S(t)$ is selected to be the sum of two sine waves, i.e.

$$S(t) = \sin(2\pi(1)t) + \sin(2\pi(4)t). \quad (17)$$

Thus, the deterministic test signal input $S(t)$ in Figure 2 consists of the sum of two sine waves with unity amplitude and fundamental frequencies at 1 Hz and 4 Hz. This is a biperiodic signal as considered by Khovanov and McClintock (2007), where it was shown that a high frequency bias can control synchronisation at the lower frequency. In a subsequent work (Khovanov 2008), it was shown that improvement of SNR can be obtained by increasing array elements and efficiency of information processing. Figure 17 shows the test signal $S(t)$ in real time and Figure 18 portrays a fast Fourier transform (FFT) of $S(t)$ indicating spectral power identified at 1 Hz and 4 Hz. The noise term $\eta(t)$ in Figure 2 is zero-mean Gaussian with variance σ^2, which will be varied. The objective is to measure only $y(t)$ in Figure 2 and attempt to identify the

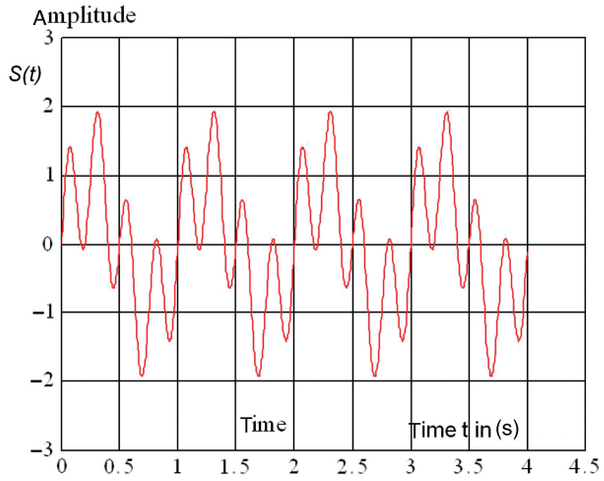


Figure 17. Test signal $S(t)$ in real time.

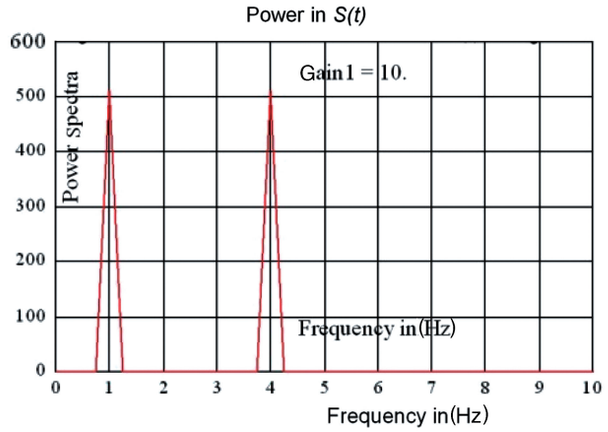


Figure 18. FFT of $S(t)$ test signal.

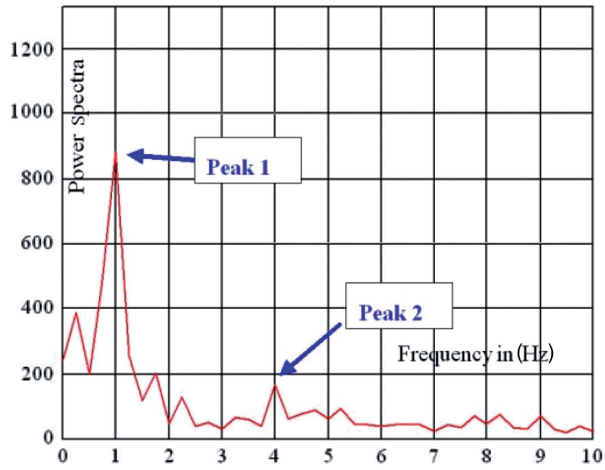


Figure 19. Spectra of $y(t)$ for $\sigma = 2000$.

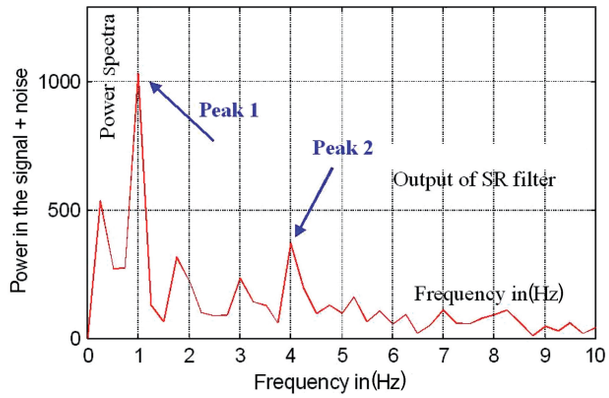


Figure 20. Spectra of $y(t)$ for $\sigma = 9000$.

original $S(t)$ signal in Equation (17). As the variance σ^2 (power) in the noise $\eta(t)$ increases, it becomes increasingly difficult to identify in the FFT of $y(t)$, where the two principal components at 1 Hz and 4 Hz appear. However, the SR filter behaves much better. To show this result for the case of the output of the SR filter, Figure 19 shows the power spectra FFT of $y(t)$ for the case $\sigma = 2000$ and Figure 20 for the case of $\sigma = 9000$. To provide a comparison to the performance of this SR filter, a unity gain block (which is the case without the SR effect) is selected as a baseline comparison and is determined by Equation (18). In all the spectra plots, the frequency resolution $\Delta f = 0.25\,\text{Hz} = 1/(\text{total real time simulation in seconds})$. Also $\Delta t = 4\,\text{s}/1024$ points.

8.1. *Unity gain block as a baseline and comparison to the SR filter*

$$y_2(t) = S(t) + \eta(t), \tag{18}$$

where, again, $\eta(t)$ is zero-mean Gaussian with variance σ^2. Figure 21 shows the power spectra FFT of $y_2(t)$ when $\sigma = 70$. From Figure 21, it is not possible to discern the frequencies of the input signal $S(t)$ at 1 and 4 Hz. Thus, the noise has completely overpowered $S(t)$, which is now significantly subthreshold. Thus without the SR filter, it is not possible to identify the input signal. To fairly compare the SR filter to a unity gain block in Equation (18), the following SNR figure of merit was selected (Gingl, Vajtai, and Kiss 2000):

$$\text{SNR} = (\text{average power in 1 Hz and 4 Hz signal})/ \times (\text{average power in adjacent spectra}). \tag{19}$$

Thus, as the noise power variance σ^2 increases in magnitude, the FFT of $y(t)$ is calculated and the average power in the frequencies 1 Hz and 4 Hz are compared to the average power in the frequencies adjacent to the spectra estimates due to noise only at adjacent frequencies (0.75 Hz, 1.25 Hz, 3.75 Hz and 4.25 Hz in Figure 21). As the SNR approaches 1.0,

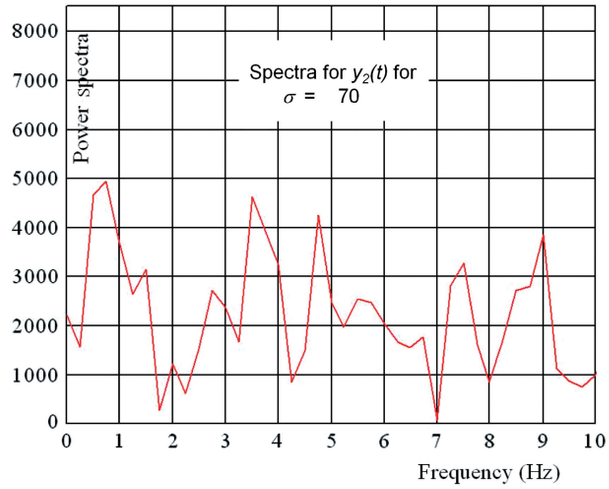


Figure 21. y_2 spectra for $\sigma = 70$.

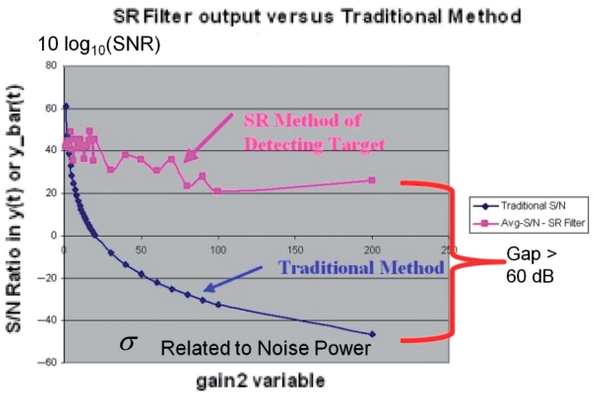


Figure 22. Comparison of SR to unity block.

it becomes virtually impossible to detect the power in the original signal $S(t)$ (cf Figure 21 for the case of $y_2(t)$) at the frequencies 1 and 4 Hz.

To evaluate the SR effect for $y(t)$ computed via Equation (1) (Figure 2) *versus* the unity gain baseline ($y_2(t)$) from Equation (18), the SNR variable of Equation (19) was determined by computer simulation as σ increases in value for both systems. Figure 22 shows this comparison. The y-axis is $10\log_{10}$ of the SNR variable. The top curve is for $y(t)$ as the output of the SR filter in Equation (1). The bottom curve is for the traditional (unity gain) of $y_2(t)$ of Equation (18). The x-axis is proportional to σ. The difference in performance (the gap in Figure 22) exceeds 60 units on a $10\log_{10}$ scale, which is equivalent to a signal-to-noise amplification gain of 10^6 and greater. Thus, the SR filter in Equation (1) and Figure 2 amplifies the SNR a factor of a million as compared to the tradition situation of unity gain. It should be cautioned, however, that one caveat of this analysis is a form of

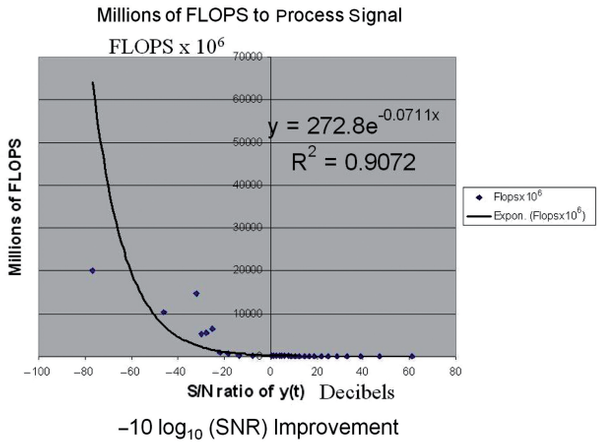


Figure 23. Millions of FLOPS *versus* SNR amplification gain.

a no-free-lunch theorem, which occurs due to computational complexity.

8.2. *A no free lunch theorem*

The example just presented seems to infer that SNR gains using SR systems (for this elementary example) may exceed 10^6 or higher using the SR effect *versus* not using it. However, there is a computational cost for this improvement of SNR that may limit its applicability. Figure 23 shows a plot of millions of computational floating point computer operations per second (FLOPS) *versus* the SNR gain in decibels ($10\log_{10}$(SNR)). It is seen that to obtain SNR amplifications of the order of 10^6 it requires 2×10^{10} FLOPS, which may induce an excessive computational burden. Thus, the limits in getting the SNR quality of the output signal improved may not be practical in certain applications. This is a topic of future research.

9. Summary and conclusions

SR is powerful effect in nonlinear systems which can, for example, significantly amplify the signal-to-noise characteristics of a deterministic input signal $S(t)$. The numerical simulation shows significant gains of SNR using an SR system compared to a unity gain with an accompanying computational cost. The nonlinear control theory aspects of this problem are examined in this article showing numerous such filters can be constructed and designed via their nonlinear dynamics and through the manipulation of equilibrium points.

Notes on contributors

D.W. Repperger received his BSEE and MSEE degrees from Rensselaer Polytechnic Institute (1968 and 1969) and his PhD in Electrical Engineering from Purdue University (1973). He was a federal employee for the US government (United States Air Force) from 1975 to 2010 working at what is now called the Air Force Research Laboratory. His early interests included optimal control and estimation theory involving numerical methods. In recent times, he worked in nonlinear control theory with applications in Biomedical Engineering. He served on a number of editorial boards as an associate editor, including the *IEEE Transactions on Control Systems Technology*, *Control Engineering Practice*, and *IEEE Transactions on Rehabilitation Engineering*, as well as several other journals. He was a Fellow of IEEE, American Institute of Medical and Biological Engineering (AIMBE), as well as several other organisations. Dr. Daniel Repperger passed away on January 3, 2010. He was an estimable man with the highest regard for science and engineering. He had an impeccable work ethic and a heart of gold. He will be greatly missed in the scientific research community.

K.A. Farris earned her Bachelor's of Science Degree in Mathematics in 2008 from the University of Arizona. She is currently a federal employee for the US government at the United States Air Force Research Lab. She is employed as a quantitative research scientist in the Human Effectiveness Directorate. Her current areas of interest are in the algorithmic development of modelling complex networks and computational methods used in Genomics and Biological studies. Kate is an active member of the Association for Women in Mathematics (AWM), the American Mathematical Society (AMS) and the Society of American Military Engineers (SAME). She previously received an award for Outstanding Leadership from the Society of American Military Engineers (SAME) in May, 2008.

References

Apostolou, N., and King, R.E. (2002), 'Design of Globally Stable Controllers for a Class of Chaotic Systems', *International Journal of Systems Science*, 33, 379–386.

Chapeau-Blondeau, F. (1997), 'Input-Output Gains for Signal in Noise in Stochastic Resonance', *Physics Letters A*, 232, 41–48.

Diong, B.M. (2004), 'Sliding-Mode Control Design for a Class of Systems with Non-matching Nonlinearities and Disturbances', *International Journal of Systems Science*, 35, 445–455.

Gingl, Z., Vajtai, R., and Kiss, L.B. (2000), 'Signal-to-Noise Ratio Gain by Stochastic Resonance in a Bistable System', *Chaos, Solitons, and Fractals*, 11, 1929–1932.

Hritonenko, N., and Yatsenko, Y. (2007), 'Bifurcations in Nonlinear Integral Models of Biological Systems', *International Journal of Systems Science*, 38, 389–399.

Khovanov, A., and McClintock, P.V.E. (2007), 'Synchronization of Stochastic Bistable Systems by Biperiodic Signals', *Physics Review E*, 76, 031122.

Khovanov, A. (2008), 'Array Enhancement of Stochastic Synchronization and Signal-to-Noise Ratio Gain in the Nonlinear Regime of Signal Transmission', *Physics Review E*, 77, 011124.

Li, Y., and Liu, Y. (1999), 'Bifurcation on Stability of Singular Systems with Delay', *International Journal of Systems Science*, 30, 643–649.

Liao, X., and Chen, G. (2003), 'On Feedback-Controlled Synchronization of Chaotic Systems', *International Journal of Systems Science*, 34, 453–461.

Loerincz, K., Gingl, Z., and Kiss, L.B. (1996), 'A Stochastic Resonator is Able to Greatly Improve Signal-to-Noise ratio', *Physics Letter A*, 224, 63–67.

Ludwig, D., Jones, D.D., and Holling, C.S. (1978), 'Qualitative Analysis of Insect Outbreak Systems: the Spruce Budworm and Forest', *Journal of Animal Ecolology*, 47, 315.

Palmer, R. (1989), 'Broken Ergodicity', in *Lectures in the Sciences of Complexity*, ed. D.L. Stein, Reading, MA: Addison-Wesley.

Repperger, D.W., Alderman, E.M., and Djouadi, M.S. (2006), US Patent Number 7,030,808 B1: Nonlinear Target Recognition, Washington, DC: U.S. Patent Trademark Office.

Repperger, D.W., Buckholtz, K., and Daniels, M. (2004), 'A Study on the Stabilization of the van der Pol Limit Cycle', *International Journal of Systems Science*, 35, 661–669.

Repperger, D.W., Phillips, C.A., Berlin, J., Neidhard-Doll, A., and Haas, M. (2005), 'Human-Machine Haptic Interface Design Using Stochastic Resonance Methods', *IEEE Transactions on Systems, Man, and Cybernetics, Part A, Humans and Systems*, 35, 574–582.

Repperger, D.W., Phillips, C.A., and Neidhard, A. (2001), 'A Study on Stochastic Resonance Involving the Hodgkin-Huxley Equations', *Proceedings of the American Control Conference*, Vol. 1, Arlington, VA, USA, 229–234.

Serletis, A., and Andreadis, I. (2000), 'Chaotic Analysis of US Money and Velocity Measures', *International Journal of Systems Science*, 31, 161–169.

Shin, M.H., and Chung, M.J. (2002), 'Parameterization and Stabilization of the Equilibrium Points of Affine Non-linear Control Systems', *International Journal of Systems Science*, 33, 301–311.

Son, Y.I., Yang, J.W., Jo, N.H., Shim, H., and Seo, J.H. (2004), 'Feedback Passivity Approach to Output Feedback Disturbance Attenuation for Uncertain Nonlinear Systems', *International Journal of Systems Science*, 35, 467–477.

Strogatz, S.H. (1995), *Nonlinear Dynamics and Chaos*, Reading, MA: Addison-Wesley.

Wang, D., and Ip, W.H. (2005), 'Ant Search Based Control Optimization Strategy for a Class of Chaotic System', *International Journal of Systems Science*, 36, 951–959.

Wei, H.L., and Billings, S.A. (2004), 'Identification and Reconstruction of Chaotic Systems Using Multiresolution Wavelet Decompositions', *International Journal of Systems Science*, 35, 511–526.

Xu, B., Jiang, Z.-P., Wu, X., and Repperger, D.W. (2009), 'Investigation of Two-dimensional Parameter-Induced Stochastic Resonance and Applications in Nonlinear Image Processing', *Journal of Physics A – Mathematical and Theoretical*, 42, 145207, 1-45207, 9.

Appendix 1: Interpreting Figures 1 and 2 in terms of statistical optimality

In Figure 1, assume the resonance curve has as the y-axis the probability of detecting $S(t)$ in the high state, when this is really true. The x-axis would be the power (variance) of a zero-mean Gaussian noise source. From Figure 3, $S(t)$ is subliminal (below the threshold) in the top graph and cannot be detected. Hence, the SR curve in Figure 1 must start at the origin. As the noise power is increased, the middle diagram in Figure 3 now becomes descriptive. Note when $S(t)$ is in the high state, it sometimes pierces through the threshold and is correctly detected as being high. This reduces the number of misses when the ground truth is that $S(t)$ is high. Thus, the SR curve rises from the origin in Figure 1 as the noise power increases. As the noise power continues to grow larger, the SR curve reaches a peak. The bottom diagram in Figure 3 is now appropriate. The detection probability has increased through the gradual reduction of misses of detecting $S(t)$ being in the high state. However, with too much noise power added, the number of false alarms ($S(t)$ is detected as high, but really it is low) increases to the point that the overall detection accuracy starts to decline. Thus, the SR curve reaches a peak in Figure 1 and then decreases back to zero as the noise power continues to increase. The benefits gained by reducing the misses in the detector are less helpful due to the increases in the number of false positives for large values of noise power. Finally, the SR curve approaches zero as the noise power becomes large, with little benefit derived from the addition of the noise.

Appendix 2: Validating Equation (14)

In Figure 15, the equation for $y(x)$ is given by

$$y(x) = rx - x^3. \tag{A.1}$$

To find the local minimum and maximum of $y(x)$, the first derivative of $y(x)$ is set to zero.

$$\frac{\mathrm{d}y(x)}{\mathrm{d}x} = \frac{\mathrm{d}}{\mathrm{d}x}(rx - x^3) = r - 3x^2 = 0. \tag{A.2}$$

Solving for x yields the two stable (symmetric) equilibrium points for the local maximum and minimum:

$$x_{\min} = -x_{\max} = -\sqrt{\frac{r}{3}}, \quad r > 0 \tag{A.3}$$

But the h value (from Figure 15) is obtained from substituting the result from Equation (A.3) into Equation (A.1) for $y(x)$ at the equilibrium point, hence:

$$h_c(r) = rx_{\max} - (x_{\max})^3 = \frac{2r}{3}\sqrt{\frac{r}{3}} \tag{A.4}$$

and it can be shown that the negative of Equation (A.4) occurs for the locus of negative equilibrium points.

Graphical and Statistical Communication Patterns of Automated Conversational Agents in Collaborative Computer-Mediated Communication Systems

John McIntire, Paul Havig, Katheryn Farris
711th Human Performance Wing / RHCVZ
Air Force Research Laboratory, WPAFB, OH USA
{firstname.lastname}@wpafb.af.mil

Lindsey McIntire
Infoscitex Corporation
Dayton, OH USA
lindsey.mcintire@wpafb.af.mil

***Abstract*—Automated conversational agents, also known as "chatbots" or "chatterbots," are computer programs used in a variety of collaborative communications systems, often for entertainment or business purposes. However, their use as malicious tools has more recently made them a growing nuisance and security concern. We present a detailed graphical and statistical analysis of communication patterns (specifically involving message sizes and inter-message delays) for improving the detection of automated conversational agents in collaborative computer-mediated communication systems.**

I. INTRODUCTION

Conversational agents are automated natural language communication programs that interface with humans via real-time text-based computer-mediated communications (CMC). These programs are also known widely as "chatbots" or "chatterbots." The use of conversational agents (hereafter referred to as "chatbots" or simply as "bots") for entertainment, education, and business communications is growing in popularity and utility [1,2,3,4]. As the technology continues to progress in sophistication, bots will undoubtedly find emerging application in many other areas.

Unfortunately, chatbots are already finding application in the Internet's underground. They are an increasing nuisance for chat, IM, and text-based CMC users, because of phishing attempts and through the spread of malware and spam [5,6,7]. More maliciously, they are being used as automated intelligence agents and scam artists, e.g., by attempting social engineering and teasing out personal information from users to be used for identity theft [8,9]. Bots can pose considerable security risks considering that they may be dispatched in virtually unlimited numbers and can work tirelessly at their goals, in contrast to individual humans attempting such feats in a person-to-person manner. To combat this threat, reliable methods for detecting bots are needed, so that malicious automated agents can be identified and removed from CMC systems.

Some previous and current attempts to stop chatbots from entering public chat rooms utilize CAPTCHAs (Automated Turing Tests) so that only humans can enter [e.g., 10], but this defense can be bypassed with human assistance. Keyword-based filtering and related spam detection methods are common but seem to be having limited success with chatbots. Gianvecchio et al. [11] suggested using metrics based upon gross communication patterns, such as communicators' message sizes and inter-message delay times. Using a large dataset of public chat transcripts, they provided extensive evidence that chatbots could be classified into as many as 14 unique behavioral types based upon these metrics. And these metrics, when coupled with pattern-matching and/or machine-learning algorithms, could be extremely effective at distinguishing between humans and bots in computer-mediated communications. However, the chat data they analyzed was limited to public chat rooms with many multiple simultaneous communicators, and was primarily focused on bots that were attempting to get users to click on hyperlinks (either posted in messages or in their online profiles). It was unclear from their work whether similar chatbot behavioral patterns would be present in other types of CMC, such as one-on-one instant messaging, or when bots are attempting to carry on lengthy, convincing conversations with humans.

We attempted to study such questions in a previous work [12], using a different (and considerably smaller) dataset. We provided supporting evidence that the gross behavioral metrics suggested by Gianvecchio et al. (message sizes, inter-message delays) could be useful for passively distinguishing between humans and bots. In this work, we present a more detailed graphical and statistical analysis of related measures using a similar but larger data set.

II. METHOD

A. Chat Transcript Data Set

As in our previous work [12], we gathered our chat data from the publicly-available transcripts of the Loebner Prize in

Artificial Intelligence [13]. The Loebner Prize is a formal public Turing Test competition in which the world's best chatbot programs compete to be the most "human" in terms of conversational capabilities. The Turing Test was suggested by British mathematician Alan Turing as a practical method for gauging the "intelligence" of an artificially-intelligent conversational agent [14]. The Test requires human judges to determine through text-based conversations whether their conversing partner is a human confederate or a computer program.

One of the great advantages of the Loebner Prize dataset is that the communicators are all unambiguously defined as either humans or chatbots; another is that an exceedingly large proportion of the conversations specifically involve chatbots, whereas other public chat datasets do not typically possess such high frequency of bot communications. Yet another advantage is that the bots are specifically trying to carry on human-like conversations with people for lengthy periods, whereas in many public data sets bots are mostly trying to convince users to click hyperlinks posted in their messages or in their online profiles [11].

A potential disadvantage of this dataset is that the conversational situation might not be considered "typical," in that most of the human judges are suspicious as to the other communicators' identities from the start, and many of the conversations might be rightly classified as "interrogations" as opposed to merely "normal, everyday conversations." Nonetheless, since both the human participants (confederates) and bots faced similar interrogations by the human judges, we do not believe this to be a fatal flaw to our analyses and interpretations; indeed, in our previous work [12], there was little difference between the two human groups (judges and confederates) on the inter-message delay and message size distribution measures. In any case, the nature of our dataset is certainly a caveat to be kept in mind when interpreting our data and drawing conclusions.

Although in our previous work [12] we used data only from the 2008 competition, in this work we analyzed five separate competitions, from the years 1996, 1997, 2004, 2005, and again from 2008. These particular years were chosen for analysis -- instead of using all 15 years of available data -- for several reasons: these years' data included text files, which allowed for ease of analysis; and these years included all the necessary and desired information for analysis of the transcripts, i.e., some years did not include clear delineation of individual communicator identities, time-stamping, structuring of conversations, etc.

This dataset captured a total of 9,206 messages in 254 brief conversations involving over 50 individual humans and 22 chatbots. It should be noted that some bots (and humans) were apparently repeat participants across two or more competition years.

B. Data Analysis Methods

From the raw transcript data, we calculated three primary metrics of interest: inter-message delay times (the time between sequential message transmissions), message size (the number of words transmitted per message), and message rate (the message size divided by the inter-message delay, in words per minute or wpm; note that this measure is related but distinct from *typing speeds* which are also typically measured in words per minute). From these metrics, we analyzed features of the human and bot distributions, including central tendencies, dispersions, skewness, kurtosis, complexity, and visual analysis of distributions. Other statistical tests (i.e., ANOVA) are performed where appropriate, using alpha=.05.

Messaging complexity was measured via two methods: normalized entropy and normalized entropy rate. *Entropy* was calculated using Shannon's information entropy approach [15], and can be interpreted in this context as the complexity of messaging behaviors. Entropy was measured by first calculating the empirical probability, $p(x_i)$, of occurrence of each inter-message delay, message size, and message rate. The probabilities for each category were then used in the Shannon entropy equation:

$$H(x_i) = -\sum(p(x_i)*\log(p(x_i)))$$

Entropy rate can be interpreted as the entropy of sequential pairs of observations. For our purposes, high entropy rates would suggest a highly complex messaging process. Similar to above, the empirical probability, $p(x_{ij})$, was calculated for each sequential pair of observations (i,j) in the inter-message delays, message sizes, and message rates. It is similar in form and calculation to the entropy calculation:

$$H(x_{ij}) = -\sum(p(x_{ij})*\log(p(x_{ij})))$$

Due to the fact that these measures are sensitive to sample size, we normalized our calculations of both the entropies and the entropy rates (giving values from 0 to 100%). Normalization was accomplished by taking the observed entropy (or entropy rate) divided by the maximum possible entropy (or entropy rate) given the sample size, n:

$$\text{Normalized } H(x_i) = -\sum(p(x)*\log(p(x)) / \log(n))$$

These normalized entropy and entropy rates provided measures of complexity that could more be easily compared across the varying the sample sizes used in these analyses.

III. Results and Discussion

A. Inter-Message Delay Times

Overall results concerning inter-message delay times are presented in Table I; further results are presented and discussed in the text and associated figures.

TABLE I. Inter-message Delays

Measures	*Humans*	*Bots*
Mean (arithmetic)	20.4	7.3
Median	17.5	4.0
Standard Deviation	13.6	9.2
Skewness	0.793	2.183
Kurtosis	-0.074	5.800
Sample size (n)	5424	3527

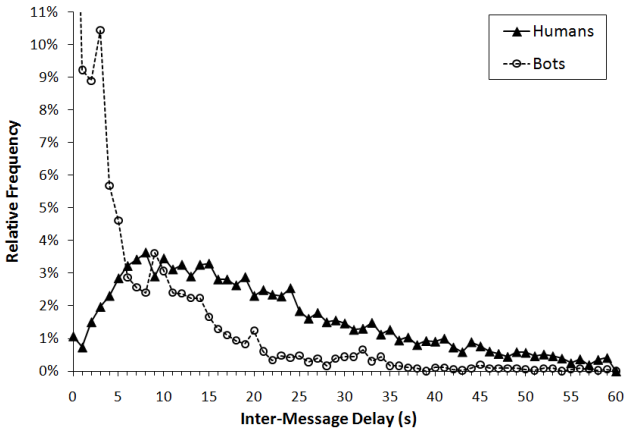


Figure 1. Inter-message delay relative frequency distributions for humans and chatbots. Inter-message delays were measured in seconds.

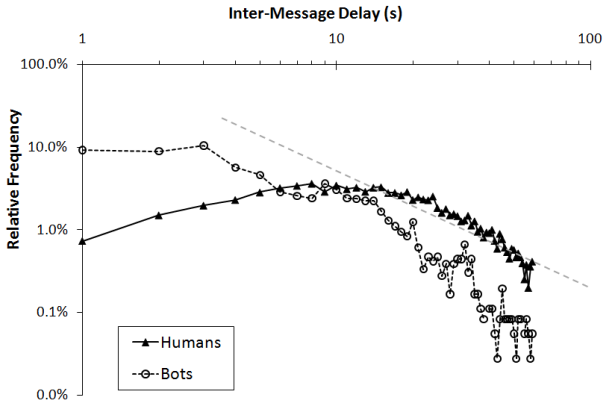


Figure 2. The same data as in Fig. 1, but plotted on logarithmic X and Y axes. The bulk of the human data points (almost 75%) fall nearly on a straight line (the dashed grey line).

1) Central Tendencies and Dispersions: Average human inter-message delay times were substantially longer in comparison to bots (20.4 versus 7.3 seconds, respectively). Median values produced a similar pattern (17.5 versus 4.0 seconds, respectively). Variability for the human inter-message delays was also noticeably larger (standard deviations of 13.6 versus 9.2 seconds).

2) Distribution Shapes: Measures of distribution shapes were also noticeably different between humans and bots. Human inter-message delay distribution skewness was measured at 0.793 (with a standard error of skewness of 0.033), while the bot distribution skewness was measured at 2.183 (with a standard error of skewness of 0.041). Human inter-message delay distribution kurtosis was measured at -0.074 (with a standard error of kurtosis of 0.066). Bot distribution kurtosis was measured at 5.800 (with a standard error of kurtosis of 0.082).

3) Graphical Analysis: As can be seen in Fig. 1, human inter-message delays follow a distinctly different distribution pattern than bots. To allow a more straightforward comparison to the results of Gianvecchio et al. [11], we also plotted the inter-message delay distributions on log-log scales in Fig. 2. The advantage of this plotting method is that a power law distribution (which we would expect to find for the humans) tends to reveal itself as a linear pattern when plotted on log-log scales. Just as Gianvecchio et al. found, the bulk of the human data points (inter-message delays of 10 s or greater comprise almost 75% of the distribution) fall nearly on a straight line in this log-log plot; the same is not quite true for the bots.

4) Discussion: Inter-message delay times seemed to offer a variety of distinguishing measures for detecting bot versus human chat communications. At least in our dataset, potentially useful metrics involving inter-message delays included: measures of central tendency, dispersion or variability, skewness, and kurtosis. These results directly support our previous work in [12], the results of Gianvecchio et al. [11], and other human chat data [16] demonstrating temporal patterns in human communication that can be used to distinguish them from chatbots. Further, these results indirectly support other research showing that humans exhibit distinct temporal patterns in virtually all forms of their communications, including asynchronous CMC such as blogging, texting, cyber forums, e-mailing, and even non-CMC including surveys and spoken communication [17].

B. *Message Sizes*

Overall results concerning message sizes are presented in Table II; further results are presented and discussed in the text and associated figures.

1) Central Tendencies and Dispersions: Average human message sizes were only slighter smaller than bots (8.4 versus 9.3 word count per message, respectively). Median values produced a similar pattern (6.0 versus 7.0, respectively). Variability, too, was very similar between humans and bots (standard deviations of 7.8 versus 8.6 words per message).

2) Distribution Shapes: Measures of distribution shapes were somewhat mixed on whether there was a noticeable difference between humans and bots. Human message size distribution skewness was measured at 3.077 (with a standard error of skewness of 0.033), while the bot distribution skewness was measured at 2.603 (with a standard error of skewness of 0.041), suggesting little difference on this measure. However, human message size distribution kurtosis was measured at 17.213 (with a standard error of kurtosis of 0.066) while bot distribution kurtosis was measured at 12.464 (with a standard error of kurtosis of 0.081), which may suggest a possible distinguishing metric.

3) Graphical Analysis: As can be seen in Fig. 3, human and bot message size distributions are highly similar and practically indistinguishable.

TABLE II. MESSAGE SIZES

Measures	*Humans*	*Bots*
Mean (arithmetic)	8.4	9.3
Median	6.0	7.0
Standard Deviation	7.8	8.6
Skewness	3.077	2.603
Kurtosis	17.213	12.464
Sample size (n)	5586	3620

TABLE III. MESSAGE RATES

Measures	*Humans*	*Bots*
Mean (arithmetic)	35.7	145.4
Median	24.0	72.0
Standard Deviation	67.4	233.4
Skewness	10.36	4.72
Kurtosis	151.46	31.90
Sample size (n)	5364	2805

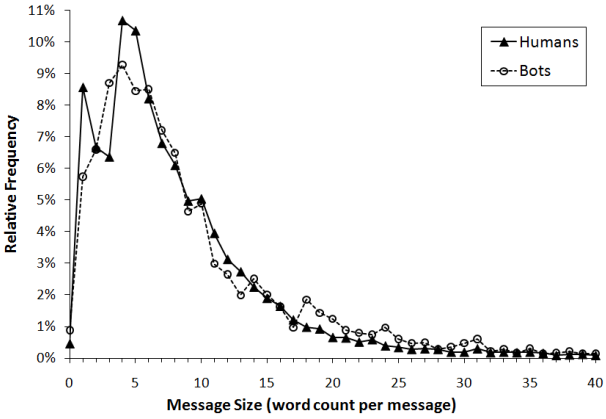


Figure 3. Message size relative frequency distributions for humans and chatbots. Message sizes were measured in word count per message.

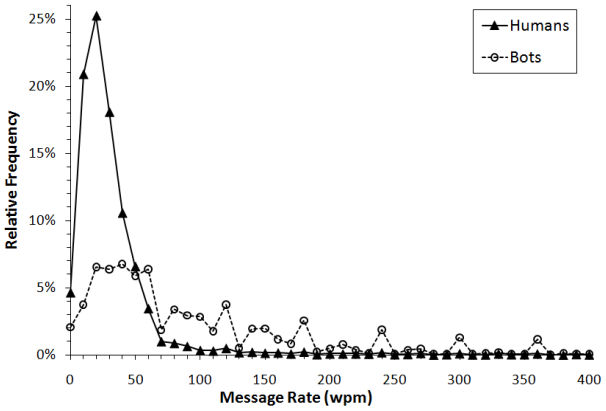


Figure 4. Message rate relative frequency distributions for humans and chatbots. Message rates were measured in words per minute (wpm).

