Contributions of the

Symposium on Network Analysis in Natural Sciences and Engineering

part of

AISB'06: Adaptation in Artificial and Biological Systems

Bristol, UK
April 5-6, 2006

Edited by
Susanne Hoche
Jane Memmott
Nick Monk
Andreas Nürnberger

Symposium Chairs:

Susanne Hoche
Dept. of Computer Science
University of Bristol, UK

Jane Memmott
School of Biological Sciences
University of Bristol, UK

Nick Monk
Comp. Syst. Biology Group
University of Sheffield, UK

Andreas Nürnberger
Information Retrieval Group
Univ. Magdeburg, Germany

Symposium Sponsor:

EU Coordinated Action - http://www.nisis.de

Program Committee:

Michael Berthold, Germany
Hendrik Blockeel , Belgium
Christian Borgelt, Germany
Nick Britton, UK
Seth Bullock, UK
Chris Cannings, UK
Nello Cristianini, USA
Marcin Detyniecki, France
Peter Flach, UK
Susanne Hoche, UK
Tamas Horvath, Germany
Johannes Jaeger, UK
Dick James, UK
Konstantin Klemm, Germany
Edda Leopold, Germany
David Lusseau, UK
Jane Memmott, UK
Nick Monk, UK
Nicolas Labroche, France
Trevor Martin, UK
Andreas Nürnberger, Germany
Richard Tateson, UK
Gilles Venturini, France

The original proceedings volume of AISB'06 that includes the contributions of this symposium appeared as:

Adaptation in Artificial and Biological Systems, Proc. of AISB'06, Volume 3, Tim Kovacs and James Marshall (eds.), Society for the Study of Artificial Intelligence and the Simulation of Behaviour, 2006 (ISBN 1 902956 96 7)

Preface

Network analysis and modeling address a wide spectrum of techniques for studying domains consisting of individuals that are linked together into complex networks. Networks refer to artificial and natural systems like communication networks, social networks and biological networks. They constitute a very active area of research in a variety of scientific disciplines, including Physics, Biology, Artificial Intelligence and Mathematics. Both graph theory and techniques recently developed for the analysis of networks provide a substantial background for studying complex network structures and dynamics in artificial and biological systems. They allow us to answer questions in common to these networks like aspects of adaptability, error and attack tolerance, complexity, community structures, and propagation patterns.

One of the key features of natural networks is their ability to adapt to changing environments, maintaining an appropriate pattern of behaviour. Examples of such adaptive capacity span the whole range of natural networks, from gene-protein interaction networks within individual cells, through physiological systems, to ecosystems.

The aim of this symposium was to provide a forum to bring together researchers in biology, computer science and related disciplines in order to discover related mechanisms in natural and artificial networks and to initiate, combine and promote research in both fields.

April 2006, The Symposium Chairs

Contents

Network Analysis and Dynamic Conflict

Ulrik Brandes*

*Department of Computer and Information Science, University of Konstanz
Box D 67, 78457 Konstanz, Germany
Ulrik.Brandes@uni-konstanz.de

Abstract

After a general introduction to the emerging field of network analysis, we will focus on the analysis of group structure in dynamic networks. By defining groups based on similarity of neighborhoods rather than local density, nodes can be associated to roles representing structural positions in a network. A recently proposed relaxation of discrete role assignment allows for varying degrees of membership in such roles, and also points to dominant representatives within roles. The utility of this new approach is demonstrated on dynamic event data extracted from news reports on conflicts that took place in the Persian Gulf and on the Balkans.

Adaptation in Biological Networks: From the genome to ecosystems

Barbara Drossel*

*Institute of Solid State Physics, Technical University of Darmstadt
Hochschulstr. 6, D-64289 Darmstadt
barbara.drossel@physik.tu−darmstadt.de

Abstract

Networks are ubiquitous in biological systems. Two widely studied examples are regulatory genetic networks and foodwebs. Both types of networks show a high degree of adaptation: Genetic networks perform reliably, even though the individual elements are affected by noise. Foodwebs retain their complex structure in spite of highly nonlinear population dynamics and changes in the foodweb composition. This talk will discuss both types of systems from the perspective of a theoretician. After a general introduction, models for both systems will be presented that capture the features that are essential for tackling the question of robustness and stability. By discussing the structure and dynamics of these networks, features that convey stability are identified. Among these are stabilizing topological elements in genetic networks and adaptive foraging behavior in food webs.

Fault tolerance and network integrity measures: the case of computer-based systems

Peter Andras*
*School of Computing Science
University of Newcastle
peter.andras@ncl.ac.uk

Olusola Idowu*
*School of Computing Science
University of Newcastle
o.c.idowu@ncl.ac.uk

Panayiotis Periorellis*
*School of Computing Science
University of Newcastle
panayiotis.periorellis@ncl.ac.uk

Abstract

Fault tolerance is a key aspect of the dependability of complex computer-based systems. Fault tolerance may be difficult to measure directly in complex real world systems, and we propose here to measure it in terms of integrity preservation of the system under the assumption of a particular fault occurrence distribution. We measure the integrity preservation ability of the system by measuring the change of structural integrity of the graph representing the system while it is exposed to random node removal according to the assumed fault distribution. We show how to use such measures to measure the integrity reservation of computer-based systems and in this way indirectly their fault tolerance. We discuss the application of the proposed method in the context of a real world example, the Linux operating system. The results indicate that integrity preservation metrics can serve as an appropriate measure of fault tolerance of complex computer-based systems.

1 Introduction

The concept of fault tolerance of complex computer-based systems, and in particular of computers and software, emerged very early in the 1950s (Lee and Anderson, 1990). It was recognized that unexpected faults may emerge in computer-based systems, and that effective dealing with such faults it is critical for highly dependable systems. Fault tolerance is a key measure of the dependability of computer-based systems (Lee and Anderson, 1990; Laprie, 1992), dependability being defined as reliability, availability, safety, security, survivability, and maintainability of a system (Aviziensis et al, 2001).

Generally systems can be perceived as a set of units that are interconnected by their actions and behaviours (von Bertalanffy, 1973). Computer-based systems can be seen as systems with units which can be computer hardware, software, humans, and possibly a variety of other machines and human artefacts containing sensors and actuators. The interconnecting actions and behaviours of these units can take the form of data entry to the computers, data communications between hardware components, data interchange and processing by software components, and display or communication of data to actuators.

An interesting issue is how to measure the fault tolerance of a computer-based system. Systematic mathematical analysis of fault tolerance of models of computer-based systems started in 1960s (Lee and Anderson, 1990). Typically fault tolerance is evaluated by full probabilistic analysis of the system, by calculating measures such as mean time to failure and mean time to repair under the assumption of a fault occurrence scenario (e.g., identical and independent fault occurrence distribution for each system component) (Amari, 2000; Chang et al, 2004; Ou and Dugan, 2003; Scerrer and Steininger, 2003).

One stream of fault tolerance research is focused on the analysis of graphs that represent computer-based systems (Billinton and Jonnavithula, 1999; Bell, 2003; Cheng and He, 2004). These works assume a fault occurrence scenario in the graph (e.g., node failure or edge failure) and measure the probability of connectedness (Beichelt and Tittmane, 1991; Elperin et al, 1991) or of having flow capacity above a given limit (Chan et al, 1997; Kishimoto, 1997) of the graph as a proxy measure for the fault tolerance of the system represented by the graph. The main drawback of these methods is that they are very computationally intensive and in many cases they are restricted to a narrow range of particular graph topologies (Al-Sadi et al, 2002; Goerdt, 2001; Goerdt and Molloy, 2003).

An alternative way to analyse the robustness of systems is to use structural graph analysis methods that reveal vulnerable components and the sensitivity to structural damage of the system (Albert et al, 1999). These methods assess the integrity of the system and the change of integrity measures after structural damage to the system in terms of structural measures, such as diameter, average minimum path length or average clustering coefficient. The underlying theoretical assumption is that system structural integrity implies functional integrity of the system (Andrews and Beeson, 2003; Ferrandi et al, 2003). This is supported by practical examples, which show that structural integrity and functional integrity of systems are strongly correlated (Albert et al, 1999; Jeong et al, 2001). Consequently, the analysis of the structural integrity of the graph representing a system by appropriate structural measures can provide indicator measures of the functional integrity of the system.

We propose in this paper the use of structural graph analysis methods to measure the integrity of computer-based systems. We measure the likely structural damage as an approximation of likely functional damage due to the presence of faults. In this way we can assess the fault tolerance of the system by measuring the likely change of structural integrity of the system.

The rest of the paper is structured as follows. Section 2 discusses system integrity measures. In Section 3 we analyse the link between fault tolerance and integrity measures. Section 4 presents an example of the application of the proposed methodology to the assessment of fault tolerance of computer-based systems. Finally, in section 5 we draw some conclusions of the paper.

2 System integrity

Systems are sets of component units interconnected by their interactions (von Bertalanffy, 1973). Component units interact by their behaviour modifying the state of the units participating in such interactions. In a stronger sense we may consider systems as only those sets of interacting component units, in which the interactions between components depend primarily on earlier interactions between system components (Andras and Charlton, 2005). We should also point out that system components may also interact with other units, which are not part of the system. Such interactions constitute the system's interaction with its environment.

The integrity of a system can be defined in functional terms as the system's ability to perform the full range of system behaviours (Ferrandi et al, 2003). The system behaviours are possible patterns of behaviours of its component units (Lee and Anderson, 1990). Some of these behaviours may have an effect on the system's environment, while others may cause only a change of the internal behaviour of the system.

Measuring functional integrity directly may be difficult, as the full range of possible system behaviour may not be known (Ferrandi et al, 2003). A way to approximate the functional integrity of a system is to measure its structural integrity ((Andrews and Beeson, 2003). In practical cases of living cells (Jeong et al, 2001), nervous systems of animals (Scannell et al, 1995), and technological systems (Albert et al, 1999) it has been shown that their functional integrity correlates strongly with their structural integrity. Measuring structural integrity is much simpler than measuring functional integrity, in the sense that it requires only the measurement of the existence of components and interactions between components, disregarding the actual functional semantics of interactions and interaction patterns.

Structural integrity measures of systems are based on the measurement of the structural integrity of the underlying graph structure of the system, which is made of nodes representing system units, and edges or arcs representing undirected or directed interactions between system units. (We consider undirected graphs only in what follows).

Simple measures of structural integrity of graphs include the diameter, the average minimum path length and the average clustering coefficient of the graph. The diameter is defined as the largest of the minimal path lengths between nodes of the graph:

$$D(G) = \max\{l(i,j) \mid i,j \in V\}, G = \{V, E\} \quad (1)$$

where V is the set of nodes and E is the set of edges of the graph, and l(i,j) is the minimal length of a path between the nodes i and j. The average minimum path length is defined as the average length of minimal paths between all pairs of nodes of the graph that can be connected (i.e. infinite length shortest paths are ignored):

$$\mu = \frac{1}{|V|^2} \sum_{\substack{i,j \in V \\ l(i,j) < \infty}} l(i,j) \quad (2)$$

where we use the same notations as above. The clustering coefficient of a node is the proportion between the number of existing edges between the

neighbours of the node and the number of all possible edges:

$$c(i) = \frac{2 \cdot |\{(i,j) \,|\, (i,j) \in E\}|}{|\{j \,|\, (i,j) \in E\}| \cdot (|\{j \,|\, (i,j) \in E\}| - 1)}, i \in V \quad (3)$$

The average clustering coefficient of the graph is the algebraic average of the clustering coefficients of all nodes:

$$\eta(G) = \frac{1}{|V|} \sum_{i \in V} c(i) \quad (4)$$

We note that the above measures evaluate somewhat different aspects of the graph integrity; none of them provides a comprehensive evaluation of the graph integrity. In order to be on the safe side in practical applications the best practice is to use such a set of simple integrity measures and evaluate the graph integrity using the resulting set of integrity measure values (i.e., by considering a vector of integrity measure values). In particular, if we need a single value measure of the graph integrity on the basis of a vector of integrity measures, the safest is to take the value indicating the greatest amount of integrity loss.

Other more sophisticated measures of graph integrity include the calculation of coefficients of the graph's characteristic polynomial, and eigenvalues of the graph's adjacency matrix. These methods can provide a full picture of the graph's integrity and in principle capture all its aspects. The disadvantage of these methods is that they are computationally very expensive, and the calculation of the required numbers may be impractical for very large graphs representing complex systems. The above introduced simple integrity measures are well correlated with the more general measures. The largest eigenvalue of the adjacency matrix is related to the density of the edges, the second eigenvalue is related to the conductivity within the network (Farkas et al, 2001). The second coefficient of the characteristic polynomial is related to the number of edges, while the third coefficient is twice the number of triangles in the network (Biggs, 1994).

An important issue regarding the use of graph integrity measures to assess the integrity of systems is that of how to actually measure the system components and their interactions. One approach can be to consider the design of the system, if this is available. (For technological systems this might often be the case.) However, this approach can lead one to fall into the trap of showing the robustness of the designed system and not of the actual system. We believe that the right approach is to measure the existing components of the real system and their existing interactions in order to assess the integrity and robustness of the actual system. However, we recognize that in some practical cases such measurements might prove to be difficult (e.g., monitoring of human – computer interactions), limiting the applicability of the structural graph analysis based assessment of system integrity evaluation.

In the case of computer-based systems we typically have a set of non-computer related units (e.g., humans, sensors), a set of hardware units making the computer hardware part of the system, and usually a very large set of software modules, constituting the units of the system. In some cases we ignore the non-computer related units and even the hardware part of the system and we focus our attention exclusively on the system made of software units. The interactions between software units take the form of data transactions between them, which can be measured by appropriate monitoring of the system (Periorellis et al, 2004).

3 Fault tolerance and integrity preservation

Faults are unexpected behaviours of system components. Faults in computer-based systems may have a number of origins; they can be classified as design faults, physical faults and interaction faults (Aviziensis et al, 2001). Faults cause errors in the system, which are deviations from the expected behaviour of the system. Errors in computer-based systems may stay latent, until they are detected, when they cause abnormal behaviour at the interface of the system with its environment (Lee and Anderson, 1990). Errors cause failures of the system, when the system is unable to perform its function correctly (Aviziensis et al, 2001).

Faults in the system may occur at various places. An important feature of faults is their occurrence within the system and their distribution at these places within the system. In many cases we may suppose that the faults may appear at any system unit according to the same occurrence distribution, no unit being more susceptible for producing faulty interactions than others (Amari, 2000). In some cases we may also use the hypothesis that the likelihood of faulty interactions is proportional with the likelihood of the unit being involved in interactions. In other cases the fault distribution may follow some peculiar well defined distribution, such as in the case of faults induced by malicious logic (e.g., attacks by hackers). The types of fault

distributions determine the fault occurrence environment of the system.

Fault tolerance is the ability of the system to maintain its functionality in the presence of active faults (Lee and Anderson, 1990). Fault tolerance is typically achieved by error detection, recovery and fault handling (Aviziensis et al, 2001). Fault tolerance of computer-based systems depends on the fault occurrence environment of the system (e.g., in presence of naturally occurring faults the system may prove sufficiently fault tolerant, while in the presence of targeted attack by hackers, it may prove fault sensitive).

In general the measure of fault tolerance of the system is a relative measure, which shows to what extent the system preserves its functionality in a certain fault occurrence environment (Lee and Anderson, 1990). To assess the fault tolerant nature of a computer-based system we need to assess the level of functionality of the system within the considered fault occurrence environment. In other words we need to evaluate the functionality preservation ability of the system. Usually some probabilistic approach is used to evaluate fault tolerance measures such as mean time to failure or mean time to repair (Lee and Anderson, 1990). These methods take into consideration the whole system resulting computationally very intensive analyses in case of large systems (Chang et al, 2004; Billington and Jonnavithula, 1999). To perform such exhaustive evaluations may prove difficult in practice, as monitoring and assessing all aspects of the functionality of the system and performing all the required calculations may be extremely time and resource consuming (Ferrandi et al, 2003). Alternative methods were proposed recently, involving game theoretic approaches (Bell, 2003), formal languages inspired analysis (Phoha et al, 2004) and structural network analysis approaches (Albert et al, 1999).

We adopt the structural network analysis approach, which has the key advantage that it implies a relatively low computational load for the evaluation of large systems. We measure the fault tolerance of the system by evaluating the ability of the system to preserve its integrity. The measure of integrity preservation is calculated by using system integrity measures based on a structural graph analysis of the graph representing the system. As structural integrity is strongly correlated with functional integrity, the structural integrity preservation measure provides a proxy measure of the functional integrity preservation measure of the system. Consequently, we can use the structural integrity measures introduced in the previous section to measure the change of the integrity of the system in a given fault occurrence environment.

To measure the effects of faults on the integrity of the system, we simulate the faults by sampling the fault occurrence distributions and then evaluating the integrity measures of the system in the presence of simulated faults. The presence of faults causes the elimination from the graph of the system of edges between nodes or of nodes of the graph. These changes happen according to the fault occurrence distributions and have the effect that the integrity measures of the system graph are modified. The expected changes in terms of integrity measures may be calculated analytically in the case of small systems or can be evaluated by numerical simulations in the case of large and complex systems. The expected changes associated with a fault occurrence environment characterise the system's integrity preservation ability and are used as an approximate measure of the fault tolerance of the system.

To show how to use the calculated integrity measures to assess the fault tolerant nature of a system we consider below a toy example. Let us consider a software system of 1000 units of which corresponding graph representation is shown in Figure 1. The system's structural integrity measures are the following: (1) diameter: $D(S)=19$; (2) average minimum path length: $\mu(S)=3.58151$; (3) average clustering coefficient: $\eta(S)=0.022702$

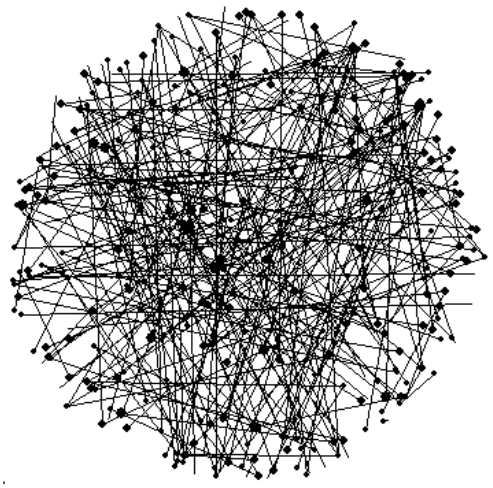


Figure 1: The graph representation of the model system with 1000 nodes. The size of the nodes indicates the number of connections of the node. Only the subset of more connected nodes and the subset of connections between these nodes are displayed to keep the figure comprehensible.

We consider a fault occurrence environment in which the faults occur with equal uniform probability (p=0.15) at each unit of the system and each fault temporarily knocks out the system unit where it occurs. To evaluate the fault tolerance of the system we perform a numerical simulation of the fault occurrences, and evaluate the integrity measures of the system for each simulation. After the simulations we calculate the average values and variances of the system integrity measures. We chose to run 20 simulations in order to get reliable estimates of mean values (the variance of the mean value calculated from n measurements is σ/square-root(n), where σ is the variance of the calculated values). The calculations after the simulations led to the values: (1) diameter: avg(D(S))=23.05, var(D(S))=4.1355; (2) average minimum path length: avg(μ(S))=3.6994, var(μ(S))=0.0397; (3) average clustering coefficient: avg(η(S))=0.023, var(η(S))=0.0014.

To evaluate the integrity preservation ability of the system we calculate first, whether the average values of system integrity measures after the simulation of faults differ significantly or not from the corresponding values calculated for the fully functional system. Next we calculate the normalized distance of the pre-damage and post-damage integrity measure values, which together with their attached statistical significance levels characterize the fault tolerance of the system. In order to be on the safe side, we choose the worst measure (i.e., the largest and most significant distance) to be the numerical evaluation of the fault tolerance of the system. In the case of the above system the normalized distances (z-score, i.e., the distance measured between the mean value and original value in units equal to the standard deviation – $(v_{original} - m)/\sigma_m$) and statistical significance levels (statistical significance levels show how likely is that the original value is the same as the estimated mean value after damage, low p-value indicates that the likelihood of them being the same is very low, or in other words the two values differ significantly) are listed in Table 1.

In the case of the above toy example we have shown how to apply in principle the proposed structural graph analysis based integrity evaluation methods to assess the fault tolerance of a computer-based system. The data shown in the last column of Table 1 shows the values of the likelihoods that original value of the integrity measure is the same as its value after the damage. The results indicate that under the above described fault occurrence environment assumption the system suffers significant damage ($p<0.01$) in terms of diameter and average shortest path length, the amount of the latter damage being more significant than the former. Considering the most significant damage (i.e., the damage in terms of average path length, $p=7.29 \times 10^{-14}$), we conclude that under the considered fault occurrence assumption the system represented by the graph suffers very significant structural and functional damage, and consequently has low fault tolerance.

Table 1: Summary of integrity measures of the system before and after damage, including the z-score for the original values considering the mean and variance of the after damage values (z-score = (original – damage mean)/(damage variance / square-root(20))), and the statistical significance level of the difference between the original values and the mean values calculated after the damage. The p values above 0.1 are omitted.

Integrity measure	Original	Damage mean	Damage variance	z-score	p-value
Diameter	19	23.05	4.135	4.379	1.2×10^{-5}
AvShortPath	3.582	3.699	0.040	13.26	7.29×10^{-14}
AvClusCoef	0.023	0.023	0.002	1.079	-

4 Application

Linux is one of the most popular operating systems, which is due to a good extent to its open source based development. It is commonly claimed that Linux is more reliable and secure than many other operating systems. An immediate question is how fault tolerant is Linux actually.

We analysed the network structure of the Linux under typical running conditions with a set of usual programs running. To perform the analysis we considered the calls between the classes present in the Linux kernel (version 2.4.19). We found 6815 classes and 19909 calls between them, by parsing the source code of the classes. The interaction network of the classes (see Figure 2) was then analysed in terms of structural network analysis.

Analysing the connectivity distribution of the processes we found that the distribution follows a power law distribution (with exponent $\gamma = -1.33$; see Figure 3) similar to the case of the Internet (Albert et al, 1999). This indicates that among the Linux classes there are relatively few very highly connected classes (which call and are called by many other classes) and many others with relatively few connections. This implies that similarly to the Internet (Albert et al, 1999) the Linux is very robust and fault tolerant if faults happen randomly following a uniform fault distribution over the processes (run-time representation of classes), while it should be very vulnerable and fault sensitive if faults are distributed such that they affect mostly the most highly-connected processes.

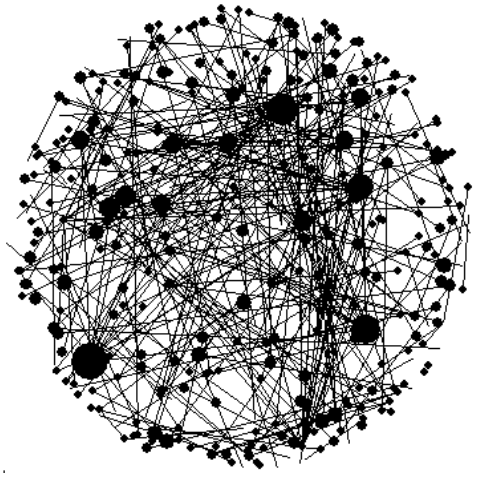


Figure 2: The graph representation of the Linux. The size of the nodes indicates the number of connections of the node. Only the subset of more connected nodes and the subset of connections between these nodes are displayed to keep the figure comprehensible.

We performed an analysis of the Linux class network to evaluate the effects of faults on integrity measures. We simulated a scenario with uniform random distribution with probability of faults at each node. We also performed a simulated a scenario when the likelihood of a node being faulty was proportional with the connectivity of the node. The analysis results are shown in Table 2 and Table 3.