4) Discussion: Analysis of message sizes offered very little in terms of distinguishing between humans and bots. This result is in contrast to our previous findings (with a smaller sample) that suggested message sizes to possibly be a distinguishing metric [12]. This result is also somewhat in conflict with the results of Gianvecchio et al [11] who found many "types" of bots whose message size distribution shapes differed widely from the observed human distributions. Even though in our sample there was a statistically significant difference between human and bot message sizes (F=34.95, p<.001), the statistical effect size was near zero (partial-eta^2=.004). Indeed, visual inspection of the relative frequency distributions of human versus bot message sizes shows extensive overlap and similarity of distribution shapes between the groups, and both distributions appear to follow an exponential decay (after excluding the initial spikes).

C. Message rates

Overall results concerning message rates are presented in Table III; further results are presented and discussed in the text and associated figures.

1) Central Tendencies and Dispersions: Average human message rates were measured at a reasonable 35.7 words per minute (median of 24.0 wpm), while average bot message rates were a blistering 145.4 words per minute (median of 72.0 wpm). Variability, too, was widely different between the two groups. The standard deviation of human message rates was surprisingly large at 67.4 wpm, but bots were even larger at 233.4 wpm.

2) Distribution Shapes: Distribution shape measurements regarding message rates suggest large differences between the groups. Human distribution skewness was measured at 10.358 (standard error of skewness was 0.033), while bot skewness was 4.720 (standard error of skewness was 0.046). Human distribution kurtosis was 151.460 wpm (standard error of kurtosis was 0.067) versus bot kurtosis of 31.901 (standard error of 0.092).

3) Graphical Analysis: As can be seen in Fig. 4, human and bot message rate distributions are distinctly different. Almost all of the observed human message rates fall below 70 wpm, with a large spike peaking at about 20 wpm. In contrast, bot message rates are scattered with a flat, wide, and irregularly-shaped distribution. All observed message rates are shown in Fig. 5, and median message rates for individual humans and bots is plotted in Fig. 6. The differences between humans and chatbots in regards to message rates are clearly evident in both of these figures.

4) Discussion: Message rate, a combined measure of inter-message delays and message sizes, seems to offer yet another way to distinguish between human and chatbot communications. Useful message rate metrics found in our

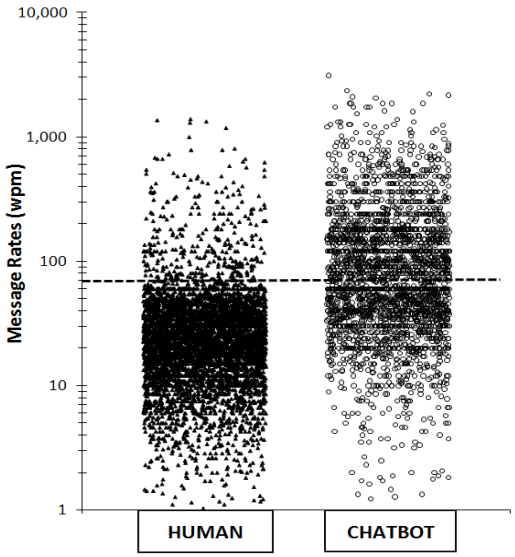


Figure 5. All observed message rates in words per minute (wpm). Notice the y-axis is a logarithmic scale to accommodate extreme observations. Each mark is a single observation for humans or chatbots. The thick dotted horizontal line indicates a cut-off message rate that captures 90.4% of the humans and excludes 60.4% of the bots (between 0 and 70 wpm, inclusive).

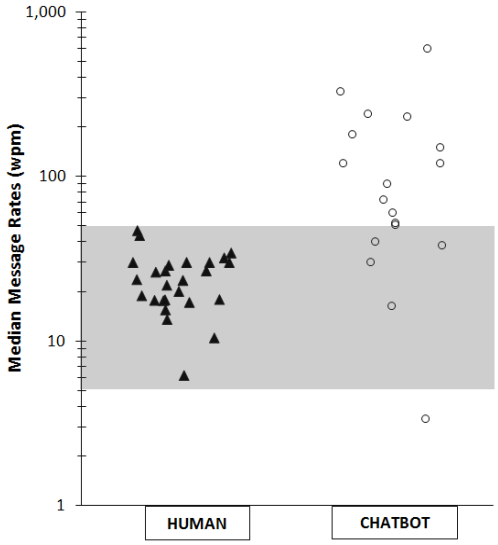


Figure 6. Median message rates per individual, in words per minute (wpm). Notice the y-axis is a logarithmic scale to accommodate extreme values. Each mark is the median of observations for an individual human or chatbot. The shaded grey region indicates the median message rates that include 100% of the humans and exclude 78% of the bots (between 5 and 50 wpm).

data include means, standard deviations, skewness, and kurtosis. Had we attempted bot detection in our dataset by using a simple rule that included only median message rates of individual communicators that fell within the typical human range (5 to 50 wpm), we would have been able to quite easily detect and remove 78% of the bots while excluding no humans. Another simple rule that excluded individual messages with rates above 70 wpm would have included 90.4% of the human messages, while excluding 60.4% of the bot messages.

D. Messaging Complexity

1) Entropy: Analysis of the normalized messaging entropies revealed some differences between human and bot communication entropy. Entropies were calculated per individual communicator. Results are shown in Table IV. As would be expected if humans demonstrated more complexity in their communications than bots, humans showed higher entropies on all three measures. The individual human entropies were 1.7 times higher than bots on the inter-message delays (Fig. 7), 1.4 times higher on message sizes (Fig. 9), and 1.3 times higher on message rates (Fig. 11). Here, the pattern of results is the same as those of Gianvecchio et al. [11], who found the messaging complexities (as measured by the corrected conditional entropies) to be higher for humans in both inter-message delays and message sizes (they did not analyze message rates).Visual analysis of these distributions (Fig. 7, 9, and 11, top row, next page) offers a clear distinction between humans and bots, particularly on the inter-message delay and message rate entropies.

2) Entropy Rate: Analysis of the normalized messaging entropy rates also revealed large differences between human and bot communications. Averages of the individual entropy rate results are shown in Table V. Humans on average exhibited substantially higher entropy rates (i.e., complexity) on all three messaging behaviors. The human inter-message delay entropy rates were on average 2.4 times higher than bots (Fig. 8), message size were 1.6 times higher (Fig. 10), and message rates 1.3 times higher (Fig. 12). These results also support the findings of Gianvecchio et al. [11], who also examined the entropy rates of inter-message delays and message sizes. Visual analysis of the distributions (Fig. 8, 10, and 12, lower row, next page) again shows clear distinctions between humans and bots.

TABLE IV. MEAN INDIVIDUAL NORMALIZED ENTROPY

Normalized Entropy (%)	*Humans*	*Chatbots*
Inter-Message Delay	80.1	46.3
Message Size	76.9	56.9
Message Rate	82.8	64.6

TABLE V. MEAN INDIVIDUAL NORMALIZED ENTROPY RATE

Normalized Entropy Rate (%)	*Humans*	*Chatbots*
Inter-Message Delay	58.0	18.4
Message Size	46.3	19.3
Message Rate	61.9	35.1

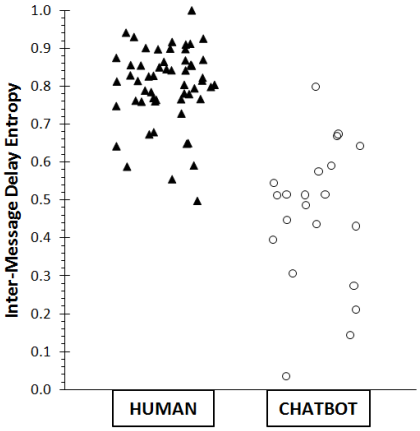


Figure 7. Individual inter-message delay entropy (normalized). Each mark is the observed entropy of an individual human or chatbot.

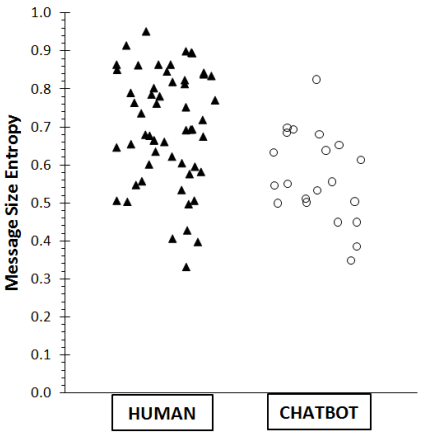


Figure 9. Individual message size entropy (normalized). Each mark is the observed entropy of an individual human or chatbot.

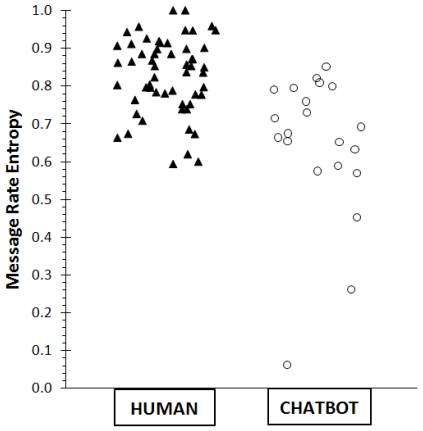


Figure 11. Individual message rate entropy (normalized). Each mark is the observed entropy of an individual human or chatbot.

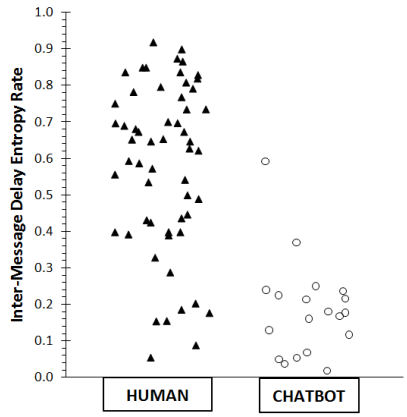


Figure 8. Individual inter-message delay entropy rates (normalized). Each mark is the observed entropy rate of an individual human or chatbot.

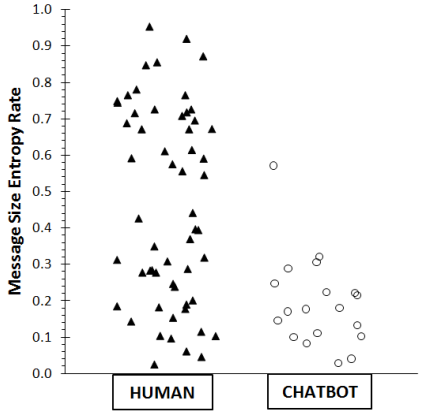


Figure 10. Individual message size entropy rates (normalized). Each mark is the observed entropy rate of an individual human or chatbot.

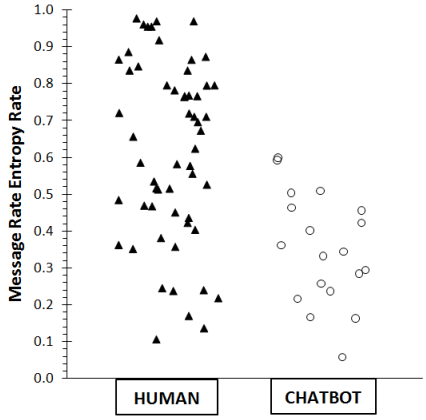


Figure 12. Individual message rate entropy rates (normalized). Each mark is the observed entropy rate of an individual human or chatbot.

IV. Conclusions and Future Work

Despite the nature of our dataset in comparison to others [e.g., 11], our results generally support the utility of chat communication metrics for distinguishing between human and bot communications. Specifically, we found that measures of central tendency, dispersion, distribution shapes, and related patterns of inter-message delays were very distinct between humans and bots. Further, we found that message rates, a derived measure from both inter-message delays and message sizes, could also be of some utility. Our entropy and entropy rate measurements suggested that individual bots generally exhibit lower complexity than individual humans in their chat communications, and this was true for inter-message delays, message sizes, and message rates. However, the statistical and distributional analysis of message sizes seemed to offer few metrics that might aid in the identification of bot communications.

Future research into differentiating metrics for humans and bots might include the analysis of keystroke dynamics, alternative measures of complexity, and other graphical and/or statistical patterns of chat, including word choice/phraseology, typographical error distributions, detailed chronemic analysis, conversational dynamics, etc. We plan on developing visualization tools for chatbot detection that specifically focus on some of the metrics discussed in this work. We are also considering the development of a set of "humanness" metrics for rating the conversational capabilities of chatbots, which may also be advantageously visualized for chat users, chat room administrators, and AI developers.

Chatbots will conceivably continue to progress in sophistication until their conversational capabilities are virtually indistinguishable from humans. In the future, we will find these automated conversational agents in a variety of unforeseen applications, in addition to their current growing applications in commerce, education, and entertainment. However, malicious chatbots can pose considerable security risks since they can be unleashed in virtually unlimited numbers and can work tirelessly on gathering intelligence, spreading malware and spam, and generally hampering

effective collaborations and communications. As bots approach the communication level of an average human, their capability to be used as malevolent tools will also grow. We hope the analysis and discussion presented herein provides improved capabilities to detect and remove malicious automated conversational agents from collaborative computer-mediated communication systems, so that such systems can be used efficiently, effectively, and securely for their intended purposes.

ACKNOWLEDGMENTS

We wish to thank David Dommett at the Air Force Research Laboratory's 711th Human Performance Wing/RHCVZ, for his assistance with the entropy and entropy rate calculations. We also wish to thank Dr. Xiaoping Annie Shen, associate professor in the Department of Mathematics at the Ohio University, for her many helpful comments and critiques on the mathematical portions of this paper.

REFERENCES

[1] S. Quarteroni and S. Manandhar, "A chatbot-based interactive question answering system," Proc. 11th Workshop on the Semantics and Pragmatics of Dialogue, Trento, Italy, pp. 83-90, May 2007.

[2] M. Leaverton, "Recruiting the chatterbots: How virtual agents make the web more human," CNET web article, Oct 2000 (accessed May 2010), available: http://www.cnet.com/4520-6022_1-102077-1.html.

[3] A. Kerly, P. Hall, and S. Bull, "Bringing chatbots into education: Towards natural language negotiation of open learner models," Knowledge-Based Systems, vol. 20, pp. 177–185, March 2007.

[4] S. Kowalski, "Two case studies in using chatbots for security training," Proc. 9th IFIP World Conference on Computers in Education, 2009.

[5] S. Thakur, "AOL no more chat room spam petition," Petition Online, online public web petition (accessed May 2010), available: http://www.petitiononline.com/chatspam/petition.html.

[6] J. Hu, "AOL: spam and chat don't mix," CNET News web article, July 2003 (accessed May 2010), available: http://news.cnet.com/AOL-Spam-and-chat-dont-mix/2100-1032_3-1024010.html.

[7] B. Krebs, "Yahoo! Messenger network overrun by bots," Washington Post web article, Aug 2007 (accessed May 2010), available: http://blog.washingtonpost.com/securityfix/2007/08/yahoo_messenger_network_overru.html.

[8] P. Naughton, "Flirty chat-room 'bot' out to steal your identity," FOXNews web article, Dec 2007 (accessed May 2010), available: http://www.foxnews.com/story/0,2933,316473,00.html.

[9] I. Fried, "Warning sounded over 'flirting robots'," Beyond Binary, CNET web article, Dec 2007 (accessed May 2010), available: http://news.cnet.com/8301-13860_3-9831133-56.html.

[10] A. Mohta, "Yahoo chat: Captcha check to remove bots," TechnoSpot Network web article, Sept 2007 (accessed May 2010), available: http://www.technospot.net/blogs/yahoo-chat-captcha-check-to-remove-bots/.

[11] S. Gianvecchio, M. Xie, Z. Wu, and H. Wang, "Measurement and classification of humans and bots in Internet chat," Proc. 17th USENIX Security Symp., San Jose, CA, July 2008.

[12] J. McIntire, L. McIntire, and P. Havig, "Methods for chatbot detection in distributed text-based communications," Proc. International Collaborative Technologies and Systems Symp., Chicago, IL, May 2010.

[13] The Loebner Prize in Artificial Intelligence: The First Turing Test, annual competition website (accessed May 2010), available: http://www.loebner.net/Prizef/loebner-prize.html.

[14] A. Turing, "Computing machinery and intelligence," in Readings in Cognitive Science: A Perspective from Psychology and Artificial Intelligence, A. Collins and E.E. Smith, Eds. San Mateo, CA: Morgan Kaufmann Publishers, Inc., 1988.

[15] C. Shannon, "A mathematical theory of communication," Bell System Technical Journal, vol. 27, pp.379-423, 623-656, July/Oct 1948.

[16] A. de Siqueira and S. Herring, "Temporal patterns in student-advisor instant messaging exchanges: Individual variation and accomodation," Proc. 42nd Hawai'I International Conf. on Systems Sciences, Los Alamitos, CA: IEEE Press, 2009. Pre-print available: http://ella.slis.indiana.edu/ ~herring/desiqueira.herring.2009.pdf.

[17] Y. Kalman, G. Ravid, D. Raban, and S. Rafaeli, "Pauses and response latencies: A chronemic analysis of asynchronous CMC," J. Computer-Mediated Communication, vol. 12, issue 1, 2006. Available: http://jcmc.indiana.edu/vol12/issue1/kalman.html.

Quantify Effects of Long Range Memory on Predictability of Complex Systems

Xiaoping Shen[1], Katheryn A. Farris[2] and Paul R. Havig[2]
[1]Department of Mathematics, Ohio University, Athens, OH 45701, USA
[2]711 HPW AFRL/RHCV, Wright-Patterson Air Force Base, Dayton, Ohio 45433, USA

***Abstract*— This paper explores the connection between uncertainty and memory effects of time series associated with complex system. Traditionally, information theory based algorithms, such as Shannon entropy and its relatives, are employed as measurements to describe uncertainty quantitatively. This study brings into focus the important role of the long range memory effects on the uncertainty measurements. The method is applicable to arbitrary complex systems. Financial data are investigated as an example. The approach provides important insights into the predictability of a complex system.**

I. Introduction

Uncertainty is pervasive in complex system. Quantify uncertainty and how it propagates through the system is major challenge in complex system modeling and calibration. Many information theory based algorithms, such as Shannon entropy and its relatives, are employed as measurements to describe uncertainty quantitatively. Some parameters or indicators are used as long range memory measurements. Hurst and Holder exponents are the most important and popular parameters. Traditionally, these parameters are studies independently. On the contrary, the study on the relations among these parameters is not commonly found in literature. This study brings into focus the important role of the long range memory effects on the uncertainty measurements. More specifically, it intended to reveal the relation between two important parameters, the Hurst exponent (measurement of global long range memory) and the Entropy (measurement of uncertainty or predictability). Fraction Browning Motion is investigated in numerical experiments. The approach provides a new view to understand the uncertainty and predictability of time series from complex systems.

The paper is organized as follows, after the brief introduction, this section is concluded by providing some necessary background in time series analysis to make the paper self-contained. Section 2 discusses the statistical parameters in modeling non stationary time series. Section 3 is devoted to the discussion of numerical simulation, and the last section provides a summery of the empirical results and future study.

II. Background

We recall brie□y the mathematical and statistical definitions and properties of the Hurst exponents, ApEn and related concepts in this section. General references on self-similar processes and long-memory processes are given in [21] and [6].

A. The Hurst exponent

An important indicator of complex systems is to be able adapt to inputs and evolve. Adaptation and evolution are characteristic of critical infrastructure systems. Self-similarity and chaos are examples are used to describe a complex system qualitatively. One example of complex systems is a the long-memory time series which was brought to the attention by Hurst [16]. It has been of great interests in the research community ever since. Hurst exponent (or parameter) H is named by Mandelbrot in honor of both Harold Edwin Hurst and Ludwig Otto Holder. Since the pioneering work of Mandelbrot and Ness[21], the fractional Brownian motion (fBm) has become widely popular in a theoretical context as well as in the practice of modeling self-similar phenomena.

H: Hurst self-similarity parameter, arises from the generalization of the Brownian motion from integer to fractional. It was interlaced to the applied statistics community by Mandelbrot and van Ness [21] in honor Hurst (the original notation used by Hurst was K) and Holder.

To begin, we define the fractional Brownian motion (fBm) and fractional Gaussian noise (fGn).

Let $(\Omega, \mathcal{F}, P)$ be a complete probability space on which all stochastic processes are defined.

Definition 1: (Fractional Brownian motion). Fractional Brownian motion (fBm) is defined by its stochastic representation

$$B_H(t) := \frac{1}{\Gamma(H+\frac{1}{2})} \int_{-\infty}^{t} (t-\tau)^{H-\frac{1}{2}} dB(t) \qquad (1)$$

where Γ is the Gamma function defined by $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt$ and B is a stochastic process, regular Brownian motion defined on $(\Omega, \mathcal{F}, P)$. H is called the Hurst parameter or exponent of B_H.

Remark 1: The stochastic representation (1) is not unique. The integral can be understood as a Lebesgue-Stieltjes integral. Notice that when H=$\frac{1}{2}$, $B_{1/2}(t) = B(t)$ is a ordinary Brownian motion.

A fractional Brownian motion $B_H = \{B_H(t), 0 \le t < \infty\}$ with Hurst exponent H, $0 < H < 1$ is uniquely characterized by the following properties:

1) $B_H(0) = 0$;
2) $B_H(t)$ is a zero-mean Gaussian process with continuous sample paths and stationary increments; and

3) The autocovariance function is given by

$$\rho_H(s,t) = EB_H(s)B_H(t) = \frac{1}{2}\{s^{2H}+t^{2H}-(t-s)^{2H}\},$$

4) $\mathrm{E}[\mathrm{B}_H^2(t)] = V_H t^{2H}$, where

$$V_H = Var(B_H) = \frac{-\Gamma(2-2H)\cos(\pi H)}{\pi H(2H-1)}.$$

A close related parameter is,

d: fractional integration parameter, arises from the generalization of the Box-Jenkins ARIMA(p,d,q) models from integer to non integer. It was introduced by Granger and Joyeux [15](1980) and Hosking [17] (1981), independently.

Definition 2: (Fractional Gaussian noise). The incremental process X=$\{\mathrm{X}_k : k = 0,1,...\}$ of fBm, $\mathrm{X}_k = B_H(k+1) - B_N(k)$ is called Fractional Gaussian noise.

X_k has following properties,

1) X_k has a standard normal distribution;
2) The autocovariance function

$$\begin{aligned}\gamma(k) &= \frac{1}{2}[|k-1|^{2H} - 2|k|^{2H} + |k+1|^{2H}] \\ &\sim H(2H-1)k^{2H-2}\end{aligned}$$

Two parameters induced from H and d are:

β: power law exponent (to be definied in next section)

For fraction Brownian motion, $\beta = 2H+1$;

For fractional Gaussian noise, $\beta = 2H-1$.

Many algorithms have been developed to compute Hurst exponents. The numerical experiments in this project are based on the built in wavelet based programs in Matlab. The following figures illustrate the property of the Hurst exponent. From Figure 1 to Figure 3, the time series are generated with prescribed Hurst exponent, H=0.2 (negative memory), 0.5 (regular Brownian motion, no memory), and 0.9 (long range memory). The Hurst exponent is an indicator of the smoothness of the time series.

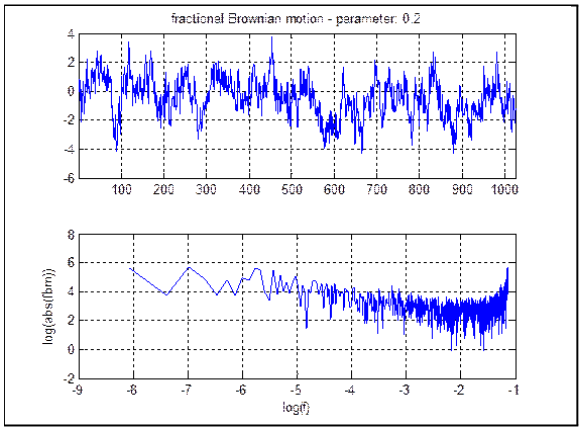


Fig. 1. Fractional Brownian motion – $H = 0.2$.

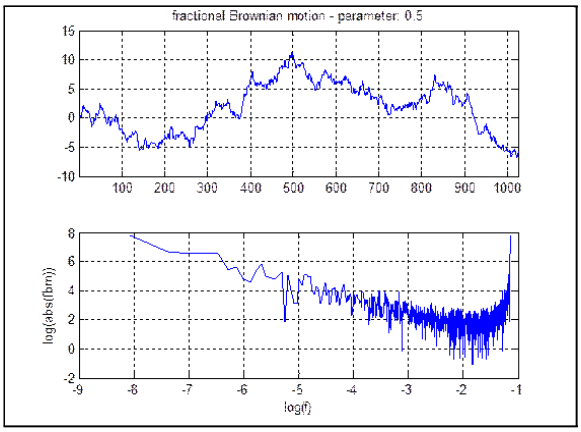


Fig. 2. Fractional Brownian motion – $H = 0.5$.

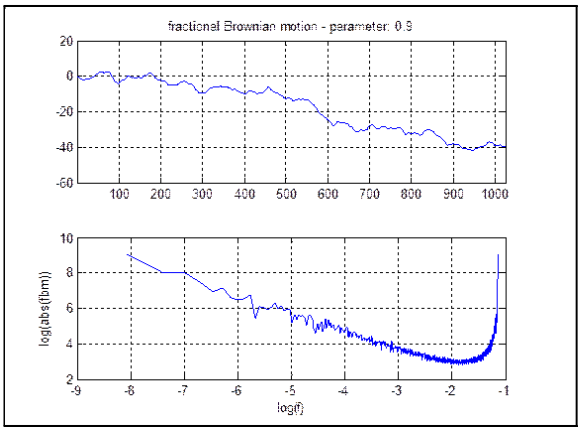


Fig. 3. Fractional Brownian motion – $H = 0.9$.

B. Entropy and its relatives

Shannon entropy [28]in its basic form is a measure of uncertainty rather than a measure of information. Mathematically, it can be defined as,

Definition 3: The entropy of a discrete random variable X is defined by

$$H(X) = -\sum_{x\in X} p(x)\log p(x) \qquad (2)$$

Similarly, entropy for continuous random variable can be defined as,

Entropy is always positive. Entropy measures the uncertainty inherent in the distribution of a random variable.

III. Numerical Simulations

A. Data description and empirical results.

We first simulated the fBm sample data using a wavelet based algorithm. The Hurst exponents are set from 0.025 to 0.975 with a step size of 0.025, total 39 samples. Each sample contains $2^{10} = 1024$ sample points. These data are stored in a matrix with size 1024 by 39.

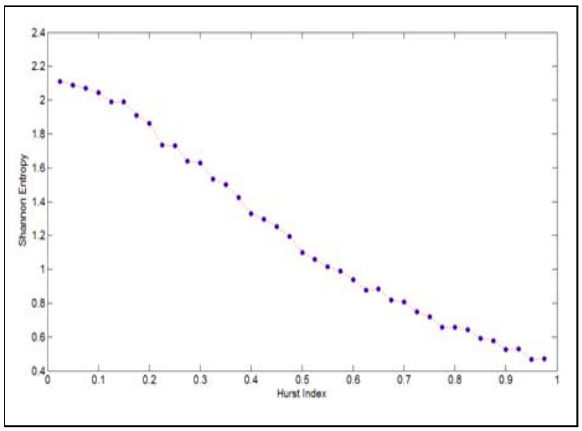


Fig. 4. Relation between Shannon entropy and Hurst exponent

1) Relation between Shannon Entropy and Hurst exponent: We then run Monte Carlo simulation for the Shannon entropies of each fBm random variable. Figure 4 shows the experiment results for the relation between Shannon entropy and Hurst exponents.

The regression results for the relation between Shannon Entropy and Hurst exponent are shown in Table 1.

Table 1. Relation between entropy and Hurst exponent

	Relation	L_2 Error
Linear	$E = -1.89\ H + 2.16$	0.44116
Quadratic	$E = 0.81H^2 - 2.70H + 2.30$	0.25608
Cubic	$E = 1.67H^3 - 1.7H^2 - 1.69H + 2.21$	0.18002

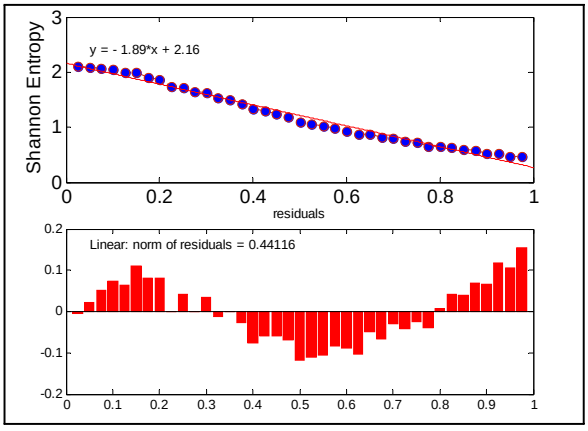


Fig. 5. Linear relatin between entropy and Hurst parameter.

These initial studies revealed two parameters, entropy and Hurst exponent are not independent.

2) Relation of Approximate Entropy and Hurst exponent: In his seminal work, S.M. Pincus, introduced the concept and algorithm of approximate entropy (ApEn) [26]. ApEn is a statistics that quantifies the randomness and quantify the degree of predictability in a time series. For instance, a low

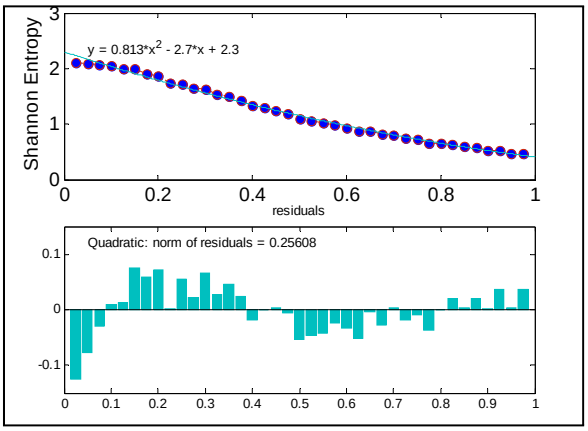


Fig. 6. Quadratic relation between entropy and Hurst parameter.

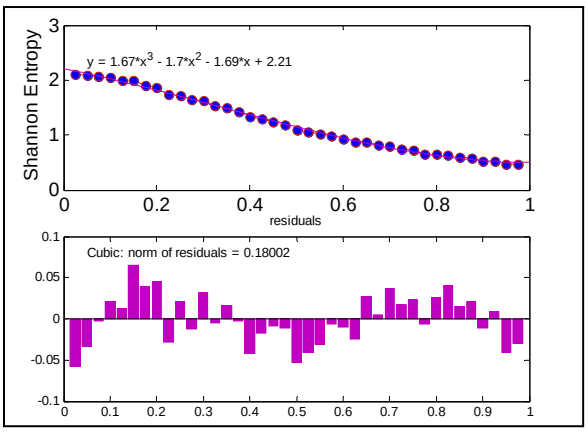


Fig. 7. Cubic relation between entropy and Hurst parameter.

ApEn indicates the time series has many repeated patterns or less randomness. ApEn can defined as,

Definition 4: Assume $X_1, X_2, ..., X_N$ are a sequence of independent and identically distributed random variables (i.i.d). Each random variable takes values in $\{1, ..., N\}$. Denote $Y(i,m) = [X(i),\ X(i+1),\ \ldots,X(i+m-1)]$ where $i = 1$, ..., $N - m + 1$,

$$C_r^m(i) \quad = \frac{1}{N-m+1}\#\{j : 1 \le j \le N-m \\ |Y(j,m) - Y(i,m)| < r\}$$

and

$$\gamma^m(r) = \frac{1}{N-m+1}\sum_{i=1}^{N-m+1} \ln(C_r^m(i))$$

The ApEn value for a finite length N is,

$$\text{ApEn}(m, r, N) = \gamma^m(r) - \gamma^{(m+1)}(r).$$

.

Notice that r is a parameter which measures the similarity of two pieces of a given time series, in particular, $r = 0$, $Y(j,m)$ and $Y(i,m)$ have identical patterns. The uniqueness of this algorithm is that it distinguishes a wide variety of

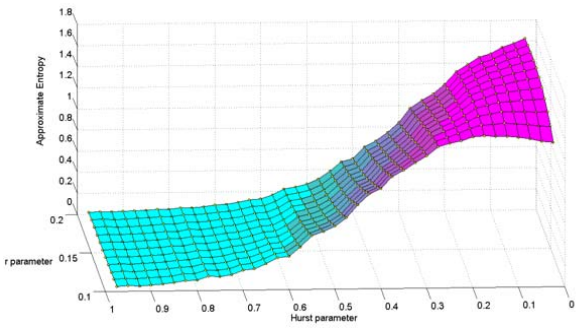


Fig. 8. Relation of the parameter r of ApEn and Hurst exponent H.

systems. In particular, for small m, estimation of ApEn(m, r) by ApEn(m, r, N) can be achieved with relatively few data points. ApEn assigns a non-negative number to a time-series with larger values corresponding to more irregularity in the data. In short, ApEn measures the logarithmic likelihood that runs of patterns which are within the tolerance window (r) for m contiguous operations that remain close to that tolerance window on subsequent incremental comparisons [26]. We employed the numerical algorithm developed in [27] (the related Matlab program is available at the Matlab Central [24])

Using Monte Carlo simulation, for $r = 0.1 : 0.01 : 0.2$, the relation ApEn(m, r, N) and the Hurst exponent is show in Figure 8.

IV. Conclusion and Future Study

To quantify uncertainty or study the predictability of complex system, we attempt to fit the associated time series to models of non stationary process with long memory. The fBm models always play a important role in the non stationary process modeling. In the paper, we report the initial results on calibration fBm and the relations between Shannon entropy, approximate entropy and the Hurst exponent of for fBms. These results will pave a way to build a semi-parametric (based on fBm) multiscale model for quantify uncertainty of complex systems.

Acknowledgment

The first author was on sabbatical leave with the Battlefield Visualization Lab, Air Force Research Lab, Wright-Patterson Air Force Base, and the research is supported in part by grant AFOSR FA8650-08-D-6801.