The results show that as we expected Linux is remarkably fault tolerant in a fault occurrence environment characterised by uniform fault distribution, while it is significantly more fault sensitive in the case of a fault distribution centred on the mostly linked processes. This suggests that indeed the common belief about the reliability and fault tolerance of Linux is well founded in case of random uniformly distributed errors, but also highlights that Linux is also a vulnerable system in case of well designed malicious attacks. (Note that the tables show that the average shortest path is decreasing after damage. This is because infinite shortest paths between nodes belonging to isolated sub-networks are ignored.)

Table 2: Summary of integrity measures of Linux before and after random damage, including the z-score for the original values considering the mean and variance of the after damage values (z-score = (original – damage mean)/(damage variance / square-root(20))), and the statistical significance level of the difference between the original values and the mean values calculated after the damage. The p values above 0.1 are omitted.

Integrity measure	Original	Damage mean	Damage variance	z-score	p-value
Diameter	44	38.85	4.869	4.729	2.25×10^{-6}
AvShortPath	12.01	11.50	0.740	3.077	0.0021
AvClusCoef	0.133	0.133	0.007	0.350	-

Table 3: Summary of integrity measures of Linux before and after targeted damage, including the z-score for the original values considering the mean and variance of the after damage values (z-score = (original – damage mean)/(damage variance / square-root(20))), and the statistical significance level of the difference between the original values and the mean values calculated after the damage. The p values above 0.1 are omitted.

Integrity measure	Original	Damage mean	Damage variance	z-score	p-value
Diameter	44	35.45	5.062	7.552	1.41×10^{-13}

AvShortPath	12.01	10.47	1.444	4.766	1.87×10^{-6}
AvClusCoef	0.133	0.132	0.005	1.018	-

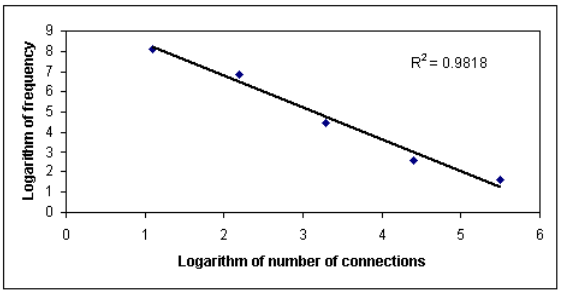


Figure 3: The distribution of connectivities in the case of the Linux classes and calls network

4 Conclusions

Fault tolerance is measurable aspect of the dependability of computer-based systems. Direct measurement of fault tolerance of large real world systems poses considerable problems, considering that most existing work is focused on exhaustive analytical evaluation of relatively simple model systems (Billinton and Jonnavithula, 1999; Bell, 2003; Cheng and He, 2004). An approach to find a proxy measure for the fault tolerance of large systems is to measure their structural integrity preservation under the assumption of a fault occurrence environment. This measure is based on the assumption that functional integrity is strongly correlated with structural integrity (Andrews and Beeson, 2003; Ferrandi et al, 2003), which is supported by experimental analysis of various real world complex systems (Albert et al, 1999; Jeong et al, 2001).

Integrity and integrity preservation of large systems can be measured by structural graph analysis of their graph representation, where nodes represent system units and edges represent interactions between system units. Simpler (e.g., average minimum path length) or more complex (e.g., eigenvalues of the adjacency matrix) measures can be used to estimate the structural integrity preservation of the system while exposed to faults occurring in accordance with an assumed fault distribution. Using these measures we can evaluate the likely amount of loss of integrity (or integrity damage) and the statistical significance of this loss.

A toy example and a real world example were presented to show the application of this fault tolerance measurement approach. These examples show that indeed the proposed methods can be applied effectively and lead to meaningful conclusions about the analysed systems. In the case of the real world example (Linux) the analysis indicates that the system is very fault tolerant under the assumption of uniform fault distribution, but that, not surprisingly, it is very vulnerable under the assumption of a fault distribution driven by properly targeted malicious interventions. However, the methods discussed here enable the extent of this effect to be analyzed in useful detail.

We believe that using relatively simple network integrity measures can simplify considerably the effective analysis of fault tolerance of large real world computer-based systems. Although these methods do not provide an exact measure of fault tolerance they provide good approximations of the actual measure. Such approximate measures can be used to rapidly determine the effects of a variety of fault occurrence environments, allowing the designers and developers of large systems to prepare appropriate defence and repair strategies to support the dependability of their system effectively and efficiently.

Acknowledgement

The authors thank Brian Randell for very helpful comments.

References

A. Aviziensis, J.C. Laprie, B. Randell. Fundamental concepts of dependability. *Newcastle University Report no. CS-TR-739* 2001.

R. Albert, H. Jeong, A.-L. Barabási. Diameter of the World Wide Web. *Nature*, 401:130-131, 1991.

J. Al-Sadi, K. Day, M. Ould-Khaoua. Unsafety vectors: a new fault-tolerant routing for the binary n-cube. *Journal of Systems Architecture*, 47:783-793, 2002.

S.V. Amari. Generic rules to evaluate system-failure frequency. *IEEE Transactions on Reliability*, 49:85-87, 2000.

REPLACE Charlton, B, Andras, P: The nature and function of management - a perspective from systems theory. Philosophy of Management 2004; 3: 3-16.

J.D. Andrews and S. Beeson. Birnbaum's measure of component importance for noncoherent systems. *IEEE Transactions on Reliability*, 52:213-219, 2003.

F. Beichelt and P. Tittmann. A generalized reduction method for the connectedness probability of stochastic networks. *IEEE Transactions on Reliability*, 40:198-204, 1991.

M.G.H. Bell. The use of game theory to measure the vulnerability of stochastic networks. *IEEE Transactions on Reliability*, 52:63-68, 2003.

N. Biggs. *Algebraic Graph Theory*. Cambridge, Cambridge University Press, 1994.

R. Billinton and S. Jonnavithula. Calculation of duration, frequency, and availability indexes in complex networks. *IEEE Transactions on Reliability*, 48:25-30, 1999.

Y. Chan, E. Yim, A. Marsh. Exact and approximate improvement to the throughput of a stochastic network. *IEEE Transactions on Reliability*, 46:473-486, 1997.

Y.-R. Chang, S.V. Amari, S.-Y. Kuo. Computing system failure frequencies and reliability importance measures using OBDD. *IEEE Transactions on Computers*, 53:54-68, 2004.

Y. Cheng and Z. He. Bounds on the reliability of distributed systems with unreliable nodes & links. *IEEE Transactions on Reliability*, 53:205-215, 2004.

T. Elperin, I. Gertsbakh, M. Lomonosov. Estimation of network reliability using graph evolution models. *IEEE Transactions on Reliability*, 40:572-581, 1991.

I.J. Farkas, I. Derenyi, A.-L. Barabasi, T. Vicsek. Spectra of Real-World Graphs: Beyond the Semi-Circle Law. *arXiv*:cond-mat/0102335, 2001.

F. Ferrandi, F. Fummi, G. Pravadelli, D. Sciutto. Identification of design errors through functional testing. *IEEE Transactions on Reliability*, 52:400-412, 2003.

A. Goerdt. Random regular graphs with edge faults: expansion through cores. *Theoretical Computer Science*, 264:91-125, 2001.

A. Goerdt and M. Molloy. Analysis of edge deletion processes on faulty random regular graphs. *Theoretical Computer Science*, 297:241-260, 2003.

H. Jeong, S.P. Mason, A.-L. Barabasi, Z.N. Oltvai. Lethality and centrality in protein networks. *Nature*, 411:41-42, 2001.

W. Kishimoto. Reliable flow with failures in a network . *IEEE Transactions on Reliability*, 46:308-315, 1997.

J.C. Laprie (ed). *Dependability: basic concepts and terminology in English, French, German, Italian and Japanese*. Wien, Springer-Verlag, 1992.

P.A. Lee and T. Anderson. *Fault Tolerance. Principles and Practice*. Wien: Springer-Verlag, 2nd ed, 1990.

Y. Ou and J.B. Dugan. Approximate sensitivity analysis for acyclic Markov reliability models. *IEEE Transactions on Reliability*, 52:220-230, 2003.

P. Periorellis, O.C. Idowu, S.J. Lynden, M.P. Young, P. Andras. Dealing with complex networks of protein interactions: A security measure. In *Proceedings of 9th IEEE International Conference on Engineering of Complex Systems (ICECCS)*, P. Bellini, S.A. Bohner, B. Steffen (eds) pp 29-36, IEEE Computer Society, 2004.

V.V. Phoha, A.U. Nadgar, A. Ray, S. Phoha. Supervisory control of software systems. *IEEE Transactions on Computers*, 53:1187-1199, 2004.

J.W. Scannell, C. Blackmore, M.P. Young. Analysis of connectivity in the cat cerebral-cortex. *Journal of Neuroscience*, 15:1463-1483, 1995.

C. Scherrer and A. Steininger. Dealing with dormant faults in an embedded fault-tolerant computer system. *IEEE Transactions on Reliability*, 52:512-522, 2003.

L. von Bertalanffy. *General System Theory: foundations, development, applications*. Harmondsworth, Penguin, 1973.

Towards Associative Information Access

Michael R. Berthold*
*Konstanz University
Dept. of Computer and Information Science
Konstanz, Germany
Michael.Berthold@uni-konstanz.de

Andreas Nürnberger†
†Otto-von-Guericke Universität Magdeburg
Faculty of Computer Science
Magdeburg, Germany
nuernb@iws.cs.uni-magdeburg.de

Abstract

We propose a framework for a unifying, associative access to distributed and heterogenous information resources. The classical index generation is replaced by a process which builds associations between existing information entities and allows for an interactive exploration of information accessible through this structure. Positive ("'this looks interesting") as well as negative ("'I know this!") user feedback allows the system to quickly narrow down on interesting pieces of information. The continuous integration of new analysis engines, added sources of information and user feedback allow the formation of a corporate wide memory and expert knowledge repository.

1 Motivation

Large corporations increasingly drown in all sorts of data and other types of information they collect. Modern storage technology essentially sets no limit to the amount of information that can be stored. The huge challenge is the problem of usage — how can users be sure that they did take into account all relevant pieces of information that relate to the current task or problem they are dealing with?

One prime example for this scenario are research departments in many pharmaceutical companies. In order to successfully develop new drugs, many different types of information need to be combined, in the end resulting in a new idea for a medication that has not been patented before, that has no dangerous side effects, or that is not, in some similar form, already being explored elsewhere. Currently this process relies heavily on experts having intuition, long years of experience and hopefully the right insights at the right time. The sources of information these experts rely on are distributed across the entire company (and some also over the entire internet): experimental protocols, patent information, scientific publications, biological information about metabolic pathways just to name a few, and not to forgot, also the colleague down the hall who would have something interesting to say but who our expert did not happen to meet at the coffee pot.

Current approaches try to address this problem by building huge information repositories based on sophisticated database technology. Associative Information Networks, as described here, aim to take an alternate approach – instead of bringing all the information together we propose to build a meta structure that points to the information and helps the user find interesting associations among different pieces of information through means of exploration and context refinement. This meta structure is continuously updated as more sophisticated methods to analyze the information sources arise. In addition, it is possible to naturally incorporate user annotations, capturing expert knowledge and feedback on the way. This process is supported by methods derived from research in the areas of data mining, information retrieval, knowledge management, network and graph theory, data visualization and human computer interaction.

2 Related Work

There has been a lot of work done in the past on the idea of associative information processing, which was in the beginning mainly motivated by the associative information processing capabilities of the human brain (see, e.g., the work of Collins and Loftus (1975)). Thus we can find methods ranging from very general neural network based approaches of Kohonen (1977, 1984), over possibilistic networks or graphs (Borgelt et al., 2000; Cao, 2000) and belonging reasoning methods (Dubois et al., 1994; Gebhardt and Kruse, 1995) to very specific ideas related to document indexing and retrieval, e.g. (Chen, 1995; Chung et al., 1998; Belew, 2000). Furthermore, also ontologies and the Sematic Web (Berners-Lee et al.,

2001) might be considered as an approach to enable linking of semantically associated information.

However, several of the earlier projects failed, since almost all of them are based on the idea that it is possible to know in advance or learn automatically an almost perfect descriptive link from (index-)keywords to documents or in-between documents. This information was then used in some kind of reasoning mechanism to retrieve relevant documents. Unfortunately, in most cases this leads to the retrieval of too few or far too many documents. A further major problem had been the poor visualization methods used.

In order to circumvent these problems, more recently, some projects started in which methods have been studied that are also able to handle more general associative networks by providing interactive visualization methods. In order to navigate and browse complex association networks powerful tools for visualizing relevant subsets for the current exploration (or search) context of the user are required. Recent commercially available approaches that try to tackle this problem are, e.g., the Personal Brain (`http://www.thebrain.com/`), a navigator for indexed data that is however only able to access documents on a local data repository and the iAS KnowledgeSuite (`http://www.knowledgesuite.de/`). The KnowledgeSuite performs a semantic text analysis and creates strong links between previously identified, named entities. In this case, however, association are originate only from primed neurons using positive activity spreading. No interactive refinement or inclusion of uncertain, imprecise information is possible.

In general one might argue that the linking of documents as proposed for the semantic web might solve the problems of linking information sources. However, in the semantic web, one is forced to either link or not link documents, where an existing link has a clear, semantically valid meaning. Even though it is in general possible to introduce mechanisms for context based links (as realized for example in topic maps, see e.g. Biezunski et al. (1999)), no mechanism for storing 'gradual' (e.g. possibilistic, probabilistic, or simply anecdotal or evidential) links between documents are implemented. Furthermore, in the semantic web the whole web is seen as the knowledge base which includes both, links *and* information chunks. In our approach we add a general layer of links over (the possibly already existing link layer within) the considered database of information entities, which could consist of information in the world wide web, a local database or even notes on a local PC. This layer allows to model a personal (or group based) view on the same information, independent of (and not conflicting with) links already present in the data. However, we can easily incorporate general concepts of the (semantic) web, like URIs and existing ontologies in order to model and exploit already available information.

Another aspect that distinguishes our approach from semantic web (or more general logic based) approaches is that we do not use reasoning mechanisms that require a consistent descriptions of relations between information chunks. The main goal of the reasoning mechanism is to detect information that is most likely interesting to the user for any reasons (may be even because its contradicting!). In contrast, the reasoning mechanism itself is able to provide an explanation why some information has been proposed.

One additional differentiator is the ability for continuous learning and updating of the underlying structure. Through integration of new analysis engines, new information sources, or also manual feedback the network continuously refines it's internal structure.

3 Associative Information Networks

3.1 Structure

Associative Information Networks (AI Net in the following) consists of nodes and labelled edges. Each node represents an entity, which can be a concept from the application area (e.g. a disease, or metabolic pathway) or a named entity, such as a gene, a protein, or a specific target. Edges represent links between these entities and are labelled with a reference to the information source(s) and information about the analysis engine that created it from these sources. In addition, each edge holds a weight, modelling the strength of association, and a label indicating the type of the edge. This way, a link can potentially also be derived from an ontology, representing semantic connections between nodes.

3.2 Learning and Refinement

In order to generate the AI Net we need to introduce nodes, and links in between them. Refinement may cause adjustment of links and addition of new nodes. There are two primary ways how both, nodes and links can be added:

- automatic generation: using analysis engines, links between existing nodes can be added or modified. Each analysis engine has a particular purpose and will, for instance, find co-occurrences of words in documents, correlations of genes in gene-expression experiments, structure-activity relationships via the analysis of cell-assay images, or connections between genes and diseases from the analysis of patent information. In comparison, this would resemble the collection and modelling of automatically derivable domain knowledge. Of course, the addition of newly developed analysis engines can continuously update the network.

- manual interaction: throughout usage of the AI Net, the user is able to manually adjust weights of links, mark links as wrong, or insert new links with annotations explaining their purpose. This interactive refinement allows to capture expert knowledge and feedback on the fly and enables the system to model expertise available within a corporation. It is, of course, crucial that this interaction is handled in an intuitive way. The user should not be required to adjust numerical weights or draw links between abstract nodes.

Adding new databases, or more generally, information sources is straightforward – as long as an analysis engine is provided that produces dependencies between entities represented by nodes, new links can easily be added. One further extension of this system would also allow to generate new nodes (and node types) by analyzing external information sources.

3.3 Link Formation: Details

As described above, links can be introduced automatically or through manual refinement. The latter process can be seen as user annotations, incorporating expert knowledge into the network and are therefore mainly an issue of user interface. In the following, we briefly outline, based on a number of examples, how the automatic generation of links and link-weights works.

- semantic links: these are strong links (usually weight $= 1.0$) which are derived from well-known structures, such as ontologies or semantic networks. Those are usually created by an expert. Semantic nets, as extracted (semi-) automatically from data will need to add a component that computes the confidence for each link and convert this to a weight.

- syntactic links: these are links that are generated by a shallow analysis of data. The most prominent example would be a text parser that converts words to stems, eliminates fill words and then produces a set of bi- or trigrams. The corresponding nodes in the AI Net will be connected by weak links. For an example of the corresponding weight computation, see below.

- anecdotal evidence: These are links set by a user, creating links for hypotheses generated by a user (or based on hear-say). Weights of such links are generally low. These links are in contrast to expert-based annotations that generally have very high weights.

- data driven links: These types of links will constitute the vast majority of network weights in most cases. They are generated automatically from data repositories. A few example (here for the context of a pharmaceutical AI Net) could be:

  Gene correlations derived from gene expression data. Links are introduced when, for example, a specific threshold θ for co-occurrence in experimental data is surpassed. The link's weight reflects the correlation strength and for more than two-dimensional correlations the corresponding multi-edges are introduced. In addition each of these links will carry an annotation pointing to the source of it's weight, in this example a link to the experiment and some meta information (threshold θ, date of analysis, reference to exact computation of weight).

  Textual analysis where co-occurrence of named entities within a specific distance (= words in between) results in a weak link to be introduced. The weight depends on distance and quality of text source.

  Links between gene and protein names derived from scientific articles based on a bigram analysis. Weights are derived from the average distance and frequency of occurrence in documents, analogous to the TFIDF-score (Term frequency / inverse document frequency).

- ontology/thesaurus links: Based on an existing ontology links will be introduced to connect entities that are related based on this ontology. This resembles a 1:1 correspondence between each link in the ontology and a link in the network. The resulting links are strong links, i.e. carry a weight of 1.0 since there is (usually) no doubt about the reliability of that particular

piece of knowledge. Otherwise it would need to be reflected in the link's weight.

Obviously many other types of links can be generated, since the underlying structure is invariant to origin or meaning of links.

3.4 Exploration: Finding interesting associations

The network's structure can be used in various ways to find potentially interesting pieces of information. Most straightforward would be the search for tightly connected other entities, such as another gene that is related to the ones the user just saw within an experiment. This can be implemented via a simple neighbor-search in the network, finding all genes that are connected to the set of "query" genes.

More powerful are, however, searches that find related pieces of information via various steps, or so-called bridge concepts. This can be implemented analogous to activity spreading methods, as known from the neural network community (Cohen and Kjeldsen, 1987). The real power, in the concept presented here, lies in the ability to perform this search interactively. Throughout the search the user can weight entities that he finds interesting positively (and the ones he does not care about negatively), instantly affecting the activation pattern and hence the associations the network proposes. Such an interactive scheme will heavily rely on a suitable visualization of the graph network (see, e.g. Chen (2004)) and appropriate adaptive user interfaces.

4 Conclusions

In this paper we have briefly presented the idea of a generalized associative information network. With this concept we try to simulate aspects of the associative capabilities of the human brain in order to support a user in gathering information about a specific problem at hand. The tool is not meant to offer problem solving capabilities, but rather to point out information pieces a user might have otherwise not had the chance to look at, be it for lack of knowledge about their existence or because of a failure to see their importance for the task at hand.

References

R.K. Belew. *Finding out About*. Cambridge University Press, 2000.

T. Berners-Lee, J. Hendler, and O. Lassila. The semantic web. *Scientific American*, May, 2001.

M. Biezunski, M. Bryan, and S.R. Newcomb, editors. *ISO/IEC 13250:2000 Topic Maps: Information Technology – Document Description and Markup Languages*. ISO/IEC, 1999.

C. Borgelt, J. Gebhardt, and Ru. Kruse. Possibilistic graphical models. In G.D. Ricci, R. Kruse, and H.-J. Lenz, editors, *Computational Intelligence in Data Mining*, pages 51–68. Springer-Verlag, Wien, 2000.

T. H. Cao. *Fuzzy Conceptual Graphs: A Language for Computational Intelligence Approaching Human Expression and Reasoning*, pages 115–120. Physica-Verlag, Heidelberg, 2000.

C. Chen. *Information Visualization*. Springer, 2004.

H. Chen. Machine learning for information retrieval: neural networks, symbolic learning, and genetic algorithms. *J. Am. Soc. Inf. Sci.*, 46(3):194–216, 1995.

Yi-Ming Chung, William M. Pottenger, and Bruce R. Schatz. Automatic subject indexing using an associative neural network. In *DL '98: Proceedings of the third ACM conference on Digital libraries*, pages 59–68, New York, NY, USA, 1998. ACM Press.

P.R. Cohen and R. Kjeldsen. Information retrieval by constrained spreading activation in semantic networks. *Information Processing and Management*, 23(2):255–268, 1987.

A.M. Collins and E.F. Loftus. A spreading activation theory of semantic processing. *Psychological Review*, 82:407–428, 1975.

Didier Dubois, Jerome Lang, and Henri Prade. Automated reasoning using possibilistic logic: Semantics, belief revision and variable certainty weights. *IEEE Trans. Data and Knowledge Engineering*, 6 (1):64–71, 1994.

Jörg Gebhardt and Rudolf Kruse. *Reasoning and Learning in Probabilistic and Possibilistic Networks: An Overview*, volume 912 of *Lecture Notes in Artificial Intelligence*, pages 3–16. Springer-Verlag, Berlin, 1995.

Teuvo Kohonen. *Associative Memory - A System Theoretic Approach*. Springer-Verlag, Berlin, 1977.

Teuvo Kohonen. *Self-Organization and Associative Memory*. Springer-Verlag, Berlin, 1984.

Observing Dynamics in Community Structures

Tanja Falkowski
Otto-von-Guericke-Universität Magdeburg
Universitätsplatz 2, 39106 Magdeburg
falkowski@iti.cs.uni-magdeburg.de

Myra Spiliopoulou
Otto-von-Guericke-Universität Magdeburg
Universitätsplatz 2, 39106 Magdeburg
myra@iti.cs.uni-magdeburg.de

Abstract

Network analysis methods are widely used to detect community structures in static graphs. Since these structures undergo changes caused by internal and external factors it is necessary to provide methods to detect and observe transitions in community structures. For this we partition the interactions of community members by time windows and aggregate them. The resulting static graphs of each interval are analyzed for sub-communities. Through this we detect if communities persist over time or undergo a transition. We briefly present an interactive software environment which supports a temporal community analysis and provides several forms of visualization and analysis settings.

1 Introduction

Communities have proven to be of strategic importance e.g. to improve knowledge sharing or to enhance customer retention. Thus, communities have been studied in many research fields. So far they are mainly regarded as a static phenomenon and aggregated data over longer periods has been used to detect communities. However, the analysis of aggregated interactions between community members has some drawbacks. Old interactions are favored over newer ones and temporal developments in the interaction behavior can not be observed. But since communities are highly dynamic social networks, observing its transitions along the time axis is an important task, e.g., to adapt community platforms in order to support community building or to improve the efficiency of communities.

Tools such as SoNIA (Moody et al., 2005) and TeCFlow (Gloor and Zhao, 2004) visualize temporal social graphs by creating movies of them. Both tools work on the vertex and edge level thus visualizing a changing behavior between single actors. We propose to analyze and visualize temporal changes on the community level to allow for an exploration of sub-group dynamics.