References

[1] H. D. I. Abarbanel, *Analysis of Observed Chaotic Data*. Springer,
[2] D. Abasolo, R. Hornero, and P. Espino, *Approximate Entropy of EEG Background Activity in Alzheimer's Disease Patients*, Vol. 15, 591-603, 2009.
[3] P. Abry and D. Veitch, Wavelet Analysis of Long-Range-Dependent, Traffic. IEEE Transactions on Information Theory, Vol. 44 (1), 2-15, 1998.
[4] P. Abry, D. Veitch and P. Flandrin, Long-Range Dependence: Revisiting Aggregation with Wavelets. Journal of Time Series Analysis, Vol. 19(3), 253-266, 1998.
[5] R.T. Baille and S-K. Chung, Modeling and forecasting from trend-stationary long memory models with applications to climatology. International Journal of Forecasting, Vol. 18, 215-226, 2002.
[6] J. Beran, *Statistics for Long Memory Processes*. Chapman & Hall/CRC Press, 1994.
[7] R. Bhattacharya, V. K. Gupta and E. Waymire, The Hurst effect under trends. Journal of Applied Probability, Vol. 20, 649-662, 1983.
[8] E.C. Cherry, *A History of the Theory of Information*, Springer, Vol. 13, 233-241, 1961.
[9] A. J. Chorin and O. H. Hald, *Stochastic Tools in Mathematics and Science*. Springer-Verlag, 2006.
[10] R. Dalhaus, Efficient parameter estimation for self-similar processes. The Annals of Statistics, Vol. 17(4), 1749-1766, 1989.
[11] P. Doukhan, G. Oppenheim and M. Taqqu, *Theory and Applications of Long-Range Dependence*. Birkhauser, 2003.
[12] P. Embrechts and M. Maejima, *Selfsimilar Processes*. Princeton University Press, 2002.
[13] R. Fox and M. S. Taqqu, Large-Sample Properties of Parameter Estimates for Strongly Dependent Stationary Gaussian Time Series. The Annals of Statistics, Vol. 14 (2):517-532, 1986.
[14] J. Geweke, and S. Porter-Hudak, The estimation and application of long memory time series models. Journal of Time Series Analysis, 4, 221-237, 1983.
[15] C. W. J. Granger, and R. Joyeux, An Introduction to Long-range Time Series Models and Fractional Differencing. Journal of Time Series Analysis, Vol. 1, 15-30, 1980.
[16] H. E. Hurst, Long-term storage capacity of reservoirs. Transactions of the American Society of Civil Engineers, Vol. 116, 770-808, 1950
[17] J. R. M. Hosking, Fractional Differencing. Biometrika, 68(1), 165-176, 1981.
[18] H. Kantz and T. Schreiber, *Nonlinear Time Series Analysis*, Cambridge University Press, 1997.
[19] S. Lu, X. Chen, J.K. Kanters, I.C. Solomon, and K.H. Chon, *Automoatic Selection of the Threshold Value r for Approximate Entropy, Biomedical Engineering*, IEEE Transactions on, Vol. 58 (8), 2008.
[20] J. Macha, *Entropy, Information and Computation*, Am. J Phys. 67, 1074-1077, 1999.
[21] B. B. Mandelbrot and J. W. van Ness, Fractional Brownian Motions, Fractional Noises and Applications, SIAM Review, Vol. 10(4), 422-437, 1968.
[22] E. J. McCoy and A. T. Walden, Wavelet Analysis and Synthesis of Stationary Long-Memory Processes. Journal of Computational and Graphical Statistics, Vol. 5(1), 26-56, 1996.
[23] Y. Meyer, F. Sellan and M. S. Taqqu, Wavelets, generalized white noise and fractional integration: The synthesis of fractional Brwonian motion, J. Fourier Analysis and Application, Vol. 5, 465-494, 1999.
[24] A. Parnandi, Approximate entropy.m, available at http://www.mathworks.com/matlabcentral/fileexchange/26546-approximate-entropy, date visited 06/06/ 2011
[25] C. Papadimitriou, K. Karamanos, F.K. Diakonos, V. Constantoudis and H. Papgeorgiou, *Entropy analysis of natural language written texts*, Physica A Vol. 389, 3260-3266, 2010.
[26] S.M. Pincus, *Approximate entropy as a measure of system complexity*, Proc. Ntl. Acad. Sci. 88, 2297-3301, 1991.
[27] J.S. Richman, J.R. Moorman, *Physiological time-series analysis using approximate entropy and sample entropy*, Am. J. Physiol. Heart Circ. Physiol. Vol. 278(6), H2039-49, 2000.
[28] C.E. Shannon, *A Mathematical Theory of Communication*, Bell System Technical Journal, Vol. 27, pp. 379-423, 623-656, 1948.
[29] M. Taqqu, V. Teverovsky, and W. Willinger, Estimators for long-range dependence: an empirical study. Fractals, 3(4), 785-798, 1995.
[30] F. Tillman and B.R. Russell, *Information and Entropy*, Springer, Vol. 13, 233-241, 1961.
[31] D. Veitch and P. Abry, A Wavelet-Based Joint Estimator of the Parameters of Long-Range Dependence. IEEE Transactions on Information Theory, Vol. 45(3), 878-897, 1999.

On Using Information-Theoretic Quantities in Characterization Dissimilarity of DNA Strings

F. Mohd-Zaid[1], X. Shen[2] and K. A. Farris[3]

April 25, 2012

Abstract. **To discern similarity and differences in partial DNA strings based on dissimilarity (distance/difference) among the various SNPs, one of the challenge aspects is to select felicitous metrics or measurements. Some of information theoretic quantities are often employed in practice. Unfortunately, certain information-theoretic variables for example, information distance and mutual information, may not yield consistent results for decision-making. In this paper, we investigate the consistency of information theoretic quantities Experiments are designed to show that the selection of measures and metrics in information-theoretic based analysis is crucial for decision-making. Future possible research directions are discussed.**

Keywords: Distance metric, information theory, DNA, SNPs

1. Problem Specification.

SNPs (single nucleotide polymorphisms) are DNA sequence variation that occurs when a single nucleotide in the genome differs. SNP arrays are a type of DNA microarray that detect SNP occurrences and act as samples of DNA strings that can be extracted from microchips (hardware) and other devices that come in contact with the DNA of living organisms. These SNP arrays do not represent a complete DNA string, which, e.g. for a human, would consist of about 3.2×10^9 base pairs of the human chromosome. A typical SNP arrays would represent a fragment of this string with a length of, perhaps, up to 500,000 base pairs. Each base pair of the human DNA may be in one of four states (A, C, T, or G). The goal is to correctly identify genetic sequences of different individuals to help classify chromosomal regions where genetic variants are shared. For crops and animals, the study of SNPs is important in fertilization and breeding. For human DNA, the extracted SNPs may define how people contract diseases and respond to certain treatments, drugs, vaccines, chemicals, pathogens and other agents.

In memory of Dr. Daniel R. Repperger. BIOCOMP'12 - The 2012 International Conference on Bioinformatics & Computational Biology, Las Vegas, USA.

1- Contact author. 711 HPW, AFRL, WPAFB, Ohio 45433-7022. Email: Fairul.Mohd-Zaid@wpafb.af.mil

2- Dept. of Math., Ohio University, Athens, Ohio 45701. Email: shenx@ohio.edu

3- 711 HPW, AFRL, WPAFB, Ohio 45433-7022. Email: Katheryn.Farris@gmail.com

SNPs may be great enablers in developing personalized medicine, but there is some controversy of how this information may be possibly abused. Some basic definitions are appropriate in this work:

Definition 1: Phenotype – Is a measure of a trait/skill of an individual.
Definition 2: Genotype – The information carried by the genes.
Definition 3: Homozygous – The chromosomes are identical in every state.
Definition 4: Heterozygous – There exists a SNP between two chromosomes.

It is noted in Definition 4, that a SNP is not a weighted difference, in the sense that no distinction has been made between the states, e.g. A and G as being further apart from A and C. Future work may weigh different pair combinations as having distances between SNPs that are predicated on which base pairs are involved. For example, in Figure (1) three DNA strings are shown which are constructed, for simplicity, from a hypothesized 8 base pair fragment of DNA. For simplicity, the notation will be used that A=1, C=2, T=3, and G = 4, for the cells (alleles) although they are categorical variables. It is seen that DNA_1 and DNA_2 differ from each other by only one base pair. However, DNA_1 and DNA_3 differ by four base pairs. In some similarity sense using a distance/difference metric then DNA_1 is closer to DNA_2 and DNA_1 is further apart from DNA_3. This paper will investigate how to characterize the distance/difference of the various SNPs to discern similarity and differences in partial DNA strings. The use of information-theoretic variables will be employed to study the use of a measure of distance of SNPs via mutual information as well as alternative means.

Since the investigation of the SNPs will clearly depend on the appropriate measure of distance/difference between candidate DNAs, the use of classical information theoretic variables will be employed. Figure (2) displays an information theory channel [1-4]. In Figure (2) – Basic elements of an Information Channel from Shannon [1]. Figure (3) is a Venn diagram of the key information-theoretic measures involving two random variables X and

Y (Cover and Thomas [2], Sheridan and Ferrell [3], Repperger, et al. [4]).

DNA_1	1	2	3	4	3	2	1	2
DNA_2	1	2	1	4	3	2	1	2
DNA_3	4	2	2	4	3	2	2	1

$distance(DNA_1\text{-}DNA_3) > distance(DNA_1\text{-}DNA_2)$

Figure (1) – Three DNA strings with Different Relative Distances.

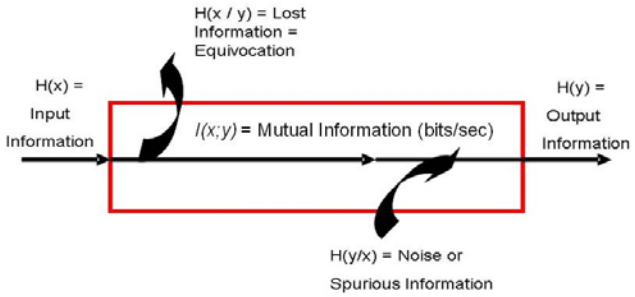
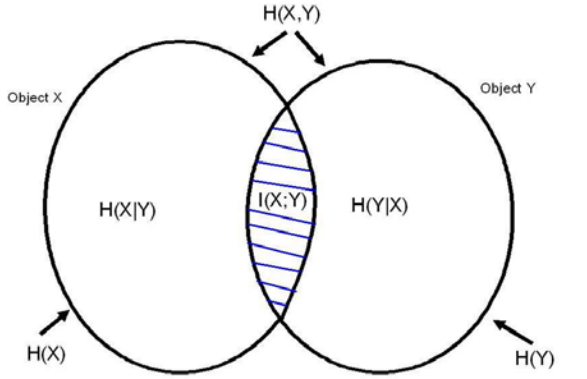


Fig. (2) – Basic elements of an Information Channel from Shannon [1]



$$H(x,y) = I(x;y) + D_R(x;y) = I(x;y) + H(x|y) + H(y|x)$$

The Information Variables in a Venn Diagram

Figure (3) – A Venn Diagram of the Key Variables

In Figure (3) the five information-theoretic quantities that describe the types of uncertainties (entropies) between the input and output elements of the information channel in Figure (2) are portrayed. Three of these five variables can be shown to be independent.

From Figures (2,3), the five basic entities of an information channel can be expressed as follows:

$H(x)$ = The input uncertainty to the channel (1)
$H(y)$ = The output uncertainty of the channel. (2)
$H(x/y)$ = Equivocation lost to the environment. (3)
$H(y/x)$ = Spurious uncertainty from the environment (4)
$I(x;y)$ = Mutual information transmitted (5)

More specifically, equations (1-5) can be better described by letting $p(.)$ represent the probability of an event. For an information channel with input symbol set in Figure (2), $x \in X$, of size n, and received symbols $y \in Y$ at the output set of size q (q may not equal n), the following entropy ($H(.)$) relationships can be defined:

$$H(x) = \sum_{i=1}^{n} p(x_i)\log_2(1/p(x_i)) \quad (6)$$

$$H(y) = \sum_{j=1}^{q} p(y_j)\log_2(1/p(y_j)) \quad (7)$$

$$H(x,y) = \sum_{i,j}^{n,q} p(x_i,y_j)\log_2(1/p(x_i,y_j)) \quad (8)$$

$$H(x/y) = \sum_{i,j}^{n,q} p(x_i,y_j)\log_2(1/p(x_i \mid y_j)) \quad (9)$$

and $$H(y/x) = \sum_{i,j}^{n,q} p(x_i,y_j)\log_2(1/p(y_j \mid x_i)) \quad (10)$$

The important relationships that pertain to the modeling of the SNPs are dependent on the key variables (1-5). From Figures (2,3) and the basic definitions (6-10), the following relationship can be shown to be true (Cover & Thomas [2]):

$$I(x;y) = H(x) + H(y) - H(x,y) \quad (11)$$

where the mutual information $I(x;y)$ also satisfies:

$$I(x;y) \geq 0 \quad (12)$$

Finally, another important variable that will be used in the sequel is the relative information distance $D_R(x;y)$:

$$D_R(x;y) = H(x/y)+H(y/x) = H(x)+H(y) - 2I(x;y) \quad (13)$$

where $D_R(x;y)$ also has a positivity property, as in equation (12):

$$D_R(x;y) \geq 0 \quad (14)$$

There are advantages the variable D_R provides over $I(x;y)$ which are known in the literature (Cover & Thomas, [2], [5], and [6]) and restated here:

Property 1: $D_R(x;y)$ is a metric; however, $I(x;y)$ is only a measure. Please see the appendix and a counter example where $I(x;y)$ fails as a metric by not satisfying the triangular inequality.

A second property can be stated as follows:

Property 2: The relative information distance metric $D_R(x;y)$ is the complement of $I(x;y)$, i.e.

$$D_R(x;y) = \bar{I}(x;y) \quad \text{or} \quad I(x;y) = \bar{D}_R(x;y) \quad (15)$$

Appendix A demonstrates this second property.

2. Methods and Technical Solutions

Contingency tables (Sheridan and Ferrell [3], Repperger, et al., [4], and Kullback [7]) will be used to formulate the SNP similarity and difference problem to utilize information-

theoretic quantities in examining the distance/difference between DNA strings.

Using the DNA strings in Figure (1), Contingency Table 1 is constructed which compares DNA_1 versus DNA_2 in terms of similarity and differences. Contingency Table 2 then compares DNA_1 versus DNA_3, and finally Contingency Table 3 compares DNA_2 versus DNA_3.

Contingency Table 1 – DNA_1 versus DNA_2

	$DNA_2 \rightarrow$			
$DNA_1 \downarrow$	A=1	C=2	T=3	G=4
A = 1	2	-	-	-
C = 2	-	3	-	-
T = 3	1	-	1	-
G = 4	-	-	-	1

Contingency Table 2 – DNA_1 versus DNA_3

	$DNA_3 \rightarrow$			
$DNA_1 \downarrow$	A=1	C=2	T=3	G=4
A = 1	-	1	-	1
C = 2	1	2	-	-
T = 3	-	1	1	-
G = 4	-	-	-	1

Contingency Table 3 – DNA_2 versus DNA_3

	$DNA_3 \rightarrow$			
$DNA_2 \downarrow$	A=1	C=2	T=3	G=4
A = 1	-	2	-	1
C = 2	1	2	-	-
T = 3	-	-	1	-
G = 4	-	-	-	1

Next, a normalized matrix is calculated based on the total number of responses in each table. The normalized matrices are summarized below for Contingency Tables 1-3.

Table 1 – Normalized

$DNA_1 \downarrow$ / $DNA_2 \rightarrow$					$H(x) \downarrow$
	2/8	0	0	0	2/8
	0	3/8	0	0	3/8
	1/8	0	1/8	0	2/8
	0	0	0	1/8	1/8
	3/8	3/8	1/8	1/8	
$H(y) \rightarrow$					

Table 2 – Normalized

$DNA_1 \downarrow$ / $DNA_3 \rightarrow$					$H(x) \downarrow$
	0	1/8	0	1/8	2/8
	1/8	2/8	0	0	3/8
	0	1/8	1/8	0	2/8
	0	0	0	1/8	1/8
	1/8	4/8	1/8	2/8	
$H(y) \rightarrow$					

Table 3 – Normalized

$DNA_2 \downarrow$ / $DNA_3 \rightarrow$					$H(x) \downarrow$
	0	2/8	0	1/8	3/8
	1/8	2/8	0	0	3/8
	0	0	1/8	0	1/8
	0	0	0	1/8	1/8
	1/8	4/8	1/8	2/8	
$H(y) \rightarrow$					

To calculate the requisite entropies, the following procedures are then employed:

Step 1: Calculate $H(x)$ across the rows and then summing down the column on the right side of the normalized matrix (cf. Table 1-Normalized).
Step 2: Calculate $H(y)$ down the columns and then summing across the row on the bottom of the normalized matrix (cf. Table 1-Normalized).
Step 3: Calculate $H(x,y)$ for all cells in the normalized matrix. Then

$$I(x;y) = H(x) + H(y) - H(x,y) \quad (16)$$

and

$$D_R(x;y) = H(x/y)+H(y/x) = H(x) + H(y) - 2\,I(xy;). \quad (17)$$

The calculations proceed as follows for Table 1, for example:

$$H(x) = -2 * (2/8)\log_2(2/8) - (3/8)\log_2(3/8) - (1/8)\log_2(1/8) = 1.9056 \text{ bits} \quad (18)$$

$$H(y) = -2*(3/8)\log_2(3/8) - 2*(1/8)\log_2(1/8) = 1.8113 \text{ bits} \quad (19)$$

$$H(x,y) = -(3/8)\log_2(3/8) - (2/8)\log_2(2/8) - (3)*(1/8)\log_2(1/8) = 2.1556 \text{ bits} \quad (20)$$

$$I(x;y) = H(x)+H(y)-H(x,y) = 1.5613 \text{ bits} \quad (21)$$

$$D_R = H(x)+H(y)-2I(x;y) = 0.594 \text{ bits} \quad (22)$$

Finally, it is noted in Figure (1) that in a distance/difference sense, it is expected that:

$$\text{dist. } (DNA_1 - DNA_3) > \text{ dist.}(DNA_1 - DNA_2) \qquad (23)$$

Table 4 summarizes these results. It is seen that dissimilarities between DNAs are generally associated with large D_R values, small $I(x;y)$ values, and larger Hamming distance values. The Hamming distance (independent of position) is defined as the percent of cells that differ in a dyadic comparison and is the gold standard in discerning differences between computer words.

Table 4 – Distances between the SNPs in Figure (1)

Distance Variable	DNA_1-DNA_2	DNA_2-DNA_3	DNA_1-DNA_3
Hamming	0.125	0.50	0.50
$I(x;y)$	1.5613	1.3113	0.9056
$D_R(x;y)$	0.5944	1.1887	1.8444

Typically a reduction in the value of D_R would be accompanied by an increase in $I(x;y)$. For the two random variable case (as shown in the appendix), it can be demonstrated that D_R and $I(x;y)$ are complements of each other (i.e. $\overline{D}_R = I(x;y)$ and $\overline{I}(x;y) = D_R$). The results of Table 4 are consistent. As the Hamming distance increases (column 2 in row 2) when compared to either column 3 or column 4, then $I(x;y)$ decreases and D_R increases, as expected. Two counter examples are now presented.

3. Empirical Evaluation

The first counter example is illustrated with Venn diagrams in Appendix A which shows that $I(x;y)$ is not consistent in discerning distance/differences between DNAs because it does not satisfy the triangular inequality.

Case 1 Counter Example with Venn Diagrams
Please see appendix A for an example using Venn diagrams and set theory. This presentation is based on geometric arguments. It is shown that $I(x;y)$ violates the triangular inequality thus does not satisfy the property of being a norm. The second counter example deals with SNPs.

Case 2 Counter Example with SNPs
To generalize the counter example, analogous to the Venn diagrams in appendix A to DNA identification, the following three DNA strings are constructed:

DNA_1	1	2	3	4	3	2	1	3	2	4
DNA_2	1	3	2	1	2	4	1	4	3	1
DNA_3	1	2	2	4	3	2	3	4	2	4

Figure (4) – Counter Example in terms of SNPs

To show that difficulties may occur by using $I(x;y)$ as well as D_R to characterize distance/difference between DNAs, the three normalized matrices resulting from the contingency tables are displayed for the counter example DNAs in Figure (4). Using similar notation, as before, Table 5 portrays

Table 5 – Normalized

DNA_1 \ DNA_2 →					$H(x)$ ↓
	2/10	0	0	0	2/10
	0	0	2/10	1/10	3/10
	0	2/10	0	1/10	3/10
	2/10	0	0	0	2/10
$H(y)$ →	4/10	2/10	2/10	2/10	

Table 6 – Normalized

DNA_1 \ DNA_3 →					$H(x)$ ↓
	1/10	0	1/10	2/10	4/10
	0	1/10	1/10	0	2/10
	0	2/10	0	0	2/10
	0	1/10	0	1/10	2/10
$H(y)$ →	1/10	4/10	2/10	3/10	

Table 7 – Normalized

DNA_2 \ DNA_3 →					$H(x)$ ↓
	1/10	0	1/10	0	2/10
	0	3/10	0	0	3/10
	0	1/10	1/10	1/10	3/10
	0	0	0	2/10	2/10
$H(y)$ →	1/10	4/10	2/10	3/10	

DNA_1 versus DNA_2, Table 6 shows DNA_2 versus DNA_3, and Table 7 illustrates DNA_1 versus DNA_3. The calculations from Tables 5-7 are summarized in Table 8. Also enclosed in this table is the calculation from the Hamming distance, which has been a traditional measure of distance between computer words [8,9].

Table 8 – Results of the Calculation of the SNPs in Figure 4

Information-theoretic Variable (bits)	DNA_1 vs. DNA_2 Table 5	DNA_1 vs. DNA_3 Table 7	DNA_2 vs. DNA_3 Table 6
H(x)	1.971	1.9710	1.9219
H(y)	1.9219	1.8464	1.8464
H(x,y)	2.5219	2.6464	2.9219
I(x;y)	1.371	1.1710	.8464
$D_R(x;y)$	1.151	1.4755	2.0755
Hamming Distance	.80	.30	.70

Note the following conclusions are reached from Table 8:
(1) The Hamming Distance (gold standard) is a well accepted metric and will be used as a baseline (ground truth) to evaluate the information-theoretic variables investigated herein.
(2) In Table 8, bottom row, comparing column 3 to column 4 (results of Table 7 versus Table 6), as the Hamming distance increased (from .3 in column 3 to 0.7 in column 4) then *I(x;y),* decreased from 1.171 to 0.8464 which was expected. Also, D_R increased, accordingly.
(3). However, when comparing column 3 to column 2 (results of Table 5 versus Table 7), when the Hamming distance increased from 0.3 to 0.8, the *I(x;y)* should have decreased, but it increased from 1.171 to 1.371, which is inconsistent. This same inconsistency occurred with the variable D_R. Thus the information variables differ in their determination of distance/difference between DNAs and are not consistent with the ordering provided by the Hamming metric.

Next a discussion is presented on the transitive property of key variables and related to the measures and metrics discussed so far involving decision making, in general.

4. Transitive Property of Measures/Metrics

From Logic: Definition: A dyadic relation R is said to be transitive in a set S if whenever a R b and b R c imply a R c. For example, the relation "is greater than or equal" satisfies the transitive property for scalar numbers.

The structure of transitivity is the mainspring of deductive reasoning. An argument is said to be deductive when the truth of the conclusion is purported to follow necessarily. Deductive reasoning is one of the two basic forms of valid reasoning. While inductive reasoning argues from the particular to the general, deductive reasoning argues from the general to a specific instance. The basic idea is that if something is true of a class of things in general, this truth applies to all legitimate members of that class. The key, then, is to be able to properly identify members of the class. Miss-classifying (or miss-categorizing) will result in invalid conclusions and affecting decision making, adversely.

One of the most common and useful forms of deductive reasoning is the syllogism. The syllogism is a specific form of argument that has three easy steps, for example

1 . Every X has the characteristic Y. This thing is X.
2. Therefore, this thing has characteristic Y.

Also, the transitive property makes elimination possible; if a R b and b R c, we can eliminate b and assert a R c.

Finally, as applied to decision making, if a decision is made that the distance/difference between two DNAs is greater for one pair as compared to another pair, then the data may be mined out if the goal was to find highly correlated DNA pairs. Using *I(x;y)* may lead to an error by mining out more correlated pairs of DNAs. If a distance metric such as the Hamming distance (as discussed in this paper) were employed, then the conclusion would not suffer from that error. As mentioned previously, the weakness of the Hamming distance is that it is a relative measure, not an absolute measure (the position of where the SNPs are is lost).

5. Significance and Impact

Decision making based on closeness as measured by distance/difference between candidate DNAs is critically important if DNA analysis is used to make accurate determinations in data. Problems of consistency are seen when selecting D_R and mutual information ($I(x;y)$), being widely used in the literature. The property that D_R is a metric and *I(x;y)* is only a measure, demands that proper decision making should be predicated on at least a good measurement tool (D_R in lieu of *I(x;y)*). Apparently D_R satisfying the triangular inequality still does not guarantee consistency in the decision making, as shown earlier, on the decision regarding simple binary choice of a string of DNA being more or less similar.

6. Future Work and New Research Directives

As mentioned previously, the classification of the similarity and differences between sample DNAs and the causality mapping between the SNP's scripts with the phenotype traits is a wide open area of research. A discussion on some of the fundamental problems in this area and possible solutions are now conducted. First some basic history is presented.

The human genome project has its early roots in the 1940's when the Department of Energy made an effort to develop new energy resources and still understand the potential health and environmental risks associated with these resources. In 2001, two publications [10, 11] described the initial sequencing and analysis of the human genome. By 2003, the sequencing was completed, two years earlier than anticipated. The generalizations are now far reaching. The DNA in each human cell is packaged into 46 chromosomes

arranged into 23 pairs. Each chromosome contains many genes (approximately 25,000 for the human genome), which are the basic physical and functional units of heredity. Genes are specific sequences of bases that encode instructions on how to make proteins. It is from the action of the proteins that the phenotype traits emerge.

Previously discussed, SNPs are variations in the DNA that may be extracted as SNP arrays by microchips or by other processes. The question arises if the sample is representative of that portion of the DNA string being relevant to the phenotype trait of interest? This is better understood from some other properties that reside within the human DNA sequencing: (1) Only about 2% of the genome actually encodes the instruction for the synthesis of proteins, (2) The human genome sequence is almost (99.9%) exactly the same in all people, and (3) particular gene sequences in animals have been associated with numerous diseases and disorders, including breast cancer, muscle disease, deafness, and blindness. For example, in a mouse, [12] cancer susceptibility can be related to new gene-mapping resources and specific genes can be indentified that concur with mice that contract the disease.

The tumor classification problem is of high interest in the field of bioinformatics [13]. The design of the candidate chips to extract the fragment DNA (SNPs) is a problem of considerable concern. Such systems are far from perfect and the environment can exert an undue influence in the process. The environment can mutate certain genes, thus producing a gene with a higher vulnerability to disease. For example, exposure to smoking is known to mutate a gene and thus produce cells that may start developing cancer. Thus if only one difference occurs in a base pair, this is still very important to capture since it may greatly influence a phenotype trait.

The future problems that may be studied in this area can be investigated in an algorithmic way on how certain SNPs signatures may result in a phenotype trait. The trait could be a "good attribute" like resistance to disease, increased strength, size, and other qualities. Alternatively, the modified SNP signature may also be a "bad attribute" including susceptibility to viruses, diseases, etc. For simplicity of discussion, the presumption will be made that the phenotype trait will exist in only two states, e.g.

Phenotype trait 1: No disease outcome (being resistant to a specific disease).

Phenotype trait 2: Being vulnerable to a specific disease.

Assume four DNA samples are taken from four individuals that equally fell into one of the two states above. Table 9 would classify the four DNA samples:

Table 9 – Four DNA samples obtained

Individual Number	No Disease State	Disease State
1	DNA_1	
2		DNA_2
3	DNA_3	
4		DNA_4

Recall that only 2% of the DNA is related to producing proteins that will affect the phenotype outcomes, then to develop the similarities and differences between the sample DNAs in Table 9, the following six steps should be conducted:

Step 1: Remove all common alleles (this includes the 98% of the DNA not associated with the protein production). Let the symbol Ω represent those common cells (alleles) that ***are not*** related to differences between the DNAs. In a set theory description, it represents the intersection of all the sample DNAs, i.e.

$$\Omega = DNA_1 \cap DNA_2 \cap DNA_3 \cap DNA_4 \quad (24)$$

Then let the underlined notation characterize that part of each DNA_i different from the common intersection of all SNPs, i.e.

$$\underline{SNP}_1 = DNA_1 - \Omega \quad (25)$$
$$\underline{SNP}_2 = DNA_2 - \Omega \quad (26)$$
$$\underline{SNP}_3 = DNA_3 - \Omega \quad (27)$$
$$\underline{SNP}_4 = DNA_4 - \Omega \quad (28)$$

Next, from Table 9, take those common SNP values for the diseased State:

Step 2: $$SNP_D = \underline{SNP}_2 \cap \underline{SNP}_4 \quad (29)$$

To characterize those common SNP values for the non diseased state:

Step 3: $$SNP_{ND} = \underline{SNP}_1 \cap \underline{SNP}_3 \quad (30)$$

Step 4: Now check if the disease and non diseased SNP portions are mutually exclusive: Is it true that:

$$SNP_D \cap SNP_{ND} = \phi \quad (31)$$

where ϕ is an empty set? If (31) is not true, then recalculate steps 1-3 until the result in equation (31) is satisfied.

Step 5: Now repeat steps 1-4 for more than two individuals.

Step 6: With a sufficient data base built up on the two classes $\{SND_D\}$ and $\{SND_{ND}\}$ predictions can then be made for individuals ***outside the data used to develop the two classes***. This will test the efficacy of this method.

References

[1] Shannon, C. E. (1949). Communications in the presence of noise, *Proceed. of the IRE*, **37**, 10-22.

[2] Cover, T. M. and Thomas, J. A. (1991). *Elements of Information Theory*, John Wiley & Sons, Inc.

[3] Sheridan, T. B. and Ferrell, W. R. (1981).*Man-Machine Systems: Information, Control, and Decision Models of Human Performance*, The MIT press, Cambridge, Mass.

[4] Repperger, D. W., Roberts, R. G., Lyons, J. B., and Ewing, R. L., "Optimization of an Air Logistics Systems via a Genetic Algorithm Model," To appear in *International Journal of Logistics Research*, 2011.

[5] J. P. Crutchfield, "Information and Its Metric," in *Nonlinear Structures in Physical Systems-Pattern Formatio Chaos and Waves*, L. Lam and H. C. Morris, Eds., L. Lam and H. C. Morris, Eds., Springer-Verlag, NY (1990), 119-130.

[6] C. H. Bennett, P. Gacs, M. Li, P. M. B. Vitanyi, and W. H. Zurek, "Information Distance," *IEEE Transactions on Information Theory*, **44**(4), July, 1998, pp. 1407- 1423.
[7] S. Kullback, *Information Theory and Statistics*, New York, Dover, 1968.
[8] W. Zhao, E. Serpedin, and E. R. Dougherty, "Inferring Connectivity of Genetic Regularity Networks Using Information-Theoretic Criteria," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*," **5**(2), April-June, 2008, pp. 262-274.
[9] A. G. O'yachkov and D. C. Torney, "On Similarity Codes," *IEEE Transactions on Information Theory*, **vol. 46**, no. 4, July, 2000, pp. 1558-1564.
[10] E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, and J. Baldwin, "Initial Sequencing and Analysis of the Human Genome," *Nature*, **409**, 2001, pp. 860-921.
[11] J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, and G. G. Sutton, "The Sequence of The Human Genome," *Science*, **291**, 2001, pp. 1304-1351.
[12] P. Demant, "Cancer Susceptibility in the Mouse: Genetics, Biology, and Implications for Human Cancer," *Nature Reviews/Genetics*, **vol. 4**, September, 2003, pp. 721-735.
[13] R. Desper, J. Khan, A. A. Schaffer, "Tumor Classification Using Phylogenetic Methods on Expression Data," *Journal of Theoretical Biology*, **228**, 2004, pp. 477-496.
[14] M. Li, X. Chen, L. Xin, M. Bin, and P. M. B. Vitanyi, "The Similarity Metric," *IEEE Transactions on Information Theory*, **50**,(12), Dec 2004, pp. 3250-3264.

Appendix A – Counter Example 1 – With Venn Diagrams

Since geometric proofs using Venn diagrams are not technically permissible (Eves, [26]), we show as a test of relationships properties 1 and 2. In this appendix, it will be stated (Cover and Thomas, [2]) that D_R is a metric and satisfies the follow following four relationships for a metric $\rho(x,y)$:

(M-1) $\rho(x,y) > 0$ if x ≠ y. (positivity) (A.1)
(M-2) $\rho(x,y) = \rho(y,x)$ (similarity) (A.2)
(M-3) $\rho(x,z) \leq \rho(x,y) + \rho(y,z)$ (triangular inequality) (A.3)
(M-4) $\rho(x,y) = 0$ if and only if x = y (A.4)

However, it is shown by a testing example below that $I(x;y)$ violates equation (A.3), i.e.

$$I(x;z) > I(x;y) + I(y;z) \quad (A.5)$$

for three random variables X, Y, and Z.

Part A – A Constructed Example to Show That Equation (A.3) is Violated:

Figure (A-1a) is presented to define areas A_1, A_2, and A_3 consistent with Figure (3):

$$H(x|y) = A_1 \quad (A.6)$$
$$H(y|x) = A_3 \quad (A.7)$$
$$I(x;y) = A_2 \quad (A.8)$$

Figure (A-1b) now generalizes this concept to three random variables X, Y, Z. In terms of the seven areas (A_1-A_7) displayed, the following relationships become generalizations of Figure (A-1a) into Figure (A-1b):

$$I(x;y) = A_2 \quad (A.9)$$
$$H(x/y) = A_1. \quad (A.10)$$

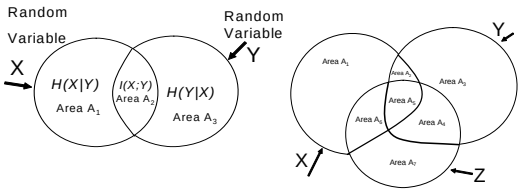


Figure (A-1a) Two Random Variables X and Y (left)
Figure (A-1b) – Three Random Variables X, Y, and Z

$$H(y/x) = A_3, \quad (A.11)$$
$$H(x|y) = A_1 + A_6, \quad I(x;y) = A_2 + A_5 \quad (A.12)$$
$$H(y|x) = A_3 + A_4, \quad I(y;x) = A_5 + A_2 \quad (A.13)$$
$$H(z|x) = A_4 + A_7, \quad I(z;x) = A_5 + A_6 \quad (A.14)$$
$$H(x|z) = A_1 + A_2, \quad I(x;z) = A_6 + A_5 \quad (A.15)$$
$$H(y|z) = A_2 + A_3, \quad I(y;z) = A_5 + A_4 \quad (A.16)$$
$$H(z|y) = A_6 + A_7, \quad I(z;y) = A_4 + A_5 \quad (A.17)$$

the left and the random variable Y to the right until:

$$A_6 > A_2 + A_4 + A_5. \quad (A.18)$$

But:
$$A_6 = I(x;z) - A_5 \quad (A.19)$$
$$A_2 = I(x;y) - A_5 \quad (A.20)$$
$$A_4 = I(y;z) - A_5 \quad (A.21)$$

Hence from (A.18):

$$I(x;z) - A_5 > I(x;y) - A_5 + I(y,z) - A_5 + A_5 \quad (A.22)$$

by construction, and

$$I(x;z) > I(x;y) + I(y;z) \quad (A.23)$$

Thus it is demonstrated that equation (A.5) is satisfied and condition (A.3) is violated.

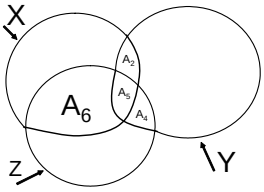


Figure A-2 – Counter example to show ($A_6 > A_2 + A_4 + A_5$)

Property 2: D_R and $I(x;y)$ are complements of each other.
It is intriguing that D_R satisfies the property of a metric but property 2 states that its complement $I(x;y)$ does not. With reference to Figure (A-1a) we wish to demonstrate that D_R and $I(x;y)$ are complements. By definition:

$$I(x;y) = H(x) + H(y) - H(x,y) \quad (A.24)$$
$$D_R(x;y) = H(x/y) + H(y/x) \quad (A.25)$$

Let S represent the entire space in Figures (3) and (A-1a). Then let e be an element of S and $S = A_1 \cup A_2 \cup A_3$ where $\cup$ indicates the union of sets. Note A_1, A_2 and A_3 are disjoint sets in Figure (A-1a). From (A.11-A.12) and Figures (5) and (A-1a) it follows that all the elements of $I(x;y)$ are in A_2 and all the elements of D_R(x;y) are in $A_1 \cup A_3$. For notational simplicity denote the complement of a set A as A', then (Eves, [26]) two cases now exist: (1) if $e \in A_1 \cup A_3$, then $e \notin A_2$, thus

$$(A_1 \cup A_3)' = (A_2) \quad (A.26)$$

or (2) if $e \in A_2$, then $e \notin A_1 \cup A_3$, thus

$$(A_2)' = (A_1 \cup A_3) \quad (A.27)$$

It then follows that if e is an element of $(A_1 \cup A_3)$ then e is an element of (A_2)', and if e is an element of (A_2), and if e is an element of $(A_1 \cup A_3)$', whence (A_2) and $(A_1 \cup A_3)$ are complements.

Runge Phenomenon: A virtual artifact in image processing

Xiaoping Shen[1], Fairul Mohd-Zaid[2] and Russell Francis[1]

April 25, 2012

***Abstract*— Interpolation using variate functions, such as polynomials and trigonometric functions are common methods used to process discrete signal samples. Since most electronic devices used for quantization provide equally spaced signal samples, these interpolation methods are often used with data which is equally spaced along the abscissa. Unfortunately these approximations often exhibit the undesirable behavior of oscillations near the boundary of the domain which cannot be removed by increasing the degree of the interpolating polynomial. Regardless of the fact that these artifacts appear in many applications, including signal and image processing, Runge phenomenon often fails to differentiate from other similar artifacts such as the more popular Gibbs phenomenon. In this article, we present numerical examples to illustrate the Runge phenomenon in the context of signal and image processing. Both simulated and field data are used in the numerical experiments.**

Keywords: Runge phenomenon, Lebesgue constant, polynomial interpolation.

2000 Mathematics Subject Classification: 65M10, 78A48.

I. Introduction

To transmit and store analog signals effectively, an analog-to-digital converter (ADC) is used to convert the domain of the signal from continuous-time to discrete-time. This follows by the reverse operation, an digital-to-analog converter (DAC) coverts the digital signal back to analog format such that the signal to be recognized by human senses or other non-digital systems. One of the central parts in both ADC and DAC is the interpolation circuits.

In practice, a wide variety of interpolation methods and techniques are in use based on different function spaces and sampling methods (the choices of nodes). Polynomials and trigonometric functions are more popular among others. Since most electronic devices used for quantization provide equally spaced signal samples, these interpolation methods are often used with data equally spaced along the abscissa. Unfortunately the interpolation approximations on equally spaced data are often exhibit undesirable behavior near the boundary of the domain especially as the degree of the approximating polynomial increases aggressively. This behavior, oscillations near the boundary which cannot be removed by increasing the degree of the interpolating polynomial, is the so called Runge phenomenon. Although these oscillations appear in many applications, yet the Runge phenomenon often fails to differentiate from other similar artifacts such as the more popular Gibbs phenomenon.

Some methods such as using lower degree spline interpolation are introduced to overcome the artifacts. The interpolating at the non uniformly distributed nodes such zeros of Chebyshev polynomials and Barycentric Lagrange interpolation are examples of classical methods [2]. There has been renewed interest in Runge phenomenon in the past a few years. Some new numerical methods are developed to overcome the phenomenon, for example, Tikhonov Regularization and spectral methods are alternatives [3] and [4]. Almost all these methods are not ADC/DAC friendly because they are based on non uniform sampling.

The primary interest of this article is twofold: 1) quantify the Runge phenomenon using Lebegue constant and construct numerical examples to illustrate its properties in the context of signal and image processing and 2) discuss new ADC and DAC friendly method.

The outline of this article is as follows: a brief review of the Runge phenomenon will be given in Section 2. In Section 3, using the Lebesgue constant as a tool to, quantify and illustrate the phenomenon for signals (one

Paper presented at the IPCV'12 - The 2012 International Conference on Image Processing, Computer Vision, and Pattern Recognition, Las Vegas, USA.

[1]Department of Mathematics, Ohio University, Athens, Ohio 45701, USA. Email: shenx@ohio.edu, rf358197@ohio.edu

[2]711 HPW AFRL/RHCV, Wright-Patterson Air Force Base, Dayton, Ohio 45433, USA. Email: Fairul.Mohd-Zaid@wpafb.af.mil (contact author).

dimensional case) and images (two dimensional case), respectively. Section 4 is devoted to construct interesting numerical examples using audio signals and images to illustrate the phenomenon in context of signal and image processing. We conclude the paper by a brief summary and a remark on our future study.

II. Background on the Runge phenomenon

To begin, we recall the well-known Weierstrass Approximation Theorem and the Lagrange interpolation.

Theorem 1: (weierstrass) Let R be a continuous function on a closed and bounded interval of R. Then f can be uniformly approximated by polynomials.

There are several different ways of proving this important theorem. A detailed proof can be found in most Real Analysis textbooks, including [1] and [19].

In the spirit of the Weierstrass Approximation Theorem, polynomial approximation plays the most important role in practice. Among many other polynomial approximation, the Lagrange interpolation are important both in theory and practice. In the one dimensional case, the Lagrange interpolation can be described as follows.

Let $X = \{x_0, x_1, ..., x_n | n \in \mathbf{N}^*\}$ such that $\forall x_i \in X$ the following holds.

$$a \leq x_0 < ... < x_{i-1} < x_i < x_{i+1} < ... < x_n \leq b$$

For any $f \in C[a, b]$, we can define the Lagrange interpolating polynomial $P_n(f; X; x)$ which agrees with $f(x_i)$ at each of the points $x_i \in X$ in the standard form [2].

$$P_n(f; X; x) = \sum_{i=0}^{n} f(x_i) L_i(X; x)$$

and

$$L_i(X; x) = \prod_{k=0, i \neq k}^{n} \frac{(x - x_k)}{(x_i - x_k)} \tag{1}$$

Carle Runge, a student of Weierstrass, later detailed a seemingly paradoxical observation [20]. He noted that the approximation of f could diverge as the degree of the polynomial approximation increased and stated conditions for this divergence. This is also true for the Lagrange interpolation. It also suffers from the Runge phenomenon. This phenomenon is the observed effect of a polynomial approximations error tending towards infinity near the extremes of the interval over which interpolation occurs as the degree of the polynomial approximation increases. A commonly used metric for measuring the accuracy of an approximation is the Lebesgue constant [8], [9] and [14]. For one dimensional polynomial approximations the Lebesgue function and constant are defined below [8] and [17].

$$\lambda_n(X; x) = \sum_{i=0}^{n} |L_i(X; x)| \tag{2}$$

$$\Lambda_n(X) = \max_{x \in [a,b]} \lambda_n(X; x) \tag{3}$$

With this groundwork, we may proceed in presenting some examples of the phenomenon. In the following sections, we will present numerical studies of the phenomenon for polynomial interpolation in one and two dimensions and observe how variations in node spacing affect the phenomenon. We will also apply these same interpolation techniques to audio data to observe how they behave in more typical situations.

III. Quantify Runge phenomenon

A. Numerical Experiments in One Dimension Space

1) Lebegue constant in one dimension case: To examine the Runge phenomenon in one dimension, we have chosen the function $\frac{1}{1+x^2}, x \in [-5, 5]$. In Figures 1 and 2 the 10^{th} and 15^{th} degree interpolating polynomials of the function with equally spaced points over the domain are shown with the corresponding Lebesgue functions. These figures show the approximation of the function diverging near the boundaries and also show that this divergence accelerates as the degree of the approximating polynomial increases. In fact, it can be shown [13] and [18] that $\Lambda_n(X) \geq C \times \frac{2^{n-1}}{n^{\frac{3}{2}}}$ for Lagrange interpolation over equally spaced nodes.

Since the Lebesgue function for Lagrange interpolation depends solely on the position of the interpolation points it is natural to ask if there is a spacing or set of points which could be used to minimize $\Lambda_n(X)$ for any given function? An answer to this question can be found in [18] where Rivlin shows that there is not a set of points which can assure uniform convergence for all continuous functions over a given domain. This fact follows from establishing a lower bound for the $\Lambda_n(X)$ which approaches ∞ as $n \to \infty$. This bound is listed in equation (**??**) and can be found in [14] and [18].

$$\Lambda_n(X) \leq \frac{2}{\pi} log(n) + c, c > 0, n = 1, 2, ..., \tag{4}$$

We can get relatively close to the lower bound by interpolating at the roots of the Chebyshev polynomial T_n. The upper bound of $\Lambda_n(T_n)$ is listed in equation (5)

$$\Lambda_n(X) \leq \frac{2}{\pi} log(n) + 1, n = 1, 2, ..., \qquad (5)$$

2) Examples: In figures 1 and 2 we present the result of Lagrange interpolation using the roots of the polynomial T_n as interpolation points. It should be noted that the roots exist within $[-1, 1]$ and we have scaled these to fall within $[-5, 5]$ while maintaining the same relative distance to one another. In table **??** we have presented approximations of Λ_n for both cases.

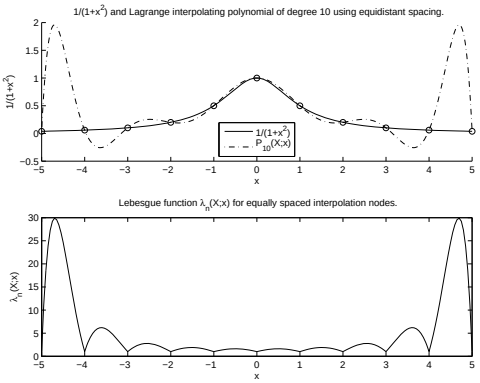


Fig. 1. The Lagrange interpolating polynomial of degree 10 for the graph of $\frac{1}{1+x^2}$ along with the Lebesgue function for this interpolating polynomial. This graph is spaced using equally spaced nodes.

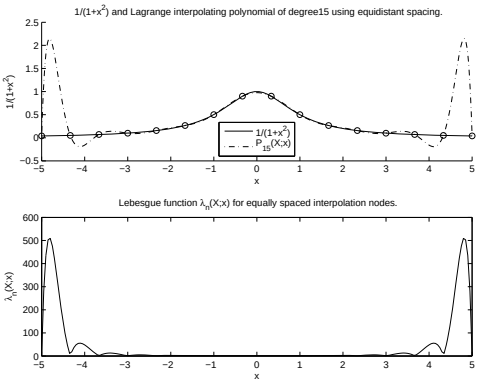


Fig. 2. The Lagrange interpolating polynomial of degree 15 for the graph of $\frac{1}{1+x^2}$ along with the Lebesgue function for this interpolating polynomial. This graph is spaced using equally spaced nodes.

(see Table 1. for details).

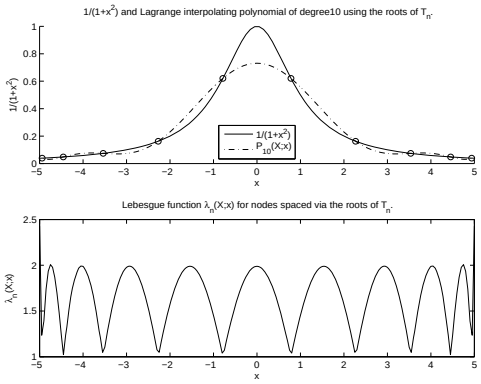


Fig. 3. The Lagrange interpolating polynomial of degree 10 for the graph of $\frac{1}{1+x^2}$ along with the Lebesgue function for this interpolating polynomial. This graph is spaced using the roots of the Chebyshev polynomial of the first type.

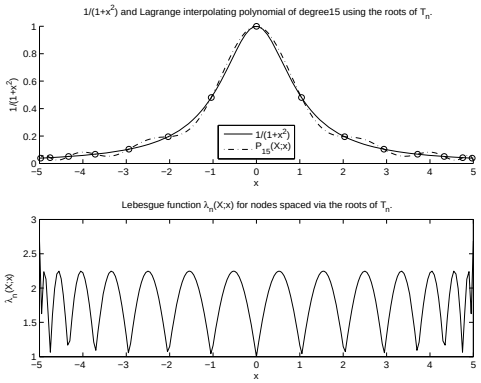


Fig. 4. The Lagrange interpolating polynomial of degree 15 for the graph of $\frac{1}{1+x^2}$ along with the Lebesgue function for this interpolating polynomial. This graph is spaced using the roots of the Chebyshev polynomial of the first type.

Table 1. Approximation the Lebesque's constant

n	Λ_n	$\Lambda_n(T_{n+1})$	$\frac{2}{\pi}\ln(n+1)+1$
1	1.00000	1.41421	1.44127
2	1.25000	1.66667	1.69940
3	1.63113	1.84776	1.88254
4	2.20782	1.98885	2.02460
5	3.10630	2.10440	2.14067
6	4.54934	2.20221	2.23880
7	6.92974	2.28702	2.32381
8	10.94565	2.36186	2.39880
9	17.84861	2.42883	2.46587
10	29.89996	2.48943	2.52655
11	51.21422	2.54477	2.58194
12	89.32491	2.59568	2.63290
13	158.10236	2.64282	2.68008
14	283.21120	2.68671	2.72400
15	512.35146	2.72778	2.76508

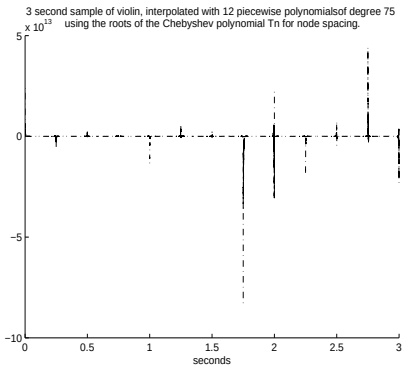


Fig. 5. A 3 second audio sample approximated using twelve 75 degree polynomials whose nodes are spaced using the roots of T_n.

B. Numerical Experiments in Two Dimension Space

In this section, we shall extend our examination of the phenomenon into two dimensions. For the examples in this section, we will be approximating $z = \frac{1}{1+x^2+y^2}, x \in [-5,5], y \in [-5,5]$. This is a natural extension of the previous formula into two dimensions.

1) Lagrange Interpolation: In this section, the examples and numerical results have been developed using a two dimensional form of the Lagrange interpolating polynomial listed in equation (6), Lebesgue function listed in equation 7 and the Lebesgue constant listed in equation 8.

In the context of image processing, samples are most commonly taken by horizontal and vertical scans. Therefore, the separable base functions formed by direct product are used most often. We will limit our discussion to the separable case. There is active research on other more exotic partitioning schemes (see [5], [21]), but these are mostly of theoretical interest.

We define the Lagrange interpolating polynomial in two dimensions by

$$P(f;x,y) = \sum_{i=0}^{n}\sum_{j=0}^{m} f(x_i,y_j)L_i(X;x)L_j(Y;y) \quad (6)$$

Where $|X| = n$, $|Y| = m$ and the terms $L_i(X;x)$ and $L_j(Y;y)$ are computed using equation (1). In the two dimensional case, our Lebesgue function will be.

$$\lambda_{n,m}(X;Y;x,y) = \sum_{i=0}^{n}\sum_{j=0}^{m} |L_i(X;x)L_j(Y;y)| \quad (7)$$

$$\Lambda_{n,m}(X;Y) = \max_{x\in X, y\in Y}(\lambda_{n,m}(X;Y;x,y)) \quad (8)$$

We will limit ourselves to presenting a numerical exposition of two different partitioning schemes. First we will present the results for an equally spaced partitioning scheme this will then be compared with spacing along the x and y axes according to the roots of the Chebyshev polynomial. This should parallel our discussion in the one dimensional case.

2) Equally Spaced Partition: The Runge phenomenon can also be observed in the two dimensional case. In the contour figures 6 and 7, where the interpolation nodes are marked with a 'x', we can observe what appears to be the approximation diverging from the actual value of the function towards the extremes of the domain with the larger polynomial approximation. This divergence is characteristic of the Runge phenomenon.

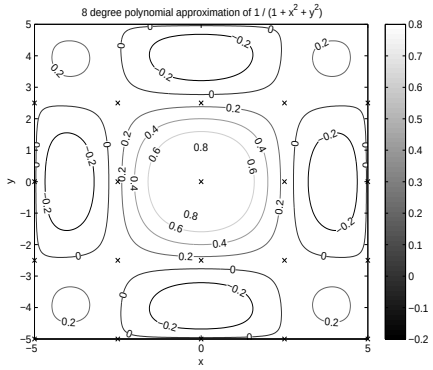


Fig. 6. A contour plot of the approximating polynomial of degree 8.

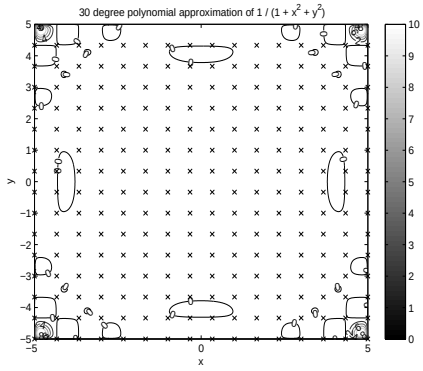


Fig. 7. A contour plot of the approximating polynomial of degree 30.

In observing the phenomenon, it is also useful to consider the Lebesgue function of the approximations. Figures 8 and 9, show the Lebesgue functions for the polynomial approximations of degree 8 and degree 14

respectively. In these figures, it can also be observed that the Lebesgue function also increases towards the edge of the domain when an equally spaced partitioning scheme is used to select the interpolation nodes.

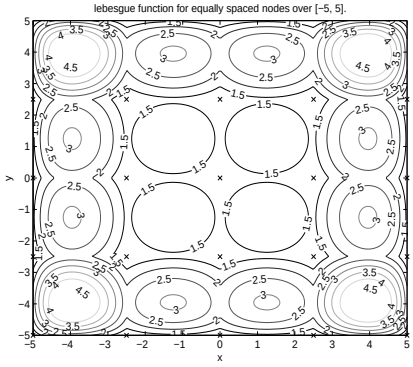


Fig. 8. A contour plot of the Lebesgue function for two dimensional Lagrange interpolation with maximum degree 8. 25 equally spaced nodes within the domain are used.

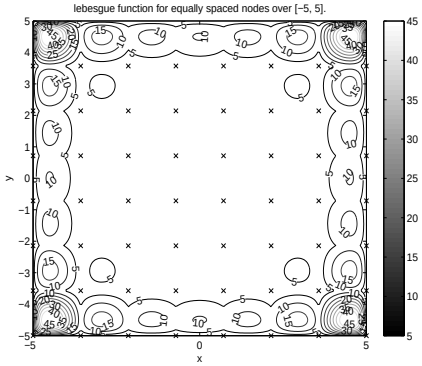


Fig. 9. A contour plot of the Lebesgue function for 2D Lagrange interpolation with maximum degree 14. 64 equally spaced nodes within our domain are used.

IV. Runge phenomenon in signal and image processing

A. Runge phenomenon in audio signals

While considering the one dimensional case, it would be interesting to consider how these methods perform with real data. To illustrate how polynomial approximation performs under real conditions we will approximate a brief audio sample using several high degree polynomials in a piecewise fashion using both equally spaced nodes and nodes spaced using the roots of the Chebyshev polynomial, T_n. The results can be seen in Figure 10 and Figure **??** respectively.

Both methods of approximation poorly approximate the original data whose range is $[-1, 1]$. It should be noted however that the approximation using the roots of T_n did perform better than the technique using equally spaced nodes, as can be observed by the maximum values attained in each figure. This seems to fit with our earlier observations that the spacing of the interpolation nodes affects the quality of the approximation.

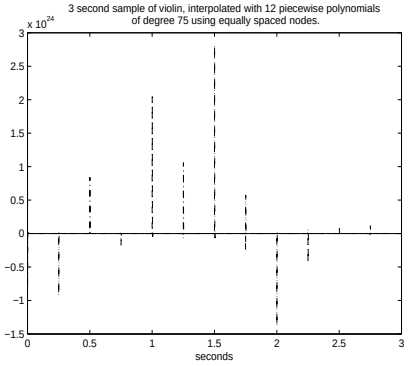


Fig. 10. A 3 second audio sample approximated using twelve 75 degree polynomials whose nodes are equally spaced.

B. Runge phenomenon in images

We illustrate sample images of Lenna with 2d Lagrange interpolation applied using blocksizes of increasing size. It shows the phenomenon pretty well especially in the blocksize 11 case. Results are shown in the following figures. Experiment results are shown in Figures 11 to 15.

Fig. 11. Runge phenomenon: block size=3.

V. Summary and future study

The approximation of variate functions through interpolation techniques is common in many disciplines.

Fig. 12. Runge phenomenon: block size=5

Fig. 14. Runge phenomenon: block size=9

Fig. 13. Runge phenomenon: block size=7

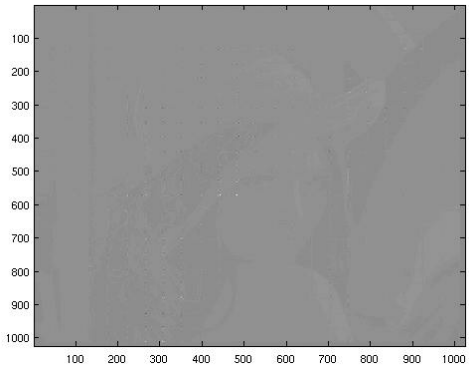


Fig. 15. Runge phenomenon: block size=11

It is also common to use the same techniques while developing models for discretely sampled data. The Lagrange interpolation is a practical method to apply and is important for the rich history it brings to the field. Unfortunately, when the interpolating nodes are equally spaced, the Lagrange interpolation technique often produces undesirable results especially when polynomials of high degree are used. In this paper, we have provided numerical examples which demonstrate the Runge phenomenon in both one and two dimensions. We also show numerically that the selection of interpolation nodes can have a dramatic affect on the quality of an approximation. Using the Lebesgue constant, we have also presented some discussion on how to quantify this phenomenon. As illustrated in section 3, the Runge phenomenon can be removed by interpolating at the zeros of Chebyshev polynomials. Since zeros of Chebyshev polynomials are distributed non uniformly, it can not be used to the data directly from ADC or DAC. Our current study are focusing on the interpolating at zeros of Slepian functions. With a connection map, the data (uniform sampled) from ADC or DAC will be mapped to non uniform distributed Slepian zeros, the interpolating polynomial is then constructed. Related the results will reported in next conference - IPCV'13.

REFERENCES

[1] R. G. Bartle. *The Elements of Real Analysis*, $2nd$ ed, John Wiley & Sons. 1976.

[2] J-P. Berrut and L. N. Trefethen. Barycentric Lagrange interpolation, SIAM Review Vol. **46** (3), 501-517, 2004.

[3] J. P. Boyd and J. Rong, On Exponentially-Convergent Strategies for Defeating the Runge Phenomenon for the Approximation of Non-Periodic Functions, Part I: Single-Interval Schemes, Communications in computational physics, Vol. 5, No. 2-4, pp. 484-497.

[4] J. P. Boyd, Defeating the Runge phenomenon for equispaced polynomial interpolation via Tikhonov regularization, Appl. Math. Lett., 5 (1992), 57–59.

[5] L. Bos, M. Caliari, S. D. Marchi, and M. Vianello. Bivariate interpolation at Xu points: Results, extensions and applications, *Electron. Trans. Numer. Anal.*, Vol. **25**, 1-16, 2006.

[6] Bos, L. Bos; Caliari, M.; Marchi, S. De; Vianello, M. A Numerical Study of the Xu Polynomial Interpolation Formula in Two Variables. Computing 76 (2006), 311-324

[7] L. Bos, S. D. Marchi and M. Vianello. On the Lebesgue constant of Xu interpolation formula, *J. Approx. Theory*, Vol. **141**, 134-141, 2001.

[8] L. Brutman. Lebesgue functions for polynomial interpolation-A survey. *The heritage of P. L. Chebyshev: a Festschrift, in Honor of the 70th birthday of T. J. Rivlin.*, *Ann. Numer. Math.* Vol. **4** (1-4), 111–127, 1997.

[9] L. Brutman and E. Passow. Rational interpolation to $|x|$ at the Chebyshev nodes, *Bull. Austral. Math. Soc.*, Vol. **56** (1), 81–86, 1997.

[10] H-M. Chen and P. K. Varshney, Registration Using Generalized Partial Volume Joint Histogram Estimation, IEEE Trans. on Medical Imaging, Vol. 22, NO. 9, 2003

[11] W. P. Dotson and J. H. Wilson, A digital-to-analog conversion circuit using third-order polynomial interpolation., NASA Techn. Rep., TR R-382, 52 p.

[12] J. Epperson. On the Runge example, *Amer. Math. Monthly*, Vol. **94** (4), 329–341, 1987.

[13] S. R. Finch. *Encyclopedia of Mathematics and its Applications: Mathematical Constants*, Cambridge University Press, 2003.

[14] R. B. Guenther and E. L. Roetman. Some observations on interpolation in higher dimensions, *Math. Comp.*, Vol. **24** (111), 517-522, 1970.

[15] J. Mason and D. Handscomb. Chebyshev Polynomials. CRC, 2003.

[16] T. J. Rivlin. *Chebyshev Polynomials: From Approximation Theory to Algebra and Number Theory*, $2nd$ ed., John Wiley & Sons. 1990.

[17] T. J. Rivlin. *An Introduction to the Approximation of Functions*, Dover, 1981.

[18] W. Rudin. *Principles of Mathematical Analysis*, McGraw-Hill, New York, 1964.

[19] C. Runge. Über empirische Funktionen und die interpolation zwischen äquidistanten ordinaten, *Zeit. Math. Phys.* Vol. **46**, 224-243, 1901.

[20] Y. Xu. Lagrange interpolation on Chebyshev points of two variables, *J. Approx. Theory*, Vol. **87**(2), 1996.

APPENDIX A

Investigation of Complex Networks and Information Theory at WPAFB

Investigation of Complex Networks and Information Theory at WPAFB

D. W. Repperger
711 HPW AFRL/RHCV
Wright-Patterson Air Force Base
Dayton, Ohio 45433

Topic 1 – Logistics System – an interesting application (performance /vulnerability).

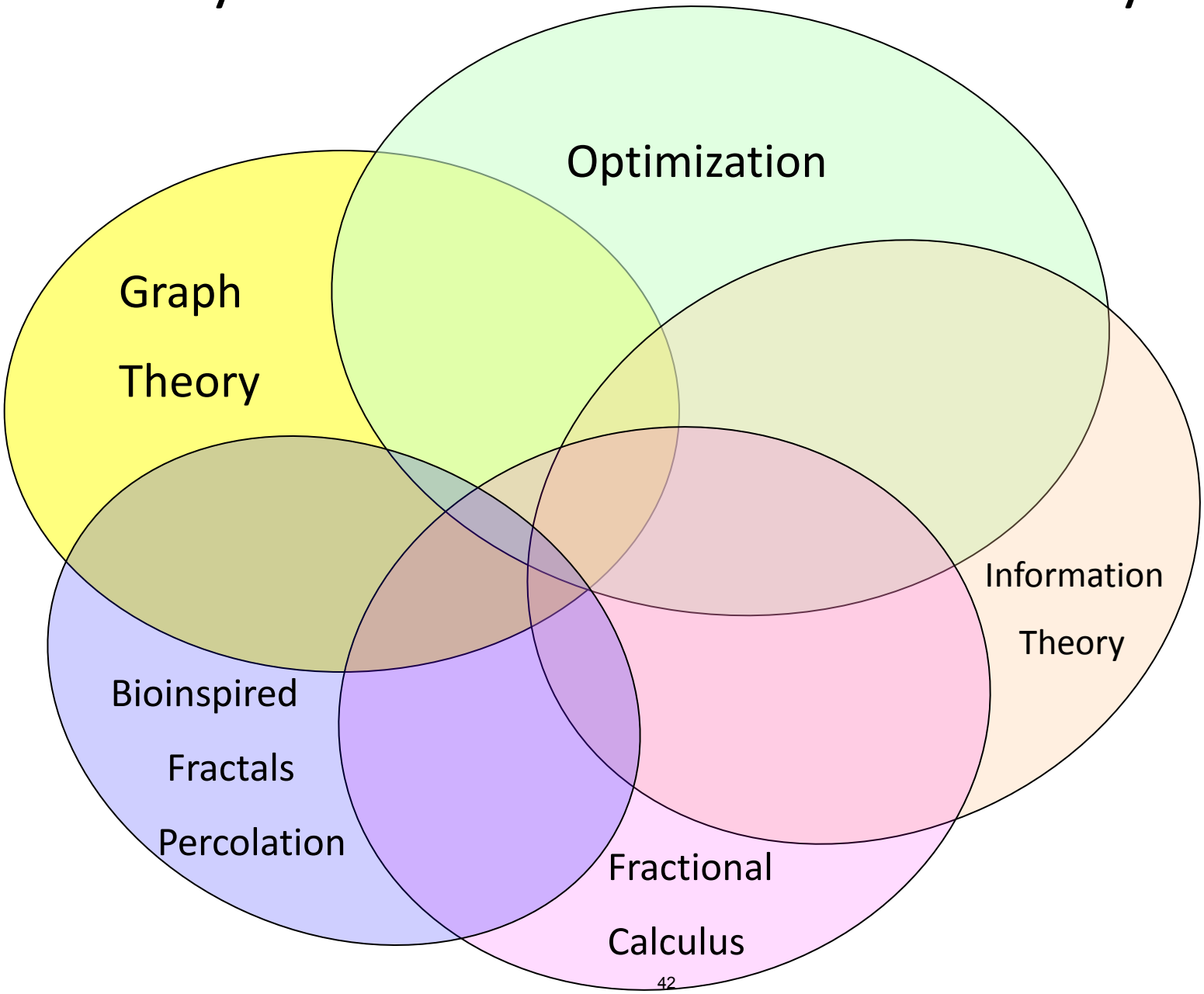
Topic 2 - Architecture – Examine Random versus Scale-free structure in networks.

Topic 3 - Fractional Gaussian and Brownian Noise Studies (design of networks).

Topic 4 - Percolation Networks – Cellular automata.

Topics 5+ - Circuits (AFIT work), Fractal Filter (show applicability).

Many Areas Interact for Network Analysis



Topic 1

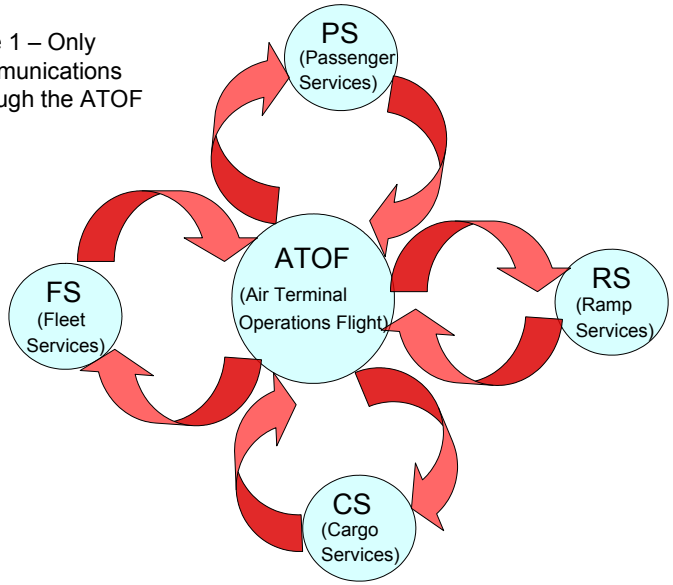
Logistics System

Step 1 – What Architecture /Structure to use

Topic 1 – Logistics System

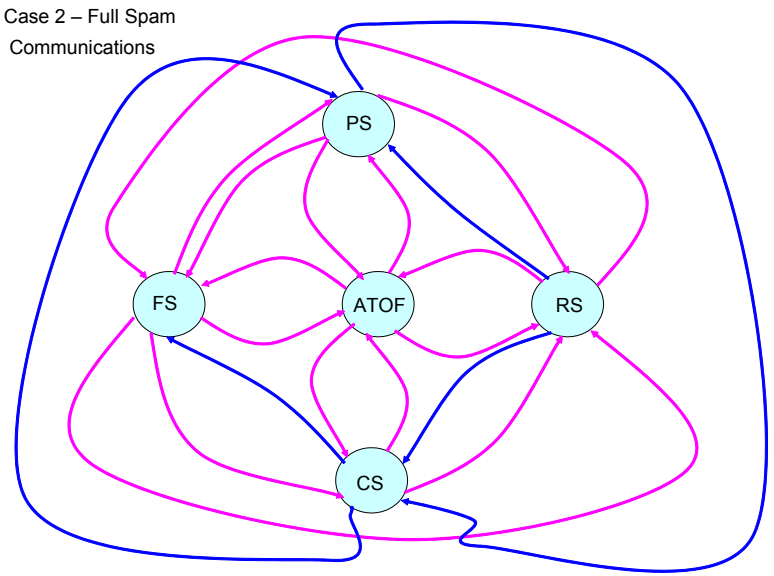
Minimum Interactions/communications

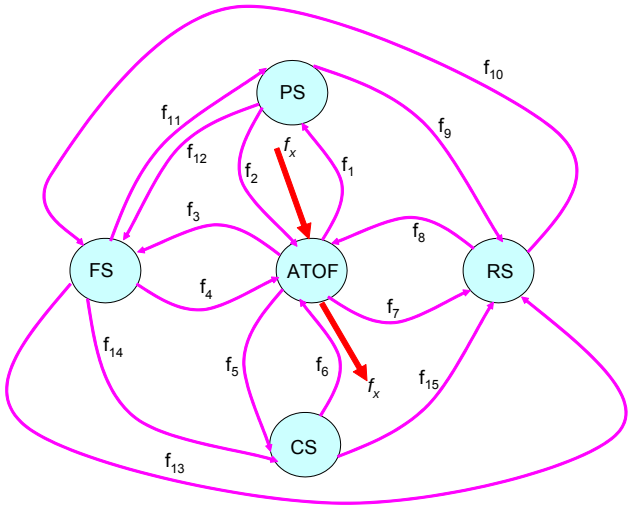
Case 1 – Only Communications Through the ATOF



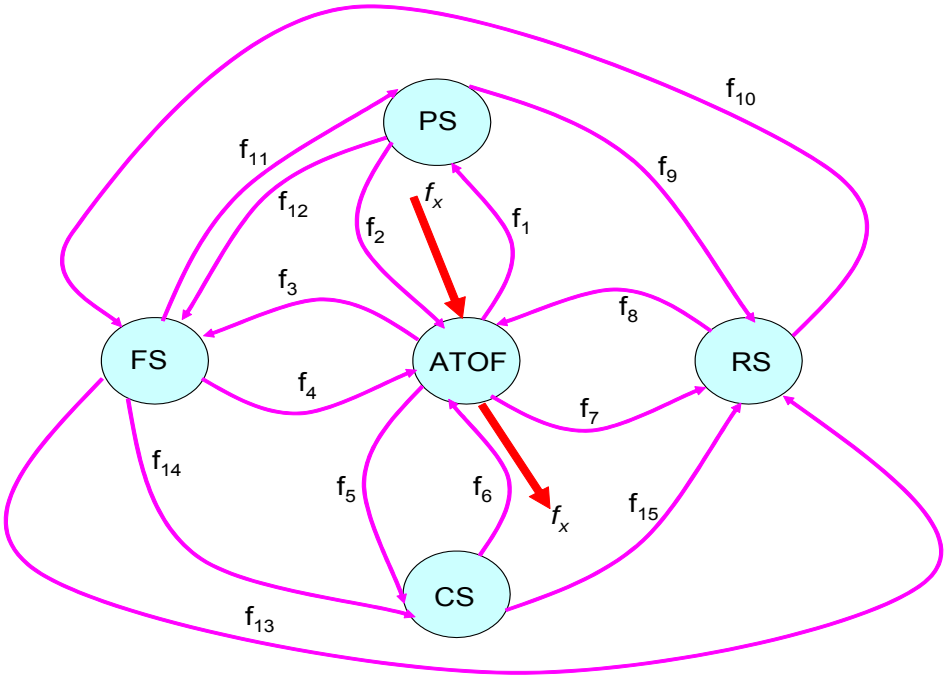
Maximum Interactions/communications

Case 2 – Full Spam Communications





A pragmatic compromise



How Constraint Equations Arise?