Therefore, we regard a community as an object that exists over time and propose a dynamic temporal observation along the time axis using sliding time windows. The communities are detected in a static representation of interactions that occur in a specified period (cf. Section 2.1). We determine the evolution of the interactions by comparing the communities in different time windows. By this we are able to detect different types of transitions a community might pass through such as a split or a merger (cf. Section 2.2). The changes in the community structure are visualized and the user can choose different analysis settings to further explore the dynamics of the community under investigation in order to detect triggers that caused a community transition (cf. Section 3).

2 Model Community Dynamics

The problem of detecting community structures in networks is of interest in social sciences as well as for many other research fields such as computer science (e.g. WWW, e-mail log files) or biology (e.g. gene or protein networks) (see, e.g., Aggarwal and Yu, 2005, Kleinberg and Lawrence, 2001, Wilkinson and Huberman, 2004). In Section 2.1 we briefly discuss the method we use to detect communities in static graphs. In Section 2.2 we describe how we apply this method to observe the dynamics of community structures.

2.1 Community Detection

First, we model the network of interactions in a way suitable to find communities. We do so by defining a graph G = (V, E), in which V denotes the set of vertices (nodes) and E the set of edges (i, j), with i,j $\in$ V. Each community member i is denoted a distinct vertex and an interaction between two members i and j, e.g., an e-mail exchange, is represented as an edge (i, j). We quantify the interaction between two members by assigning a weight $w(i, j)$ to

the edge (*i*, *j*). Appropriate weights are "number of messages exchanged" or the "total length of all messages measured in characters".

Next, we decompose the graph into communities. We define a (sub) community as a subset of vertices within a graph with a high degree of interaction among the participants. We apply a hierarchical divisive clustering approach that divides the graph by the iterative removal of edges. The edges that are removed should be those that do not contribute to a community.

To determine the edge to be removed in each iteratio we use the *edge betweenness* score proposed by Girvan and Newman (2002). The betweenness of an edge is the number of shortest path between pairs of vertices that run along it. It is based on the assumption, that the few edges between communities have more "traffic", as, e.g., an information flow between vertices in two communities has to travel along these edges. The hierarchical clustering algorithm iteratively removes the edges with the highest edge betweenness score. We apply this method to a *multigraph* as described by Newman (2004) to include weighted edges. Each edge betweenness value is divided by the edge weight. Therefore, the edge betweenness value between two very connected pairs is lowered so that rather weak connected pairs are separated faster than strong connected ones. The algorithm has a high complexity due to the recalculation of the edge betweenness in each iteration – $O(m^2n)$, where m is the number of edges and n the number of vertices. However, it is applicable for small networks with up to a few thousand vertices.

The results of the hierarchical clustering are presented in a dendrogram, a tree diagram, which illustrates the community structure of the graph (see Figure 1). Since we have no a priori knowledge about the number of communities that exist in a network, we need an indicator on where to partition the dendrogram to obtain a meaningful network partition. For this purpose, we use the quality function proposed by Newman and Girvan (2004) to determine the best dendrogram cut which is based on the concept of modularity. The quality function Q is defined as:

$$Q = \sum_i e_{ii} - \sum_{ijk} e_{ij} e_{ki} = Tr(e) - \left\| e^2 \right\| ,$$

where e_{ii} is the fraction of edges in the original network that connect two vertices inside the community i and e_{ij} the fraction of edges that connect vertices in community i to those in community j. ||x|| indicates the sum of all elements in x. Q has a value between 0 and 1. Values above 0.3 appear to indicate a significant community structure. Values approaching 1 indicate a strong community structure.

2.2 Community Transitions

Since the interactions between participants and the set of participants are not static but change over time, we use the representation of the network but consider the graph as dynamic. Vertices as well as edges appear and disappear from the graph through time. We define the dynamic graph g_t as a graph which consists of all vertices and edges that are active in the interval t. If all interactions would be aggregated over time to G by summing up all g_i all information about the temporal development would be lost. Therefore, we define g_t as a sliding time window over time interval t that spans a set of interactions. In other words, all interactions that take place in this interval t are aggregated to g_t.

After defining the interval we can partition the graph over time into equidistant time slots, each slot starting when the last slot finished. This modus is called a *non-overlapping sliding window*. An *overlapping sliding window* partially overlaps with the prior window. The degree to which it overlaps must be defined. We apply an overlapping window since it smoothes out the gaps that sometimes occur between two intervals.

Each window is considered a static representation of the network in the chosen interval. At first we apply the community detection mechanism as described in the previous section to obtain a community structure for g_t. $C^{g_t} = \{c_1^{g_t}, ..., c_n^{g_t}\}$ is the set of communities that are detected in g_t. To determine whether a community persists over time we must be able to assess if a community $c_i^{g_x}$ is the same as a community $c_j^{g_y}$. Qualitatively we would define that a community in a subsequent interval is the same, if the characteristic features are similar. In the easiest case we would say that this is the set of participants. We therefore define that community $c_i^{g_x}$ and community $c_j^{g_y}$ are the same if they share a given percentage of members. The appropriate percentage depends on the community type, the type of relations and the intent of the observer. If a community e.g. consists of a small set of very active core members and a high number of less connected members that often change, the percentage should be rather small. Otherwise, the community might not be considered the same just because many of the other "uncharacteristic" members changed, even though the most active core members are still detected as a community in different intervals.

Besides deciding whether two communities are the same, which would mean that a community persists over time, we can also observe if a community merges with another community or if it splits into separated communities. These developments can be

triggered by internal factors such as a change in leadership as well as external factors, e.g. advertisements. The challenge is to determine the factors that positively trigger the community development to offer appropriate organizational and technological infrastructures.

3 Temporal Community Analysis

To track the development of established communities and to visualize the transitions we developed a software environment that supports temporal graph analysis based on the community detection methods described in Section 2. In the following section, we briefly describe the functionality of the software.

3.1 Visualizing Community Structures

The hierarchical divisive clustering algorithm is used as described in Section 2.1 to find the communities in each interval. The results are displayed in a dendrogram. The user can experiment on the impact of different clusterings on the quality measure by moving a slider as can been seen in Figure 1.

Figure 2 has two areas: a list of all detected communities in the left window and the curves on the right. The horizontal axes represent the respective time windows.

The lowermost curve displays the total number of interactions between the chosen group and the total number of interactions of all group members with other participants of other communities. It can be seen that the chosen group is only active for about 6 weeks, but some members have an active relation with external participants. The middle and the uppermost diagram show how similar the internal community interaction behavior is over time. In the middle diagram the vertical axis depicts the correlation distance as a similarity measure for the groups in different periods. It can be seen that the group shows up in two time windows with almost the same members but the structure changes very quickly and the group disappears. In the uppermost diagram the y-axis displays the Euclidian distance as a similarity measure. The more similar a group interaction in two periods, the lower is the value of the Euclidian distance. For both measures we compare the similarity between the chosen interval and all other intervals and for two succeeding intervals. The statistics can for example be used to find point in time where the interaction behavior of a group changes compared to previous intervals. If several groups show similar behavior, this might be an indication for a change in the overall community structure.

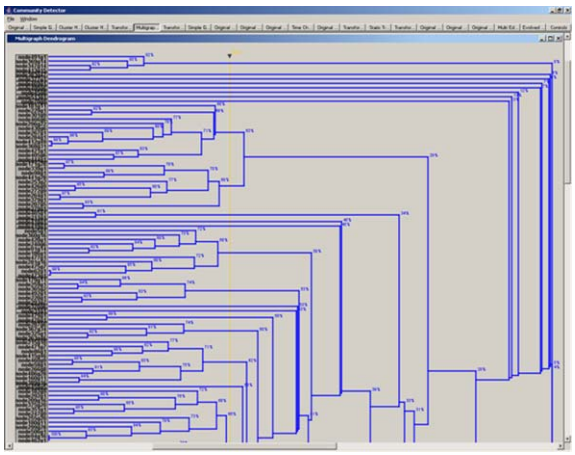


Figure 1: Clustering results in a dendrogram

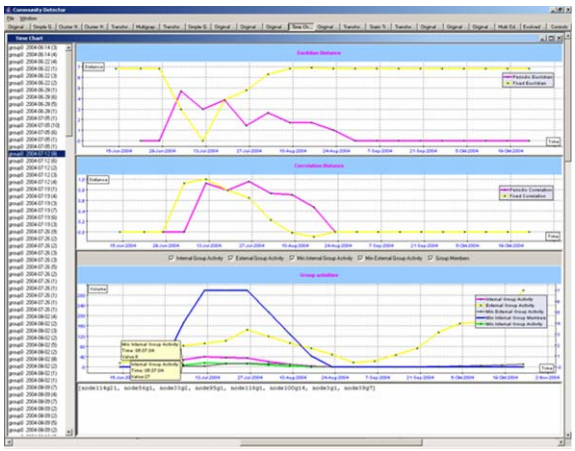


Figure 2: Statistics for temporal development for a chosen community

3.2 Visualizing Community Dynamics

In Figure 3 we see a static representation of a temporal community evolution. In this visualization, each detected community is represented as a vertex. The size of the vertex corresponds to the size of the community. Vertices that are connected by an edge are similar. Communities with the same members over several periods are positioned closer in the graph whereas communities with no members in common are more separated from each other. Furthermore, the different colors help to distinguish between similar communities and those that are not.

The user can choose for how many periods the community must at least exist to be displayed. If a long period is chosen, the user obtains only long-term community whereas in another case it might be of interest to find only short-term communities. Another slider for the time distance defines how continuous the communities are connected, separating communities by a maximum distance. Furthermore, one can define the observation period and filter the vertices so that the communities are displayed only in a selected period. The described properties can be used to filter communities so that the graph only shows data that is useful for a current analysis.

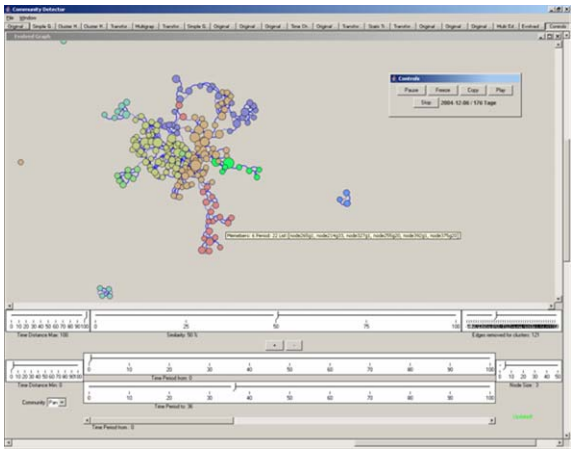


Figure 3: Visualization of communities based on Euclidian distance

Note that in the obtained graph in Figure 3 the temporal development can not be observed, as the communities are only displayed according to their similarity. In a next step, the filtered and clustered data is copied to a community history view, which allows seeing temporal developments by using the coordinates from the graph and putting the vertices on the horizontal axis according to the period they appear in (see Figure 4). The position transformation allows tracking the development along the time axis. Each community is now represented as a rectangle where its height corresponds to the size of the community. All communities that are considered as similar according to the actual settings are connected by edges and have the same color.

The left side of the screenshot in Figure 4 shows all communities over time in an overview window. The x-coordinate of each community is the same as in Figure 3. The y-coordinate is determined by the interval in which the community was detected. Thus, communities on the left are observed in earlier periods than those on the right.

The analyzed community shows different developments. In the lower part of the left side we can see

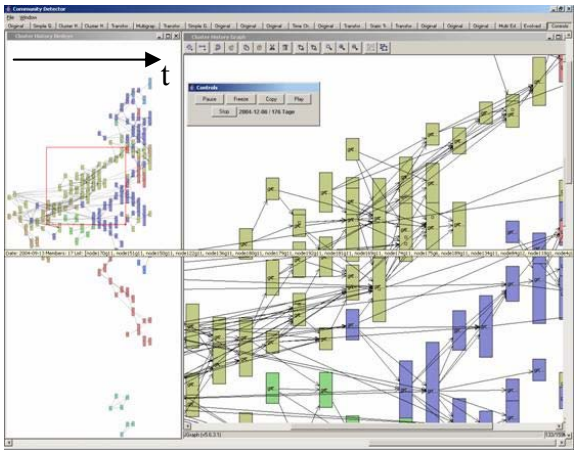


Figure 4: Community History View

an insolated community in light blue and one in red. Both existed over just a few periods. Some members of the red community joined another community which is shown in dark blue. In the cutout view on the right we can furthermore observe a community in yellow that existed over a longer period but died at some time. We can see that smaller communities merge to a bigger one but split very fast to several smaller communities. Points in time when these changes occur might be of interest for the user, as they indicate an external or internal event that influenced the community development. These effects might then be deliberately used to improve the establishment of communities.

4. Summary

We presented a software platform which allows for the analysis of the temporal development of communities. The gained insights can be helpful to understand community transition types and its triggers. This knowledge can be used to provide an appropriate technological as well as organizational infrastructure to foster community building.

References

Charu C. Aggarwal and Philip S. Yu. Online Analysis of Community Evolution in Data Streams. *Proceedings of SIAM International Data Mining Conference*, 2005.

Michelle Girvan and Mark E. J. Newman. Community structure in social and biological networks. *PNAS*, 99(12): 7821-7826, 2002.

Peter A. Gloor and Yan Zhao. TeCFlow - A Temporal Communication Flow Visualizer for Social Networks Analysis. *CSCW'04 Workshop on Social Networks*, ACM, 2004.

Jon M. Kleinberg and Steve Lawrence. The Structure of the Web. *Science*, 294(5548): 1849-1850, 2001.

James Moody, Daniel Mc Farland and Skye BenderdeMoll,. Dynamic Network Visualization. *American Journal of Sociology*, 110(4): 1206-1241, 2005.

Mark E. J. Newman. Analysis of weighted networks. *Physical Review*, E 70 (056131), 2004.

Mark E. J. Newman and Michelle Girvan. Finding and evaluating community structure in networks. *Physical Review*, E 69 (026113), 2004.

Dennis M. Wilkinson and Bernardo A. Huberman. A method for finding communities of related genes. *PNAS*, 10(1073), 2004.

Protein interaction network analysis reveals the importance of proteins with low degree of connectivity in neurodegenerative diseases

Joaquín Goñi[⋆§]
[⋆]Center Appl. Med. Res. (CIMA)
University of Navarra
31080, Pamplona (Spain)
jgoni@unav.es

Francisco J. Esteban[†]
[†]Dept. Exptal. Biology
University of Jaén
23071, Jaén (Spain)
festeban@ujaen.es

Nieves Vélez de Mendizábal[⋆‡]
[‡]Dept. Comput. Sci. Artif. Intell.
University of the Basque Country
20018, San Sebastián (Spain)
nvelez@unav.es

Jorge Sepulcre[⋆]
jsepulber@unav.es

Sergio Ardanza-Trevijano[§]
[§]Dept. Physics Appl. Mathematics
University of Navarra.
31080, Pamplona (Spain)
sardanza@unav.es

Pablo Villoslada[⋆]
pvilloslada@unav.es

Abstract

Recent developments have meant that network theory is making an important contribution to the study of the topology of biological networks such as protein interaction (PI) networks. The identification of differentially expressed genes in DNA-microarrays experiments is a known source of information regarding the main molecular pathways involved in disease. Thus, considering PI analysis and gene expression studies together may provide us with a better understanding of the topological aspects of multifactorial neurodegenerative diseases such as Multiple Sclerosis (MS) and Alzheimer Disease (AD). The aim of this study was to assess whether degree of connectivity is a key property that differentiates between implicated (seed-proteins) and non-implicated nodes (neighbors) in MS and AD. We used PIs experimentally validated and an interaction distance threshold from each seed group to obtain two networks (one MS-network and one AD-network). Results revealed for both diseases a lower degree of connectivity in seed proteins than in their neighbors in each PI network. Furthermore, we found that the MS-network and AD-network included multiple pathways and followed a very similar exponential degree distribution but with different clustering coefficient behavior. These findings reinforce the multifactorial nature of both diseases and might lead to new therapeutic strategies.
keywords: Alzheimer Disease, differentially expressed genes, microarrays, Multiple Sclerosis, network topology

1 Introduction

The structural and functional relationships underlying the organization of living systems imply the need to coordinate molecular interactions, principally those involving gene expression and protein activity. Although the genome is (almost) the same in each cell population of a given organism, dynamic changes in gene expression and thus in the protein content depend on the functional state of the cell (Lodish, 2002).

Genome-wide expression profiles using DNA microarrays, together with the development of bioinformatics approaches (Villoslada, 2006), enable us to model both genetic and protein interaction (PI) networks and thus to understand how a biological network operates (Xia, 2004). From a systems point of view (Hiesinger, 2005), the arrangement of biomolecular networks from gene expression data based on known interactions permits to understand the basic mechanisms upon which the complexity and adaptability of a living cell is founded. This information also helps us to decipher processes involved in illness, for instance the molecular heterogeneity of cancer (Rhodes, 2005). However, and consistent with the model of multifactorial diseases, it is difficult to find genes that account for direct genotype-phenotype correlation (Gunsalus, 2005). Thus, network modelling and topological analysis may provide additional knowledge about common properties

of genes and proteins involved in many severe diseases of multifactorial nature, where the cause of the pathogenesis does not depend on the malfunction of a single gene or protein. In addition, from a medical point of view, systems biology approaches to complex diseases might represent a standpoint to identify new therapeutic targets. In this case, the analysis of genes, proteins and pathways interactions might suggest common properties of the best candidates to be targeted by therapy. In addition, the understanding of the emergent properties of a system might allow the identification of new targets that will not be captured with a molecular approach (Kitano, 2004).

Multiple Sclerosis (MS) is a chronic inflammatory and neurodegenerative disease of the central nervous system (CNS) (Steinman, 2001). Its etiology remains elusive, but the interplay between environmental and genetic factors is ultimately thought to be critical to the development of the disease. MS is considered to be an autoimmune disease because of the presence of inflammatory infiltrates in the brain, in absence of infection, and its association with HLA alleles, among other factors (Oksenberg, 2001). The chronic inflammatory activity within the CNS is the main mediator of tissue damage, even in the late neurodegenerative stage of the disease, which includes widespread demyelination and axon loss (Bruck, 2005). In addition to the autoimmune processes, MS also has a neurodegenerative component whereby axons and neurons are lost through unknown processes in the late chronic stages of the disease. Several lines of evidences suggest that dying-back degeneration of demyelinated axons is the most important factor in MS neurodegeneration (Imitola, 2006). MS is a multifactorial disease in which many pathways of the immune system and CNS are involved (Fernald, 2005). Current therapies ameliorate in part the inflammatory process, but more effective therapeutical approaches are required to completely stop disease progression and prevent neurodegeneration.

Alzheimer Disease (AD) is the most common neurodegenerative disease, representing one of the biggest unmet needs in modern medicine (Walker, 2004). AD is characterized by the loss of neurons in association with the presence of oxidative stress, axonal dystrophy, mature senile plaques and neurofibrillary tangles (Cummings, 2004). A set of mutations in genes involved in amyloid beta and tau pathways have been associated with hereditary AD and in conjunction with neuropathological findings, the amyloid and tau hypothesis for the pathogenesis of AD has been put forward. However, current evidence suggests that sporadic AD is a multifactorial disease in which many pathways are involved (Cummings, 2004). Because the available therapies are only symptomatic (Scarpini, 2003) and considering the epidemic proportions of this disease in western countries, the development of new therapies to stop its progress is a major health priority.

In order to understand more about the basis of neurodegenerative diseases, the aim of this study was to assess the degree of connectivity between proteins whose genes were differentially expressed in MS and AD and their protein neighbors. In short, we tested whether the degree of connectivity is a property that differentiates between implicated (seed-proteins) and non-implicated nodes (neighbors). We also studied the topological properties of both MS-network and AD-network with a special focus on degree and clustering coefficient distributions.

2 Materials and Methods

2.1 Definitions

Some definitions were introduced to better explain the development of our topological studies. There are tree concept definitions for each disease and two general terms.

-MS seed-proteins: proteins whose genes were differentially expressed in previous microarray studies of MS (Bomprezzi, 2003).

-MS-neighbors: nodes selected as consequence of adding experimentally validated interactions starting from MS seed-proteins.

-MS-network: network that includes MS seed-proteins, MS-neighbors and their interactions.

Thus, MS-network nodes can be partitioned into two groups: MS seed-proteins and MS-neighbors.

-AD seed-proteins: proteins whose genes were differentially expressed in previous microarray studies of AD (Walker, 2004).

-AD-neighbors: nodes selected as consequence of adding experimentally validated interactions starting from AD seed-proteins.

-AD-network: network that includes AD seed-proteins, AD-neighbors and their interactions.

Thus, AD-network nodes can also be partitioned into two groups: AD seed-proteins and AD-neighbors.

-Disease-networks: this term is used to refer to both MS-network and AD-network.

-Degree: The so-called degree of connectivity. In this paper, it represents the number of experimentally validated interactions (links) that are connecting one protein (node) to its neighbors.

Table 1: Genes identified in MS expression profile study by Bomprezzi (2003).

Gene name	Brief protein description	k^*
JUN	Transcription factor	88
HSPA1A	Heat shock protein	66
BCL2	Apoptosis regulator	60
ZAP70	Tyrosine kinase	32
ATM	Serine kinase	26
SPTAN1	Spectrin α chain	24
MADH7	Mothers ag. decapentaplegic	18
ITGA6	Integrin alpha-6	17
TRAC	T-cell receptor region	14
HLA-DRA	MHC class II antigen	13
SCYE1	Multisynthetase complex	11
PAFAH1B1	PAF acetylhydrolase	11
SCYA3	Cytokine A3 precursor	10
IL7R	IL-7 receptor precursor	8
DNAJA1	DnaJ homolog	8
XPC	DNA-repair protein	7
SEC34	Golgi complex component	7
PPP2R5C	Ser/Thr phosphatase	4
DNTT	Nucleotidylexotransferase	4
TIMP1	Metalloproteinase inhibitor	3
SPTBN1	Spectrin β-chain	3
SERPINH2	Sphingosine kinase 2	3
TNFRSF7	TNF receptor precursor	3
GOLGA4	Golgi autoantigen	2
PTP4A1	Tyr-phosphatase	2
ZNF148	Zinc finger protein	2
CCR7	C-C chemokine receptor	2
IKKE	Inhib. NFκ-B kinase	1
NKTR	NK-tumor protein	1
DPPIV	Seprase	1
CSK2	Cyclin kinase subunit	1
DGKA	Diacylglycerol kinase	1
PIK3R4	PI-3-kinase	1
BRF1	Butyrate response factor	0
CD83	CD83 antigen precursor	0
BAZ2B	Bromodomain	0
TTC3	Tetratricopeptide protein	0
ZNF43	Zinc finger protein	0
9235	NK cells protein precursor	0
IFI30	Thiol reductase precursor	0
PDE7A	cAMP-phosphodiesterase	0
SLC35A1	CMP-sialic acid transporter	0
TCF7	Transcription factor	0
MAL	T-lymphocyte maturation	0
H1F2	Histone H1.2	0

*Degree (see section 2.3).

It is important to remark that we did not consider neighbors as new proposals for proteins implicated in disease, but they were taken to capture the network context where seed-proteins were involved.

2.2 Gene expression data

2.2.1 Multiple Sclerosis

For MS-network construction and analysis, we selected seed-proteins from previously published data (Bomprezzi, 2003) listed in Table 1. These included 45 genes differentially expressed in peripheral blood mononuclear cells from 24 MS patients with respect to 17 controls, identified by using cDNA microarrays.

2.2.2 Alzheimer Disease

The set of selected seed-proteins for AD-network modelling and analysis is listed in Table 2. It contains the gene products of 37 differentially expressed genes detected elsewhere (Walker, 2004) using cDNA-microarrays from postmortem cerebral RNA extractions in 5 normal and 4 clinically diagnosed AD patients.

2.3 Network modelling

Starting from seed-proteins involved in MS, we obtained a PI network (MS-network) throughout the interactions of these proteins. Figure 1 shows a general scheme of the approach performed in this paper. We considered a minimum of one thousand neighbors as an appropriate size to analyze the network context where the seed-proteins were involved. Hence, we expanded each disease-network until the one thousand nodes mark was reached. A depth-2 configuration allowed us to obtain 1127 neighbors in the MS-network, which included proteins directly interacting with seed-proteins. We applied the same approach using AD seed-proteins, obtaining 331 neighbors using depth-2, and 1640 neighbors using depth-3 expansion (which includes direct and with one intermediary interactions). Thus, we used depth-3 for AD-network as expanding threshold.

The growth of each network was carried out with `STRING` database (von Mering, 2003, 2005)[1]. The parameters used as criteria for network growing in the `STRING` database were: *active prediction method*: experiments; *confidence score*: 0.7 -high confidence-; *network depth*: 2-3; and *edge scaling factor*: 80%. This configuration involves just those experimental

[1] http://string.embl.de/

Table 2: Genes identified in AD expression profile study by Walker (2004).

Gene name	Brief protein description	k^*
EEF1A1	Elongation factor	70
B2M	β-2-μglobulin precursor	17
GRP58	Disulfide isomerase	16
CLU	Clusterin	14
DTNA	Dystrobrevin α	14
CD81	Surface antigen	13
HLA-B	Histocompatibility antigen	12
ATF4	Transcription factor	11
KRT8	Keratin, cytoskeletal	11
APLP1	Amyloid-like precursor	8
C4B	Complement C4 precursor	8
RAPD1GDS1	Stimulatory GTP exch.	7
183	Angiotensinogen precursor	6
CDC10	Septin 7	6
RANGAP1	GTPase-activating protein	6
FTL	Ferritin light chain	5
NEDD5	Septin 2	5
HBB	Hemoglobin beta chain	4
DMPK	DM protein-kynase	3
GSTM2	Glutathione S-transferase	3
PRDX1	Peroxiredoxin	3
MT1G	Metallothionein	2
P60201	Proteolipid protein	2
PLEKHB1	Evectin	2
HBG1	Hemoglobin epsilon chain	1
HBG2	Hemoglobin epsilon chain	1
IGHM	Ig α-1 chain C region	1
LIMS2	Senescent antigen-like	1
10099	Tetraspanin 3	0
ADD3	Gamma adducin	0
CHN2	β-chimaerin	0
OSBPL3	Oxysterol binding protein	0
PCL1	PrenylCys oxidase prec.	0
PCSK1N	Proprotein convertase inh.	0
PTS	6-pyruvoyl THB-synthase	0
RPL31	60S ribosomal protein L31	0
TU3A	TU3A protein	0

*Degree (see section 2.3).

evidences of interactions with high confidence, which were extracted from the database as valid links for each PI network. A detailed description of each parameter can be found elsewhere (von Mering, 2003). We did not consider neither the direction of each protein interaction nor self interactions.

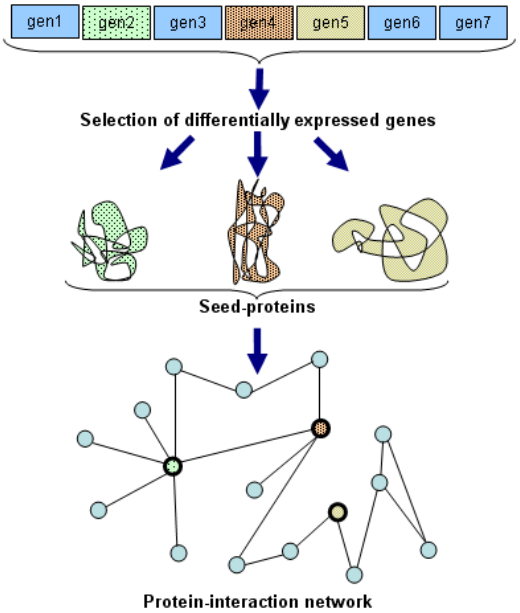


Figure 1: General scheme of the approach for each disease.

2.4 Topological analysis

We analyzed the degree distribution *P(k)*,

$$P(k) = \frac{n(k)}{N}\,, \tag{1}$$

where *k* is the number of links connected to a given node and *n(k)* the number of nodes with degree *k*.

In order to assess the degree distribution, a power law approximation (Barabási, 1999),

$$P(k) \sim k^{-\gamma}\,, \tag{2}$$

was first studied plotting *P(k)* versus non-zero *k* in log-log scale; the so-called frequency-degree. The scaling exponent γ was obtained from the slope absolute value of the least-squares fit.

The clustering coefficient for a node i with k_i neighbors, $C_i(k_i)$, represents the ratio of the number of actual connections between the neighbors of node *i* to the number of possible connections:

$$C_i(k_i) = \frac{2n_i}{k_i(k_i-1)}\,, \tag{3}$$

where n_i is the number of links connecting the k_i neighbors of node *i* to each other (Nacher, 2004).

The average clustering coefficient (i.e., the clustering of nodes with respect to *k*) provides information about the modular organization of networks (Almaas, 2006),

$$C(k) = \frac{\sum_{i:k_i=k} C_i(k_i)}{n(k)} . \quad (4)$$

The rank-degree distribution, *Rank(k)*, represents the number of nodes with a degree greater than *k*,

$$Rank(k) = n(K > k) = \sum_{k'=k+1}^{m} n(k') , \quad (5)$$

with *m* being the maximum degree found (Tanaka, 2005b).

It is important to notice that *Rank(k)* provides precise information to elucidate whether degree distribution is better explained either as a power law or as an exponential distribution (Tanaka, 2005b).

P(k), *C(k)* and *Rank(k)* were calculated from the `STRING` files (simple tab delimited flatfiles), which contain all the nodes and interactions obtained using the methodology described in section 2.3.

2.5 Gene Ontology

In order to assess the multifactorial character of the biological pathways in which seed-proteins were involved following the *Gene Ontology*[2] (GO), we studied those pathways arising when genes corresponding MS seed-proteins or AD seed-proteins were compared to complete human proteome (Swissprot identifiers). This approach was carried out using the `FatiGO` web tool (Al-Shahrour, 2004)[3].

2.6 Statistical analysis

Frequency-degree linear approximation were carried out with the least squares fitting method. Rank-degree curves were also fitted as a second order exponential decay. To compare the degree distribution between all the network nodes and the seed-nodes, we performed an ANOVA test using `Statgraphics Plus 5.1` software. The regression lines were compared using the *Comparison of Regression Lines Analysis Dialog Box*, which automatically constructs the necessary indicator variables for comparing two or more simple regression models. Finally, we used a U Mann-Whitney test to compare degree between seed-proteins and neighbors for each disease.

The level of significance was set at $p<0.01$. P-values associated to pathways under `FatiGO`[3] analysis were corrected by false discovery rate multiple comparison method.

[2]http://www.geneontology.org/

[3]http://fatigo.bioinfo.cipf.es/

3 Results

3.1 Multiple sclerosis

The MS-network is shown in Figure 2. This map contains 1172 nodes, including 45 seed-proteins and 1127 neighbors. Twelve seed-proteins had no links (i.e., no experimental evidence of interactions), and 8 nodes (including 1 seed-protein) formed an independent small fully interlinked net. The degree corresponding to each MS seed-protein node is listed in Table 1.

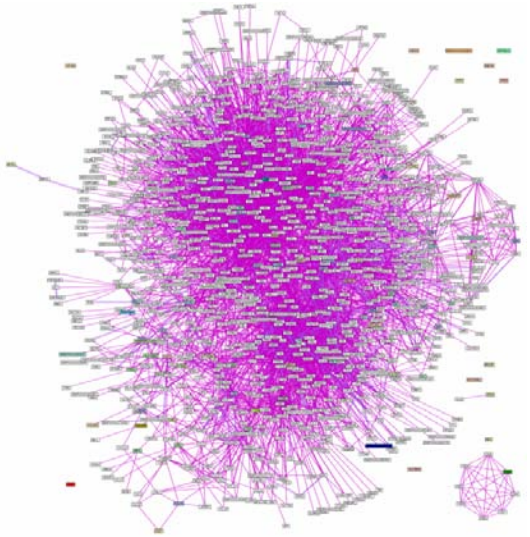


Figure 2: Protein interaction of the Multiple Sclerosis network (MS-network).

The frequency-degree distribution, the clustering coefficient and the rank-degree distribution of MS-network nodes are shown in Figure 3. We found a dependence of *P(k)* respect to *k* in both MS-network nodes and MS seed-proteins. According to equation (2), the linear approximations (slope = -1.48 and slope = -0.43) explained these dependencies as power laws $P(k) = k^{-1.48}$ and $P(k) = k^{-0.43}$ respectively. Furthermore, clustering coefficient also followed a power law respect to *k*, with (γ = 0.68). On the other hand, rank-degree plot revealed that degree distribution was better explained as exponential ($R^2 = 0.99$) instead of linear ($R^2 = 0.90$).

The comparison of MS-network and MS seed-proteins regression lines provided a statistically significant difference ($p<0.001$). Furthermore, degree group comparison indicated that seed-proteins degree was significantly lower than MS-neighbors degree ($p=0.003$).

Finally we assessed whether some functional pathways were overrepresented in the genes set corre-

sponding to MS seed-proteins. We found that none of the 22 biological modules detected (GO level 3) were statistically overrepresented.

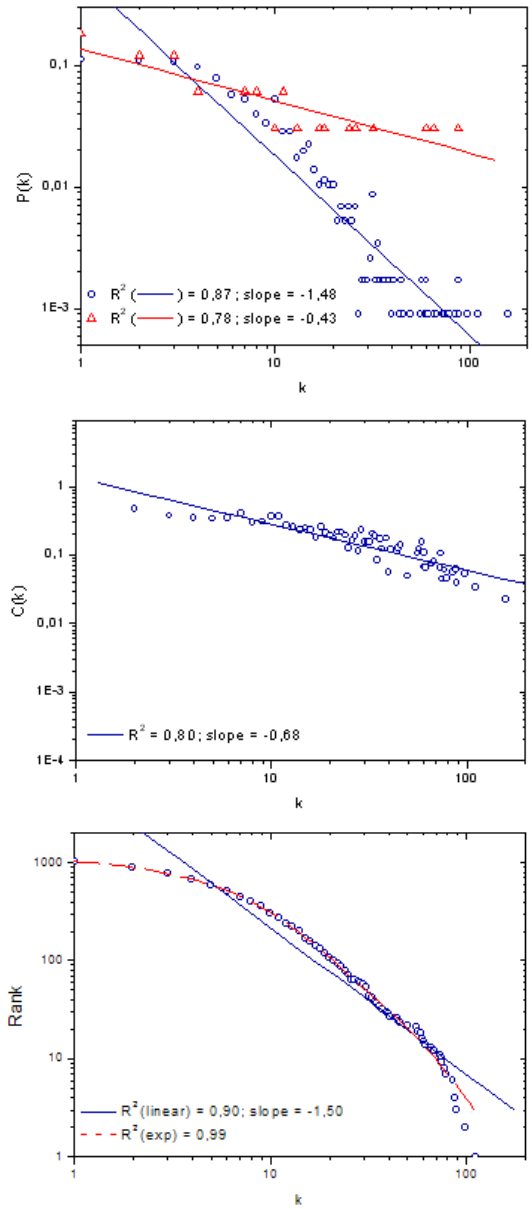


Figure 3: Topological analysis of the MS-network (log-log plots). Top: frequency-degree distribution of MS-network nodes (blue-circles) and MS seed-proteins (red-triangles). Middle: clustering coefficient distribution of MS-network nodes. Bottom: rank-degree distribution of MS-network nodes.

3.2 Alzheimer Disease

The AD-network contains 1687 nodes, including 47 seed-proteins and 1640 neighbors (figure 4). Table 2 includes the degree of connectivity for each seed-node.

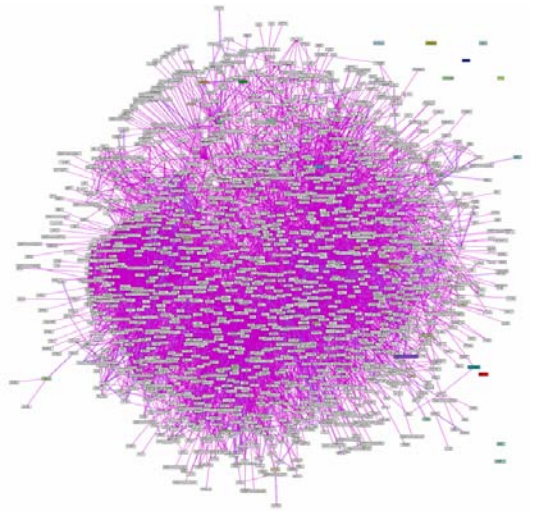


Figure 4: Protein interaction of the Alzheimer Disease network (AD-network).

The frequency-degree distribution, the clustering coefficient and the rank-degree distribution of AD-network nodes are shown in Figure 5. We found a dependence of *P(k)* respect to *k* in both AD-network nodes and AD seed-proteins. According to equation (2), the linear approximations (slope = -1.58 and slope = -0.36) explained these dependencies as power laws $P(k) = k^{-1.58}$ and $P(k) = k^{-0.36}$ respectively. On the other hand, clustering coefficient resulted to be independent from *k* (γ = 0.28 and $R^2 = 0.08$). Finally, rank-degree plot revealed that degree distribution was better explained as exponential ($R^2 = 0.99$) instead of linear ($R^2 = 0.81$).

The comparison of AD-network and AD seed-proteins regression lines provided a statistically significant difference ($p<0.001$). Furthermore, degree group comparison indicated that seed-proteins degree was significantly lower than AD-neighbors degree ($p=0.005$). These results were very similar to the obtained in the MS study in section 3.1.

Finally we assessed whether some functional pathways were overrepresented in the genes set corresponding to AD seed-proteins. We found that none of the 9 biological modules detected (GO level 3) were statistically overrepresented.

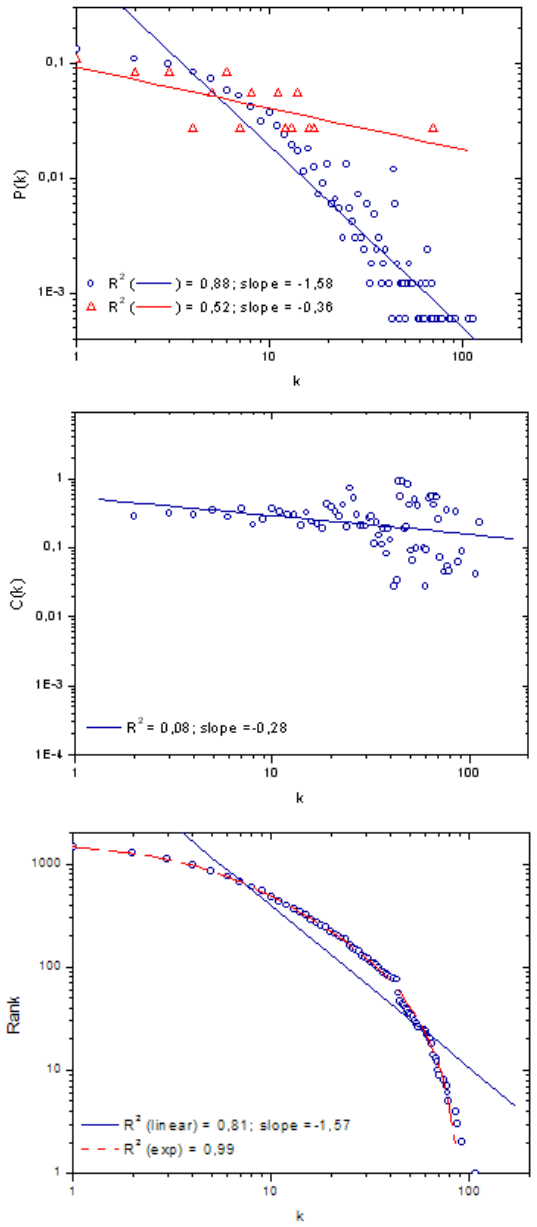


Figure 5: Topological analysis of the AD-network (log-log plots). Top: frequency-degree distribution of AD-network nodes (blue-circles) and AD seed-proteins (red-triangles). Middle: clustering coefficient distribution of AD-network nodes. Bottom: rank-degree distribution of AD-network nodes.

3.3 Common characteristics between MS and AD networks

As indicated in Tables 1 and 2, we found a low degree in seed-proteins respect to the degree of its PI neighbors in both diseases, with only 3 (MS) and 1 (AD) highly connected seed-proteins (k>60). In addition, direct interactions between seed-proteins were very low: 4 direct interactions in MS-network, and 5 in AD-network. There were 586 common proteins to MS and AD networks. In order to detect the possible topological distribution relationship between both neurodegenerative disorders, the disease networks and the seed-proteins sets were independently compared through regression line analysis (Figure 6), and no significant differences between slopes were detected neither between disease-networks (p=0.31) nor between seed-proteins (p=0.52).

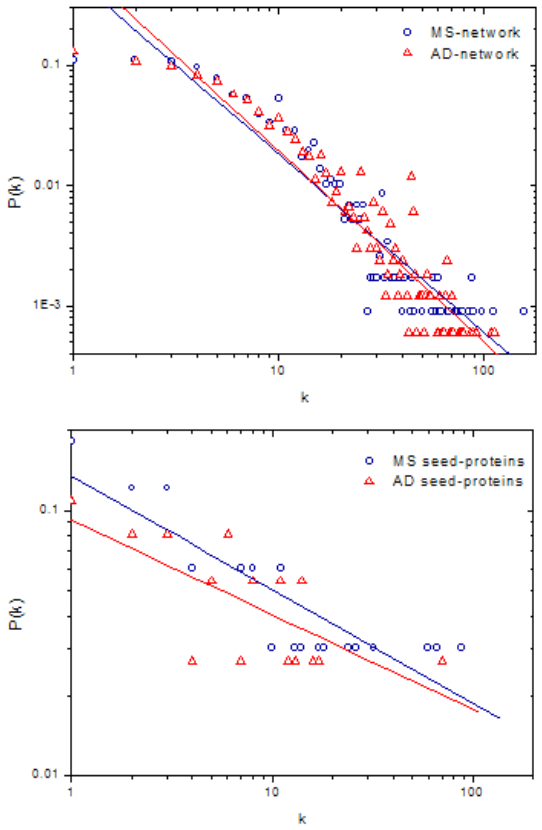


Figure 6: Linear regression comparison between disease-networks and seed-proteins.

4 Discussion

4.1 Topology of MS and AD networks

Network theory can provide a useful tool to study the complexity of neurodegenerative diseases. In the present study we report a novel approach to study PI networks, based on the products of differentially expressed genes of MS and AD. The network growth was carried out expanding the network through experimentally validated protein interactions.

Network stability, dynamics and function is generally characterized by determining the topology of the map, i.e., the configuration of its nodes and the connecting edges (Han, 2005). One of the most succeeding features analyzed has been the degree distribution, and whether or not it followed a power law; the so-called scale-free property. Roughly speaking, these studies were aimed to characterize the properties of real networks in basis on their topological features. For example, networks with a SF topology are known to be resistant to random failure and vulnerable to targeted attack, specifically against the most connected nodes (hubs). However, it has recently been shown that metabolic networks are supported by different modular scales, with a power law degree distribution of the global system but an exponential behavior in modules. As such they are described as scale-rich (SR) networks (Tanaka, 2005a).

Our results reveal that MS-network and AD-network may be better adjusted to exponential than to power law distributions. The power law behavior of *P(k)* together with the dependence of *C(k)* versus k, point to a scale-free distribution (SF) with an inherent modularity of the MS-network. However, the exponential behavior of the rank distribution agree with a SR -instead of a SF- topology. The particular degree distribution of the seed-proteins involved in MS also showed a power law topology. In the case of AD, γ exponent values of AD-network nodes and AD seed-proteins were very similar to the obtained in MS-network nodes and MS seed-proteins respectively. Although a constant behavior of *C(k)* in AD-network also point to a scale-free distribution (SF), the exponential behavior of the rank distribution agree with a SR topology. In this sense, recent studies support that other biological networks, such as the complete human PI map, filtered yeast interactome dataset and metabolic networks, could be better explained as SR networks (Tanaka, 2005a,b). These networks would not be so hub-dependent as SF are, and could be formed by exponential subnetworks and critical nodes that might not be so k-dependent (according with both MS-network and AD-network). In addition, the clustering coefficient dependence of k point to a inherent modularity of the MS-network, but not completely reaching the characteristic hierarchical network. Hierarchical modularity is detected by the scaling of the clustering coefficient, which should follow $C(k) \sim k^{-1}$ (i.e., a straight line of slope -1 on a log-log plot) (Barabási, 2004).

On the other hand, both the MS and AD network may be considered as proportional SR samples of the complete human protein-interaction map recently studied (Tanaka, 2005b). This is particularly the case if we take into account that the rank-degree distribution is exponential and the frequency-degree distribution is linear, with very similar exponents. Regarding the identification of common properties among those genes involved in neurodegenerative disorders, these facts provided very interesting results under a topological analysis: multiple pathways affected by proteins with low degree, following very similar SR distributions (Tanaka, 2005a).

4.2 Seed-proteins connectivity

During the last decade, network studies have been applied to biological data bearing in mind that the degree of connectivity is a key property of any network. The most common approach to identify key nodes consists of obtaining networks from high throughput data and having obtained the network, searching for the highest connectivity nodes (hubs). The underlying assumption was that these hubs could be critical to explain pathogenesis of diseases.

Our study was performed from a novel viewpoint. Starting from critical nodes (in terms of differentially expressed genes) we analyzed whether the connectivity was higher or lower than the connectivity of the PI neighbors. Thus, we found that seed-proteins connectivity was lower than PI neighbors degree. To our knowledge, these properties have not been reported before and they situate seed proteins in peripheral regions of the network, distributed along several pathways that could be involved in disease. Somehow this enters into conflict with the hub relevance hypothesis (Jeong, 2001), at least in these two neurodegenerative diseases. Therefore our results support the application of strategies other than those previously applied, whereby only hubs that could compromise the robustness of networks were generally searched (Barabási, 2004).

The etiology of MS and sporadic AD still remains elusive and many environmental and genetic factors have been proposed. In this sense, our results strongly support the multifactorial nature of these diseases, due to the fact that many pathways participates somehow in both diseases, any of them being predominant.

4.3 Etiopathogenic and therapeutical implications

We can consider complex diseases as an evolutionary stage in which the pathogenesis process hijacks the robustness of the biological pathways; this may be followed by cascading failures in such pathways (Ki-

tano, 2004, 2006). In this sense, it may be necessary to target many of the pathways involved, although following a systems biology rationale based on the dynamics and topology of the networks involved. The aim of this therapy would be to drive those pathways to a non-pathological state or at least, to a less deleterious state.

The topological implications of the claimed SF properties of biological networks suggests that the best therapeutic targets, in order to modify the network behavior, should be the genes (or proteins) corresponding to hubs in the network. However, our findings suggest that low connected proteins might be more appropriate therapeutic targets, at least in neurodegenerative diseases, than hyper-connected ones.

The fact that in both diseases (MS and AD) and in two different tissues analyzed (blood and cerebral tissue), seed-proteins were low connected nodes taking part in many different pathways, strengthen the multifactorial pathogenesis of neurodegenerative diseases. Our results suggest that in order to modify the disease course we need to target many genes or proteins in several pathways. Another reason why hubs might not be good therapeutic targets is because their critical role in the network modules might prevent them from fluctuating substantially. For the same reason, we can speculate that networks will poorly tolerate the modification of hub behavior without spreading such changes along the network and, in this way, inducing significant side effects.