ATOF: $f_x + f_2 + f_4 + f_6 + f_8 = f_1 + f_3 + f_5 + f_7 + f_x$

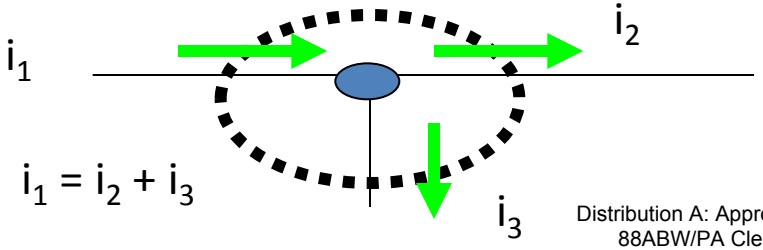
PS: $f_{11} + f_1 = f_9 + f_2 + f_{12}$

RS: $f_{15} + f_7 + f_9 + f_{13} = f_{10} + f_8$

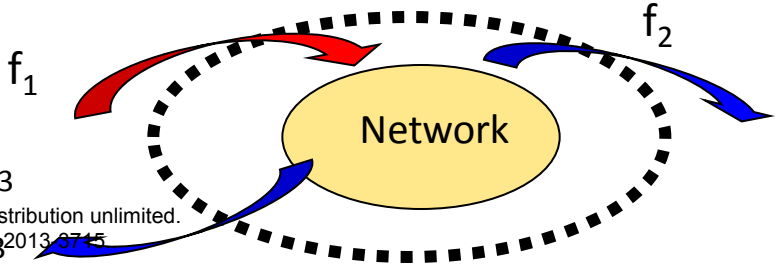
FS: $f_{10} + f_{12} + f_3 = f_{14} + f_{13} + f_4 + f_{11}$

CS: $f_5 + f_{14} = f_{15} + f_6$

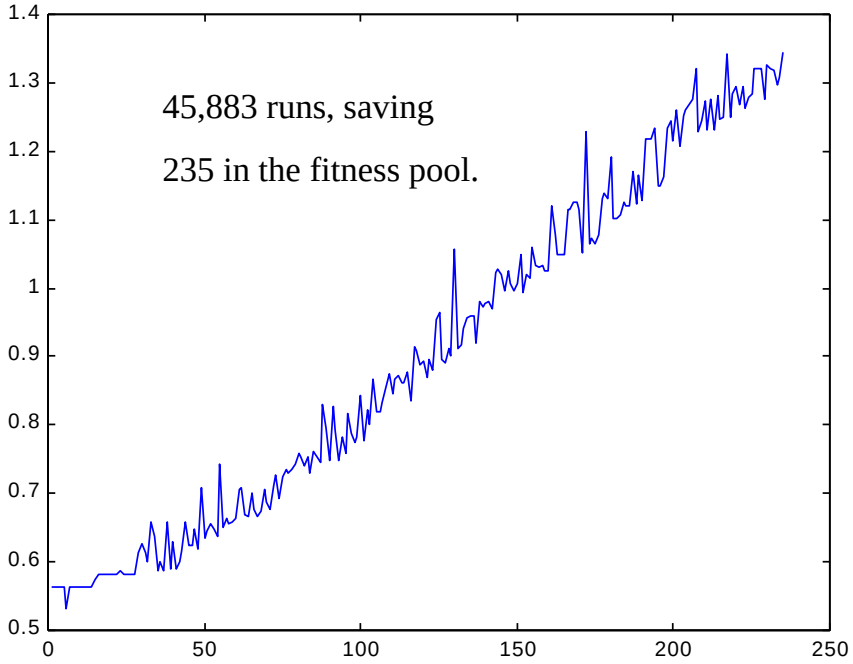
Cut set in circuit theory



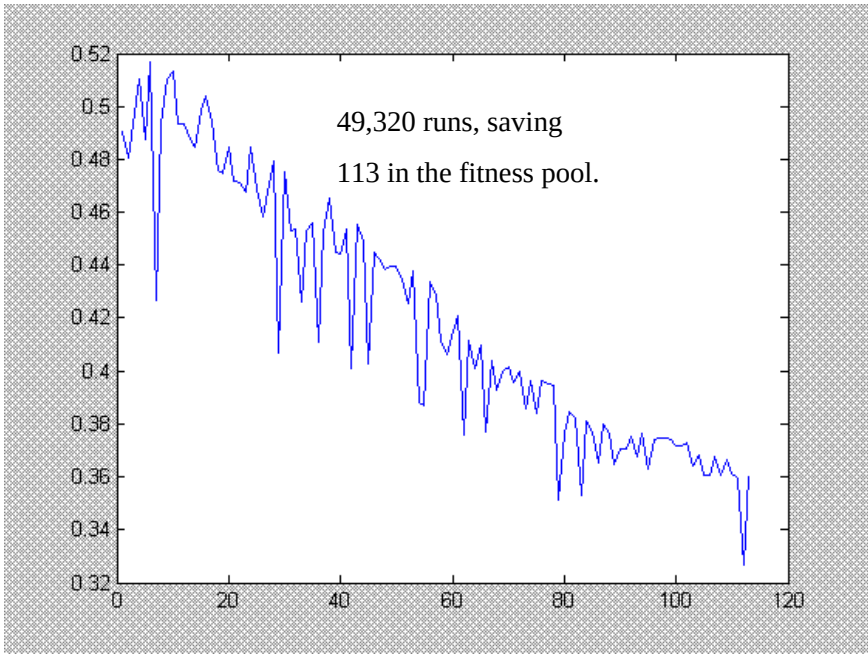
Cut set in Graph Theory



Maximize Mutual Information



Minimize Mutual Information



Optimize via Genetic Algorithms

$$I(x;y) = H(x) + H(y) - H(x,y)$$

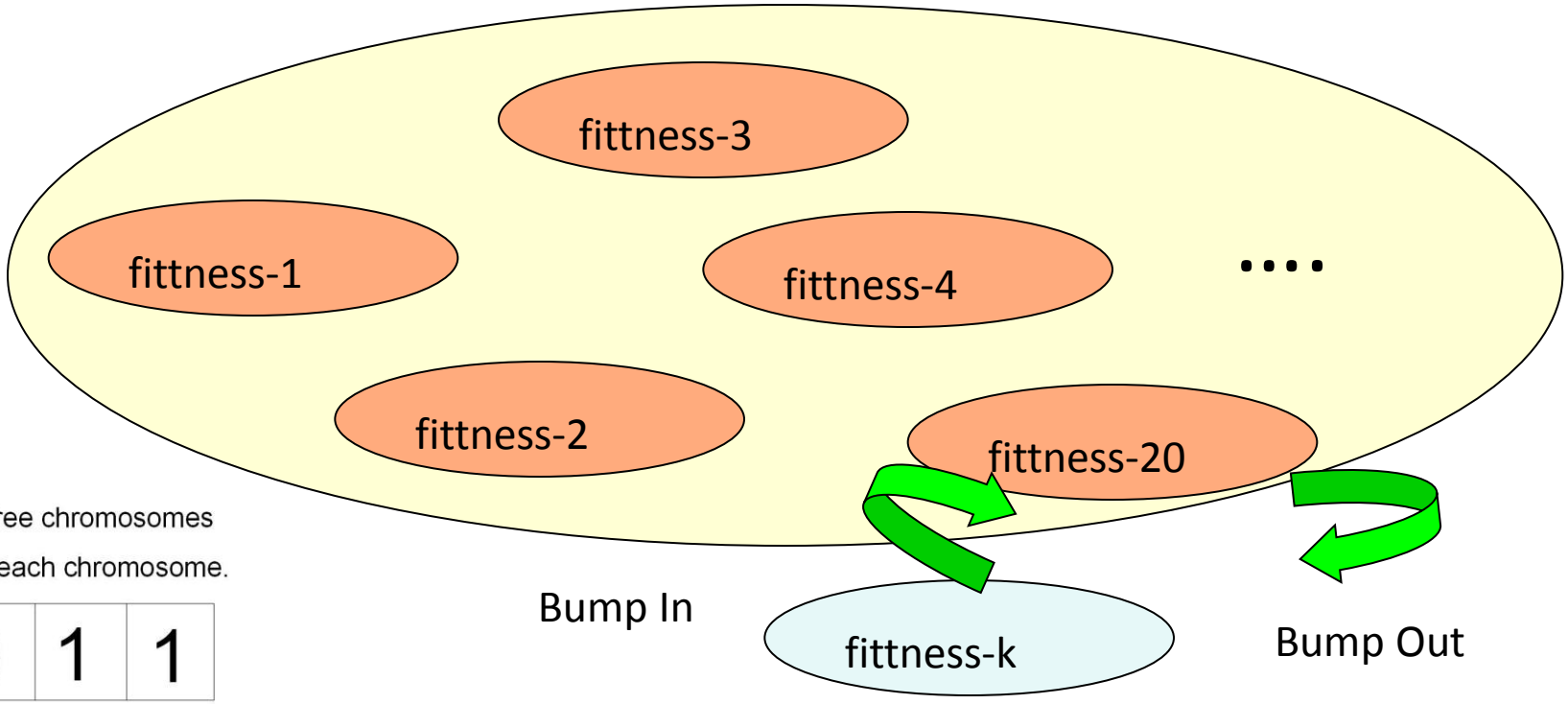
$$H(x) = \sum_{i=1}^{n} p(x_i)\log_2(1/\,p(x_i))$$

$$H(x,y) = \sum_{i,j}^{n,q} p(x_i, y_j)\log_2(1/\,p(x_i, y_{46j}))$$

$$H(y) = \sum_{j=1}^{q} p(y_j)\log_2(1/\,p(y_j))$$

(see reference)

How the Optimization is Conducted (Elite Pool)



j = 1, …, 11 free chromosomes

3 bit word for each chromosome.

j = 1	0	1	1
j = 2	0	0	1
…			
j = 11	1	0	1

(8^{11} possibilities, NP Hard)

Fig. 9 Configuration for the Chromosome

Facts:

15 unknown flows.

5 constraints (4 independent).

15-4 = 11 independent flows to determine.

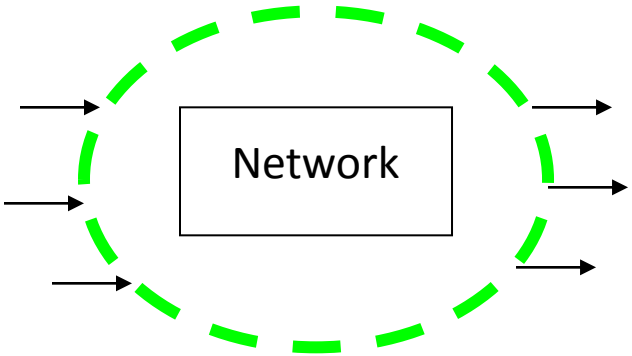
Possibilities = $(2^3)(2^3)(2^3)\ldots(2^3) = 8^{11}$

N-P Hard

Let *T* = cut set flow, let *W* be the MI = *I*(x;y).

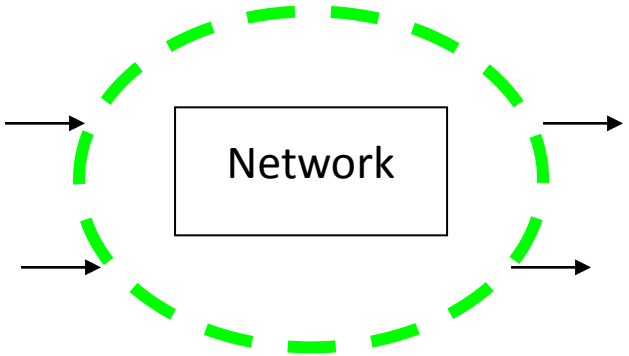
$$S_W{}^T := \lim \frac{\frac{\Delta T}{T}}{\frac{\Delta W}{W}} = \frac{\partial T}{\partial W}\frac{W}{T}$$

Maximum Flow



Cut set: flows in = flows out
= 10 units

Minimal Flow



Cut set: flows in = flows out
= 1 unit

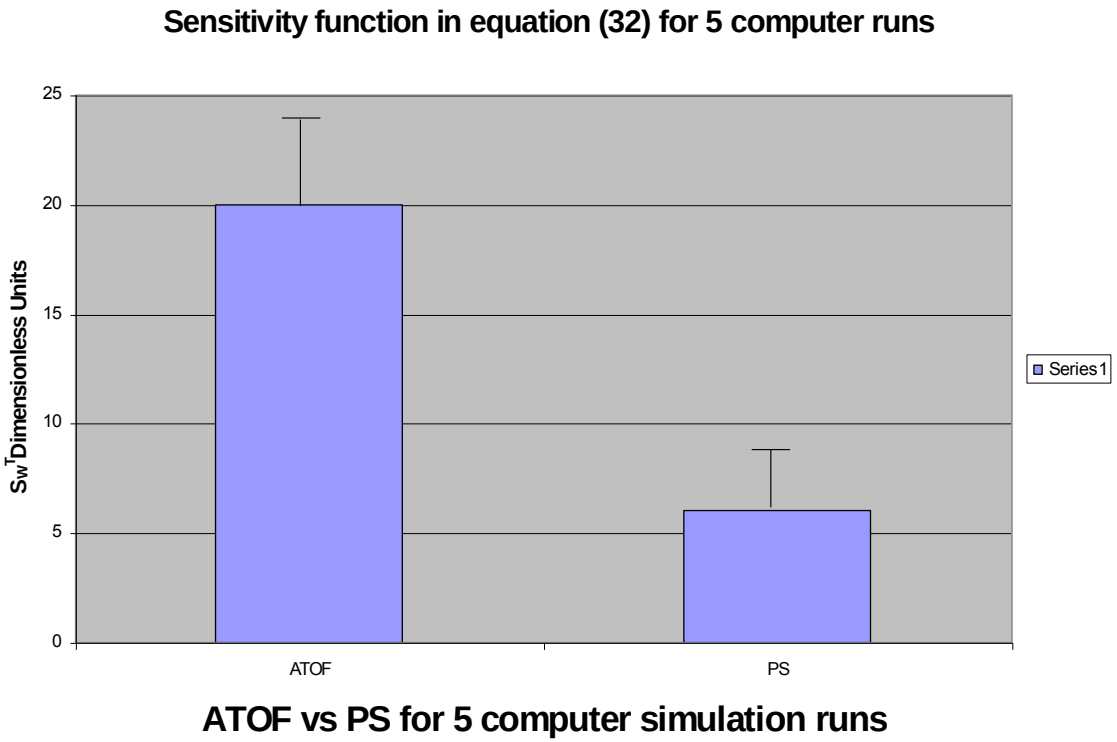
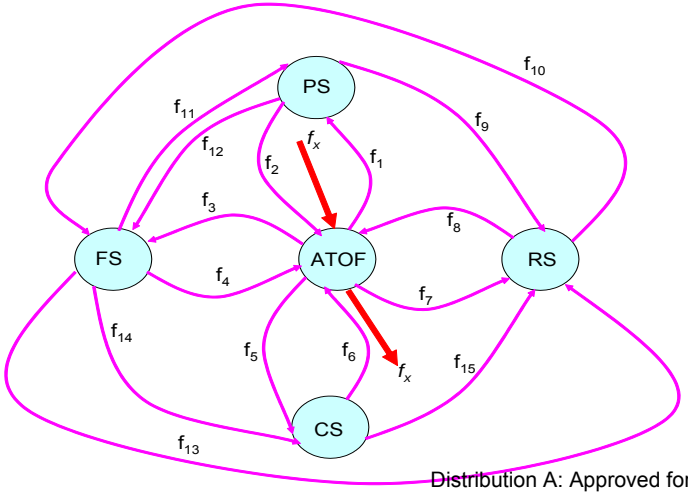


Figure (12) –The sensitivity Function defined in equation (32) for ATOF vs PS

Step 5 – Compare two paradigms



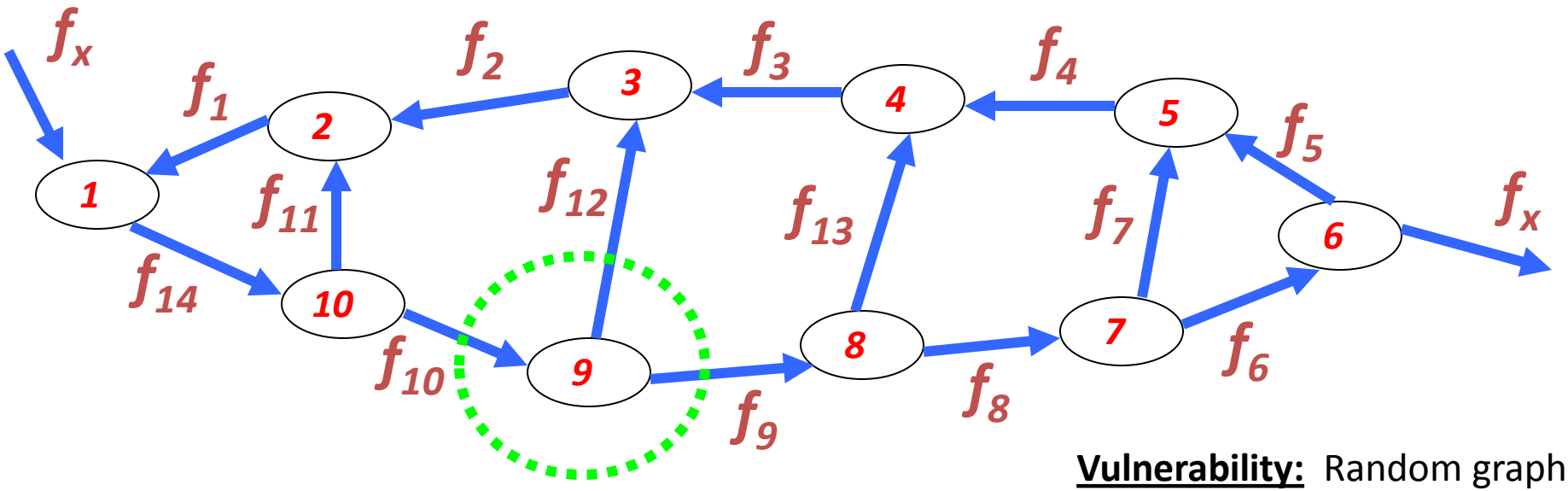
(Logistics' Journal Paper)

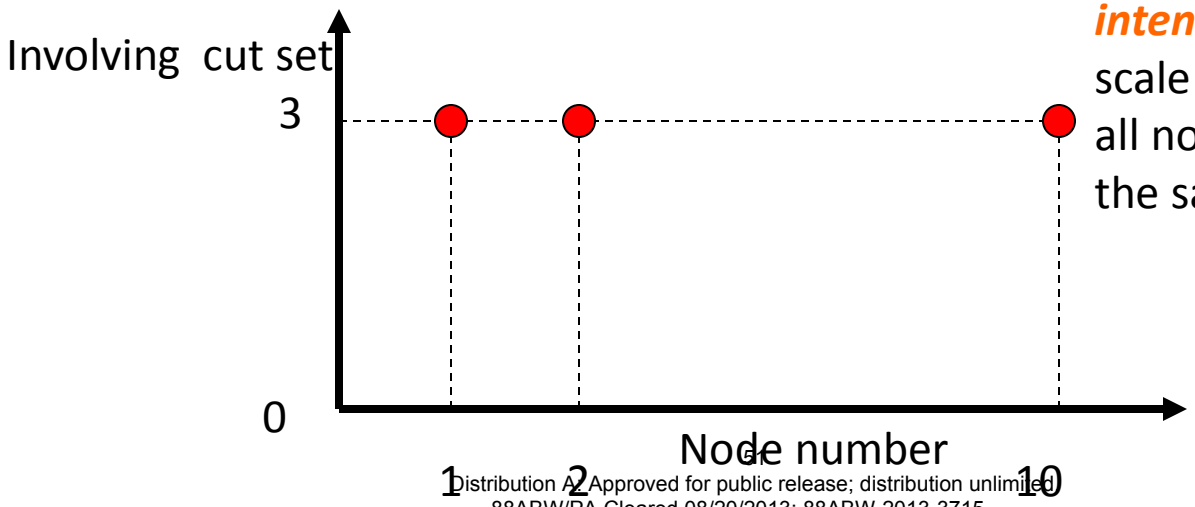
Topic 2 - Architecture Investigation

Random Graph (Gaussian, thin tail)

(Highways on a map)

Erdös-Renyi

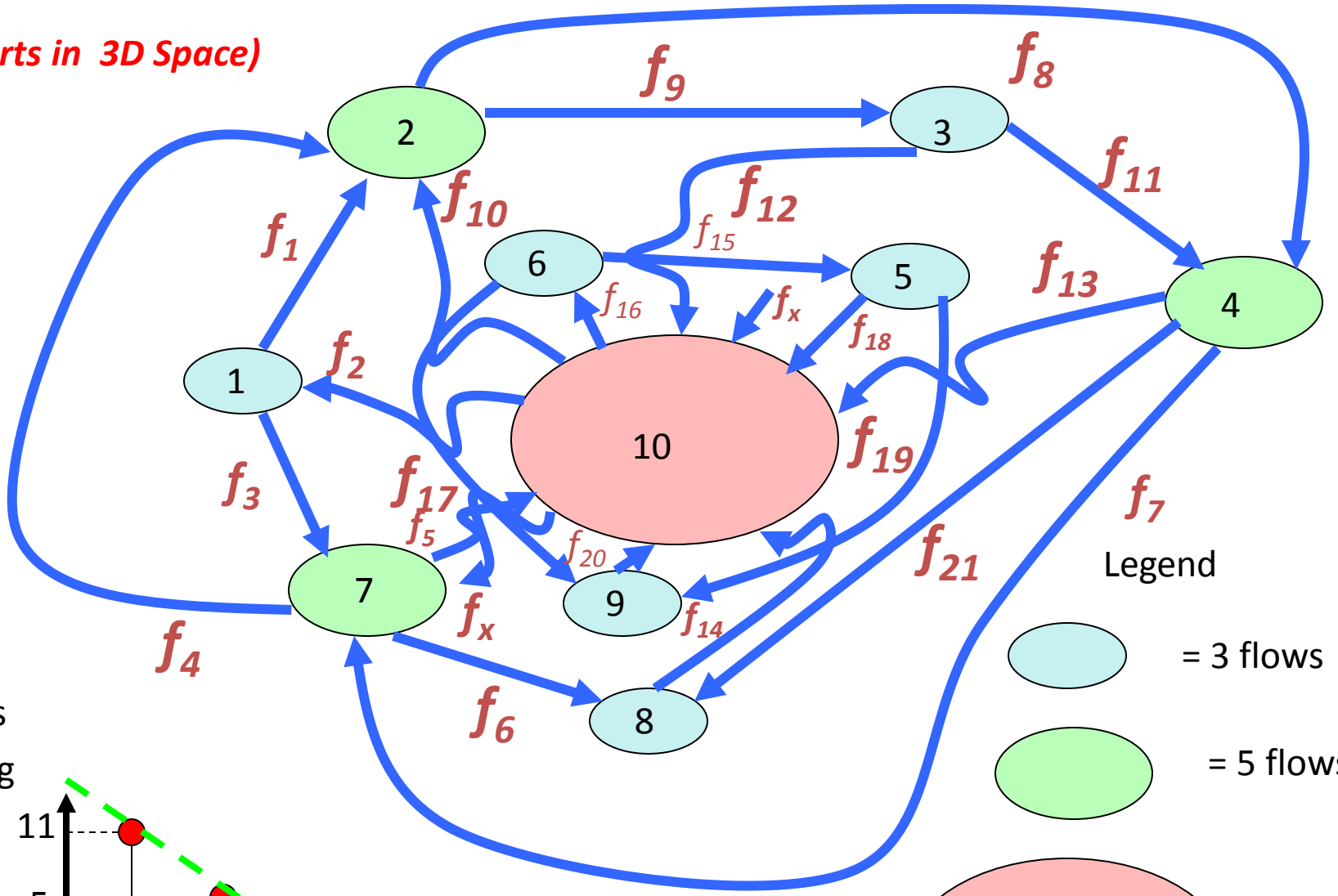




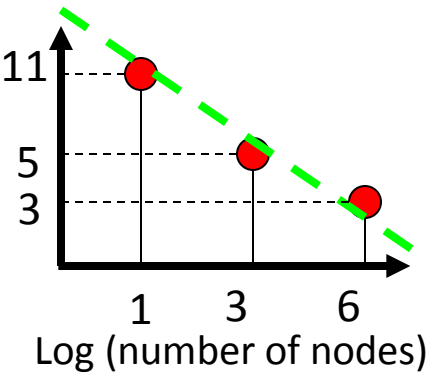
<u>Vulnerability:</u> Random graphs are more robust under ***intentional attacks*** versus scale free. For ***random attacks*** all nodes should have about the same level of robustness.

Scale Free Graph (Power Law, Fat Tail)

(Airports in 3D Space)

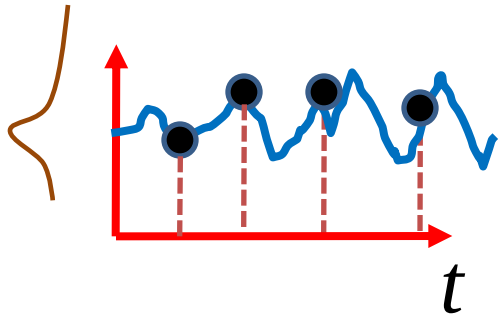


Log (∑ flows Involving cut set)



Vulnerability: Scale free are more robust under ***random attacks*** but very vulnerable under ***intentional attacks***

Consider a time series signal $p(t)$

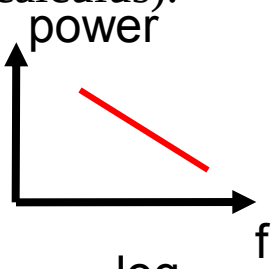


Statistical Properties
(e.g. μ and σ)

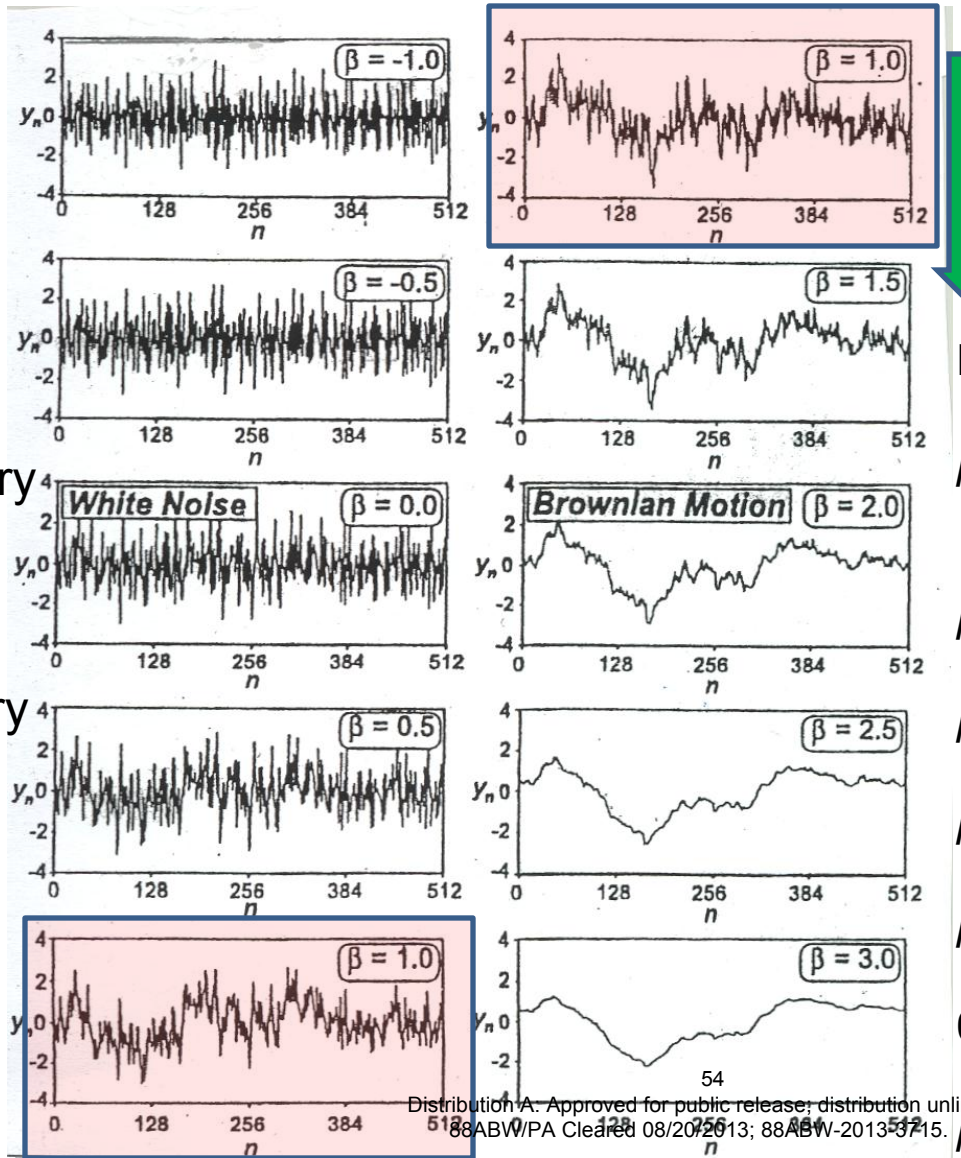
Information Properties
(e.g. persistence, memory, correlation, approximate entropy)

Topic 3 – Fractional Gaussian and Brownian Noise Studies (fractional calculus).

$$B(t) = \int \zeta(t)\,dt = \frac{d^{-1}}{dt^{-1}}\,\zeta(t)\ ,\quad P(t) = \text{pink noise } (1/f) = \frac{d^{-\frac{1}{2}}}{dt^{-\frac{1}{2}}}\,\zeta(t)$$



Fractional G. Fractional B.



Anti Persistent

Stationary

Non stationary

Weakly Persistent

Strongly Persistent

Non stationary

β = - slope

Rosetta Stone for Fractional G and B:

$\beta = 0 \Rightarrow$ Gaussian Noise, zero persistence, stationary

$\beta = 2 \Rightarrow$ Brownian Noise

$\beta = 1 \Rightarrow$ Pink Noise (1/f noise)

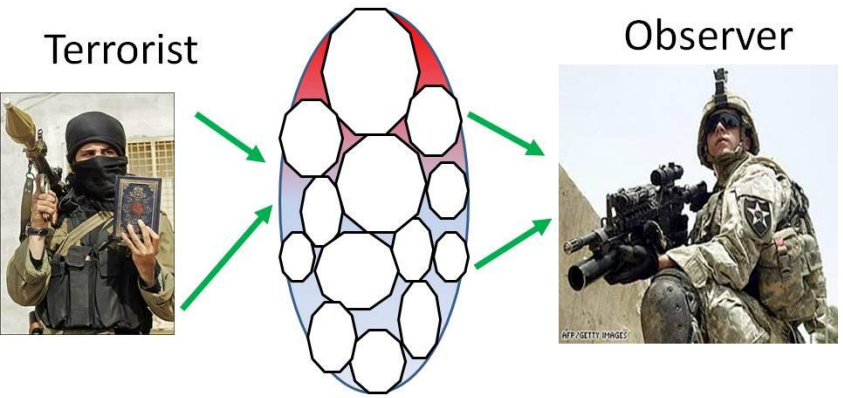
$\beta > 0 \Rightarrow$ Non stationary

$\beta > 1 \Rightarrow$ Strongly Persistent

$0 < \beta < 1 \Rightarrow$ Weakly Persistent

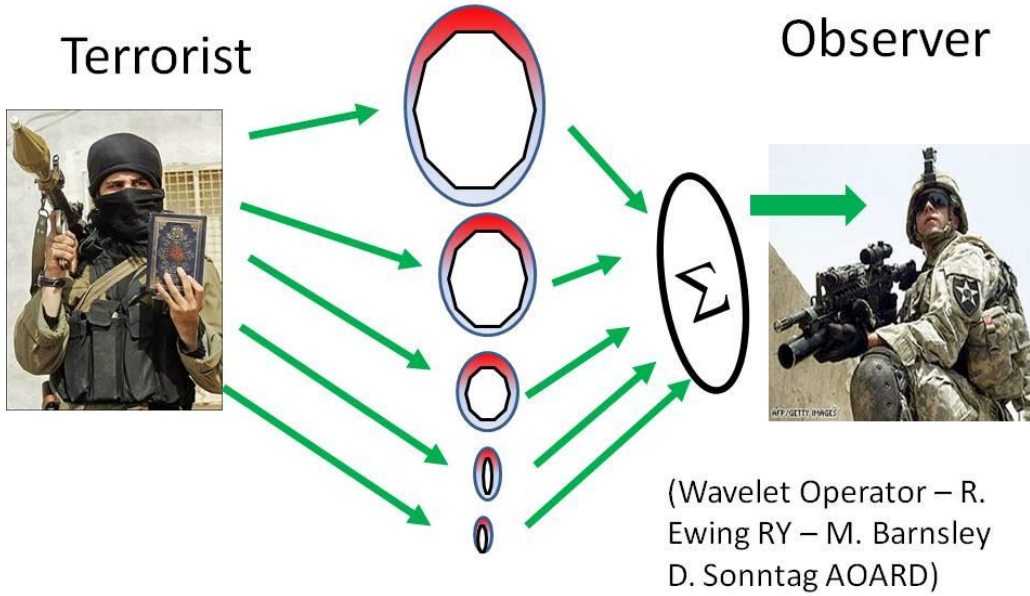
$\beta < 0 \Rightarrow$ Anti Persistent

The Fractal Filter Concept – M. Barnsley



One Filter

The Fractal Filter Concept – M. Barnsley



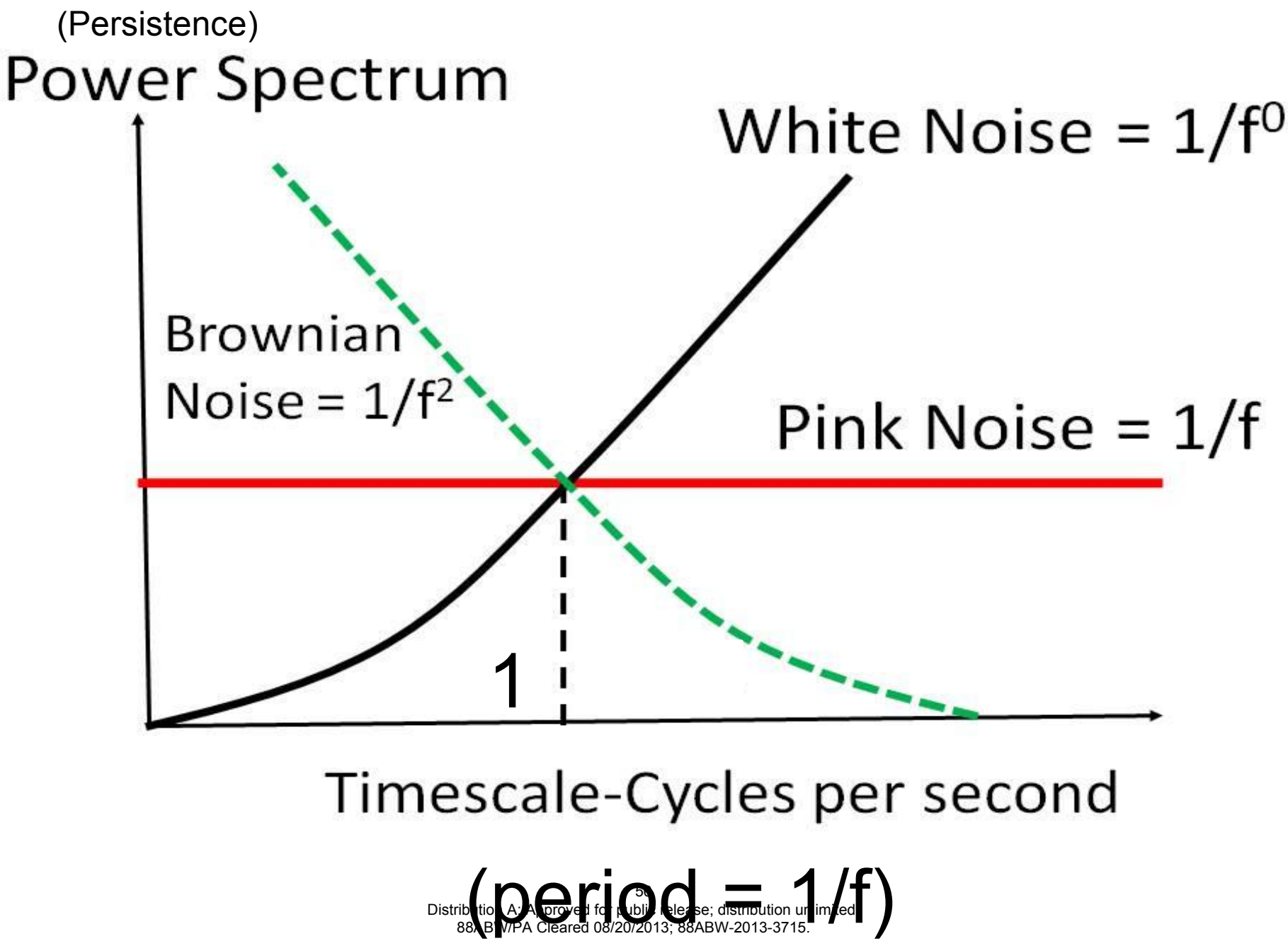
(Wavelet Operator – R. Ewing RY – M. Barnsley D. Sonntag AOARD)

Multiple Filters that are Fractal

Multiple Partners: D. Sonntag AOARD, M. Barnsley

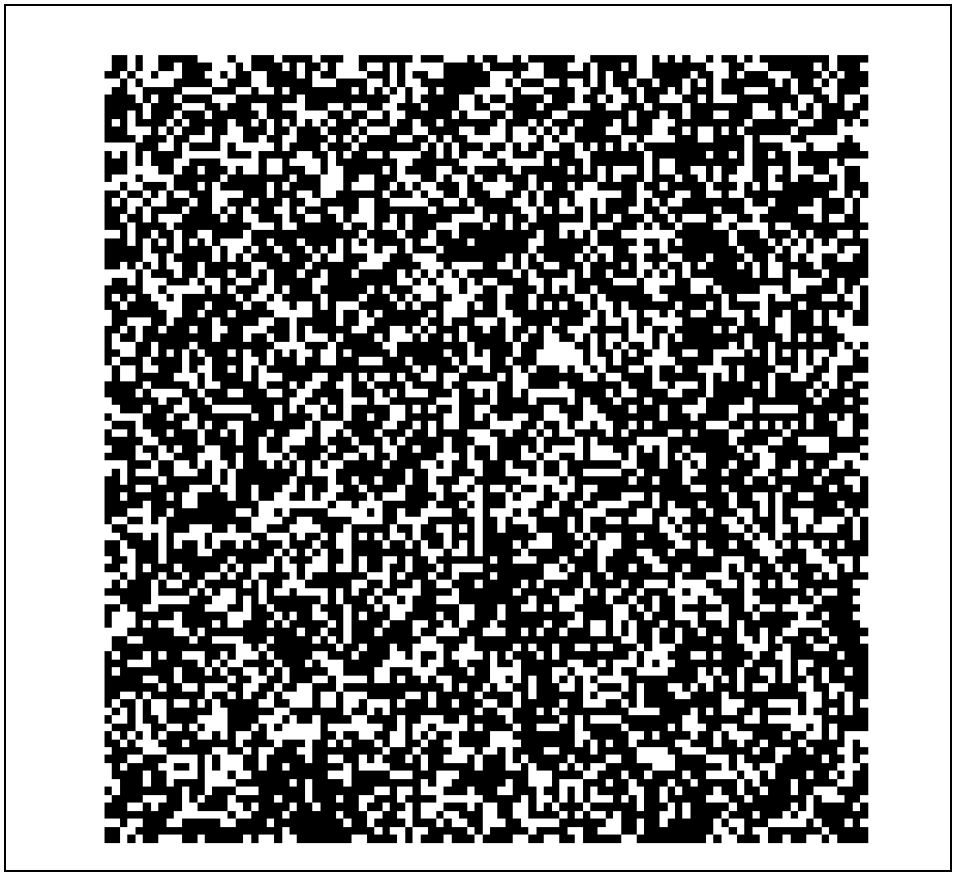
Points: Make a network (fractal geometry) ***resemble 1/f dynamics. No characteristic scale ⇒ No inherent limits of resolution.***

Why? Present values proportional to ***recent and far history equally***.