4.4 Conclusions

The results presented in this paper indicate that both neurodegenerative diseases (MS and AD) share as common characteristics the low degree of seed-proteins and the degree-distribution similarities found between disease networks, even though many different pathways are involved depending on the disease. These findings locate seed-proteins mainly in peripheral regions of the PI map (in terms of degree), involved in many pathways (as FatiGO and low direct interactions results indicated) and integrated in two subnetworks (respect to Human complete proteome network) with very similar exponential degree distributions but with different modular organization. In addition, and as stated before, no significant biological process were overrepresented to the seed-proteins of MS and AD analyzed when compared to the whole human genome. This results are likely to be explained as a consequence of the multifactorial nature of both diseases.

Acknowledgements

We thank Cristian von Mering and Iván Martínez for helping us with the use of STRING database, Marc Sefton for English revision and Ricard Solé and Bartolomé Bejarano for their comments and suggestions. J.G, N.VdM. and J.S. are fellows supported by *Navarra Government* (Spain), *Basque Country Government* (Spain) and *Spanish Ministry of Health* (CM05/00222), respectively. This work was partially supported by *Junta de Andalucía* (CVI-302; to F.J.E.), *Spanish Ministry of Science and Technology* (BFM2003-05878 to S.A.T.) and *Spanish Ministry of Health* (FIS: 04/1445; to P.V.).

References

E. Almaas and A.L. Barabási. Power Laws in Biological Networks. *Power Laws, Scale-free Networks and Genome Biology*, E.V. Koonin, G. Karev and Y. Wolf (Eds.), Springer, New York (in press) 2006.

F. Al-Shahrour, R. Díaz-Uriarte and J. Dopazo. FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics,*, 20(4):578–580, 2004.

A.L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

A.L. Barabási and Z.N. Oltvai. Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, 5(2):101–113, 2004.

R. Bomprezzi, M. Ringner, S. Kim, M.L. Bittner, J. Khan, Y. Chen, A. Elkahloun, A. Yu, B. Bielekova, P.S. Meltzer, R. Martin, H.F. McFarland and J.M. Trent. Gene expression profile in multiple sclerosis patients and healthy controls: identifying pathways relevant to disease. *Human Molecular Genetics*, 12 (17):2191–2199, 2003.

W. Bruck. The pathology of multiple sclerosis is the result of focal inflammatory demyelination with axonal damage. *Journal of Neurology*, 252 (Suppl.5):v3–v9, 2005.

J.L. Cummings. Alzheimer disease. *New England Journal of Medicine*, 351(1):56–67, 2004.

G.H. Fernald, R.F. Yeh, S.L. Hauser, J.R. Oksenberg JR and S.E. Baranzini. Mapping gene activity in complex disorders: Integration of expression and

genomic scans for multiple sclerosis. *Journal of Neuroimmunology*, 167(1-2):157–169, 2005.

K.C. Gunsalus, H. Ge, A.J. Schetter, D.S. Goldberg, J.D. Han, T. Hao , G.F. Berriz, N. Bertin, J. Huang, L.S. Chuang, N. Li, R. Mani, A.A. Hyman, B. Sonnichsen, C.J. Echeverri, F.P. Roth, M. Vidal, F. Piano. Predictive models of molecular machines involved in Caenorhabditis elegans early embryogenesis. *Nature*, 436(7052):861-865, 2005.

J.D. Han, D. Dupuy, N. Bertin, M.E. Cusick and M. Vidal. Effect of sampling on topology predictions of protein-protein interaction networks. *Nature Biotechnology*, 23(7):839–844, 2005.

P.R. Hiesinger and B.A. Hassan. Genetics in the age of systems biology. *Cell*, 123 -1173-1174, 2005.

J. Imitola, T. Chitnis, S.J. Khoury. Insights into the molecular pathogenesis of progression in multiple sclerosis: potential implications for future therapies. *Archives in Neurology*, 63(1):25–33, 2006.

H. Jeong, S.P. Mason, A.L. Barabási and Z.N. Oltvai. Lethality and centrality in protein networks. *Nature*, 411(6833):41–42, 2001.

H. Kitano, K. Oda, T. Kimura, Y. Matsuoka, M. Csete, J.Doyle and M. Muramatsu. Metabolic syndrome and robustness tradeoffs. *Diabetes*, Suppl 3: S6–S15, 2004.

H. Kitano and K. Oda. Robustness tradeoffs and host-microbial symbiosis in the immune system. *Molecular Systems Biology*, 2(1):msb4100039-E1–msb4100039-E10, 2006

H. Lodish, A. Berk, S.L. Zipursky, P. Matsudaira, D. Baltimore and J. Darnell. *Molecular Cell Biology (4^{th} Ed.)*, W.H. Freeman and Company, New York, 2000.

J.C. Nacher, N. Ueda, T. Yamada, M. Kanehisa and T. Akutsu. Study on the clustering coefficients in metabolic network using a hierarchical framework. *International Workshop on Bioinformatics and Systems Biology*, Poster Abstracts:34–35, 2004

J.R. Oksenberg, S.E. Baranzini, L.F. Barcellos and S.L. Hauser. Multiple sclerosis: genomic rewards. *Journal of Neuroimmunology*, 113(2):171–184, 2001.

D.R. Rhodes and A.M. Chinnaiyan. Integrative analysis of the cancer transcriptome. *Nature Genetics*, 37(Suppl):S31–S37, 2005.

E. Scarpini, P. Scheltens, H. Feldman. Treatment of Alzheimer's disease: current status and new perspectives. *Lancet Neurology*, 2(9):539–547, 2003.

L. Steinman. Multiple sclerosis: a two-stage disease. *Nature Immunology*, 2(9):762–764, 2001.

M. Stetter, B. Schürmann, and M. Dejori. Systems level modeling of gene regulatory networks. In: *Artificial Intelligence Methods and Tools for Systems Biology*, W. Dubitzy and F. Azuaje (*Eds.*), Springer, The Netherlands, 2004.

R. Tanaka. Scale-rich metabolic networks. *Physical Review Letters*, 94(16):168101, 2005.

R. Tanaka, T.M. Yi and J. Doyle. Some protein interaction data do not exhibit power law statistics. *FEBS Letters*, 579(23):5114–5144. 2005.

P. Villoslada and J. Oksenberg. Neuroinformatics in clinical practice: are computers going to help neurological patients and their physicians?. *Future Neurology*, 1(2):1-12, 2006.

C. von Mering, M. Huynen, D. Jaeggi, S. Schmidt, P. Bork and B. Snel. STRING: a database of predicted functional associations between proteins. *Nucleic Acids Research*, 1(31):258–261, 2003.

C. von Mering, L.J. Jensen, B. Snel, S.D. Hooper, M. Krupp, M. Foglierini, N. Jouffre, M.A. Huynen and P. Bork. STRING: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Research*, 1(33):D433–D437, 2005.

P.R. Walker, B. Smith, Q.Y. Liu, A.F. Famili, J.J. Valdes, Z. Liu and B. Lach. Data mining of gene expression changes in Alzheimer brain. *Artificial Intelligence in Medicine*, 31(2):137–154, 2004.

Y. Xia, H. Yu, R. Jansen, M. Seringhaus, S. Baxter, D. Greenbaum, H. Zhao and M. Gerstein. Analyzing cellular biochemistry in terms of molecular networks. *Annual Review of Biochemistry*, 73: 1051–1087, 2004.

Analytical and numerical results for entrainment in large networks of coupled oscillators.

Peter Grindrod*
*Lawson Software
Magdalen Centre, Oxford Science Park, Oxford, UK
Peter.Grindrod@Lawson.com

Zhivko Stoyanov†
†Department of Mathematical Sciences, University of Bath
Bath BA2 7AY
mapzvs@maths.bath.ac.uk

Abstract

In this paper we consider large networks of coupled oscillators. We choose to illustrate this using a general class of range dependent networks where the pairwise coupling is a probabilistic function of distance (range) between the nodes, and each node represents an oscillator with its own intrinsic phase and natural frequency of oscillation. Range dependent networks exhibit the "small world" phenomenon, being effectively superpositions of many networks each operating over different range lengths. We provide an asymptotic analysis in terms of a network coupling parameter that gives a simple analytic description of the coupled dynamics and which agrees well with numerical experiments.

1 Introduction

The emergent behaviour of populations of dynamical systems brought about by local pairwise (weak) coupling is of interest both from the point of view of the group dynamics and the theory and characterisation of the underlying networks. Perhaps the simplest nontrivial examples are coupled oscillatory systems where a local "diffusive" type of coupling gives rise to spatio-temporal patterns such as localised non planar waves, target patterns, or spiral waves. When the underlying network allows "longer range" couplings also even simple entrainment phenomena are not straightforward. By entrainment (or synchronisation) of a system of oscillators, we mean a state of the system, in which all oscillators move together as one with a possible difference in their phases, which remains constant for large time. This is a key concept in the understanding of self-organisation phenomena of coupled oscillators (see, for example, (K84)).

In (K75) Kuramoto considered networks of oscillators, in which the coupling between every pair of oscillators was identical. Although simple at a glance, his model was hard to analyse but due to his ingenious heuristics and assumptions, he was able to derive some properties about the system he considered.

In this paper we analyse entrainment in large networks of coupled oscillators (see, for example, (S00), (AS04)). We choose to illustrate this using a general class of range dependent networks where the pairwise coupling is a probabilistic function of distance (range) between the nodes, and each node represents an oscillator with its own intrinsic phase and natural frequency of oscillation. Range dependent networks exhibit the "small world" phenomenon, being effectively superpositions of many networks each operating over different range lengths (see, for example, (G02)). We provide an asymptotic analysis in terms of a network coupling parameter that gives a simple analytic description of the coupled dynamics and which agrees well with numerical experiments.

2 Twin coupled oscillators

First, consider the simplest case of two coupled oscillators:

$$\dot{\theta}_1 = \lambda_1 + \varepsilon A_{12} \sin(\theta_2 - \theta_1)$$
$$\dot{\theta}_2 = \lambda_2 + \varepsilon A_{21} \sin(\theta_1 - \theta_2)$$

which has state space the torus with coordinates $\theta_i \mod (2\pi)$ for $i = 1, 2$. Here the A_{jk} are nonnegative coupling coefficients; ε is a nonnegative overall "strength" parameter to scale the coupling; and the $\lambda_i > 0$ represent the uncoupled frequencies of the

separate oscillators. Setting $\phi = \theta_2 - \theta_1$ we obtain a single equation for the phase difference:

$$\dot{\phi} = \lambda_2 - \lambda_1 - \varepsilon(A_{12} + A_{21}) \sin \phi, \qquad (1)$$

which is integrable and so a closed form solution is available. However, qualitative information about the oscillation can be obtained directly from (1). First note that the frequencies become entrained for large time (with ϕ tending to a stable rest point) if and only if ε is such that

$$|\lambda_2 - \lambda_1| < \varepsilon(A_{12} + A_{21}).$$

If this condition does not hold one of the oscillators repeatedly "laps" the other.

3 N oscillators coupled via a directed graph

Let us generalise the above situation to N coupled oscillators. We shall think of them as vertices connected by a directed graph with entraining couplings defining the non negative weights of directed edges. Each oscillator is represented by a single phase variable, $\theta_i \mod (2\pi)$, having a natural, uncoupled frequency: whilst each coupling term, say from oscillator k acting on oscillator i, affects to increase or retard the rate of increase of the phase of oscillator i, so as to approach the phase of oscillator k. The state space for the full coupled system is an N dimensional torus with coordinates $\theta_i \mod (2\pi)$ for $i = 1, \ldots, N$. Specifically, we consider the following system on the N-torus:

$$\dot{\theta}_i = \lambda_i + \varepsilon \sum_{k=1}^{N} A_{ik} \sin(\theta_k - \theta_i), \quad i = 1, \ldots, N. \qquad (2)$$

Introduce the $n \times n$ coupling matrix A with zeros on the diagonal and jk^{th} component A_{jk}, which represents the weight of the coupling, or edge, from vertex j to vertex k. The parameter ε is a nonnegative overall "strength" parameter to scale the impact of A; and the $\lambda_i > 0$ represent the uncoupled frequencies of the separate oscillators.

Our interest is in whether and how the oscillators can become entrained with one another, for large time; producing a baulk oscillation, with their phases moving together, possibly separated by a constant set of phase shifts. Like the simple twin-oscillator case this behaviour depends upon the strength and nature of the couplings as well as the distribution of their natural frequencies.

3.1 No Baulk Oscillations for small ε

For any i and j we have

$$\dot{\theta}_i - \dot{\theta}_j = \lambda_i - \lambda_j +$$

$$\varepsilon \left(\sum_{k=1}^{N} A_{ik} \sin(\theta_k - \theta_i) - \sum_{k=1}^{N} A_{jk} \sin(\theta_k - \theta_j) \right).$$

The left hand side of this equation must vanish when oscillators i and j are entrained (that is when their phases differ by a constant amount through time). Set

$$\varepsilon^* = \max_{1 \le i,j \le N} \frac{|\lambda_i - \lambda_j|}{\sum_{k=1}^{N} (A_{ik} + A_{jk})}. \qquad (3)$$

Then if $\varepsilon < \varepsilon^*$, $\dot{\theta}_i = \dot{\theta}_j$ is impossible for at least one pair of oscillators and there can be no baulk oscillation. Note this condition is necessary and sufficient for no baulk oscillation to exist when $N = 2$.

3.2 Asymptotic Analysis of Baulk Oscillations for large ε

We seek an asymptotic solution, valid in the limit of large ε, representing a baulk oscillation, so that for some function, $\theta_0(t)$ say, we have

$\theta_i(t) = \theta_0(t) +$ an ε-dependent phase shift for oscillator i

for each $i = 1, \ldots, N$.

Setting $\vec{\theta}(t) = (\theta_1(t), \theta_2(t), \ldots, \theta_N(t))^T$, $\vec{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_N)^T$ and $\vec{1} = (1, 1, \ldots, 1)^T \in \mathbb{R}^N$, we shall seek a solution which is in the form of a baulk oscillation (that is, all phases entrained) where the phase shifts are represented by a regular expansion in inverse powers of ε:

$$\vec{\theta}(t) = \theta_0(t)\vec{1} + \frac{1}{\varepsilon}\vec{\theta}_1 + \frac{1}{\varepsilon^2}\vec{\theta}_2 + \mathcal{O}\left(\frac{1}{\varepsilon^3}\right). \qquad (4)$$

Here $\vec{\theta}_1$ and $\vec{\theta}_2$ are vectors orthogonal to $\vec{1}$, so that the individual phase shifts are distinct from the baulk oscillation term.

Substituting (4) into (2) and expanding out the sine terms, we obtain

$$\dot{\theta}_0\vec{1} = \vec{\lambda} + \triangle\vec{\theta}_1 + \frac{1}{\varepsilon}\triangle\vec{\theta}_2 + \mathcal{O}\left(\frac{1}{\varepsilon^2}\right). \qquad (5)$$

Here $\triangle$ denotes the "Laplacian" matrix associated with the network coupling matrix A (replacing the zeroes on the diagonal of A with the negative of the corresponding row sums):

$$\triangle = A - \text{diag}(A\vec{1}).$$

The Laplacian matrix $\triangle$ contains information about the connected nature of the network: it is of huge importance in graph theory (B95). It is easy to see that zero is an eigenvalue of $\triangle$ with multiplicity equal to the number of distinct connected sub networks. Without loss of generality we shall assume zero is a simple eigenvalue - otherwise we may consider each connected sub network separately. In that case $\triangle\vec{1} = \vec{0}$.

Let e denote the corresponding left unit eigenvector: $e^T\triangle = 0^T$. Then pre-multiplying (5) with e^T we have

$$\dot{\theta}_0 e^T\vec{1} = e^T\vec{\lambda} + \mathcal{O}\left(\frac{1}{\varepsilon^2}\right), \qquad (6)$$

which determines $\theta_0(t)$. (In the case when $\triangle$ is a symmetric matrix the term $\mathcal{O}\left(\frac{1}{\varepsilon^2}\right)$ in the right hand side of (6) vanishes.) Then to $\mathcal{O}(1)$ and $\mathcal{O}\left(\frac{1}{\varepsilon}\right)$ we have $\vec{\theta}_1$ and $\vec{\theta}_2$ respectively, determined as the unique solutions, orthogonal to $\vec{1}$, of the matrix equations:

$$\left(\frac{e^T\vec{\lambda}}{e^T\vec{1}}\right)\vec{1} - \vec{\lambda} = \triangle\vec{\theta}_1, \qquad \triangle\vec{\theta}_2 = \vec{0}. \qquad (7)$$

First note that $\vec{\theta}_2 = \vec{0}$. Next, we may write

$$\vec{\theta}(t) = \left(t\left(\frac{e^T\vec{\lambda}}{e^T\vec{1}}\right) + C\right)\vec{1} + \frac{1}{\varepsilon}\vec{\theta}_1 + \mathcal{O}\left(\frac{1}{\varepsilon^3}\right), \qquad (8)$$

where C is a constant, and $\vec{\theta}_1$ can be found by solving (7) in the subspace orthogonal to $\vec{1}$.

Hence by calculating e^T, the left eigenvector of $\triangle$ and solving for $\vec{\theta}_1$ from (7), we can use (8) to estimate the behaviour of the oscillators for large coupling parameter ε. In fact the experiments in the next section show that ε needs not be too large. Indeed, for values of ε not too much greater than ε^*, (8) provides an accurate representation of the behaviour of the system.

Finally, we note that for the network considered here the second eigenvalue of $\triangle$ is small (equalling -0.01528) with corresponding eigenvector $\vec{v}$, often called the Fiedler vector (F75). Hence $\vec{\theta}_1$ will typically be rich in the direction of $\vec{v}$. Now $\vec{v}$ is often used to explain certain network features (for example, clustering) and this suggests that the Fiedler vector $\vec{v}$ might also provide information to help understand different features in the solutions of (2).

4 Numerical Example

Example: We take $N = 100$, A a symmetric random range dependent matrix with values lying between zero and 0.96, and the λ_i as independent uniformly distributed random numbers within the interval $[0.5; 1.5]$. Then by direct calculation, $\varepsilon^* = 0.47884$.

In this case $e^T = \frac{1}{\sqrt{N}}\vec{1}$ and so we have from (6)

$$\dot{\theta}_0 = \frac{1}{N}\sum_{i=1}^{N}\lambda_i =: \hat{\lambda}.$$

Hence (8) gives

$$\theta_i(t) = \hat{\lambda}t + C + \frac{1}{\varepsilon}\theta_1^{[i]} + \mathcal{O}\left(\frac{1}{\varepsilon^3}\right),$$

where $\theta_1^{[i]}$ denotes the i^{th} component of $\vec{\theta}_1$, and

$$\theta_i(t) - \theta_j(t) = \frac{1}{\varepsilon}(\theta_1^{[i]} - \theta_1^{[j]}) + \mathcal{O}\left(\frac{1}{\varepsilon^3}\right). \qquad (9)$$

In Figure 1 we plot the phase differences, $\theta_i(t) - \theta_1(t)$ for $i = 2, \ldots, 100$, obtained directly from the numerical solution, for $t \in [0; 50]$, for various values of ε ($\varepsilon = 0.5, 0.6, 0.8, 2.0, 5.0, 10.0$).

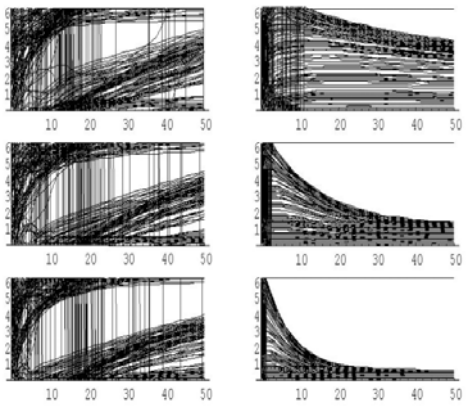


Figure 1: Plot of $\theta_i - \theta_1$, for $i = 2, \ldots, 100$, versus time t, for $\varepsilon = 0.5, 0.6, 0.8, 2.0, 5.0, 10.0$.

The entrainment as ε increases is clearly seen in Figure 1. Indeed, for $\varepsilon = 2.0$ the system settles to baulk oscillation before $t = 250$. In Figure 2 we compare the values of $\theta_i(t) - \theta_1(t)$ obtained by numerical solution with $\frac{1}{\varepsilon}(\theta_1^{[i]} - \theta_1^{[1]})$ in order to test the validity of (9), and hence the validity of the asymptotic analysis leading to equation (8). Clearly, even for ε not so large there is very good agreement between the asymptotic expression and numerical experiment, with the maximum error being around 1.3×10^{-4}.

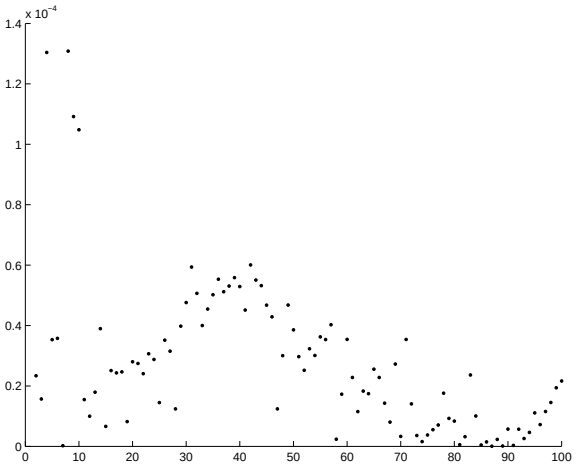


Figure 2: In this Figure we plot, for $i = 2, \ldots, 100$, the absolute value of the difference between $\theta_i(t) - \theta_1(t)$ (obtained by numerical solution of (2)) and $\frac{1}{\varepsilon}(\theta_1^{[i]} - \theta_1^{[1]})$, see (9). Here $\varepsilon = 2$ and $t = 250$.

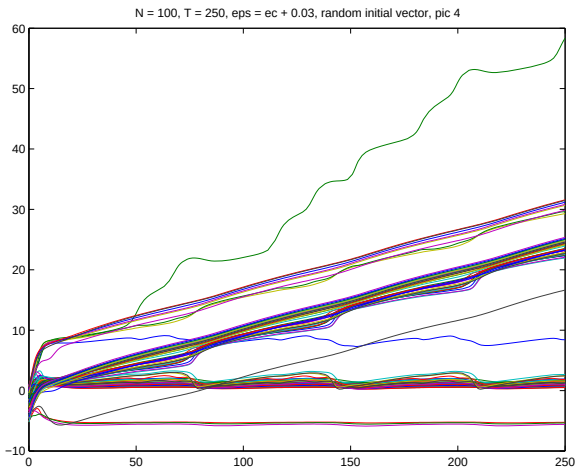


Figure 3: Plot of $\theta_i - \theta_{i_0}$ versus time, for $\varepsilon = \varepsilon^* + 0.03$.

Lastly, in Figure 3 we show the solution behaviour for the system with $\varepsilon = \varepsilon^* + 0.03$ for random starting values. This Figure represents the plot of the terms $\theta_i(t) - \theta_{i_0}(t)$, where $t \in [0; 250]$ and i_0 is such that $\lambda_{i_0} \leq \lambda_i$ for $1 \leq i \leq 100$. In our simulation in Figure 3 we observe that there are two clusters of oscillators entrained with θ_{i_0} and two other clusters which drift away from them. There is an "extreme" oscillator, which is not entrained to any of the groups, and two other oscillators, which seem to be attracted by the clusters of oscillators.

References

[AS04] D. Abrams and S. Strogatz, "Chimera States for Coupled Oscillators", Phys. Rev. Let., 93, (17), (2004), 174102.

[B95] B. Bollobas, "Random Graphs", Academic Press, New York (1995).

[F75] M. Fiedler, "A property of eigenvectors of nonnegative symmetric matrices and its applications to graph theory", Czech. Math. Journal, 25, (1975), pp 619-633

[G02] P. Grindrod, "Range dependent random graphs and their application to modelling small world proteomic data sets", Phys. Rev. E, 66, (2002), 0667021-0667027.

[K75] Y. Kuramoto, "Self-entrainment of a population of coupled nonlinear oscillators", H. Araki (ed.) "International Symposium on Mathematical Problems in Theoretical Physics", Lecture Notes in Physics, 39, 420-422, Springer, New York, (1975).

[K84] Y. Kuramoto, "Chemical Oscillators, Waves, and Turbulence", Springer Verlag, Heidelberg, (1984)

[S00] S. Strogatz, "From Kuramoto to Crawford: exploring the onset of synchronization in populations of coupled oscillators", Physica D 143 (2000) 1-20.

Scale-free structure is not the best for the shortest path length and the robustness

Yukio Hayashi[⋆†]
[⋆]Japan Advanced Institute of Science and Technology
1-1, Asahidai, Nomi, Ishikawa, Japan
yhayashi@jaist.ac.jp

Jun Matsukubo[†]
[†]Kitakyusyu National College of Technology
5-20-1 Shii, Kokuraminami, Fukuoka, Japan
jmatsu@kct.ac.jp

Abstract

We study geographical networks on a planar space, and show that the cutoff of degree improves the path length and the tolerance to failures and attacks. We also compare them with the randomly rewired non-geographical versions. These results are useful for constructing sensor or ad hoc networks.

1 Introduction

In complex network science, the topological structure called *small-world* or *scale-free* attracts interdisciplinary research fields, since it has been commonly found in many social, biological, and technological systems (Barabási, 2002). The heterogeneous structure with many low degree nodes and a few hubs has good properties in the meanings of economical and efficient communication by small number of hops in a connected network with a few links (Cancho and Solé, 2003) and the robustness against failures (Albert and Barabási, 2000). Moreover, the restriction of link lengths has been observed, e.g. Internet at both router and Autonomous System levels (Yook et al., 2002), road networks, and flight-connection in a major airline (Gastner and Newman, 2006) on geographical spaces. Recent studies of scaling relation between the path lengths and network size get much attention with the statistical physics approach.

In this paper, considering geographical networks for urban planning, electric circuits, distributed robots, sensor networks, communication networks, and so on, we investigate the effects of geographical structures on the path length and the robustness in a family of the planar networks: *Random Apollonian* and *Delaunay triangulation*. The planarity is important not only to avoid interference of the wireless beam, or to construct communication lines on the surface of earth, but also to design efficient routing methods taking into account the graph properties such as spanner. In particular, online routing algorithms (Bose and Morin, 2004) that guarantee delivery of messages using only local information about positions of the source, destination, and the adjacent nodes to a current node in the routing have been developed for planar networks.

2 Geographical Networks

2.1 Planar triangulation

Planar triangulation is a mathematical abstraction of sensor or ad hoc networks, in which the positions of nodes are temporarily fixed as base stations of backbone networks. Thus, the mobility of node is out of our scope to simplify the discussion. In computer science, it is well-known that Delaunay triangulation is the optimal planar triangulation in some geometric criteria (Imai, 2000), and widely used in practical applications for facility location and computer graphics. Moreover, there is an inclusion relation: nearest neighbor graph $\subset$ relative neighbor graph $\subset$ minimum spanning tree $\subset$ Gabriel graph $\subset$ Delaunay triangulation (Kranakis and Stacho, 2006). One of the fundamental techniques for equipping such properties is diagonal flipping. In a Delaunay triangulation, diagonal flips are globally applied to the triangles until the minimum angle of triangles is not increased by the exchange of links in a quadrilateral. Such global process is unsuitable for dynamically constructed networks. In contrast, a random Apollonian network can be generated by local procedures for the subdivision of a randomly chosen triangle at each time step in the evolution of network.

Thus, we investigate the communication efficiency measured by the average distance (defined by the sum of link lengths on a path) or hops on the optimal paths and the robustness of connectivity in the typical planar network models: random Apollonian network in

complex network science, Delaunay triangulation in computer science, and a modification to bridge them.

2.2 Delaunay-like scale-free network

Although random Apollonian networks have the several advanced scale-free properties and the small-world effect with a small diameter of graph (Zhou et al., 2005), some long-range links naturally appear near the boundary edges. To reduce the long-range links, we propose a modified model (Hayashi and Matsukubo, 2005) as follows. The main idea is based on a strategy for connecting nodes in distances as short as possible by adding with the diagonal flips in a Delaunay triangulation.

Step 0: Set an initial planar triangulation in a space.

Step 1: Select a triangle at random and add a new node at the barycenter. Then, connect the new node to its three nodes. Moreover, by iteratively applying diagonal flips, connect it to the nearest node (or more than one of the neighbor nodes) within a radius defined by the distance between the new node and the nearest node of the chosen triangle.

Step 2: The above process is repeated until the required size N is reached.

We call our model RA+NN(one/all) that means the combination with the triangulation in Random Apollonian and the rewiring to the one or all Nearest Neighbors within a radius as the localization.

Fig. 1 illustrates the linking procedures by iterative diagonal flips: in a quadrilateral that consists of the shaded triangles, the long-range (crossing) link is diagonally exchanged to the red link for maximizing the minimum angle of triangles. The dashed lines are new links from the barycenter, and form new five triangles with contours in the left of Fig. 1; The intersected solid lines with dashed ones are removed after the 2nd flips.

Fig. 2 shows the topological characteristic that our model has the intermediate structure between random Apollonian networks and Delaunay triangulations. We can see a heterogeneous structure with dense and sparse parts: the dense-get-denser may be corresponded to the subdivision of a service area according to the increasing of population with preference of aggregation. As shown in Fig. 3, we find that the degree distributions follow a power-law: $k^{-\gamma_{RA}}$ in random Apollonian networks (marked by circles), log-normal: $\exp((\ln k - \mu)^2/2\sigma^2)$ in Delaunay triangulations (triangles), and power-law with exponential cutoff: $k^{-\gamma}\exp(-ak)$ in our models (pluses and crosses).

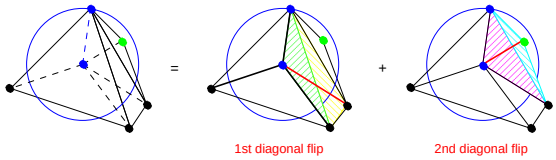


Figure 1: Linking procedures in a Delaunay-like scale-free network. The intersected lines are exclusive in each shaded quadrilateral.

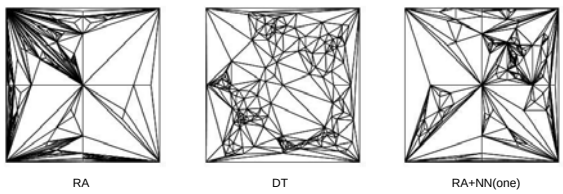


Figure 2: Examples of the geographical networks. RA: random Apollonian, DT: Delaunay triangulation, and RA+NN: our Delaunay-like scale-free network.

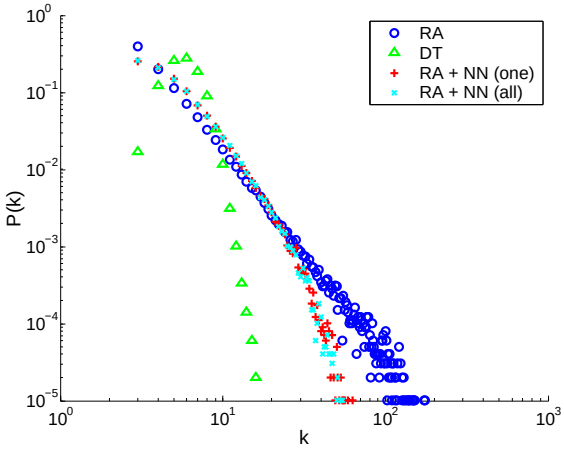


Figure 3: Degree distribution $P(k)$.

3 Path Length and Robustness

3.1 Weak disorder

In the studies of the optimal path in disordered complex networks (Braunstein et al., 2003; Kalisky et. al,

2005), each link length is associated with a weight assumed by $\exp(\delta\varepsilon)$, where the parameter δ controls the strength of disorder, and ε is a random number taken form a uniform distribution between 0 and 1. As a network approaches the strong disorder limit at $\delta \to \infty$, only the longest link becomes dominant in the shortest path length defined by the smallest sum of link lengths on a path between two nodes. At the limit, the scaling relations of the average shortest path length $\langle D \rangle \sim N^{1/3}$ for $\gamma > 4$ and $\langle D \rangle \sim N^{(\gamma-3)/(\gamma-1)}$ for $3 < \gamma \leq 4$ has been theoretically predicted (Braunstein et al., 2003) from the percolation on scale-free networks (Cohen et. al, 2002). Although the relation is unknown for $2 < \gamma \leq 3$ because of the singularity in the analysis at $\gamma = 3$, $\langle D \rangle \sim (\ln N)^{\gamma-1}$ has been also numerically suggested (Braunstein et al., 2003).

However, the assumption of length distribution may be violated on a geometric space, in addition the strong disorder limit is an extreme case. Thus, to investigate the strength of disorder in random Apollonian networks, Delaunay triangulations, and the proposed models, we compare the length distributions. Fig. 4 shows the distribution $P(l_{ij})$ of link length l_{ij} in each network. The dashed lines with an equal gap from top to bottom are corresponded to the distributions of weight $2\exp(\delta\varepsilon)/\exp(\delta)$ for $\delta = 1, 2, 4, 8, 16$, respectively. The factor $2/\exp(\delta)$ is due to the normalization for the maximum length of the boundary edges of the initial rectangle (see Fig. 2). We find that random Apollonian networks and RA+NN(one/all)s have weak disorder with small δ (Kalisky et. al, 2005), while Delaunay triangulations have a slightly broad range of disorder as similar to the exponential decay in the domestic airline flight-connection (Hayashi, 2006).

3.2 Weak small-world effect

We investigate the average distance of path length $\langle D \rangle$ on the shortest paths, the distance $\langle D' \rangle$ on the paths of the minimum hops, the average number of hops $\langle L \rangle$ on these paths, and the number of hops $\langle L' \rangle$ on the shortest paths between any two nodes in the geographical networks. The average means a statistical ensemble over the optimal paths in the above two criteria (w.r.t distance and hop) for networks in randomly generated 100 realizations at each size N. Figs. 5(a)(b) show that RA+NN(one) has the shortest distance and the intermediate number of hops in a weak small-world effect, which means the $\langle L \rangle$ is slightly larger than $O(\ln N)$ known as the effect in a scale-free network without geographical structure.

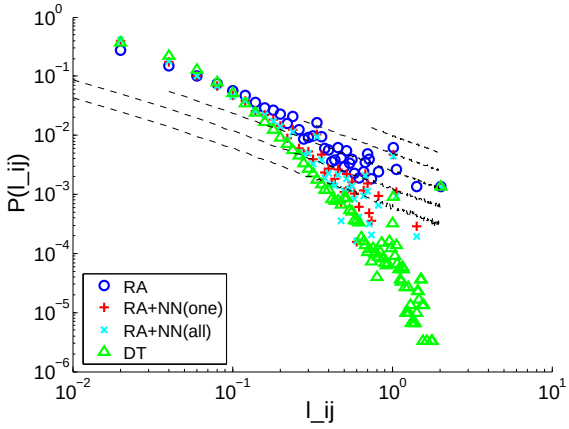


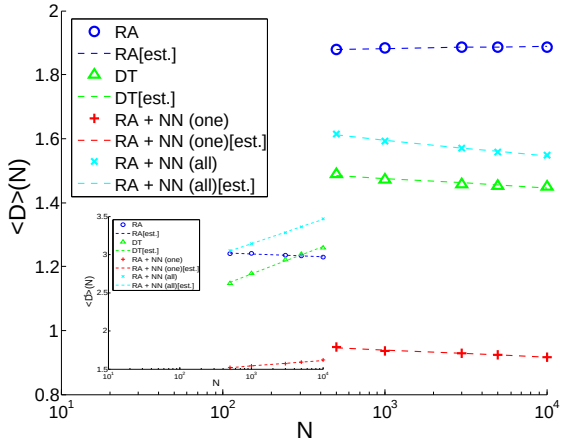
Figure 4: The distribution of link lengths with weak disorder.

Note that the shortest path and the path of the minimum hops may be distinct, these measures are related to the link cost or delay and the load for transfer of a message. It is better to shorten both the distance and the number of hops, however their constraints are generally conflicted, indeed, see Fig. 5.

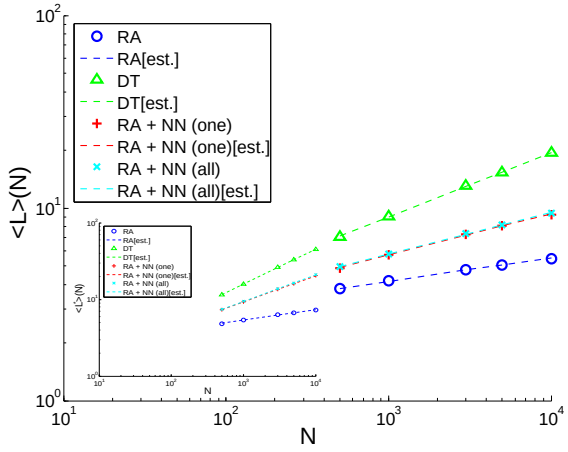
As shown in Table 1, we find the scaling relations. We remark that the values of β_d and $\beta_{d'}$ differ from $\gamma_{RA} - 1 \approx 2$ numerically suggested at the strong disorder limit (Braunstein et al., 2003), although the values of β_l and $\beta_{l'}$ are relatively close to it. In addition, the values of α_l and $\alpha_{l'}$ are close to $1/3$ predicted at the limit (Braunstein et al., 2003) for the Erdös-Rényi model as the classical random network and the Watts-Strogatz model as a small-world network. The nearest α_l in Delaunay triangulations is probably caused by that the lognormal degree distribution resembles the unimodal shapes in Erdös-Rényi and Watts-Strogatz models rather than a power-law.

Table 1: Estimated values of the exponents in the forms $\langle D \rangle \sim (\ln N)^{\beta_d}$, $\langle D' \rangle \sim (\ln N)^{\beta_{d'}}$, $\langle L \rangle \sim (\ln N)^{\beta_l}$, $\langle L' \rangle \sim (\ln N)^{\beta_{l'}}$, $\langle L \rangle \sim N^{\alpha_l}$, $\langle L' \rangle \sim N^{\alpha_{l'}}$ by a mean-square-error method.

model	β_d	β'_d	α_l	α'_l	β_l	β'_l
RA	0.012	-0.039	0.121	0.136	0.920	1.036
DT	-0.068	0.416	0.332	0.455	2.525	3.452
RA+NN (one)	-0.080	0.151	0.213	0.341	1.622	2.587
RA+NN (all)	-0.106	0.320	0.216	0.346	1.641	2.628



(a) average distance



(b) average num. of hops

Figure 5: The shortest distance and the intermediate number of hops in our model (marked by red pluses). The dashed lines correspond to the estimations in Table I. Insets show $\langle D' \rangle$ and $\langle L' \rangle$ on the paths of the minimum hops and the shortest, respectively.

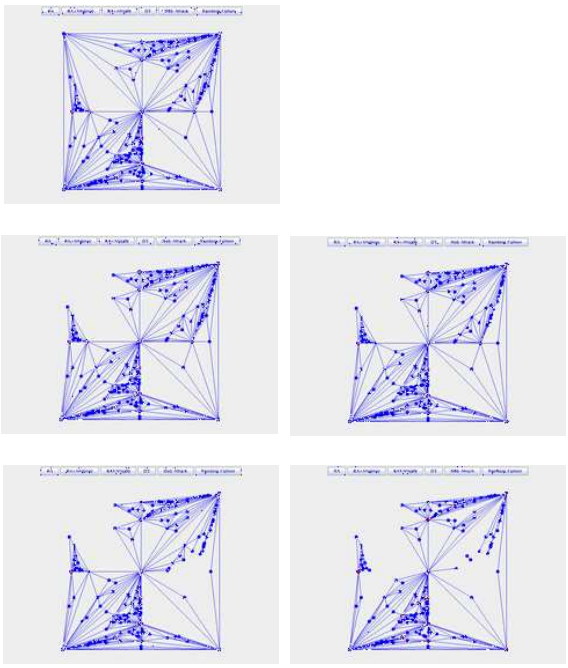
3.3 Tolerance to failures or attacks

The fault tolerance and attack vulnerability are known as the typical scale-free properties (Albert and Barabási, 2000), however the geographical effect on them are unknown. We compare the tolerance of connectivity in the giant component of the geographical and the non-geographical rewired networks with the same degree distribution (Maslov et al., 2004), when a small fraction f of the nodes is removed.

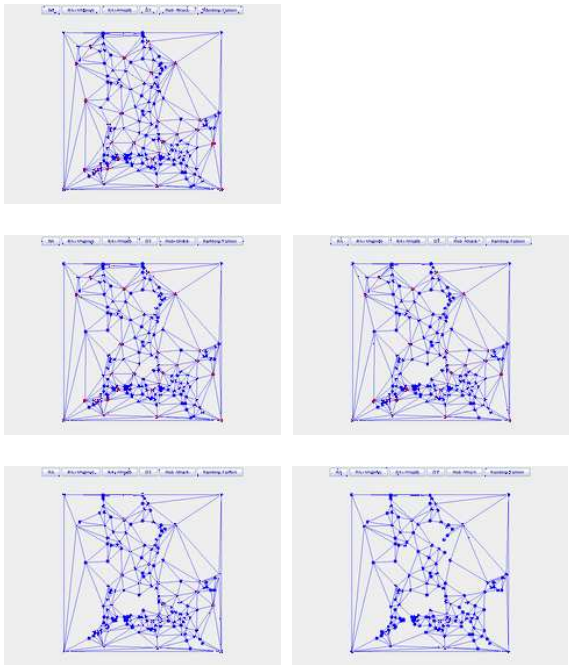
Figs. 6 and 7 show examples of random failures in the geographical networks at a small size $N = 200$ to visualize them. In the similar results, each initial component remains without isolated clusters. On the other hand, Figs. 8 and 9 show examples of targeted attacks to hubs. The random Apollonian network is the most vulnerable with many isolated clusters since the star-like stubs at the four corners and the center nodes of the initial rectangle are disconnected, while the Delaunay triangulation is relatively robust without such structure.

We investigate these differences quantitatively. The following results are obtained from the averages over 100 realizations at a size $N = 1,000$. We should remark that all networks have the same average degree $\langle k \rangle = 2(3N-7)/N = 5.986$ and the minimum degree $k_{min} = 3$. Therefore, we investigate the tolerance at the same level with the total number of links $N \times \langle k \rangle / 2$. Fig. 10(a) shows the relative size S/N for the fraction of random failures in random Apollonian networks, Delaunay triangulations, and our models, where S denotes the size of giant component. Fig. 10(b) show the robustness of connectivity in the rewired networks, whose high tolerance is similar to Barabási-Albert model (Albert and Barabási, 2000) without geographical structure. As the geographical effect, it becomes weaker in the order of random Apollonian networks, RA+NN(one/all)s, and Delaunay triangulations with degree distributions from a pure power-law to the strong cutoff. These results are not contradictory to the theoretical prediction under the power-law degree distribution with exponential cutoff (Callaway et al., 2000), since the average degree $\langle k \rangle$ is not constant but smaller as the cutoff is stronger; the connectivity is weaker in sparse networks, however the corresponding strength of cutoff is in the inverse order of random Apollonian networks, RA+NN(one/all)s, and Delaunay triangulations.

Against the attack on hubs selected in the decreasing order of degrees, Figs. 11(a)(b) show the improvements in RA+NN(one/all)s from the extremely vulnerable random Apollonian networks. By the geographical effect, each network also becomes more vulnerable than the rewired version. In other words, the improvement by rewiring is consistent with recent results for an Inet-generated graph as a modeling of the Internet (Beygelzimer et al., 2005), although it includes another bias effect of removing links from higher degree nodes. Note that the weakly inhomogeneous Delaunay triangulation is different from a homogeneous random network, which has the same behavior against the failures and the attacks at a fraction of removed nodes (Albert and Barabási, 2000).

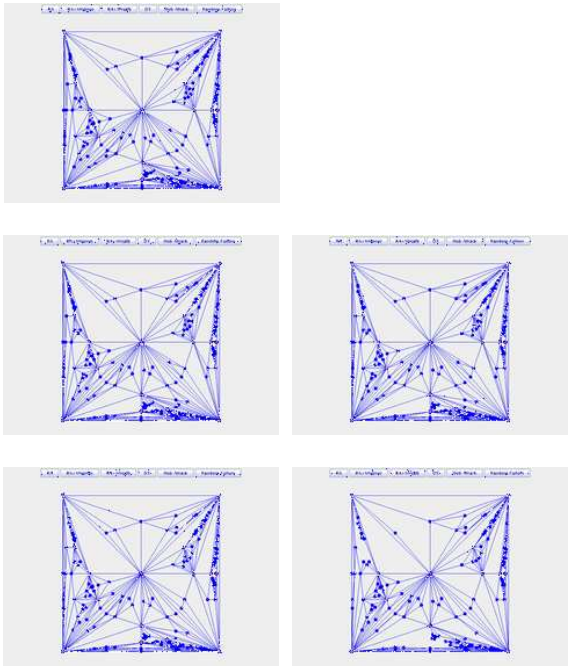


(a) RA

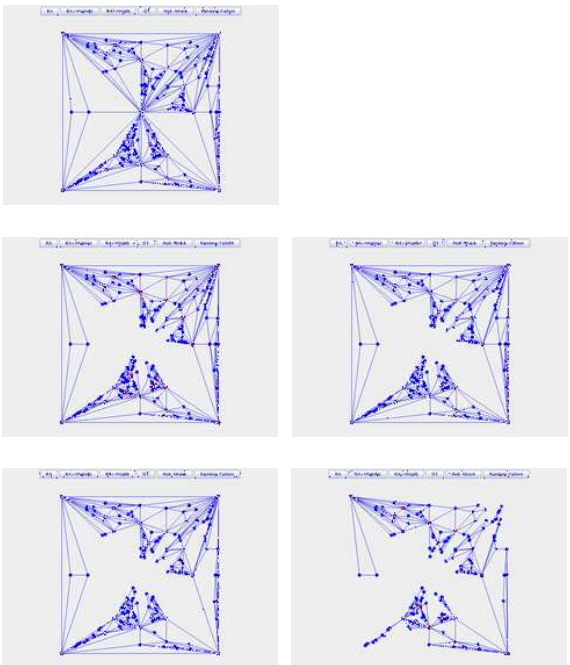


(b) DT

Figure 6: Progress of disconnection by random failures of 0, 4, 8, 16, 32 nodes from top-left to down-right for (a) random Apollonian network and (b) Delaunay triangulation.