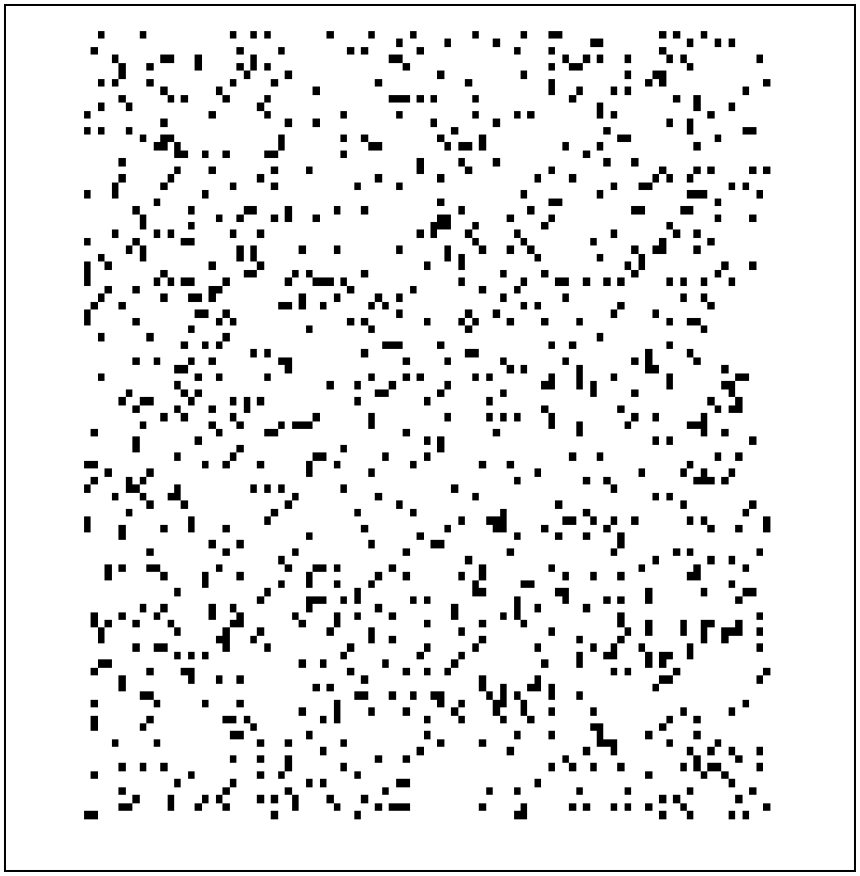


Topic 4 – Percolation Networks – Cellular Autonoma

9,999 Points



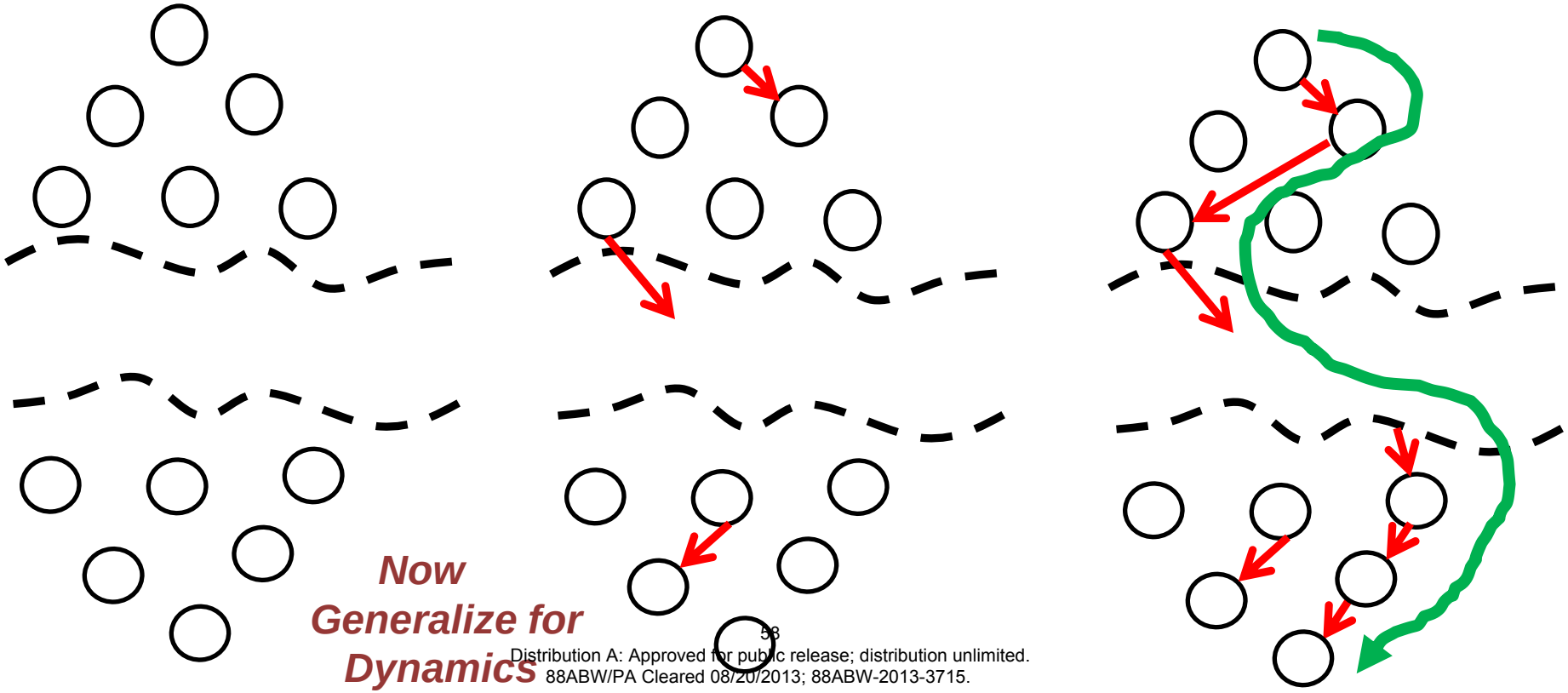
1,111 Points



How to use this *m.o.* to Analyze Networks?

(Extant work on Dynamics)

Step 1: Draw the architecture of the network (Barnsley support set).
Step 2: Populate the links randomly. (Monte Carlo)
Step 3: Test for percolation.
Step 4: When percolation occurs, continue link population.
Step 5: Continue until two or more percolation paths occur.
Step 6: With multiple percolations, now start removing nodes/links.
Step 7: If percolation disappears, then a critical node/link is identified.
Step 8: If percolation does not stop, then this is not a critical node/link.



Now Generalize for Dynamics

Work at AFIT – Lt. Col T. J. McDonald, Dr. Y. Kim

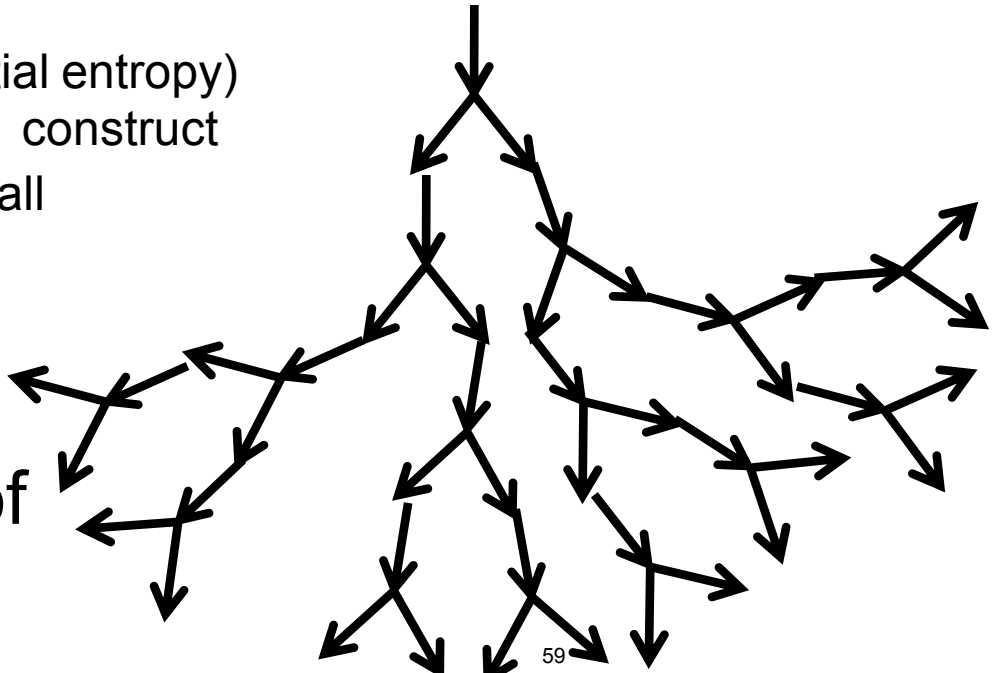
Given an electrical circuit (white/black box), can we study its time history and identity its function? (Dynamic electronic footprint problem).

Can the functionality of a circuit be hidden by adding nodes, links, etc. to disguise?

Foreign countries reverse engineer circuits/chips/software. How do we protect?

One solution: Make the reverse engineering N-P hard (maximum entropy).

Maximize (spatial entropy)
Shannon-Fano construct
Let p = sum of all
final nodes.



Number of questions = 2^p

Summary and Conclusions

Topic 1 – Logistics System.

Topic 2 – Architecture Issues (random vs scale-free graphs).

Topic 3 – Why are scale free systems in nature so optimal?

Topic 4 – Percolation Networks – Vulnerability, discover solutions, dynamics.

Topic 5 – Circuits/software at AFIT, Master's students ongoing.

APPENDIX B

Fractional Calculus – A New Paradigm for Understanding Complex Systems

Fractional Calculus – A New Paradigm for Understanding Complex Systems

D. W. Repperger[1], K. E. A. Farris[1] , R. Bradford[1]
[1]- 711 HPW, Air Force Research Laboratory, RHCV, WPAFB, Ohio 45433

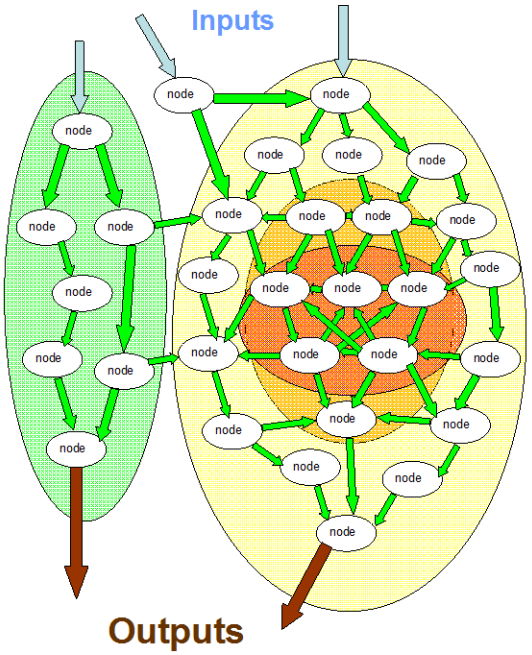


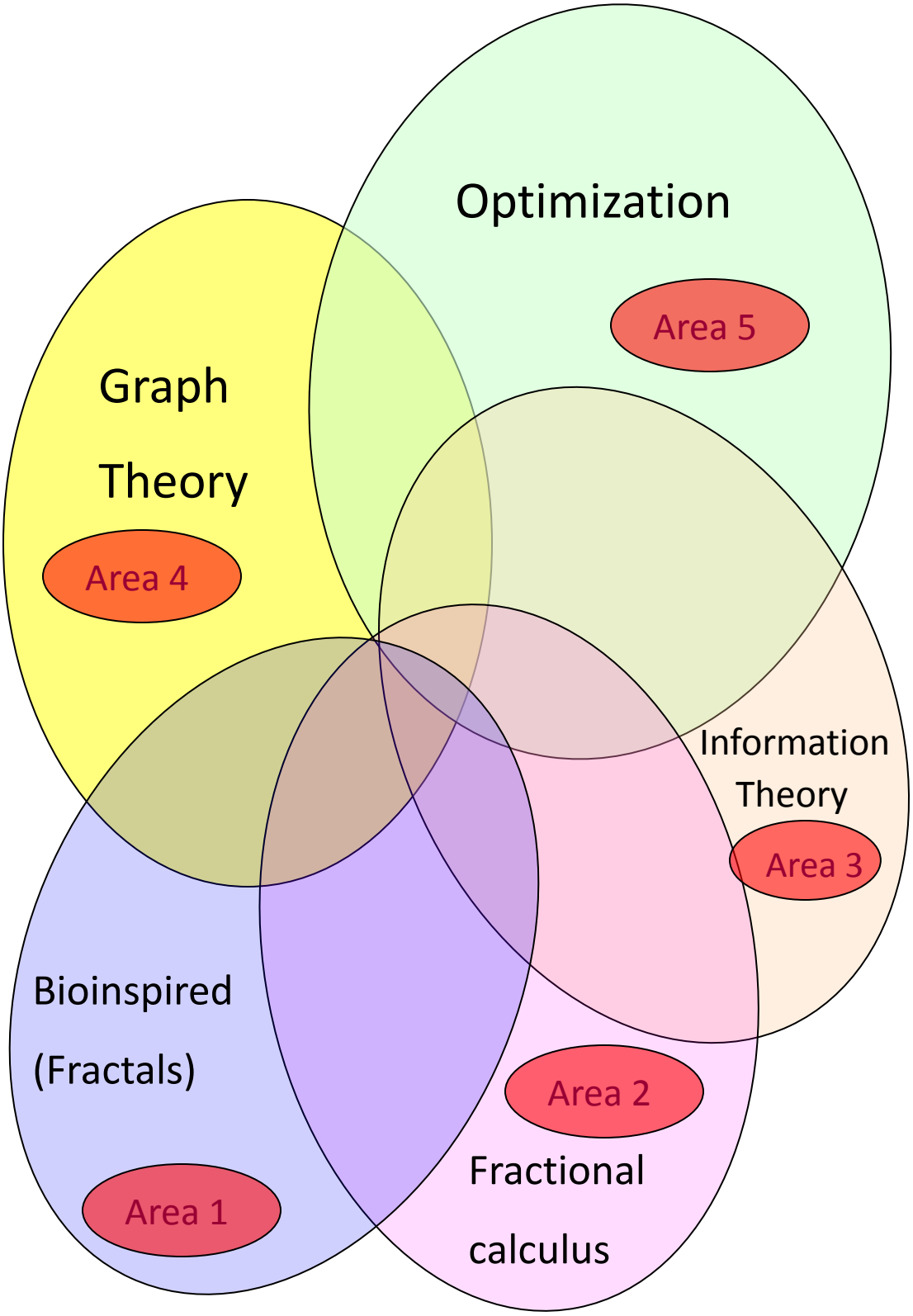
Figure 3 – The Original Network-Centric Distributed System

. Originally we are interested in flow performance of networks. Topics such as performance and vulnerability are important.

.From nature we borrow concepts on optimal flow with the side advantage of robustness.

. Fractional Calculus is a paradigm to investigate these structures and involves fractals, also ubiquitous in nature.

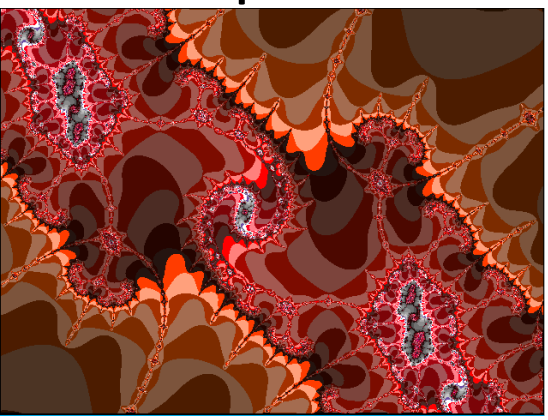
Part 1-B- Background Material

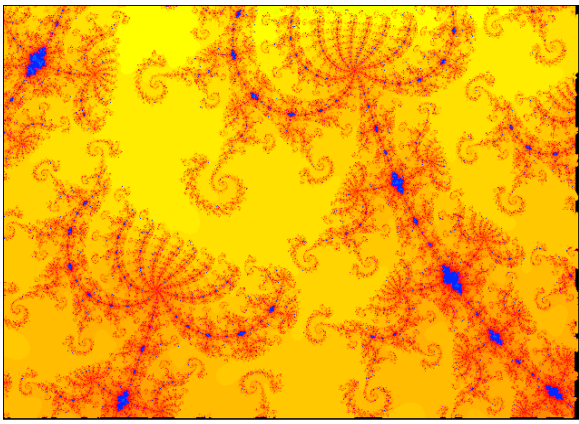


Part 1-B- Background Material

Area 1

Bioinspired - Fractals

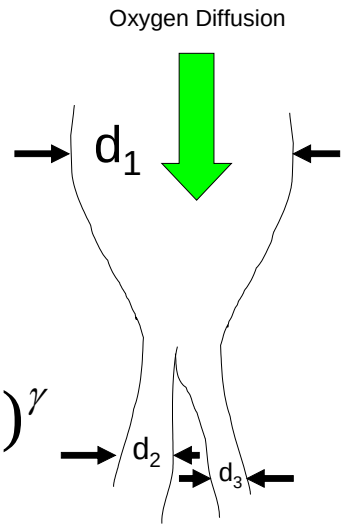




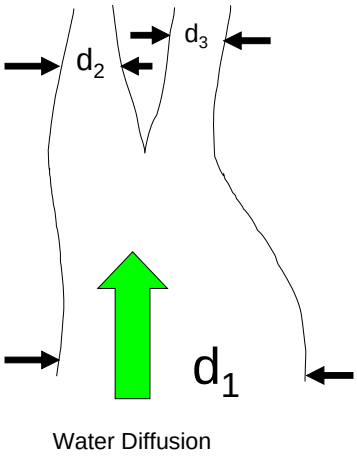
$$\frac{\partial^2 u(x,t)}{\partial x^2} = a^2 \frac{\partial u(x,t)}{\partial t}$$

$$(d_1)^\gamma = (d_2)^\gamma + (d_3)^\gamma$$

$\gamma = 2.5??$



Lung

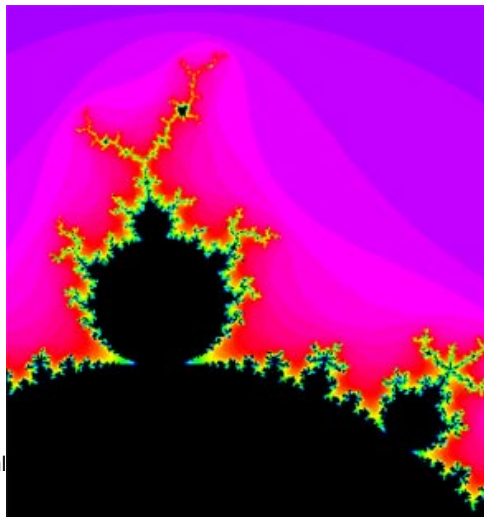


Tree

Fractional Dimensions are ***NOT***

Minimum energy –

They are ***Optimal*** for Diffusion



Part 1- Background Material

Area 1

Bioinspired - Fractals

. The Latin ***fractus*** = “broken” or “fractured”

. Fractals – scale free (self-similar), irregular overall length scales. (self similar means the structure is invariant to change in scale). ***Forever continuous but nowhere differentiable.***

. Fractals may have ***infinite circumference but finite area***.

. Fractals can have ***finite volume and infinite area.***

. A fractal can be defined in the sense of a recursive equation:

$$z_{n+1} = f(z_n)$$

. This is, apparently, the ***optimal way*** to distribute flow.

. Non Euclidean Geometry.

. Fractal examples (trees (branches), rivers, lighting bolts, cells, lung passageways, blood vessels, leaf patterns, cloud surfaces, molecular trajectories, neuron firing patterns, etc.).

Fractals – Lets Review the Area

Area 1

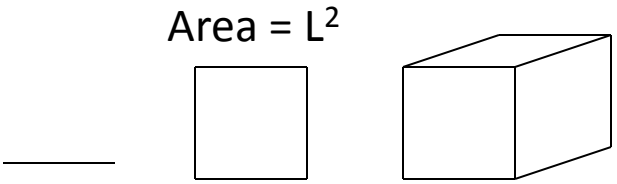
B. Mandelbrot (1960,s) asked the question: “How long is the coastline of Britain?”

(Suppose we measured the coastline with a ruler that got smaller and smaller?)

A fractal has statistical self-similarity(power law, self affine).

A fractal has N identical parts with scale factor L.

The Fractal dimension is



$$(\text{Measurement}) = L^{D}$$

$$\text{implies} \quad \log(\textit{Measurement}) = D\,(\log(L))$$

$$D = \frac{\log(\textit{Measurement})}{\log L} \neq \text{Integer}$$

Fractals – Lets Review the Area

Area 1

$$D = \frac{\log(Measurement)}{\log L}$$

$$(\text{Measurement}) = L^{D}$$

$$L \alpha A^{1/2} \alpha V^{1/3}$$

For irregular surfaces, we can define:

Let N = the number of divisions of fixed length.

Let r = length of a ruler.

$$D = \frac{\log(Total\ Length)}{\log(1/r)} \text{ as } r \to 0$$

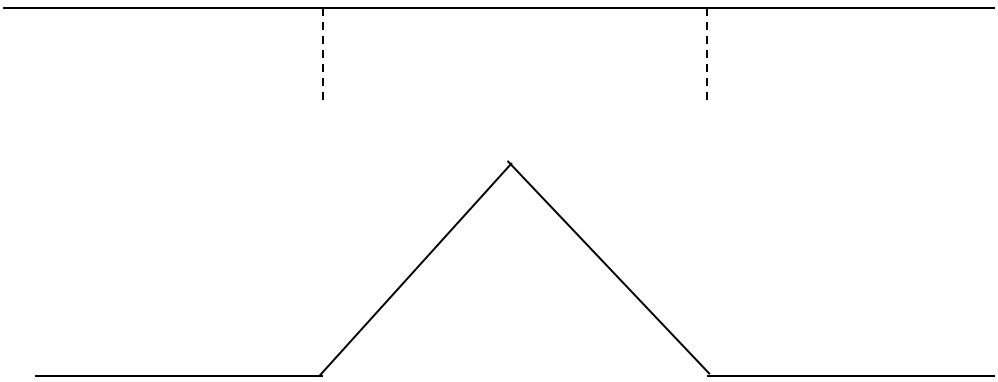
Fractals – Lets Review the Area

Area 1

$$D = \frac{\log(Measurement)}{\log L}$$

Total Length = L^D where $1 < D < 2$

Koch Snowflake

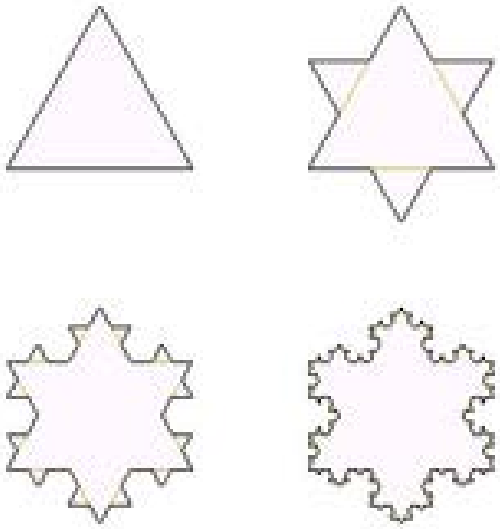


Length = 4 = measurement

Projection = topological dimension = 3

$$D = \frac{\log(4)}{\log(3)} = 1.26185...$$

Fractals – Lets Review the Area Different versions of the Koch snowflake.



Area 1

Finite

Area

Circumference

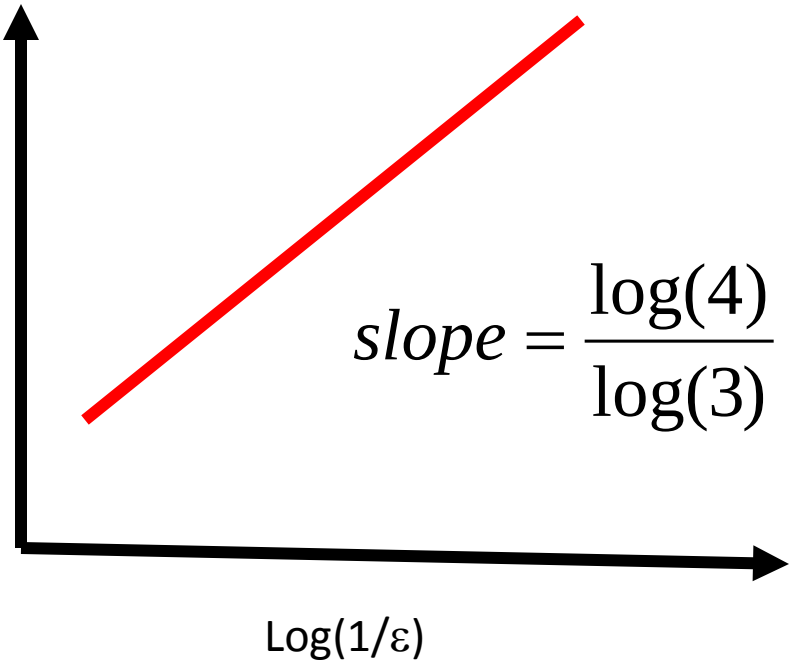
= total length

= $(4/3)^n$

lim (total length) $\rightarrow \infty$

$n \rightarrow \infty$

Power law



Biofractals

21 orders of magnitude

Microbe = 10^{-13} g

Whale = 10^{8} g

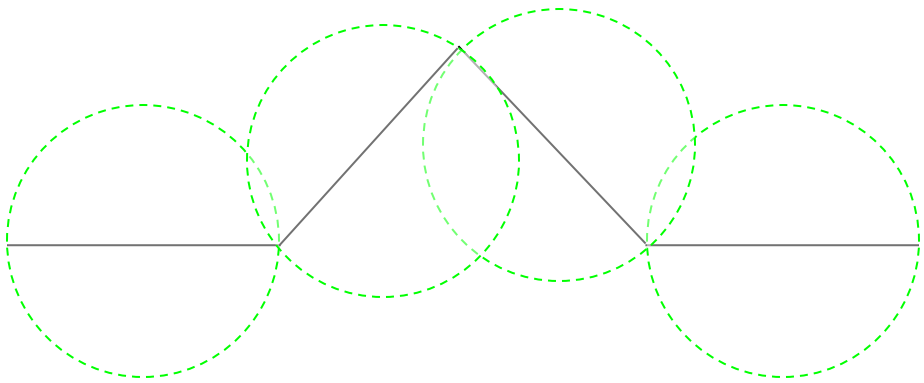
Fractals – Lets Review the Area.

Area 1

$$D = \frac{\log(Measurement)}{\log L}$$

How to determine Measurement?

We "cover" with boxes or disks.

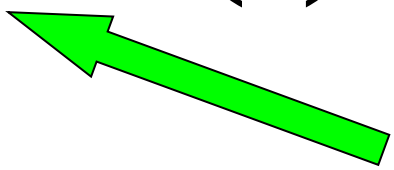


Fractional Calculus – Main Points

(non Euclidean geometry)

Area 2

$$\frac{d^n y}{dt^n} = u(t)$$



(Notation invented by Leibniz)

What can *n* be?

Answer:

(In 1695, L'Hopital asked Leibniz, suppose n= ½?)

n = integer = 1, 2, 3 4,

n = negative integer = -1, -2, -3

n can be a non integer, n = ½, 5/6.

n can be a negative non integer, n = -.6, -3.4,

n can be irrational: $n = \sqrt{2}$

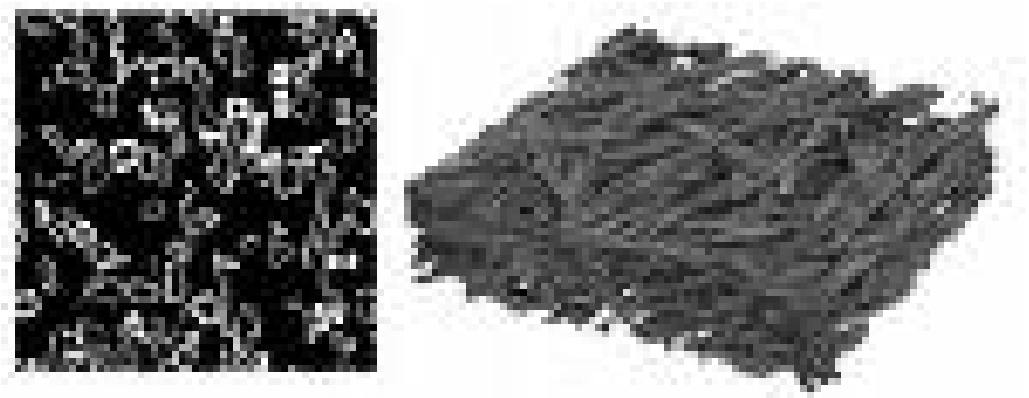
n can be a complex number: $n = \sqrt{-1}$

Fractional Calculus – Main Points

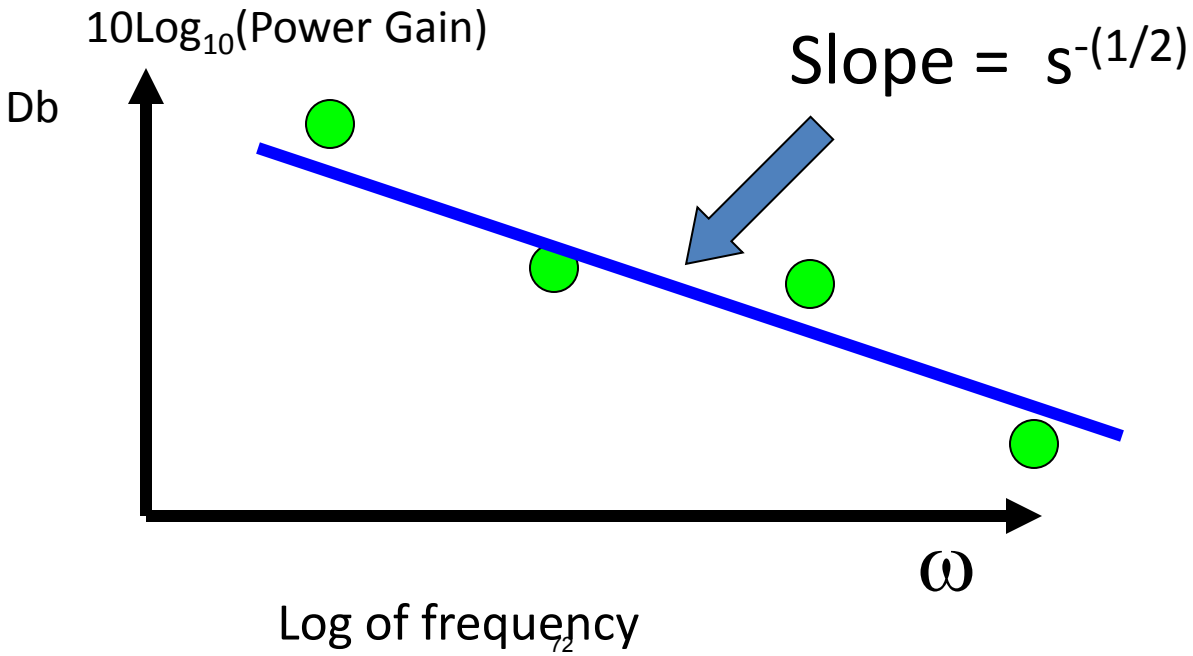
(non Euclidean geometry)

Area 2

Why Study Fractional Calculus?



Composite Materials



Fractional Calculus –Main Points

Area 2

Why use Fractional Calculus?