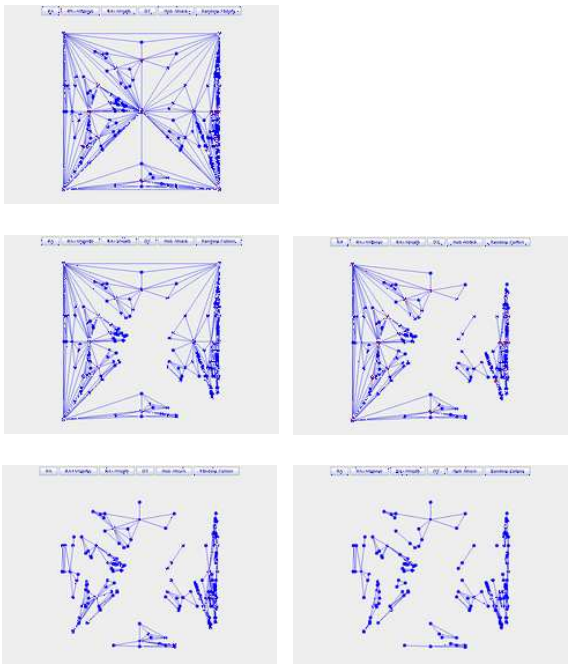


(a) RA+NN(one)

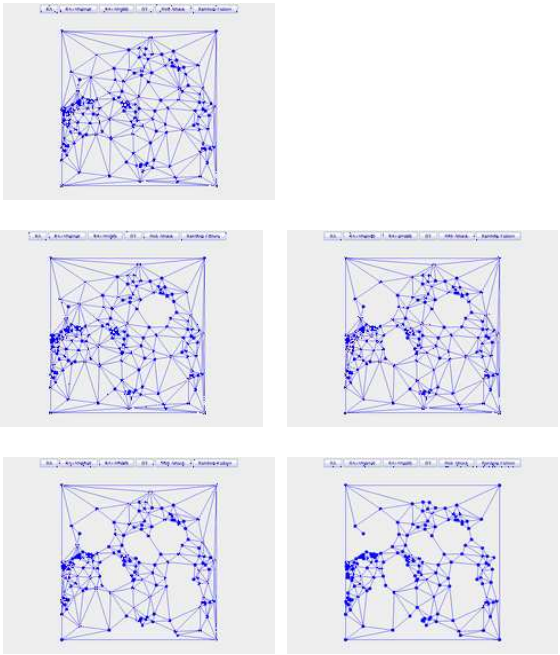


(b) RA+NN(all)

Figure 7: Progress of disconnection by random failures of 0, 4, 8, 16, 32 nodes from top-left to down-right for our models: (a) RA+NN(one) and (b) RA+NN(all).

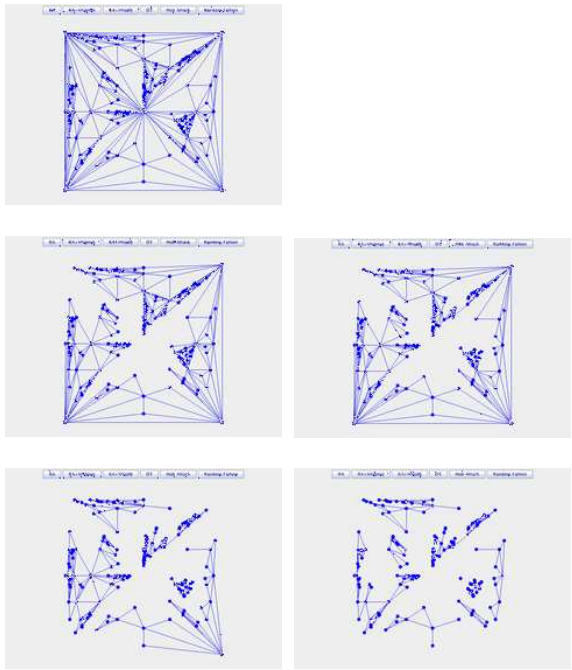


(a) RA

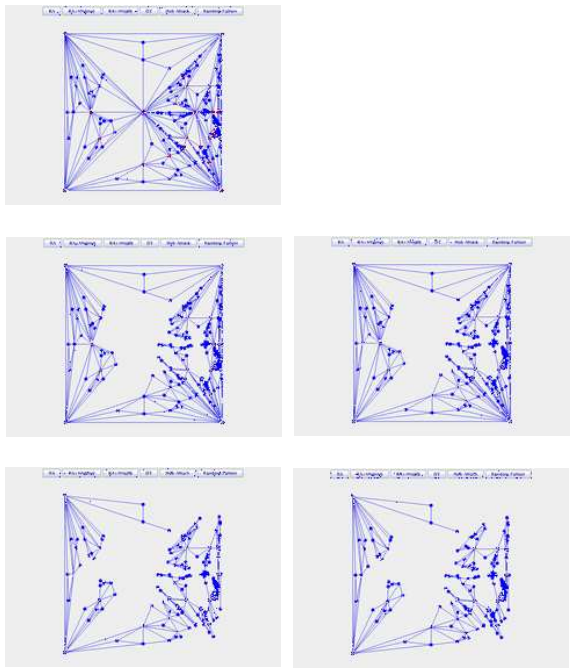


(b) DT

Figure 8: Progress of disconnection by targeted attacks on 0, 2, 4, 8, 16 nodes in decreasing order of degrees from top-left to down-right for (a) random Apollonian network and (b) Delaunay triangulation.

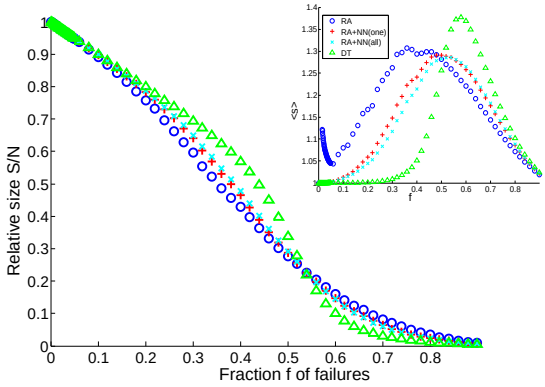


(a) RA+NN(one)

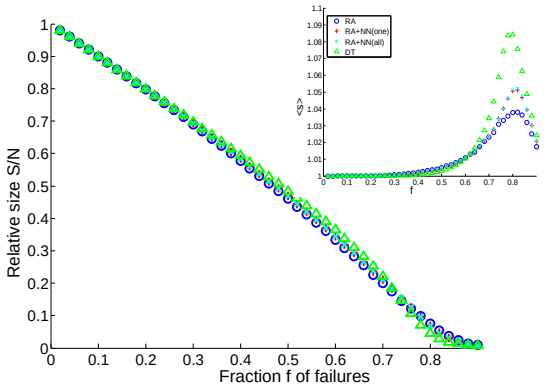


(b) RA+NN(all)

Figure 9: Progress of disconnection by targeted attacks on 0, 2, 4, 8, 16 nodes in decreasing order of degrees from top-left to down-right for our models: (a) RA+NN(one) and (b) RA+NN(all).

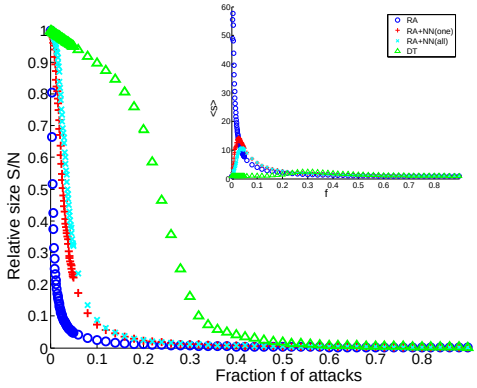


(a) geographical nets

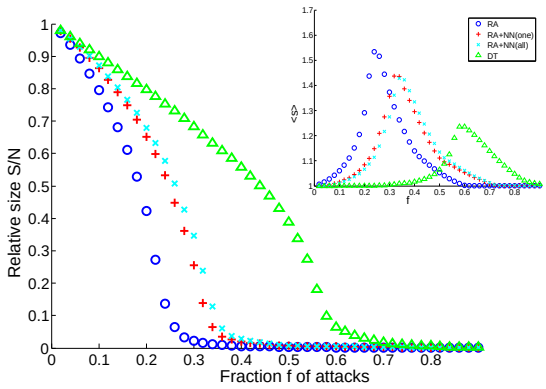


(b) randomly rewired nets

Figure 10: Relative sizes S/N of the giant component against random failures in the geographical and the rewired networks. Inset show the average size of isolated clusters except the giant component. At the peak, the giant component disappears.



(a) geographical nets



(b) randomly rewired nets

Figure 11: Relative sizes S/N of the giant component against attack on hubs in the geographical and the rewired networks.

4 Conclusion

We investigate the effect of geographical structure on the path length and the robustness of connectivity, focusing on a family of planar networks called random Apollonian network and Delaunay triangulation for communication systems. To reduce long-range links, we propose a modified model whose degree distribution follows a power-law with exponential cutoff. We find the weak disorder in the distributions of link lengths, and suggest the scaling relations of the shortest path length $\langle D \rangle \sim (\ln N)^{\beta_d}$ and of the minimum hop $\langle L \rangle \sim N^{\alpha_l}$ as similar to the case at the strong disorder limit (Braunstein et al., 2003). From the simulations, we conclude that random Apollonian networks have a path connected by a few hops but the path length becomes long including some long-range links, while Delaunay triangulations have a zig-zag path connected by many hops but each link is short. Instead of the superior geometric properties (Imai, 2000), Delaunay triangulations are no longer optimal in this criteria of the minimum hops. Our model is totally balanced: the shortest path length is the best, while the number of hops is the intermediate.

Moreover, we find that the tolerance to failures and attacks is weakened by the geographical effect. In particular, random Apollonian networks with a pure power-law degree distribution are extremely vulnerable. Although Delaunay triangulation is the most robust in these models, only it requires global configuration procedures that is unsuitable for ad hoc communication. Thus, there is a trade-off between the localization and the robustness. We will further investigate the above effect in more wide classes related to a family of scale-free networks.

References

R. Albert and A.-L. Barabási. Error and attack tolerance of complex networks. *Nature*, 406:378–382, 2000.

A.-L. Barabási. *Linked: The New Science of Networks*. Perseus, 2002.

Beygelzimer et al. Improving network robustness by edge modification. *Physica A*, 357:593–612, 2005.

P. Bose and P. Morin. Online routing in triangulations. *SIAM J. of Computing*, 33(4):937–951, 2004.

L.A. Braunstein et al. Optimal paths in disordered complex networks. *Physical Review Letters*, 91 (16):168701, 2003.

D.S. Callaway et al. Network robustness and fragility: Percolation on random graphs. *Physical Review Letters*, 85(25):5468–5471, 2000.

R.F.i Cancho and R.V. Solé. Optimization in complaex networks. In M. Rubi R. Pastor-Satorras and A. Diaz-Guilera, editors, *Statistical Mechanics in Complex Networks*, Chapter 6. Springer, Berlin, 2003. Lecture Notes in Physics **625**.

R. Cohen et al., Structural properties of scale-free networks. In S. Bornholdts and H.G. Shuster, editors, *Handbook of Graphs and Networks*, Chapter 4. Wiely-VCH, New York, 2002.

M.T. Gastner and M.E.J. Newman. The spatial structure of networks. *Eur. Phys. J. B*, 49(2):247–252, 2006.

E. Kranakis and L. Stacho. Routing and Traversal via Location Awareness in Ad Hoc Networks. In A. Boukerche, editor, *Handbook of Algorithms for Wireless Networking and Mobile Computing*, Chapter 8. Chapman & Hall/CRC, 2006.

S. Maslov, K. Sneppen and A. Zaliznyak. Detection of topological patterns in complex networks: correlation profile of the internet. *Physica A*, 333:529–540, 2004.

Y. Hayashi and J. Matsukubo. Scale-free networks on a geographical planar space for efficient ad hoc communication. In *Proc. of NOLTA*, 118–121, 2005.

Y. Hayashi, A Review of Recent Study of Geographical Scale-Free Networks. to appear in *IPSJ Journal*, Special Issue on Network Ecology, 47(3):2006 or *arXiv:physics/0512011*, 2005.

K. Imai. Structures of triangulations of points. *IEICE Trans. on Infor. and Syst.*, 83(3):428–437, 2000.

T. Kalisky, et al., Scaling of optimal-path-lengths distribution in complex networks. *Physical Review E*, 72: 025102, 2005.

S.-H. Yook, H. Jeong and A.-L. Barabási. Modeling the internet's large-scale topology. *PNAS*, 99(21): 13382–13386, 2002.

T. Zhou, G. Yan and B.-H. Wang. Maximal planar networks with large clustering coefficient and power-law degree distribution. *Physical Review E*, 71:046141, 2005.

A Discovery Method of Research Communities

Ryutaro Ichise*
*Intelligent Systems Research Division,
National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-ku,
Tokyo 101-8430, Japan
ichise@nii.ac.jp

Hideaki Takeda†
† Research Center for
Testbeds and Prototyping,
National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-ku,
Tokyo 101-8430, Japan
takeda@nii.ac.jp

Taichi Muraki‡
‡TriAx Corporation
4-29-3 Yoyogi, Shibuya-ku,
Tokyo, 151-0053 Japan
muraki@triax.jp

Abstract

Since research trends can change dynamically, researchers have to keep up with new research trends and undertake new research topics. Therefore, research communities for new research domains are important. In this paper, we propose a method to discover research communities. The key feature of our method is a network model of papers and a word assignment technique for the communities obtained. We show the performance of the proposed method using experiments with real world data.

1 Introduction

As information technologies progress, we can obtain research information faster then before. However, technologies covering a wide area can change just as rapidly. Therefore, all researchers must not only continuously follow new trends of research but also investigate new research topics. When we undertake new research topics, we need to know the research communities of researchers with the same research topic or same interest. As a result, we need an effective community mining method for finding them. In order to find research communities, we usually use bibliography information. The methods include co-citation analysis (Small, 1973; Chen and Paul, 2001) and bibliographic coupling (Kessler, 1963). Although these methods are very useful for analyzing research topics from the global viewpoint of all bibliography data, we cannot always understand what the discovered communities represent. CiteSeer (CiteSeer.IST, 2004) and Google Scholar (Google, 2004) are able to handle research communities from a micro viewpoint because they handle co-author and citation information from bibliographies and use the information for individual researchers. Although these systems are good for finding local communities involving an author, they are not suitable for finding research communities close to the author. Börner et al. (Börner et al., 2005) proposes to use co-author networks to find research communities by using weighted graphs. Their system uses heuristics to separate communities without interaction. Ichise et al. (Ichise et al., 2005) proposed a community mining method based on the interaction of users. Although the proposed system of Ichise et al. supports community mining for both a global view and a local view with several mining indexes, it does not identify the research topics of the communities obtained. In this paper, we propose a method to discover research communities with identified topics.

The present paper is organized as follows. In Section 2, we discuss the proposed method for research community mining. In Section 3, we describe the experimental evaluation of our method and then discuss the results. Finally, in Section 4, we present our conclusions.

2 Research Community Discovery

2.1 Network Model of Research Community

Although several network models using bibliographies to represent research communities have been proposed (Ichise et al., 2005), in this paper, we focus on the co-author relationships of a research paper to find the research communities. First, we assume a simple paper model. This model consists of keywords and author names. In this case, we can consider an author's work on a research topic by noting the keywords. As a result, authors who write a paper collaboratively share the same interest, represented by the keywords. If we consider the authors as nodes and the keywords as edges, we can represent the bibliog-

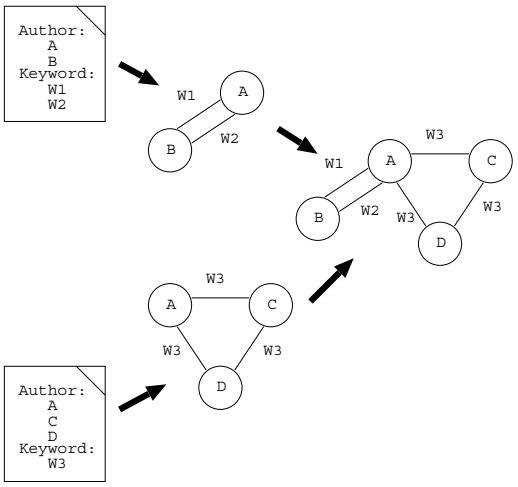


Figure 1: Network model of researchers.

raphy information as researcher networks.

Let us explain our model using an example. Assume that we have two papers, as shown on the left side of Figure 1. One was written by two authors, A and B, and has two keywords, W_1 and W_2. Another was written by three authors, A, C and D, and has the keyword W_3. We can compose graphs of the authors and edges from the two papers, as shown in the center of Figure 1. Then, the joint representation generated from the two bibliographies of the two papers is shown on the right side of Figure 1.

As one can see, we can obtain a labeled graph from the bibliography data with our modeling. Then, the next question is how do we discover research communities from this graph. We define a research community as a cluster which is densely connected by the same research interest or topic. Therefore, the research communities we want to obtain are clusters which have their edges labeled by the same keywords. Since our network model provides the research keywords on the edges, we can obtain the research communities by eliminating the edges of no interest to the system user. In other words, after the user specifies the research keywords, most of the edges which are not labeled by the specified keywords can be deleted. This process produces the research communities. For example, when the user specifies W_3 for the networks in Figure 1, the edges of W_1 and W_2 are eliminated. As a result, researcher B is isolated from the graph and we can find the research community consisting of researchers A, C and D.

2.2 Keyword Assignment for Communities

Since the clusters obtained by our method are only connected by user-specified relationships, we can consider each cluster as a research community. However, each cluster does not have its own property or identification. In other words, if the user does not have enough knowledge about the researchers, the user may not understand the meaning of the communities because there is no information about them. In order to solve this problem, we propose a method of assigning keywords for each obtained community.

In our paper model, the papers written by the authors in each community have keywords. If some words appear often in such papers, we can consider these words as a property for the community. However, if we simply counted the occurrences of the keywords in these papers, the relationships between keywords would be lost. In order to avoid this problem, we consider frequent keywords as units of the papers. The algorithm is follows:

1. Select papers, which are written by the authors in the community, from a paper database. Note that the papers are selected for each user. For example, if a paper is written by two authors in the same community, the paper is selected twice.

2. The selected papers are analyzed by the Apriori algorithm (Agrawal and Srikant, 1994). In this process, the keywords in a paper are treated as an item, and the papers are treated as transactions.

As a result, we can obtain word pairs for each community. We assign those word pairs as the property of the community.

3 Experiments

3.1 Bibliography Data

In order to evaluate our method, we conducted experiments using actual bibliography data. In the present study, we used the CiNii database (NII, 2004) to obtain bibliography information. We used 128,000 records, 90,000 records, 358,000 records, and 519,000 records for the paper, researcher, author and co-author, respectively. The author entries denote the number of authors for each paper. For example, the record is 3 when three researchers write a paper collaboratively. The co-author entries denote the number of combinations of authors for a paper. For example, the record is counted as ${}_4C_2 = 6$ for

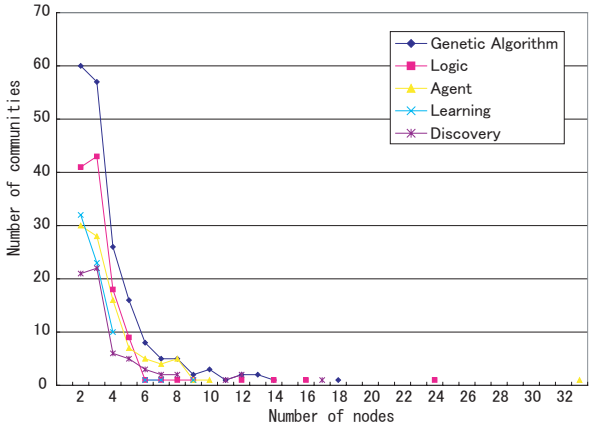


Figure 2: Number of discovered communities.

a paper when the paper is written by four authors. It was necessary to have keywords for our paper model. Therefore, we used the words in the title as keywords.

3.2 Experimental Results

The co-author network tends to have large clusters. In fact, the network constructed by all the bibliography data consists of a few large clusters and many small clusters. The number of clusters for all the data is shown in Table 1.

However, our method can successfully split a large cluster into readable research communities. We utilized five words to show the discovered communities. The five words are as follows: genetic algorithm, logic, agent, learning and discovery. We counted the number of nodes and clusters for each word. Figure 2 illustrates the result. The horizontal axis denotes the number of nodes and the vertical axis denotes the number of communities. As you can see from Figure 2, our method successfully discovers readable-sized communities. In addition, communities related to particular topics of interest to the user can be mined by our method.

Next, in order to evaluate our method for qualitative aspect, we analyzed the communities obtained by our method. The communities were constructed using the word ”Discovery”. Although our method discovered many communities, we selected five communities shown in Figure 3. Since our bibliography data mainly included papers written in Japanese, the communities obtained also consisted of Japanese researchers. The assigned keywords for each community are shown in Table 2. The longest pair of keywords, which frequently appeared in the top three,

Table 2: Keywords obtained for the topic ”discovery”.

Community ID	Keywords
No. 1	{discovery} {algorithm} {special issue}
No. 2	{Japanese poem, similarity} {poem, similarity, extraction} {English sentence, technology}
No. 3	{heuristics, method} {database, heuristics} {knowledge, exception, discovery}
No. 4	{definition, occurrence, lambda calculus} {logic, program} {unification, extension}
No. 5	{*th, workshop} {scientific, discovery} {*th, report}

were selected for the table.

Community No. 2, No. 3, No. 4 and No. 5 represent research groups in universities. Most of the assigned words for the communities are valid. However, our method assigns meaningless words such as ”workshop” and ”report”. In our future work, we plan to develop a method to suppress the assignment of meaningless words. We believe such words can be identified by a simple method of retrieving stop words. Community No. 1 is the largest community in Figure 3. Since Prof. Setsuo Arikawa at Kyushu University is one of the most famous Japanese scientists in the ”discovery” domain, he bridged several communities. For example, the bottom left part of the community is the research community at Kyushu University, and the upper left part of the community is a community in machine learning. As a result, our system assigns general words for this community such as ”discovery” and ”algorithm.” In addition, the system cannot generate a long pair of words for this community. In our future work, we will develop a method to identify a person who is a bridge between different communities. Incidentally, the word ”special issue” was assigned in Community No. 1 because the community members edit a special issue for a journal.

4 Conclusion

In this paper, we propose a research community mining method. The key feature of our research is the modeling of papers and researchers. This modeling enables us to eliminate the edges in large clusters.

Table 1: Distribution of clusters for all the bibliography data.

Number of Nodes	76375	39	28	27	26	...	10	9	8	7	6	5	4	3	2	1
Number of Clusters	1	1	1	1	1	...	31	39	57	64	143	214	328	646	1054	7173

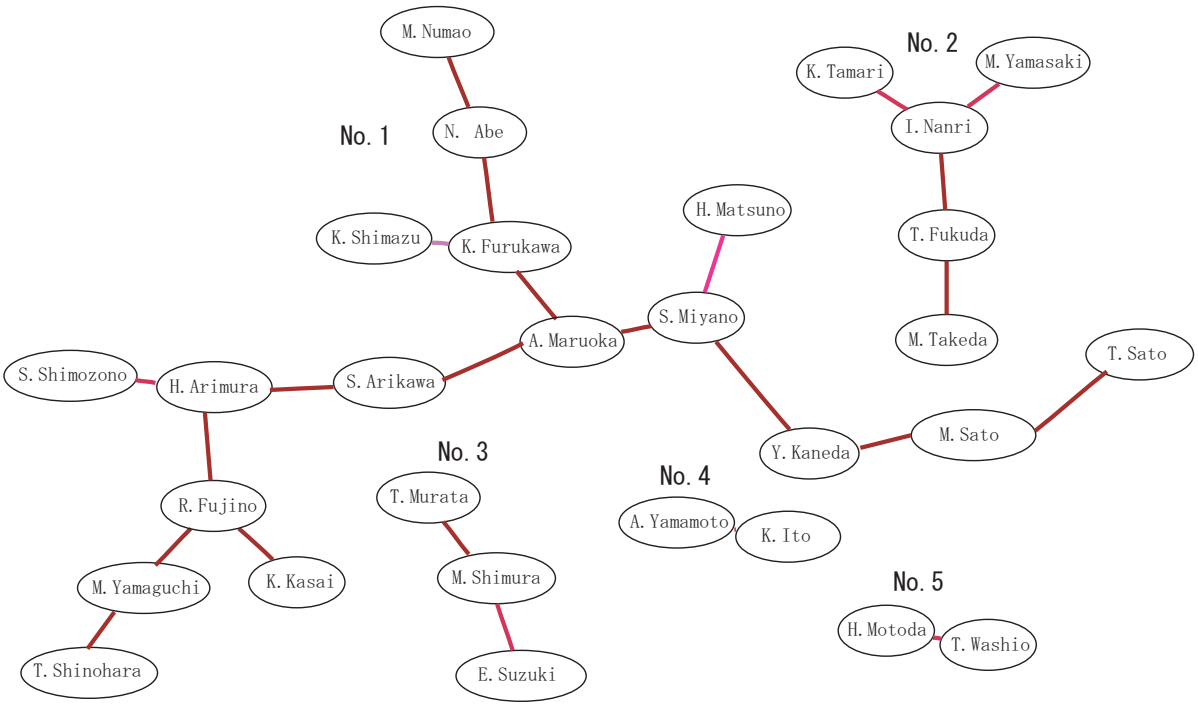


Figure 3: Research communities of discovery in Japan.