(1) It can deal with functions that are forever continuous and nowhere differentiable (fractals).

(2) It has the property of self similarity (scale invariance)

$$\frac{d^{5/2}(\alpha y)}{d(\alpha t)^{5/2}}+\frac{d^{3/2}(\alpha y)}{d(\alpha t)^{3/2}}+\frac{d^{1/2}(\alpha y)}{d(\alpha t)^{1/2}}=\frac{d^{3/2}(\alpha u)}{d(\alpha t)^{3/2}}+\frac{d^{1/2}(\alpha u)}{d(\alpha t)^{1/2}}$$

$$\frac{d^{q} f(bx)}{[dx]^{q}} = b^{q}\frac{d^{q} f(bx)}{[d(bx)]^{q}}$$

(3) It is also of the form:

$$z_{n+1} = f(z_n)$$

(Iterated function theory).

(4) It can also solve partial differential equations:

$$\frac{\partial^2 u(x,t)}{\partial x^2} = a^2\frac{\partial u(x,t)}{\partial t}$$

Fractional Calculus

Area 2

An Easier Way to View the Self Similarity Property

A power law $f(x) = x^a$ has the property that that the relative change in

$$\frac{f(kx)}{f(x)} = k^a$$

Is independent of x

In this sense, the functions lacks characteristic scale (scale free or scale invariant). Let us evaluate $\frac{f(kx)}{f(x)}$

Let $x = y^a$

Then

$$\frac{f(kx)}{f(x)} = \frac{(ky)^a}{y^a} = k^a \frac{y^a}{y^a} = k^a$$

Note: no dependence on x

Fractional Calculus –Main Points

(310 year old area). Non Euclidean

Area 2

Common Properties

(1) Scale Invariance – Self Similarity.

$$\frac{d^q f(bx)}{[dx]^q} = b^q \frac{d^q f(bx)}{[d(bx)]^q}$$

(2) $$\frac{f(kx)}{f(x)} = \frac{(ky)^a}{y^a} = k^a \frac{y^a}{y^a} = k^a$$

(3) Solves Systems in Nature (Diffusion equation).

(4) Let $m = \beta$

$$\frac{d^\beta}{dx^\beta} x^m = \frac{\Gamma(m+1)}{\Gamma(m-\beta+1)} x^{m-\beta}$$

Fractional Calculus –Main Points

(310 year old area). Non Euclidean

Area 2

Common Properties

(1) Scale Invariance – Self Similarity.

$$\frac{d^q f(bx)}{[dx]^q} = b^q \frac{d^q f(bx)}{[d(bx)]^q}$$

(2) $$\frac{f(kx)}{f(x)} = \frac{(ky)^a}{y^a} = k^a \frac{y^a}{y^a} = k^a$$

(3) Solves Systems in Nature (Diffusion equation).

(4) Let $m = \beta$

$$\frac{d^\beta}{dx^\beta} x^m = \frac{\Gamma(m+1)}{\Gamma(m-\beta+1)} (1)$$

Fractional Calculus –Main Points

Area 2

(Solution of the Diffusion Equation)

$$\Gamma(z) = \int_0^\infty e^{-u} u^{z-1} du, \qquad \Gamma(1) = 1, \Gamma(z+1) = z\Gamma(z),$$

Thus: $\Gamma(z+1) = z!,$ $\Gamma(\frac{1}{2}) = \sqrt{\pi}$

Step 1 – Derivatives in x^m

$$\frac{d}{dx} x^m = m x^{m-1}, \qquad \frac{d^\beta}{dx^\beta} x^m = \frac{m!}{(m-\beta)!} x^{m-\beta}$$ but β may not be an integer

$$\frac{d^\beta}{dx^\beta} x^m = \frac{\Gamma(m+1)}{\Gamma(m-\beta+1)} x^{m-\beta}, \qquad \frac{d^{\frac{1}{2}}}{dx^{\frac{1}{2}}} x^1 = \frac{\Gamma(1+1)}{\Gamma(1-\frac{1}{2}+1)} x^{1-\frac{1}{2}} = \frac{2}{\sqrt{\pi}} x^{\frac{1}{2}}$$

This now ***generalizes*** for derivatives in e^{ax}

$$D^v e^{ax} = a^v e^{ax}$$

(v not an integer)

Generalizations to functions that can be written in a power series:

$$f_1(t) = \sum_{n=0}^{q} a_n + b_n x^n$$

Generalizations to functions that can be written in an exponential series:

$$f_2(t) = \sum_{n=0}^{q} a_n + b_n e^n$$

$$e^{i\theta} = \cos(\theta) + i\sin(\theta)$$

Euler's Law:

$$\cos(\theta) = \frac{e^{i\theta} + e^{-i\theta}}{2}$$

Fractional Calculus –Main Points

(Solution of the Diffusion Equation)

Area 2

Step 2 – Laplace Transform

$$F(s) = L[f(t)] = \int_0^\infty f(t)e^{-st}dt$$

Then: which holds if

$$e^{-\alpha t} \mid f(t) \mid \leq M < \infty$$

$$L^{-1}[F(s)] = f(t)$$

$$L^{-1}\left(\frac{1}{s^{1+\beta}}\right) = \frac{t^{\beta}}{\Gamma(\beta+1)}, \beta > -1$$

$$L^{-1}\left[\frac{1}{s^{\frac{1}{2}}}\right] = \frac{t^{\frac{-1}{2}}}{\Gamma\left(\frac{1}{2}\right)} = \frac{1}{(\sqrt{\pi})t^{\frac{1}{2}}}$$

Step 3 - Diffusion Equation:

$$\frac{\partial^2 u(x,t)}{\partial x^2} = a^2 \frac{\partial u(x,t)}{\partial t}$$

$$U(x,s) = L[u(x,t)] = \int_0^\infty e^{-st}u(x,t)dt$$

$$L\left[\frac{\partial u}{\partial t} - \frac{1}{a^2}\frac{\partial^2 u}{\partial x^2}\right] = sU(x,s) - f(x) - \frac{1}{a^2}\frac{\partial^2 U}{\partial x^2} = 0$$

$$U(x,s) = Ae^{xas^{\frac{1}{2}}} + Be^{-axs^{\frac{1}{2}}} = \frac{1}{a^2 2\sqrt{s}} \int_{-\infty}^{\infty} e^{-\sqrt{s}|x-\tau|} f(\tau)d\tau$$

$$u(x,t) = \frac{1}{2\sqrt{\pi t}} \int_{-\infty}^{\infty} e^{-\frac{-(x-\tau)^2}{4t}} f(\tau)d\tau$$

Summary and Conclusions

. Investigation of flow performance of networks. Has initialed these efforts. Networks in nature have excellent flow performance and are very robust.

. We look at fractal structures for insight into this problem.

. Related to fractals is the Fractional calculus, which explains the scale free effects seen in nature in a mathematical sense.

APPENDIX C

Sparse Statistical Data Analysis Based on the L_1=norm = Case Study

Sparse statistical data analysis based on the L_1-norm - case study

Xiaoping A. Shen

http://www.ohio.edu/people/shenx

Department of Mathematics
Ohio University
Athens, Ohio 45701

Sponsored AFOSR FA8650-08-D-6801

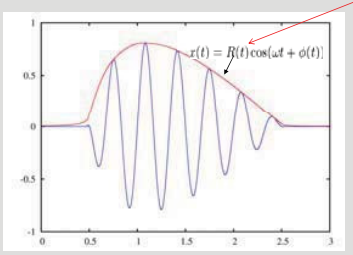
OHIO

Outline

- Background – communication channel characterization
- The basic idea - non parametric (kernel density estimator)
- Kernel selection – Slepian function
- Bandwitth selection – L_1 optimization
- Numerical examples.

OHIO

Basics of channel characterization – identify probability density function of signal envelope



OHIO

Estimation of fading envelopes

Parametric estimation Fading envelope usually assumed to follow certain probability distributions. Such fading models include:

- Rayleigh Distribution

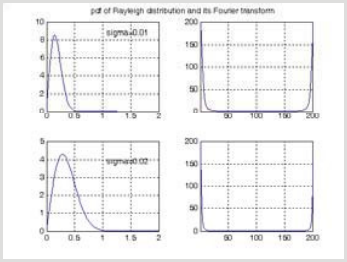
$$f(r) = \frac{r}{\sigma^2}\exp\left(-\frac{r^2}{2\sigma^2}\right) \quad \text{for } 0 \le r \le \infty$$
$$0 \quad \text{for } r < 0$$

- Rician Distribution:

$$f(r) = \frac{r}{\sigma^2}\exp\left(-\frac{(r^2+A^2)}{2\sigma^2}\right)I_0\left(\frac{Ar}{\sigma^2}\right) \quad \text{for } A \ge 0, r \ge 0$$
$$0 \quad \text{for } r < 0$$

OHIO

The pdf of Rayleigh distribution and its spectral



OHIO

An example of no-Line-of-sight propagation communication channel -- Manhattan

Rayleigh fading channel

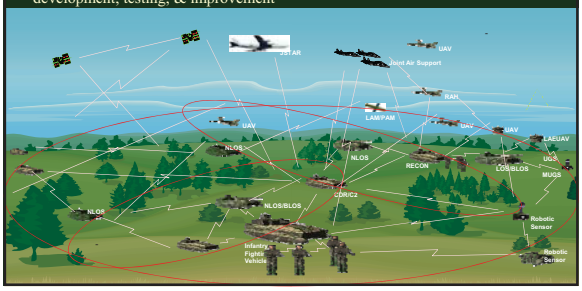
OHIO

An example of line of sight propagation communication channel -- Alaska

OHIO

Network Centric Warfare- Digital Battlefield Instrumentation

- Technically NCW will present unseen communications and analysis challenges and complexities.
- NCW Instrumentation must be equally as robust to allow the capture of the data at each node in the battlefield network in support of NCW analysis development, testing, & improvement



Challenge in NCW Data collection methods and processing

- Data manually colleted to create final reports
- Limited Real-time feedback
- How many nodes need to be monitored to have a large enough sample of the data?
- Where do you monitor -TOC, Relays, Weapons (Radio, INC, LAN), UAV's ?

OHIO

Partial list of radios in the battle environment

- **SINCGARS** > 30-88 MHZ - 16 KBPS digital voice, data 75 BPS & 16 KBPS
- **Have Quick II** (UHFAM/FM/PSK) > 225-400 MHZ - Analog & 16 KBPS digital voice, data 75 BPS to 16 KBPS
- **EPLRS** > 420-450 MHZ. Data mode 57 KBPS VHSIC SIP and 228 KBPS VECP
- **Wideband Networking Waveform** > Expected 2 MHZ to 2 GHZ up to 5 MBPS networked throughput
- **LINK 16** > 960 to 1215 MHZ. - voice 2.4 and 16 KBPS, data 28.8 KBPS to 1.137 MBPS
- **LINK 11/TADIL** > 2-30 MHZ and 225-400 MHZ - data 1364 and 2250 BPS. VHF FM - - 30-88 MHZ - analog and digital voice at 16 KBPS
- **VHF ATC Data Link** > 2-30 MHZ. – analog, voice & data at 300, 600, 1200 & 1800 BPS
- **VHF/UHF FM LMR** > 136-174 MHZ, 380-512MHZ, and 764-869 MHZ analog & digital voice at 16 KBPS

OHIO

Partial list of radios in the battle environment

- **Soldier Radio and WLAN** > 1.755-1.850 GHZ. - digital 16 KBPS voice, data at 1 MBPS.
- **MUOS** > 240-320 MHZ - data at 2.4, 9.6, 16, 32 and 64 KBPS.
- **Cellular Radio** > 824-894 MHZ, 890-960 MHZ and 1850-1990 MHZ - voice and data at 10 KBPS nominal, 3G rates up to 144/384 KBPS and 2 MBPS.
- **Mobile Satellite Service (MSS)** > 1.61-2 GHZ - digital voice at 2.4 to 9.6 KBPS.
- **Integrated Broadcast Service-Module (IBS-M)** > 225-400 MHZ - data at 2.4, 4.8, 9.6 and 19.2 KBPS.
- **VHF ATC Data Link (NEXCOM)** > 118-137 MHZ - digital voice at 4.8 KBPS, data at 31.5 KBPS.
- **UHF AM/FM PSK** > 225-400MHZ and 225-450 MHZ – analog, digital voice and data rates up to 16 KBPS.
- **Link 4A** > 225-400 MHZ. - data at 5 KBPS.
- **Link 11B** > 225-400 MHZ - data at 600, 1200, and 2400 BPS.
- **DWTS** > 1350-1850 MHZ – analog, digital voice and data at rates of 144, 256, 288, 512, 576, 1024, 1152, 1544, 2048 and 2304 KBPS.

OHIO

Channel characterization in NCW

- Relay on **infrequently** transmitted channel state information (CSI) feedback data
- Historical data may not be available
  - for a specific location to be analyzed;
  - extensive relevant for all stochastic variables.

The typical situation

- available, relevant data are too sparse to provide a good basis for probability assessment
- parametric estimation is not always valid. The rigid distributional assumptions constrain the data to fall in a given parametric family.
- heavy noise

OHIO

The idea – kernel density estimation

- choose kernel with good sampling property and
- use L_1 measure control bandwidth selection

OHIO

Examples of kernel density estimation

- The naïve estimator:

$$\hat{f}_m(x) = \int_{-\infty}^{\infty} q_m(x,y) f^*(y) dy = \frac{1}{N}\sum_{i=1}^{N} q_m(x, X_i)$$

where $q(x,y) = \chi(y - |x|)$

- Orthogonal kernel estimator: Replace χ by reproducing kernel of an orthogonal system (for example, father wavelet)

$$q_m(x,y) = \sum_n \varphi(2^m x - n)\overline{\varphi(2^m y - n)}$$

OHIO

Kernel density estimation based on Slepian function

Advantages:

Higher convergence rate

comparing to non wavelet density estimators;

comparing to positive density estimators (such Gaussian)

- Avoiding Gibbs phenomenon comparing to non positive density estimators.

Remark:

- The kernel is positive definite kernel
- It is not positive kernel
- It is locally positive (this improve the convergence rate while positive interval can be chosen to meet a particular need.)

OHIO

The energy concentration problem
(Slepian, Landau, and Pollak at Bell Lab, 1960s)

- **Define**

$$\underbrace{\alpha^2(\tau; f) \equiv \frac{\int_{-\tau}^{\tau} |f(t)|^2 \, dt}{\int_{-\infty}^{\infty} |f(t)|^2 \, dt}}_{\text{Time concentration index}}, \quad \text{and} \quad \underbrace{\beta^2(\sigma; f) \equiv \frac{\int_{-\sigma}^{\sigma} |\hat{f}(\omega)|^2 \, d\omega}{\int_{-\infty}^{\infty} |\hat{f}(\omega)|^2 \, d\omega}}_{\text{frequency concentration index}}$$

$0 \le \alpha^2(\tau), \ \beta^2(\sigma) \le 1$

-

OHIO

The energy concentration problem

Consider the function space B_σ^2 (the Paley - Wiener space)

- **Q1.** How large $\alpha^2(f;\tau)$ can be ?
- **Q2.** Which σ bandlimited function possesses the largest time concentration index?
- Mathematically, find the solution to the optimization problem:

$$\max_{f \in B_\sigma^2} \alpha^2(\tau; f)$$

OHIO

Solution of the concentration problem
-- Slepian's lucky accident

- The energy concentration problem can be rewritten as the eigenvalue problem of an integral operator,

$$S_{\sigma,\tau}(\phi)(t) = \lambda\phi(t),$$

where $S_{\sigma,\tau}(\phi)(t) := \int_{-\tau}^{\tau} \phi(x) \frac{\sin c(t-x)}{t-x} dx, \ \ c = \sigma\tau$

- The integral operator is commute with a second order differential operator
-

$$S_{\sigma,\tau} P_{\sigma,\tau} = P_{\sigma,\tau} S_{\sigma,\tau}$$

where $P_{\sigma,\tau}(\phi)(t) := (1-t^2)\frac{d^2\phi}{dt^2} - 2t\frac{d\phi}{dt} - \sigma t^2 \phi$

OHIO

Prolate functions $\{\phi_{n,\sigma,\tau}(t)\}_{n=0}^{\infty}$ are the solution to the concentration problem.

Their associated eigenvalues are the corresponding time concentration index.

$$\phi_{0,\sigma,\tau}\ \phi_{1,\sigma,\tau}\ \phi_{2,\sigma,\tau}\ \cdots\ \phi_{n,\sigma,\tau}\cdots$$

$$\downarrow\quad\downarrow\quad\downarrow\quad\ldots\quad\downarrow\ \ldots$$

$$\lambda_0 > \lambda_1 > \lambda_2 > \ldots > \lambda_n > \ldots > 0$$

$$\| $$

$$\alpha(\phi_{0,\sigma,\tau};\tau)$$

OHIO

Properties of PSWFs

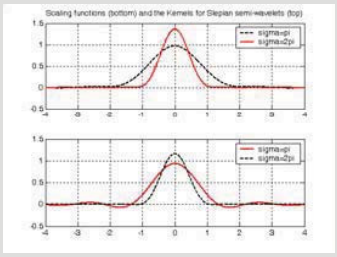
- **"Step function" property of the associated eigenvalues**

  For each $0 < \delta,\ \varepsilon < 1$, there exits γ, such that for $\sigma\tau > \gamma$, $\lambda_k < \varepsilon$ when $k \geq 2(1+\delta)\sigma\tau$, $\lambda_k > 1-\varepsilon$ when $k \leq 2(1-\delta)\sigma\tau$,

$$\underbrace{\lambda_0 \quad \lambda_1 \quad \ldots \quad \lambda_{[2\sigma\tau]}}_{\text{Close to 1}} \quad \underbrace{\lambda_{[2\sigma\tau]+1} \quad \ldots\ldots\ldots}_{\text{close to 0}}$$

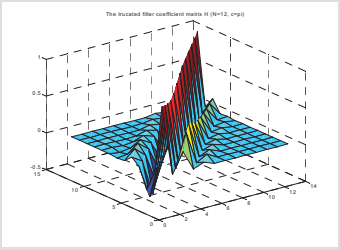
OHIO

The kernel and its associated Slepian function

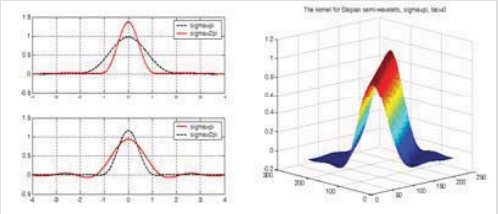


OHIO

Truncated Reproducing kernel

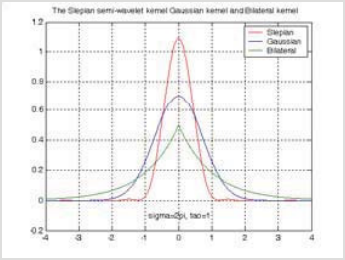


OHIO



OHIO

The Slepian kernel, Bilateral kernel, and Gaussian kernel



OHIO

Properties

1. The kernel $K_m(x,t)$ is positive definite convolution kernel satisfying

$$K_m(x,t) = K_m(x-t,0) > 0$$

for $|x-t| \le \frac{1}{2^{m+1}}$

2. The density estimator is asymptotically unbiased

OHIO

Computational matters – the estimator is simply a convolution of two sequences

- The linear kernel density estimator is defined as

$$p^{l}_{m,N}(x) = \frac{1}{N}\sum_{k=1}^{N} K_m(x, X_k),$$

where

$$K_m(x,t) = \sum_{n=-\infty}^{\infty} \frac{\phi_m(x-n2^{-m})\phi_m(t-n2^{-m})}{2^m |\hat{\phi}_m(0)|^2}$$

- Its truncated version can expressed as:

where

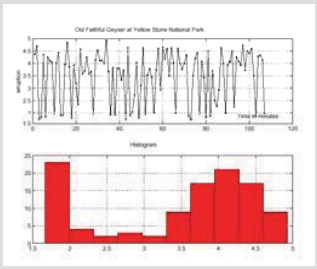
$$p^{l}_{m,N,K_1,K_2}(x) = \frac{C_{m,\tau}}{N}\sum_{k=1}^{N} d_n b_n,$$

$$b_n = \sum_{k=1}^{N} \varphi_{0,2^m\pi,\tau}(X_i - n2^{-m}),$$

$$d_n = \sum_{n=K_1}^{K_2} \varphi_{0,2^m\pi,\tau}(X_i - n2^{-m}), \quad C_{m,\tau} = \frac{\lambda_{0,2^m\pi,\tau}}{2\tau\varphi^2_{0,2^m\pi,\tau}(0)}$$

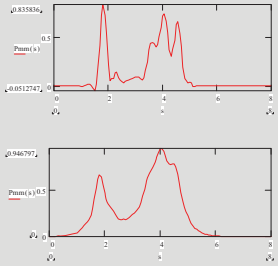
OHIO

The eruption length (in minutes) of the Old Faithful geyser at Yellowstone National Park (top) and its histogram estimator (bottom).



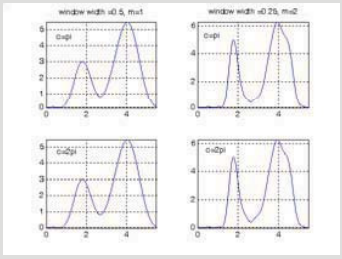
OHIO

The wavelet kernel estimator (Db4) (top) and associated positive kernel estimator (bottom)



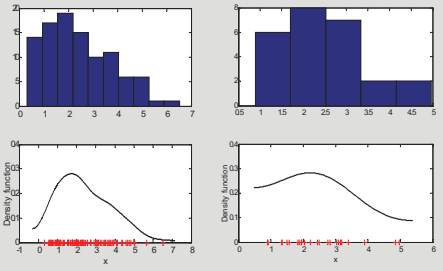
OHIO

Density estimate for the Old Faithful geyser data by using the Slepian semi-wavelet kernel, window width 0.5 (left) and window width 0.25 (right).



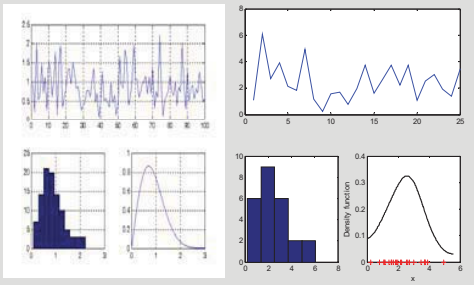
OHIO

Kernel density estimation using Gaussian kernel



OHIO

Density estimation for Rayleigh random sample (size=100)



OHIO

References

1. Slepian, D.; Pollak, H. O., (1961), Prolate spheroidal wave functions, Fourier analysis and uncertainty, I, *The Bell System Technical Journal 40, 43-64.*
2. Landau, H. J. ; Pollak, H. O. (1961), Prolate spheroidal wave functions, Fourier analysis and uncertainty, II, *The Bell System Technical Journal 40, 65-84.*
3. Landau, H. J. ; Pollak, H. O. (1962), Prolate spheroidal wave functions, Fourier analysis and uncertainty, III, *The Bell System Technical Journal 41, 1295-1336.*
4. Slepian, D. (1964), Prolate spheroidal wave functions, Fourier analysis and uncertainty, IV, *The Bell System Technical Journal 43, 3009-3058.*
5. Slepian, D. (1983), Some comments on Fourier analysis, uncertainty, and modeling, SIAM Review 25, 379-393.
6. Courant, R.; Hilbert, D., (1953) Methods of mathematical physics, New York, Interscience Publishers, 1953-62.

1. Xiao, H; Rokhlin, V.; Yarvin, N. (2000), Prolate spheroidal wave functions, quadrature and interpolation, *Inverse Problems 17 (2001), no. 4, 805--838.*
2. G. Beylkin and L. Monzon, On Generalized Gaussian Quadratures for Exponentials and their Applications, *Preprint, Dec. 2000.*

1. Walter, G. G; Shen, X., (2001), Sampling with Prolate Spheroidal Functions, submitted to *the Journal of Sampling Theory in Signal and Image Processing.*
2. Walter, G. G; Shen, X., (2002), Wavelets based on Prolate Spheroidal Functions, submitted to the *Journal of Fourier analysis and applications*.
3. G. G. Walter and X. Shen. Wavelet-like behavior of Slepian functions and their use in density estimation. *Communication in Statistics-Theory and Method*, Vol. **34** (3), 687-711, 2005.
4. X. Shen. Slepian semi wavelets and their use in density estimation for wireless signals, *Journal* of *Mathematical Study*, Vol. **40** (6), 117-131, 2007.

APPENDIX 5

Shannon Entropy and Relatives: A Brief Review and New Development

Shannon entropy and relatives: A brief review and new development

In memory of Dr. Daniel W. Repperger

Annie X. Shen[1], Catherin R. Farris[2], David J. Rieksts[1] and Paul R. Havig[2]

[1] Department of Mathematics, Ohio University, Athens, Ohio 45701
[2] 711 HPW AFRL/RHCV, Wright-Patterson Air Force Base, Dayton, Ohio 45433-7022

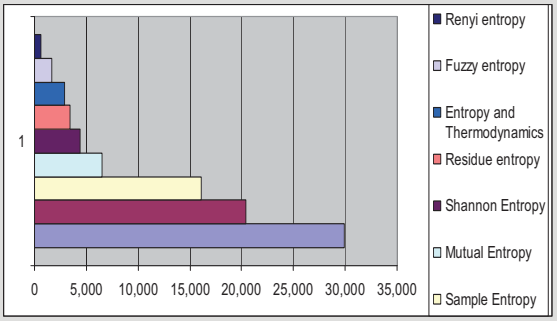
Sponsored by FA8650-08-D-6801

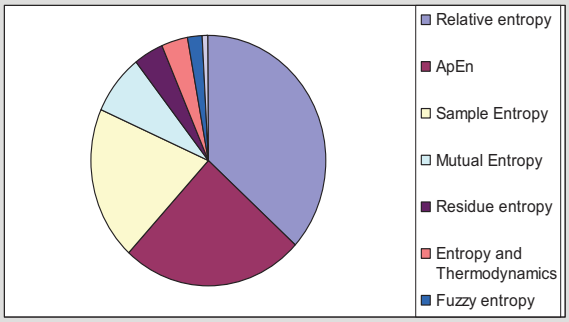
Outlines

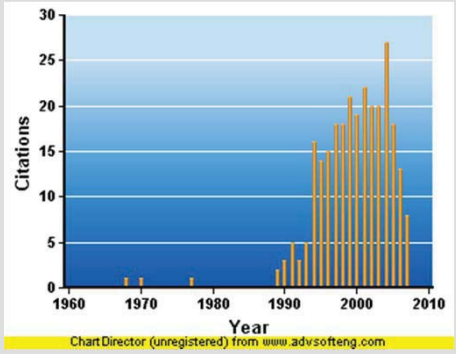
- Statistics - What happened in the past?
- Basic Mathematical Concepts – What is entropy? What is maximum entropy? Other relative concepts.
- The problems we are interested
- The problems with the problems
- A new conjecture
- Numerical experiments

A search under keyword, wavelet entropy, at google.com on July 14 returned about 1,370,000 results (0.09 seconds). The following is search results at citeseer

Keywords	Number of published papers (since 1948-
Entropy	34,406
Shannon Entropy	4423
ApEn	20,365
Sample Entropy	16,109
Mutual Entropy	6,472
Entropy and Thermodynamics	2,862
Residue entropy	3,440
Renyi entropy	573
Fuzzy entropy	1664
Relative entropy	29,863







Basic Mathematical Concepts

- **Definition 1**. The entropy of a discrete random variable X is defined by

$$H(X) = -\sum_{x \in X} p(x) \log p(x)$$

- **Remark**. (1) Entropy is always positive. (2) Entropy measures the uncertainty inherent in the distribution of a random variable.

Basic Mathematical Concepts - continue

- **Definition 2** The joint entropy H(X,Y) of a pair of discrete random variables with a joint distribution p(x,y) is defined as

$$H(X,Y) = -\sum_{x \in X} \sum_{y \in Y} p(x,y) \log p(x,y)$$

- The chain rule for joint entropy

$$H(X|Y) = H(X) + H(Y|X)$$

Basic Mathematical Concepts - continue

- **Definition 3**. The conditional entropy H(X|Y) is defined as

$$H(X|Y) = -\sum_{x \in X} p(x) H(Y|X = x)$$

- Remark. The conditional entropy of Y conditional on X refers to the average entropy of Y conditional on the value of X, averaged over all possible values of X.

Basic Mathematical Concepts - continue

Definition 4. The relative entropy between two probability distributions is given by

$$D(p||q) = \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)}$$

Remarks on relative entropy

1. The relative entropy is a “measure” of something like the distance between two different probability distributions!
2. Comparing two probability distributions using relative entropy.
3. The relative entropy is not symmetric!

Basic Mathematical Concepts - continue

Definition 5. The mutual information I(X,Y) measures how much (on average) the realization of random variable Y tells us about the realization of X, i. e., how by how much entropy of X is reduced if we know the realization of Y.

$$I(X;Y) = H(X) - H(X|Y)$$

Remark. Mutual information is of the most important information concepts for biology. For example,

Example

The mutual information between a cue and the environment indicates us how much on average the cue tells us about the environment. The mutual information between a spike train and a sensor input tells us how much the spike train tells us abut the sensory input. If the cue is perfectly informative - if it tells us everything between cue and environment if simply the entropy o the environment:

$$I(X;Y) = H(X) - H(X|Y) = H(X) - H(X|X) = H(X)$$

Remarks. (1) The mutual information between a random variable and itself is simply its entropy.
(2) Mutual information is symmetric; X tells exactly as much about Y as Y tells us about X.

The problems

Problem 1. Find an effective model of monitoring network volatility and uncertainty using the concept of entropy.

Problem 2. Quantify the volatility and uncertainty. Namely, numerically and timely detect and identify network volatility and uncertainty.

Problem 3. Build a multiscale self-adaptive model for forecasting based available data.

Progress of the project

- Current status of the project – solving problem 1.
- Conjecture and theme.

  We believe there is not a single parameter, for example, the following parameters are the most popular ones:
  (1) Lyapunov exponent
  (2) entropy – uncertainty
  (3) Hurst – long range memory (global property)
  (4) Holder- continuity (local property)

Our goal it to build a multiparametric model

Toward the goal of multiparametric model

Discover relations among parameters - Numerical examples to illustrate the relations between approximate entropy (ApEn) and parameter Hurst.

Before we show the numerical results, we recall the

- definition for Hurst exponent or parameter and
- the concept ApEn; and
- Why we use ApEn (not entropy)

About Hurst exponent and fractional Brownian motion

- Hurst exponent H is named the Hurst parameter by Mandelbrot in honor of both Harold Edwin Hurst and Ludwig Otto Holder [3]
- the fractional Brownian motion (fBm) has become widely popular in a theoretical context as
- compute Hurst exponents. The numerical experiments in this project are based on the built in
- programs in Matlab
- Fractional Brownian motion

Approximate Entropy (ApEn)

Suppose there is a sequence $S_N = (S(1), S(2),\ldots S(N))$. The variables *m* used to represent the length of the subsequence (patterns) to be compared. Let $p_m(i)$ denote subsequence starting at S(i) of length m. Further, the set of all subsequences of length m is $P_m = \{ p_m(1), p_m(2), p_m(3), \ldots , p_m(N – m + 1)\}$.

(1) r - measure the similarity of two subsequences. In particular, the subsequences pm(i) and pm(j) are similar if, $|S(i + k) - S(j + k)| < r, \square 0 \le k < m$

(2) $C_{im}(r) = n_{im}(r)/ (N – m + 1)$ can be thought of as the percentage of patterns in P_m that are similar to $p_m(i)$

Define: ApEn(SN, m, r) = ln(Cm(r)/ Cm + 1(r)).

Approximate entropy (ApEn)

A technique that was developed for determining the predictability of a sequence, such as a time series.

ApEn

Randomness

The more the repeated patterns the time series has, the lower its ApEn would be.

The less the patterns the time series has, the larger the ApEn would be.

Numerical simulations

Notes:

1. Fractional Brownian motion is simulated using matlab
2. A matlab program (modified program from matlabl central) used to compute the ApEn

Pseudo-code for ApEn

(A matlab program written by Avinash Parnandi and is available at MatlabCentral is based on these codes)

INPUT

n - the length of the pattern (template)

r - the similarity measure (matching tolerance),

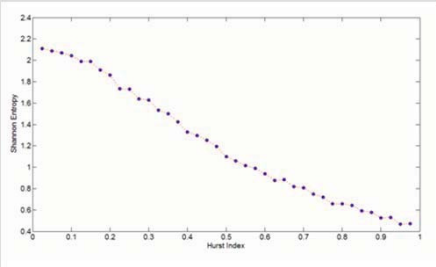
a - the input data vector.

OUTPUT apen is the approximate entropy of the sequence given the input parameters.

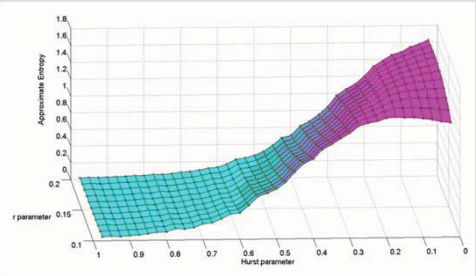
continue

Relation between Shannon entropy and Hurst exponent

(Integrate Monte Carlo and bootstrap technique)



Relation of the parameter r of ApEn and Hurst parameter



Summary and future work

Summary. In the talk, we presented some initial study to reveal relations among the most important parameters (illustreated by using ApEn and Hurst). These efforts are made to solve the problems 1.

Future work – Quantify the volatility and uncertainty. Namely, numerically and timely detect and identify network volatility and uncertainty.

High goal. Build a multiscale self-adaptive model for monitoring and forecasting based available data.

References

[1] L. Arnold And V. Wihstutz, Lyapunov exponents: A survey, in Lyapunov Exponents, L. Arnold and V. Wihstutz, eds., Lecture Notes in Mathematics 1186, Springer-Verlag, Berlin, New York,
Heidelberg, 1986, pp. 1-26.

[2] Ben Saïda, Ahmed, Using the Lyapunov Exponent as a Practical Test for Noisy Chaos (March 10, 2007). Available at SSRN: http://ssrn.com/abstract=970074

[3] B. Mandelbrot and J. Van Ness. Fractional brownian motions, fractional noises and applications. SIAM Rev., 10:422–437, 1968.

[4] J. Beran. *Statistics for long-memory processes*. Chapman & Hall/CRC, 1994.

OHIO UNIVERSITY

APPENDIX E

Digital Signal Representation with Slepian Series

Digital signal representation with Slepian series

Xiaoping A. Shen

http://www.ohio.edu/people/shenx

Department of Mathematics
Ohio University
Athens, Ohio 45701

Sponsored by ONR and AFOSR FA8650-08-D-6801

OHIO

Outline

- Introduction to Energy concentration problem and its solution
  -Slepian functions
- An hierarchical system system
- Properties of the system.
- Numerical examples.
- References

OHIO

The Paley-Wiener theorem

Assume that $h \in L^2(R)$ is analytic. If there are positive constants K and σ so that for all $z \in C$,

$$|h(z)| \le K \exp(\sigma |z|), \qquad (*)$$

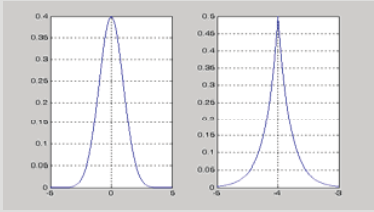
then the Fourier transform of h, $\hat{h} \in L^2(-\sigma,\sigma)$ and

$$h(z) = \int_{-\sigma}^{\sigma} \hat{h}(x)\exp(-2\pi i x z)dx \qquad (**)$$

A function satisfies (*) is said to be exponential type.
A function satisfies (**) is said to be σ – bandlimited
$B_\sigma^2 =$ collection of all analytic function of exponential type (or σ bandlimited functions) in L^2

OHIO

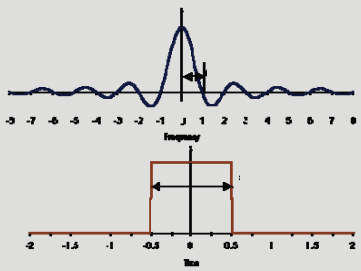
Examples - Gaussian kernel and bilateral kernel



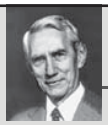
$$k_1(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} \qquad k_2(t) = \frac{1}{2} e^{-\frac{|t|}{2}}$$

OHIO

The sinc function



OHIO



Shannon Sampling Theorem

If f is π bandlimited (analytic of exponential type), then the formula:

$$f(x) = \sum_{k\in Z} f(k) \frac{\sin \pi(x-k)}{\pi(x-k)}$$

is exact.

OHIO

The restriction of the Heisenberg uncertainty principle

(Harmonic analysis) A non-trivial function can not have time limiting and frequency limiting properties simultaneously.

$2\sigma\tau$ – Theory

When $\sigma\tau$ is large, the space of signals *approximately*, of duration τ and bandwidth σ has dimension $2\sigma\tau$ *approximately*!

OHIO

The energy concentration problem (Slepian, Landau, Pollak at Bell Lab, 1960s)

Define

$$\alpha^2(\tau; f) \equiv \frac{\int_{-\tau}^{\tau} |f(t)|^2\, dt}{\int_{-\infty}^{\infty} |f(t)|^2\, dt}, \quad \text{and} \quad \beta^2(\sigma; f) \equiv \frac{\int_{-\sigma}^{\sigma} |\hat{f}(\omega)|^2\, d\omega}{\int_{-\infty}^{\infty} |\hat{f}(\omega)|^2\, d\omega}$$

Time concentration index — frequency concentration index

$0 \le \alpha^2(\tau),\ \beta^2(\sigma) \le 1$

OHIO

The energy concentration problem

Consider the function space B_σ^2 (the Paley - Wiener space)

Q1. How large $\alpha^2(f;\tau)$ can be ?

Q2. Which σ bandlimited function possesses the largest time concentration index?

Mathematically, find the solution to the optimization problem:

$$\max_{f \in B_\sigma^2} \alpha^2(\tau; f)$$

OHIO

Solution of the concentration problem -- Slepian's lucky accident

The energy concentration problem can be rewritten as the eigenvalue problem of an integral operator,

$$S_{\sigma,\tau}(\phi)(t) = \lambda\phi(t),$$

where $S_{\sigma,\tau}(\phi)(t) := \int_{-\tau}^{\tau} \phi(x) \frac{\sin c(t-x)}{t-x} dx, \quad c = \sigma\tau$

The integral operator is commute with a second order differential operator

$$S_{\sigma,\tau} P_{\sigma,\tau} = P_{\sigma,\tau} S_{\sigma,\tau}$$

where $P_{\sigma,\tau}(\phi)(t) := (1-t^2)\frac{d^2\phi}{dt^2} - 2t\frac{d\phi}{dt} - \sigma t^2 \phi$

OHIO

Prolate functions $\{\phi_{n,\sigma,\tau}(t)\}_{n=0}^{\infty}$ are the solution to the concentration problem.

Their associated eigenvalues are the corresponding time concentration index.

$$\phi_{0,\sigma,\tau}\ \ \phi_{1,\sigma,\tau}\ \ \phi_{2,\sigma,\tau}\ \cdots\ \phi_{n,\sigma,\tau}\cdots$$

$$\downarrow\quad\downarrow\quad\downarrow\quad\cdots\quad\downarrow\ \cdots$$

$$\lambda_0 > \lambda_1 > \lambda_2 > \ldots > \lambda_n > \ldots > 0$$

$$\| $$

$$\alpha(\phi_{0,\sigma,\tau}; \tau)$$

OHIO

The solution of the concentration problem

More specifically,

$\phi_{0,\sigma,\tau}(t)$ has maximum time concentration index, its associated eigenvalue is the possible largest concentration index, that is

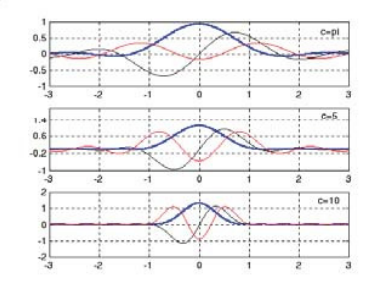
$$\lambda_{0,\sigma,\tau} = \alpha(\tau; \phi_{0,\sigma,\tau}) = \max_{f \in B_\sigma^2} \{\alpha^2(\tau, f)$$

$\phi_{k,\sigma,\tau}(t)$ is the function with maximum time concentration index among all the functions that orthogonal to $\phi_{k-1,\sigma,\tau}(t)$ with its associated eigenvalue

$$\lambda_{k,\sigma,\tau} = \alpha(\tau; \phi_{k,\sigma,\tau})$$

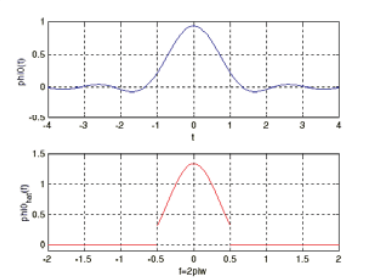
OHIO

Prolate (Slepian) functions $\{\phi_{n,\sigma,1}\}$



OHIO

A Fourier pair of pi bandlimited PSWF



OHIO

Properties of PSWFs

1. Double orthogonality

$$\int_{-\infty}^{\infty} \phi_{n,\sigma,\tau,}(t)\phi_{k,\sigma,\tau,}(t)dt = \delta(n-k)$$
$$\int_{-\tau}^{\tau} \phi_{n,\sigma,\tau,}(t)\phi_{k,\sigma,\tau,}(t)dt = \lambda_n \delta(n-k)$$

2. Discrete orthogonality

$$f(t) = \sum_{n=0}^{\infty}\{\sum_{k=-\infty}^{\infty} \varphi_n(k)f(k)\}\varphi_n(t).$$

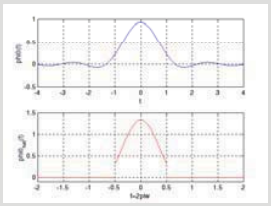
3. $\{\phi_{n,\sigma,\tau}\}$ **is an orthogonal basis of** B_σ^2, and $L^2[-\tau,\tau]$

OHIO

Properties of PSWFs -continue

3. Duality of Fourier Transform.

$$\hat{\phi}_{n,\sigma,\tau,}(\omega) = \sqrt{\frac{2\pi\tau}{\sigma\lambda_{n,\sigma,\tau}}}\phi_{n,\sigma,\tau,}(\frac{\tau\omega}{\sigma})\chi_\sigma(\omega)$$



A Fourier pair of pi bandlimited Slepian function

OHIO

Properties of PSWFs - continue

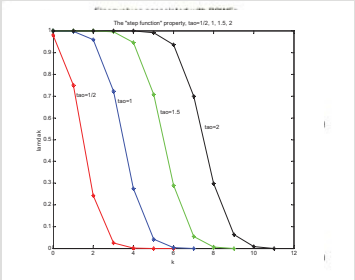
4. "Step function" property of the associated eigenvalues

For each $0 < \delta,\ \varepsilon < 1$, there exits γ, such that
for $\sigma\tau > \gamma$, $\lambda_k < \varepsilon$ when $k \geq 2(1+\delta)\sigma\tau$, $\lambda_k > 1-\varepsilon$
when $k \leq 2(1-\delta)\sigma\tau$,

$$\underbrace{\lambda_0 \quad \lambda_1 \quad .... \quad \lambda_{[2\sigma\tau]}}_{\text{Close to 1}} \quad \underbrace{\lambda_{[2\sigma\tau]+1} \quad .....}_{\text{close to 0}}$$

OHIO

Step function property



OHIO

Citation Slepian's paper

Courtesy of ResearchIndex

OHIO

Motivation

Recall spectral approximations

- For periodic problems: Fourier approximation
- For non-periodic problems: Chebyshev, Legendre approximation

For analytic functions: Both approximation assume exponential accuracy

For piecewise analytic functions,

O(N) globally

O(1) at near discontinuity – known as Gibbs phenomenon

OHIO

The problem we wish to solve:

- Assume a piecewise band-limited signal f is given by its equally spaced samples

$$\left\{ f(-1+\frac{1}{2^{M-1}}k) \right\}_{k=0}^{2^M}$$

The data may be damaged by a uniformly distributed noise. That is, what we have is

$$\left\{ f(-1+\frac{k}{2^{M-1}}) + \varepsilon_k \right\}_{k=0}^{2^M}$$

where $\{\varepsilon_k\}$ is uniformly distributed random error.

- We wish to develop a robotic algorithm to represent the given signal in an effective way, $x, y = 0,...,2^M$

OHIO

Piecewise band-limited functions

Definition (Piecewise band-limited function on [a, b])

Let $f \in L^2[a,b]$ If there exists a partition $[a,b] = \cup_{i=0}^{n-1}[a_i, a_{i+1}]$ such that

$$f(x) = f_i(x), \ x \in [a_i, a_{i+1}), i = 0,...,n-1$$

where the analytical extension F_i, associated with f_i, belong to a Paley-Wiener space $B_{\sigma_i}^2, i = 0,...,n-1$

Remark Notice that, we have

$$f_i \in B_\sigma^2, \sigma = \max_{0 \le i \le n-1} \{\sigma_i\}$$

OHIO

The hierarchical system

Definition (The hierarchical basis on [-1,1]) Consider the decomposition

$$[-1,1] = \bigcup_{j=1}^{k} [-1+\frac{k-j}{2^m}, -1+\frac{k+1-j}{2^m}), 1 \le k \le 2^m$$

We define

$$\phi_{m,n,k}(x;c) = \begin{cases} (-1)^{n(k+1)} 2^{\frac{m}{2}} \phi_n(2^m x - 2k + 2^m + 1; c), \\ -1+\frac{k-1}{2^{m-1}} \le x < -1+\frac{k}{2^{m-1}}, m = 0,1,..., n = 0,1,..., k = 1,2, ..., 2^m \end{cases}$$

Here

m : the level index

n : the label of nth eigenfunct ions

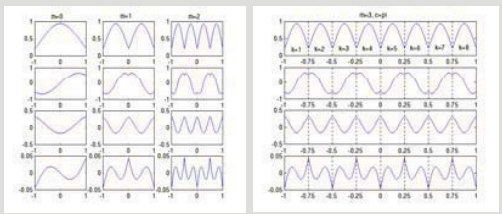
k : the piecewise index

c : the bandwidth of the original Slepian functions $\varphi_{n,\sigma,1}(t)$

$\phi_{n,\sigma,\tau}(x;c) = \varphi_{n,\sigma,1}(x)\chi_{[-1,1]}(x)$

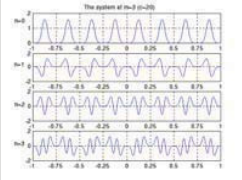
OHIO

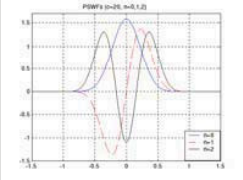
Examples of the system



OHIO

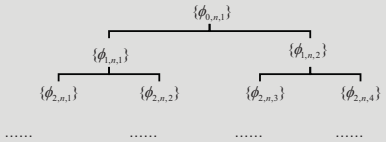
Another example





OHIO

Structure of the binary system



OHIO

The hierarchical system

The function defined as

$$\Phi_{m,n}(x;c) = \begin{cases} \phi_{m,n,k}(x;c), x \in [-1+\frac{k-1}{2^{m-1}}, -1+\frac{k}{2^{m-1}}), k = 1,...,2^m \\ 0 \end{cases}$$

is a piecewise band-limited function with each piece has same approximately bandwidth $2^m c$.

Denote

$$V_m^n = span\{\phi_{m,l,k}(x;c), l = 0,1,2,...,n, k = 1,2,...,2^m n\}$$

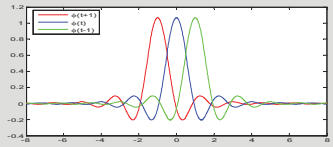
The space V_m^n consists of piecewise band-limited functions with dimension $2^m n$

OHIO

Raised cosine function

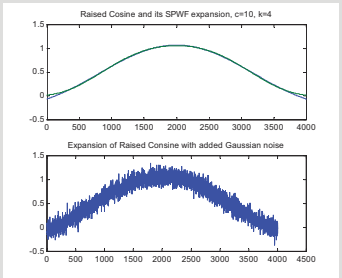
$$\phi(x) = \frac{\sin[\pi(1-\beta)x] + 4\beta x\cos[\pi(1+\beta)x]}{\pi x[1-(4\beta x)^2]}$$

$$\psi\left(t+\frac{1}{2}\right) = \frac{\sin[\pi(1+\beta)x] - 4\beta x\cos[\pi(1-\beta)x]}{\pi x[(4\beta x)^2-1]} - \frac{\sin[2\pi(1-\beta)x] + 8\beta x\cos[2\pi(1-\beta)x]}{\pi x[(8\beta x)^2-1]}$$



OHIO

Example. A signal with Gaussion noise



OHIO

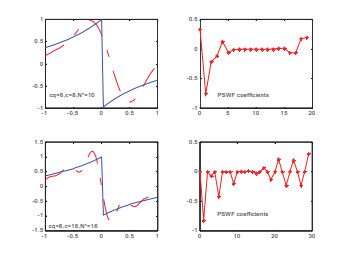
The problem with discontinuity

Thus, in approximating a piece of a non analytic function by band-limited functions, we exchange goodness of fit in analytic intervals with wildness of behavior outside this interval

– Slepian

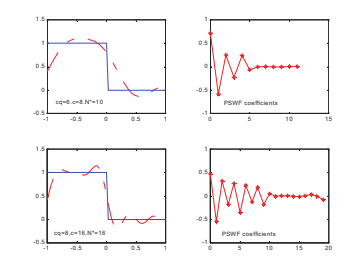
OHIO

Orthogonal expansion using Slepian functions – A piecewise band-limited function



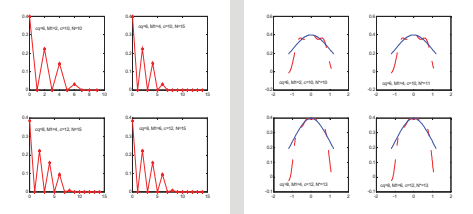
OHIO

Orthogonal expansion using Slepian functions – Heaviside function



OHIO

Orthogonal expansion using Slepian functions – Gaussian kernel $G(t)=\frac{1}{\sqrt{2\pi}}e^{\frac{-t^2}{2}}$



OHIO

Concluding remarks

- In this research, we are looking for a better representation (or model) for discretized piecewise band-limited functions with attempting to recover this type of signal via their spectral expansion coefficients effectively.
- A hierarchical system based on Slepian functions are introduced.
- The system has some peculiar properties which are useful in signal analysis for a particular type of signals.
- The system shows promising potential in certain applications, such as noised signal representation, segmentation, detection, etc.

OHIO

References

1. Slepian, D.; Pollak, H. O., (1961), Prolate spheroidal wave functions, Fourier analysis and uncertainty, I, *The Bell System Technical Journal 40, 43-64.*
2. Landau, H. J. ; Pollak, H. O. (1961), Prolate spheroidal wave functions, Fourier analysis and uncertainty, II, *The Bell System Technical Journal 40, 65-84.*
3. Landau, H. J. ; Pollak, H. O. (1962), Prolate spheroidal wave functions, Fourier analysis and uncertainty, III, *The Bell System Technical Journal 41, 1295-1336.*
4. Slepian, D. (1964), Prolate spheroidal wave functions, Fourier analysis and uncertainty, IV, *The Bell System Technical Journal 43, 3009-3058.*
5. Slepian, D. (1983), Some comments on Fourier analysis, uncertainty, and modeling, SIAM Review 25, 379-393.
6. Courant, R.; Hilbert, D., (1953) Methods of mathematical physics, New York, Interscience Publishers, 1953-62.

1. Xiao, H; Rokhlin, V.; Yarvin, N. (2000), Prolate spheroidal wave functions, quadrature and interpolation, *Inverse Problems 17 (2001), no. 4, 805--838.*
2. G. Beylkin and L. Monzon, On Generalized Gaussian Quadratures for Exponentials and their Applications, *Preprint, Dec. 2000.*

1. Walter, G. G; Shen, X., (2001), Sampling with Prolate Spheroidal Functions, submitted to *the Journal of Sampling Theory in Signal and Image Processing.*
2. Walter, G. G; Shen, X., (2002), Wavelets based on Prolate Spheroidal Functions, submitted to the *Journal of Fourier analysis and applications*.
3. Shen, X.; Saito, N., (2002) A hierarchical system based on Slepian functions, ONR report.

APPENDIX F

Complexity Analysis and Information Visualization

Complexity Analysis and Information Visualization

March 5, 2010

Air Force Research Lab
Wright-Patterson AFB, Ohio

Paul R. Havig, PhD,
and Kate A. Farris
711th Human Performance Wing
Air Force Research Laboratory

Overview

- Air Force Research Lab
- Complex Networks Overview
- Information Visualization Overview
- Where we are now. Where we are headed.
- Faculty and Internship Opportunities
- Summary
- Questions

Air Force Research Laboratory

•Technology Directorates

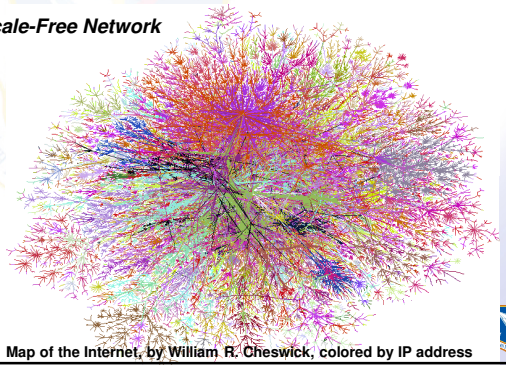
- Air Vehicles
- Directed Energy
- *711 Human Performance Wing*
- Information
- Materials and Manufacturing
- Munitions
- Propulsion
- Sensors
- Space Vehicles
- Air Force Office of Scientific Research

• 6 Locations

- Washington DC
- Mesa, AZ
- San Antonio, TX
- Rome, NY
- *Dayton, OH*
- Albuquerque, NM

Complex Networks

Scale-Free Network



Map of the Internet, by William R. Cheswick, colored by IP address

Complex Networks

• "More studies that assess the effectiveness of such [network] approximations – and provide concrete, empirically validated guidelines for practice within particular problem domains – would be a welcome addition to the literature."

– Carter T. Butts, *Revisiting the Foundations of Network Analysis*, Science, Vol 325, 24 July 2009

Objective

• Objective: To develop a concrete, empirically validated "measuring stick" that can better asses the performance to vulnerability tradeoff of a computer network.

• Four fundamental assumptions:

1. Assess what it is that is interacting
2. Qualify the nature of the interaction
3. Quantify the time scale on which that interaction takes place
4. Is the network static or dynamic?

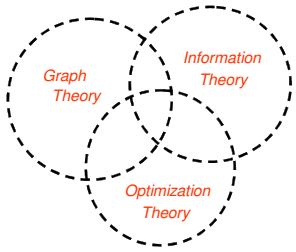
Underlying Assumptions

a. Exactly what is interacting?
   a. Nodes and Links
   b. Node = Webpage
   c. Link = Transmission Between Websites (accomplished via wires, cables and wireless access)
b. Nature of interaction?
   a. Research, E-mail, Blogging, etc.
c. Time scale which interaction occurs?
   a. Milliseconds
d. Static or Dynamic?
   a. Dynamic Network, constantly changing

7

Fusing Three Fields

• Characterize network performance and vulnerability through information, optimization and graph theories.

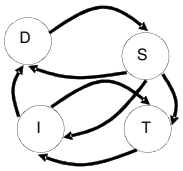


Combining Three Major Disciplines

8

Definition of graphs (networks)

• Network (graph) is a set of nodes connected by edges
• Nodes: Vertices D, I, S and T
• Edges: ID, DS, SD, IT, SI, ST and TI



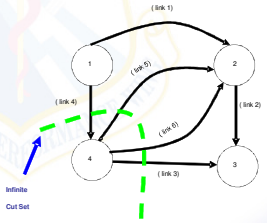
Simple Combat Network

J. Cares, *Distributed Network Operations: The Foundations of Network Centric Warfare*

9

Graph Theory

• Basic Graph Theory Concepts
• Digraph (Directed Graph)



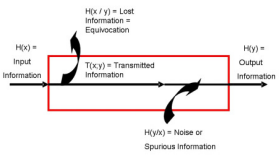
Network with Cut Set and Flow/Capacities Measures

Visual rendering designed by D. W. Repperger, 1942-2010

10

Information Theory

• Information Channel:



Basic Elements of the Information Channel

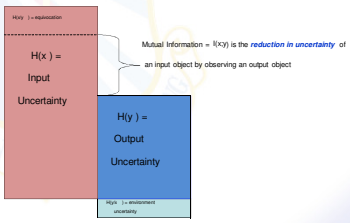
T. B. Sheridan and W. R. Ferrell, *Man-Machine Systems: Information, Control, and Decision Models of Human Performance*, The MIT press, Cambridge, Massachusetts, 1981.

• Variables:

1. *H(x) – input uncertainty*
2. *H(y) – output uncertainty*
3. *H(x/y) – equivocation lost to environment*
4. *H(y/x) – spurious uncertainty*
5. *I(x;y) – mutual information transmission*

11

Mutual Information

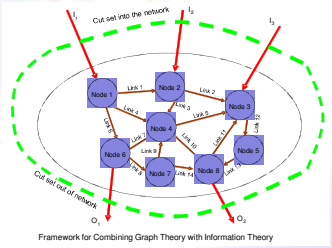


Distinguishing Mutual Information I(x;y) from the Uncertainty Variables H(.)

12

Final Picture

- Converting a Complex Network into a framework that is possible to use information-theoretic methods to calculate mean flows that occur within an ellipse.



Framework for Combining Graph Theory with Information Theory

Visual rendering designed by D. W. Repperger, 1942-2010

13

Where we are now. Where we're headed.

- Where we are now:
  - Objective: To develop a concrete, empirically validated "measuring stick" that can better asses the performance to vulnerability tradeoff of a computer network.
  - Theoretical framework has been developed
- Where we are headed:
  - Run JAVA-based computer simulations
  - Compare results of simulations against theoretical framework
  - Re-test simulations if necessary

14

Introduction

- Main points
  - 3D (metrics)
  - Interaction (tangible user interfaces)
  - Information Visualization (new directions)

15

3D Metrics

- Why metrics?
  - Not apples to oranges
- A taxonomy of metrics
  - Subjective
  - Subjective-Objective (Performance)
  - Objective

16

3D Metrics

- Subjective (currently running study)

1	2	3	4	5
Standard Better		No difference		New display superior

Using the scale above rate the new display for the following questions:

1) Which display looks sharper?
2) Which display has brighter colors?
3) Which display is more comfortable to view?
4) Which display do you prefer?

17

3D Metrics

- Performance
  - Why buy these expensive "toys"?
  - When to use? (Should we use?)
    - Issues (e.g., health)
    - Why not just use a PS3?
      - Is 2.5D good enough?
  - Can we make a style guide?
    - AFOSR research
    - Killer apps?

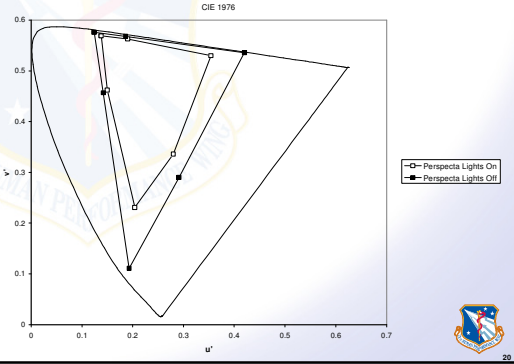
18

3D Metrics

- Objective metrics
  - System
    - Find it on a spec sheet
  - Measurable
    - Find/verify in the laboratory

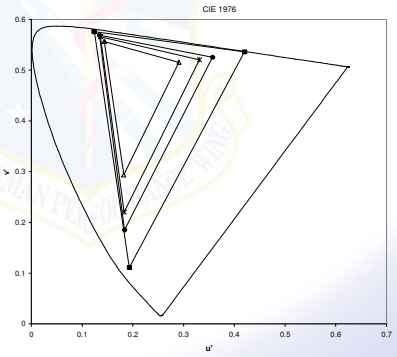
19

3D Metrics



20

3D Metrics



21

Tangible User Interfaces

22

Tangible User Interfaces

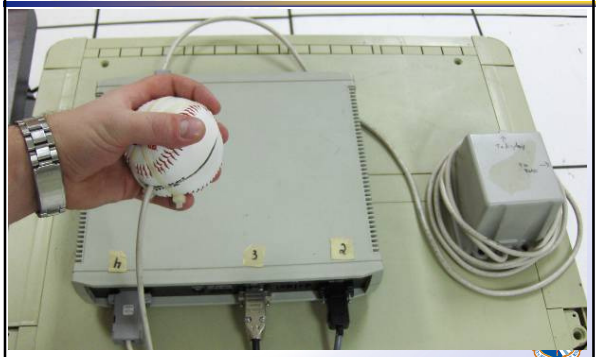
- Ok for a 2D environment (Word, Excel, etc.)
- Not for a 3D (nor 2.5D) (Gaming, CAD, etc.)
  - 2D ≠ 3D!
- Tangible User Interfaces (TUIs)
  - 6 degrees of freedom (DOF)
  - Spatial versus desktop TUIs
- Is the keyboard and mouse *that* bad?

23

TUIs

24

TUIs



25

TUI Comparisions

TUI	Pros	Cons
Keyboard and Mouse	1. Years of training with mouse to control applications. 2. For 2D tasks it is great. 3. Little to no additional training required.	1. For 3D tasks involving 6D motion, additional keyboard keys needed. 2. Requires two hands. 3. Basic mouse is only 2 DOF. 4. May not be intuitive to new users.
Sandio Tech 3D Mouse	1. Inherits the things mice are good for. 2. The extra controls take only 5-10 minutes to understand. 3. 6 DOF with one hand.	1. It was easy to rotate or translate the reverse directions, so some confusability between directions. 2. Unintuitive. 3. Cannot vary speed of motion.
3D Connections	1. Only a single knob to manipulate. 2. Once learned, is nice for quick zooming in/out or moving viewpoint from one location to another. 3. 6 DOF with one hand.	1. Troubles with cross talk between pushing left / right and tilting left / right. 2. Getting "lost" or dis-oriented. 3. Not entirely intuitive. 4. Co-axial rotation and translation can bleed without a proper dead-zone and careful device manipulation.
In-House TUI	1. Very intuitive mapping 2. Rotation still a problem; difficult to return to "home"; might want a "standard axis" around which to rotate that can be toggled on/off (i.e., North/South) 3. Tangible, real-world motion capture	1. Only partially mapped inputs, so cannot control everything like the other input devices. 2. Utilizes a magnetic field to function. Device is sensitive to fields created by ferrous metals and other nearby electronic devices. 3. Short range - Tracker must stay within 1 meter of emitter to get accurate readings.

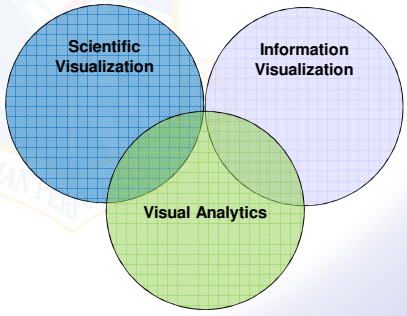
26

Conclusions and Future Directions

- **What we have**
  - **3D does help**
    - **Where is the killer app?**
  - **TUIs are needed**
    - **But…keyboards are quite helpful!**
- **Where do we want to go now**
  - **Deep dives and blue skies!**
  - **Laboratory upgrade**

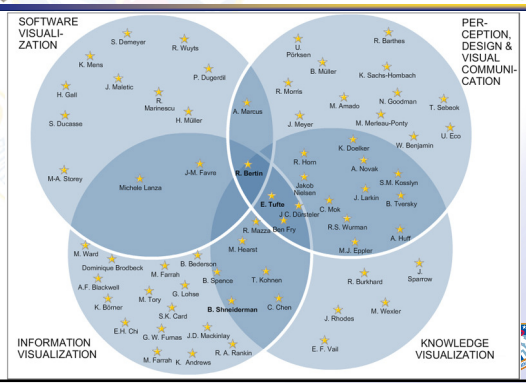
27

Relationship Between Fields



28

Key Players



29

Faculty and Internship Opportunities

- **Summer Faculty Fellowship Program (SFFP)**
- **Student Temporary Employment Program (STEP)**

30

Summary

- Basic research in complex network dynamics
- Building research in information visualization
- Connecting math to pictures

31

Questions???

- Contact Info: Paul.Havig@wpafb.af.mil
  Katheryn.Farris@wpafb.af.mil

32

APPENDIX G

Visualizing Complexity of Sunspot Cycles Via Wavelet Transform

Abstract. Dark spots, some as large as 50,000 miles in diameter, move across the surface of the sun, contracting and expanding as they go. These strange are known as sunspots [1] (see **Figure 1**). Sunspots generate magnetic regions on the Sun with strengths thousands of times stronger than the Earth's physically damage satellites and pose a threat to astronauts. This study employed wavelet analysis, a powerful mathematical tool, to reveal and visualize the data. **The long term goal of this research is to discover the underlying mathematical theory behind the data analysis and visualization associat integrated complex networks to improve human effectiveness and performance capability via effective information visualization**.

1. Background.

Wavelet analysis is a new mathematical tool which can be used to extract information from data, such as audio signals and images. As opposed to Fourier bases, wavelet bases are **localized** with **multiscale**. The fast Wavelet Transform (FWT) is faster than the fast Fourier transform (FFT). Wavelet filters can be finite impulse response (FIR) or infinite impulse response (IIR). Mathematically, just like a tuning fork physically resonates with a signal of its specific tuning frequency, the wavelet filters will resonate if the unknown signal contains information of similar frequency (see [2] and the references there in). **Figure 2** shows three different types of wavelet families – the first two are FIR and the last one is IIR.

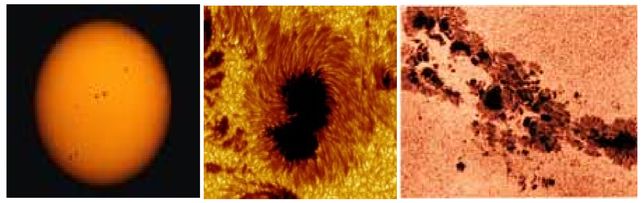


Figure 1. Sunspots and the region around sunspots (*Courtesy National Solar Observatory).*

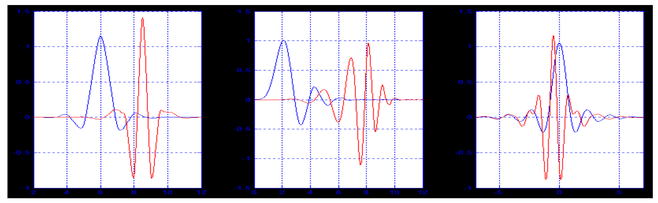


Figure 2. *– From left: father and mother of Coiflets (coif3), Daubechies (db4), and the Meyer wavelet*

2. Initial Study.

2.1 Data description. The International Sunspot Number (ISSN) is compiled by the Solar Influences Data Analysis Center in Belgium. NOAA sunspot number is compiled by the US National Oceanic and Atmospheric Administration. The data are the monthly ISSN averages and standard deviation derived from ISSN from January 1749 to June 2010 (a total of 3138 months).

2.2 Basic statistical properties. Basic statistical properties of the data are shown in **Figure 3**. The descriptive statistics show that the average sunspots number is 51.9733 per month with standard deviation (std) 44.3406. The largest sunspot number, 253, occurred in October 1957. The 75th percentile of the sunspot number is 76.6. Recently, fro 50.4153 with std=46.0629. Since 2000, the largest sunspot number, 170, occu histogram (middle panel) suggested, the normality test indicates (right panel) that t normal distribution. The statistical findings and observations lead to the following

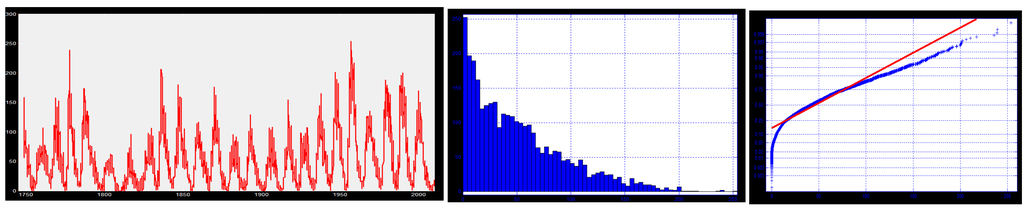


Figure 3. Monthly average sunspot numbers. 1749 – 2010 (left). Frequency histogram (middle) and normality test.

Conjectures. 1. Sunspot number is a random variable following a long tailed distribution. 2. Sunspot number has fractal structure with self similarity.

A parametric test indicates that the Sunspot number possesses the Hurst number 0.27 approximately (below normal human Heartbeat interval EEG data). We then apply **wavelet decomposition** (as **multiple lenses** using Coiflets) to see the signal at different resolutions and better understand the data structure (results shown in **Figure 4**). Motivated by the initial study, the following method is designed to provide visual evidences for the conjectures.

3. The Method. **Step 1**. Develop reference signals. The selected reference signals are, the famous fractal curve - von Koch curve (Hurst exponent H=1.7381 and the fractional dimension D=1.2619) and an electric consumption signal which does not have fractal structure (see **Figure 5**). **Step 2**. Perform continuous wavelet transform (CWT) with the 1-D graphical tool in Matlab Wavelet Toolbox©. The wavelet coefficients of the Koch curve show a repeating pattern on many scales (see **figure 4**, the vertical axis is scale measure) while the coefficients of the other signal does not. **Step 3**. Perform CWT of sunspots as described for the reference signals. Results are shown in Figure 6, from left top corner, the CWT based on Daubechies wavelet (N=8), the CWT based on Coeiflets and zoom in pictures. The graphical results suggest that the conjecture holds for the current sunspot data. It is worth mentioning that Wavelet decomposition is very well adapted to the study of the fractal properties of signals and images. In particular, when the characteristics of a fractal evolve with time and become local (multifractal), wavelets are especially suitable tool for practical analysis and generation. Readers are referred to reference [3] for more details of the study.

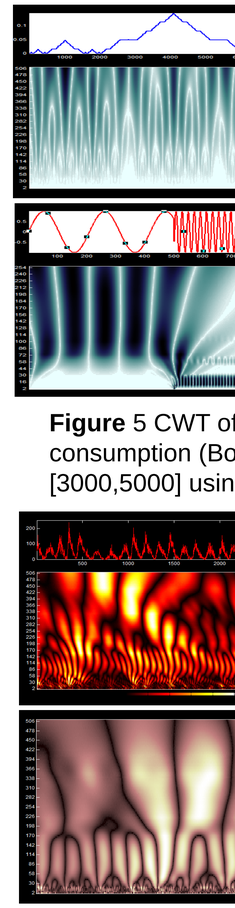


Figure 5 CWT of consumption (Bo [3000,5000] usin

Figu

References

[1] National Aeronautics and Space Administration, http://solarscience.msfc.nasa.gov, date visited 08/04/2010.
[2] G. G. Walter and X.Shen. *Wavelets and Other Orthogonal Systems*, 2nd Ed., Studies in Advanced Mathematics, xx+370 pp, Chapman 5848-8227-1.
[3] X. Shen. Visualizing information of engineered complex systems, working paper, 2010.