In addition, the modeling can also help to retrieve communities for particular topics. We also propose a method to assign a word to each cluster. We implemented our method and show how to investigate bibliography data with our system. The experimental results show that the performance of our method looks promising.

References

Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules. In *Proc. 20th Int. Conf. Very Large Data Bases, VLDB*, pages 487–499. Morgan Kaufmann, 1994.

Katy Börner, Luca Dall'Asta, Weimao Ke, and Alessandro Vespignani. Studing the emerging global brain: Analyzing and visualizing the impact of co-authorship teams. *Complexity*, 10(4):58–67, 2005.

Chaomei Chen and Ray J. Paul. Visualizing a knowledge domain's intellectual structure. *Computer*, 34 (3):65–71, 2001.

CiteSeer.IST. Scientific literature digital library, 2004. http://citeseer.ist.psu.edu/.

Google. Google scholar, 2004. http://scholar.google.com/.

Ryutaro Ichise, Hideaki Takeda, and Kosuke Ueyama. Community mining tool using bibliography data. In *Proceedings of the 9th International Conference on Information Visualization*, 2005.

M. M. Kessler. Bibliographic coupling between scientific papers. *American Documentation*, 14(1): 10–25, 1963.

NII. Cinii (citation information by national institute of informatics), 2004. http://ci.nii.ac.jp/.

Henry Small. Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society of Information Science*, 24:265–269, 1973.

Clustering Coefficients for Weighted Networks

Gabriela Kalna[†]
[⋆]Department of Mathematics,
University of Strathclyde,
Glasgow G1 1XH, UK.
aas04105@maths.strath.ac.uk

Desmond J. Higham[⋆]
[†]Department of Mathematics,
University of Strathclyde,
Glasgow G1 1XH, UK.
aas96106@maths.strath.ac.uk

Abstract

The clustering coefficient has been used successfully to summarise important features of unweighted, undirected networks across a wide range of applications. Recently, a number of authors have extended this concept to the case of networks with non-negatively weighted edges. After reviewing various alternatives, we focus on a definition due to Zhang and Horvath that can be traced back to earlier work of Grindrod. We give a natural and transparent derivation of this clustering coefficient and then analyse its properties. One attraction of this version is that it deals directly with weighted edges and avoids the need to discretise, that is, to round weights up to 1 or down to 0. This has the advantages of (a) retaining all edge weight information, and (b) eliminating the requirement for an arbitrary cutoff level. Further, the extended definition is much less likely to break down due to a 'divide-by-zero'. Using our new derivation and focussing on some special cases allows us to gain insights into the typical behaviour of this measure. We then illustrate the idea by computing the generalised clustering coefficients, along with the corresponding weighted degrees, for pairwise correlation gene expression data arising from microarray experiments. We find that the weighted clustering and degree distributions reveal global topological differences between normal and tumour networks.

1 Introduction

Many complex data sets have natural representations as networks. Summarising, comparing, categorising and modelling these data sets are important activities that are taking place simultaneously across a wide range of disciplines (Newman, 2003). It is accepted that typical real-life networks are neither random graphs in the classical Erdös-Rényi sense nor regular lattices (Watts and Strogatz, 1998). Various quantities can be computed in order to characterise a network; most prominently the concepts of *path-length*, *degree* and *clustering coefficient* have proved extremely useful.

Watts and Strogatz (1998) coined the phrase *small world network* to describe the commonly occurring situation where a sparse network is highly clustered (like a regular lattice) yet has small pathlengths (like a random graph). Since that landmark paper, many complex networks have been analysed and labelled as small worlds.

Similarly, the so-called *scale–free* property of the degree distribution, (Barabasi and Albert, 1999; Newman, 2003), has become accepted as a hallmark of many real data sets, although there is now some doubt as to its true prevalence (Khanin and Wit, to appear; Pržulj et al., 2004).

Both the small world and scale–free properties have been widely studied for unweighted, or binary, undirected networks. In the case of more general weighted edges it is of course possible to create a binary network by normalising, imposing a cutoff and rounding to 0 and 1 (Rougemont and Hingamp, 2003). However, it is our tenet that the original weights should be respected where possible. Recently, a number of authors have attempted to generalise the clustering coefficient concept to the case of weighted edges (Barrat et al., 2004; Lopez-Fernandez et al., 2004; Onnela et al., 2005; Zhang and Horvath, 2005), producing a range of possible definitions.

We present here a natural and transparent derivation of a clustering coefficient for weighted graphs. The resulting definition coincides with those in (Grindrod, 2002; Zhang and Horvath, 2005) and hence we argue for the use of this Grindrod-Zhang-Horvath clustering coefficient as a generalised measure of clustering. We believe that this measure, along with the corresponding weighted degree distribution, gives an informative high-level picture that can be used for classifying, comparing and modelling

weighted networks, just as in the unweighted case. We do some analysis to provide insight into the usefulness of this clustering coefficient, and then show some results for gene expression microarray data.

Many methods for microarray data analysis monitor differences in the expression of genes under various experimental conditions: normal/tumour (Chen et al., 2002), multiclass cancers (Golub et al., 1999; Ramaswamy et al., 2001), treatment/survival (Segal, 2005). Pair-wise gene expression correlation has long been used to predict relationships between genes. Recently, gene co-expression networks have emerged (Stuart et al., 2003; Zhang and Horvath, 2005) connecting genes with high correlation. However, despite the fact that genome-wide gene expression data sets are available, their full potential is often not used and information from only a subset of genes, usually with highest variation, is extracted. Hence, we view these weighted networks as ideal candidates on which to apply the new clustering coefficient framework.

Using available microarray data we construct two distinct gene coexpression networks that represent normal and tumour states. We examine weighted clustering coefficients and weighted degree distributions of these networks with the aim of finding tumour-related differences. We emphasize that our aim is to characterize overall network topology rather than to categorize individual genes or samples.

The rest of this article is organised as follows. In section 2 we start with the binary definition of clustering coefficient and list some generalisations that have been proposed for weighted networks. In section 3 we give a natural derivation that leads to the Grindrod-Zhang-Horvath definition, and show how this can be easily computed via matrix products. We then use some simple examples to explore the properties of this coefficient. In section 4 we give some realistic computations on pairwise correlation networks arising from microarray data.

2 Clustering Coefficient and its Generalisations

Consider an undirected graph with normalised weights $0 \leq w_{ij} \leq 1$ between nodes i and j. In the binary case $w_{ij} \in \{0, 1\}$ the clustering coefficient, or curvature, for node k is defined as

$$\mathrm{clust}(k) := \frac{t}{v(v-1)/2}, \qquad (1)$$

where v is the number of immediate neighbours of node k, and t is the number of triangles incident to node k (Rougemont and Hingamp, 2003; Watts and Strogatz, 1998). In words, $\mathrm{clust}(k)$ answers the question "given two nodes that are both connected to node k, what is the likelihood that these two nodes are connected to each other?" It is straightforward to see that the definition breaks down when $v < 2$, that is, node k has less than two immediate neighbours, and otherwise $0 \leq \mathrm{clust}(k) \leq 1$.

Recently, a few different extensions of the clustering coefficient to the general weighted case have emerged. In Lopez-Fernandez et al. (2004) the weighted clustering coefficient for node k is defined as

$$\mathrm{wclust}_{\mathrm{LF}}(k) := \frac{\sum_{i \neq j \in N(k)} w_{ij}}{v(v-1)},$$

where the term $\sum_{i \neq j \in N(k)} w_{ij}$ can be seen as the total weight of relationship in the neighbourhood $N(k)$ of node k.

Barrat et al. (2004) introduced a measure of clustering that combines topological information with the weight distribution of the network

$$\mathrm{wclust}_{\mathrm{B}}(k) := \frac{1}{s(v-1)} \sum_{i,j} \frac{(w_{ki} + w_{kj})}{2} a_{ik} a_{kj} a_{ij}.$$

Here $s = \sum_j w_{kj}$ denotes the weighted degree of node k and a_{ij} is an element of the underlying binary adjacency matrix. The normalisation factor $s(v-1)$ ensures that $0 \leq \mathrm{wclust}_{\mathrm{B}}(k) \leq 1$. This definition of weighted clustering coefficient considers only weights of edges adjacent to node k but not the weights of edges between neighbours of the node k.

Onnela et al. (2005) took into account weights of all edges: adjacent to node k and between-neighbours. They considered weights $0 \leq w_{ij} \leq 1$ and replaced the number of triangles t in (1) with the sum of triangle intensities

$$\mathrm{wclust}_{\mathrm{O}}(k) := \frac{2 \sum_{i,j} (w_{ik} w_{kj} w_{ij})^{1/3}}{v(v-1)}.$$

We remark that the three clustering coefficient definitions above suffer from the drawback that they require an underlying binary network to be generated; if this is not available as a separate set of data, then presumably it must be obtained by discretizing the weighted edges. Hence, as in the case where the original binary definition is used for weighted networks (Rougemont and Hingamp, 2003), they are dependent upon some thresholding parameter. Further, they break down in the case where the number of binary neighbours, ν, is less than 2.

A definition that uses only the network weights was proposed by Zhang and Horvath (2005)

$$\mathrm{wclust}_{\mathrm{HZ}}(k) := \frac{\sum_{i\neq k}\sum_{j\neq i,j\neq k} w_{ki}w_{ij}w_{jk}}{(\sum_{i\neq k} w_{ki})^2 - \sum_{i\neq k} w_{ki}^2}. \quad (2)$$

The numerator in (2) was obtained by finding a lower bound for the denominator, this ensuring that $\mathrm{wclust}_{\mathrm{HZ}}$ is in the range $[0, 1]$.

We also mention that rather than one clustering coefficient per node, Schank and Wagner (2005) presented a single weighted clustering coefficient for the whole network as

$$\mathrm{wclust}_{\mathrm{S}} := \frac{1}{\sum_v w(v)} \sum_v w(v)c(v).$$

Here $c(v)$ is a clustering coefficient for node v and $w(v)$ a weight function. One of possible choices of weight function is the weighted degree.

3 Weighted Clustering

3.1 Definition and Properties

Consider now an undirected weighted network of M nodes that is fully connected with weights $0 \leq w_{ij} = w_{ji} \leq 1$ between nodes i and j and $w_{ii} = 0$. Some simple algebra allows the binary clustering coefficient (1) to be rewritten as

$$\mathrm{wclust}(k) := \frac{\sum_{i=1}^{M}\sum_{j=1}^{M} w_{ki}w_{kj}w_{ij}}{\sum_{i=1}^{M}\sum_{j=1,j\neq i}^{M} w_{ki}w_{kj}}. \quad (3)$$

This formula directly extends to the real value case where $w_{ij} \geq 0$ and hence gives a natural definition for weighted networks. We also mention that the same formula was used in Grindrod (2002) in the context where w_{ij} represents the probability of an edge between nodes i and j in a random network model. Closer inspection shows that the formula (3) has a simple interpretation that is analogous to that of the binary case:

$$\tfrac{1}{2}\sum_{i=1}^{M}\sum_{j=1}^{M} w_{ki}w_{kj}w_{ij}$$

is a reasonable measure of how many "big triangles" involve node k and

$$\tfrac{1}{2}\sum_{i=1}^{M}\sum_{j=1 j\neq i}^{M} w_{ki}w_{kj}$$

says how many "big pairs of neighbours" there are. It is easy to verify that (3) retains the property $0 \leq \mathrm{wclust}(k) \leq 1$.

Computationally, note that the numerator of (3) is

$$\begin{aligned}\tfrac{1}{2}\sum_{i=1}^{M} w_{ki}\sum_{j=1}^{M} w_{kj}w_{ij} &= \tfrac{1}{2}\sum_{i=1}^{M} w_{ki}(W^2)_{ki} \\ &= \tfrac{1}{2}(W^3)_{kk}\end{aligned}$$

and the denominator is

$$\begin{aligned}\tfrac{1}{2}\Big(\sum_{i=1}^{M}\sum_{j=1}^{M} w_{ki}w_{kj} &- \sum_{i=1}^{M} w_{ki}^2\Big) \\ &= \tfrac{1}{2}\big((e^T w_k)^2 - ||w_k||_2^2\big).\end{aligned}$$

Here, $(W^p)_{ij}$ denotes the (i, j) element of the pth power of W, w_k denotes the kth row (or column) of W and e denotes the vector with all elements equal to one. Hence, a neater representation of (3) is

$$\mathrm{wclust}(k) = \frac{(W^3)_{kk}}{(e^T w_k)^2 - ||w_k||_2^2}, \quad (4)$$

which shows that the weighted clustering coefficient can be computed across all nodes in $O(M^3)$ operations. The formula (4) also makes it clear that (3) is entirely equivalent to the Zhang-Horvath definition (2).

Having derived this definition from what we believe to be a natural and informative viewpoint, we now attempt to gain further insights by focussing on particular types of weighted network.

3.2 Limit Forms of Clustering

We now zoom to a particular node K of a graph and explore its weighted clustering coefficient (3) in specific cases. Starting with a binary network $w_{ij} \in \{0, 1\}$ we replace zero weights with a small weight $0 < \epsilon << 1$ (weak connections) and unit weights with $1 - \epsilon$ (strong connections). Thus, we are dealing with fully connected graph.

(A) In the first case, let node K have $m > 1$ strong and $n > 1$ weak connections to other nodes in the graph. Then there are (a) $m(m-1)/2$ strong-strong, (b) mn strong-weak and (c) $n(n-1)/2$ weak-weak neighbour pairs. Let there be r, s and u strong edges between neighbours in cases (a), (b) and (c) respectively. It is easy to show that equation (3), for $\epsilon \to 0$, results in $\mathrm{wclust}(K) = 2r/m(m-1)$. In words, r strong triangles are built over $m(m-1)/2$ strong neighbour pairs. Thus the weighted clustering coefficient (3) approaches the binary value (1).

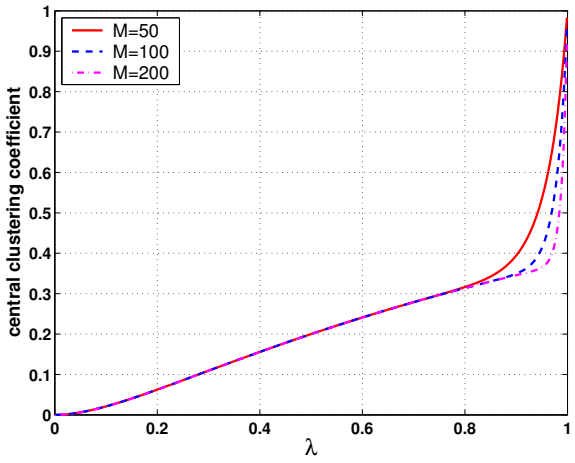


Figure 1: Clustering coefficient of the central node in the weighted graph defined by (5).

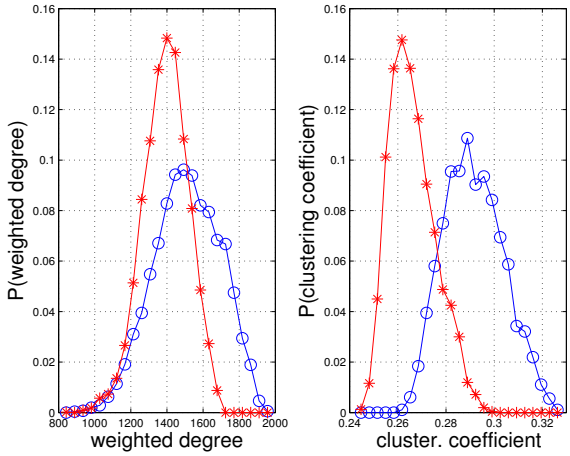


Figure 2: Probability of weighted degree (left) and curvature (right). Breast cancer: normal (circles) and tumour (stars).

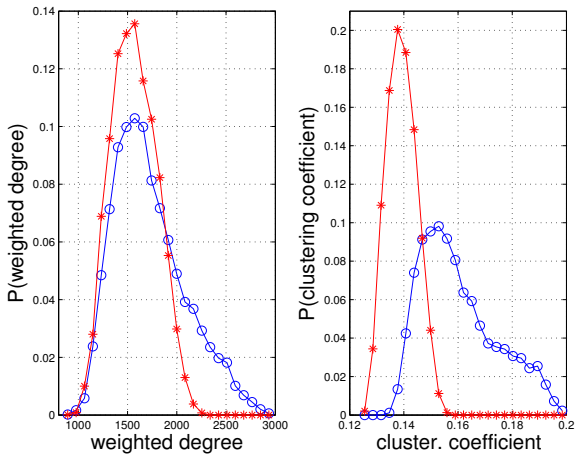


Figure 3: Probability of weighted degree (left) and clustering coefficient (right). Liver cancer: normal (circles) and tumour (stars).

(B) In the second case we consider the marginal setting $v = 1$: node K has strong connection, $1-\epsilon$, only to one node P and n weak, ϵ, connections to all other nodes in a complete graph. Then n out of all possible neighbour pairs involve the strong edge between nodes K and P and $n(n-1)/2$ pairs are formed by n weak edges adjusted to node K. Between-neighbour edges will be again strong or weak. Let there be r strong edges with one end in node P and s strong edges between "weakly" connected neighbour nodes of node K. Then from (3) we get $\mathrm{wclust}(K) = (r\epsilon(1-\epsilon)^2 + (n-r)\epsilon^2(1-\epsilon) + s\epsilon^2(1-\epsilon) + (n(n-1)/2 - s)\epsilon^3)/(n\epsilon(1-\epsilon) + n(n-1)\epsilon^2/2)$. This expression results in r/n for $\epsilon \to 0$. In words, we can get to r out of n "weakly" connected neighbours of the node K through the strong edge KP and strong edges connecting node P with these r nodes. It is clear that $\mathrm{wclust}(k) = 1$ only if $r = n$, that means there is a strong edge between P and all nodes weakly connected to K. Because $\mathrm{wclust}(k) = 0$ if $r = 0$, the strong edge between nodes K and P is the only edge involving node P. That means this edge would be separated from the graph in the corresponding discretised network.

Case **B** reveals an important advantage of the generalised definition (3). It continues to provide useful information in the small ϵ regime where *any discretization process based on thresholding to a binary network would result in* $v = 1$ *and hence an undefined clustering coefficient in (1).*

3.3 Uniform Connectivity

Another case where the clustering coefficient simplifies arises when node K has equal weights with all other nodes: $w_{Kj} = \text{constant}$ for all $j \neq K$. In this case we have

$$\mathrm{wclust}(K) = \frac{\sum_{i=1}^{M}\sum_{j=1}^{M} w_{ij}}{(M-1)(M-2)} \approx \frac{\sum_{i=1}^{M}\sum_{j=1}^{M} w_{ij}}{M^2},$$

and we see that $\mathrm{wclust}(K)$ then reflects the average connectivity between the other nodes in the network.

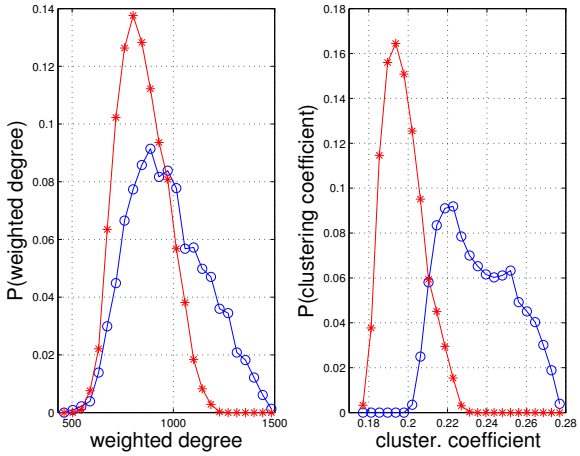


Figure 4: Probability of weighted degree (left) and clustering coefficient (right). Lymphoma: normal (circles) and tumour (stars).

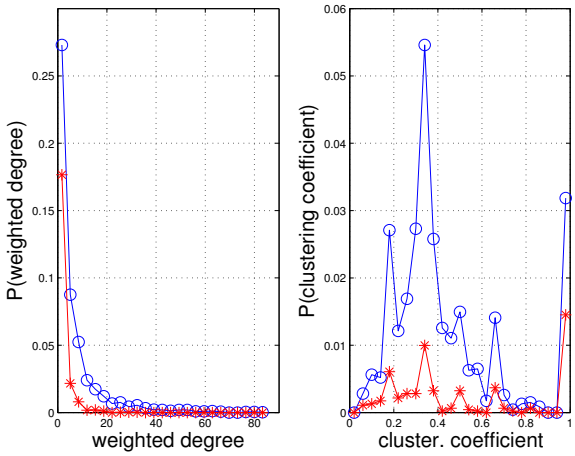


Figure 5: Probability of weighted degree (left) and clustering coefficient (right). Threshold 0.8. Lymphoma: normal (circles) and tumour (stars).

3.4 Range Dependent Weights

The concept of a *range-dependent weighted random graph*, or RENGA, was introduced and analyzed by Grindrod (2002) and further studied by Higham (2005). We may adapt this idea to the case of non-random range-dependent weights. Suppose that the nodes are ordered $1, 2, 3, \ldots, M$ and that the connectivity weight decays as a function of lattice distance. To be specific, we let

$$w_{ij} = w_{ji} = \lambda^{|i-j|}, \tag{5}$$

for some $\lambda \in [0, 1]$. At one extreme, $\lambda \approx 0$, there are no edges after discretising to a binary network, and hence the traditional clustering coefficient is undefined. At the other extreme, $\lambda = 1$, all edges are present after discretising to a binary network, and hence the traditional clustering coefficient is 1 for each node. In Figure 1 we use networks of size $M = 50, 100, 200$ and compute the generalised clustering coefficient (3) for the central node, $k = \frac{1}{2}M$, as λ ranges from 0 to 1. Note that the definition (3) makes sense for any $\lambda > 0$. We see that the clustering coefficient approaches the value zero as λ approaches zero from above; this is perfectly reasonable behaviour. Further as λ increases away from zero, the clustering coefficient monotonically increases, and it matches the binary value of 1 at $\lambda = 1$. Overall, the generalised version provides a natural, informative interpolation of the classical clustering coefficient.

4 Microarray Illustration

We now examine the distribution of the clustering coefficient (3) in practice, along with that of the corresponding weighted degree, using pairwise correlation networks arising from cDNA microarray data. Most importantly, we would like to explore differences in character of weighted degree and clustering coefficient distributions of two different networks: normal and tumour.

The initial gene expression data arising from cDNA microarray experiments is a rectangular $M \times N$ matrix A of log-transformed ratios a_{ij} of $i = 1, \ldots, M$ genes in a set of $j = 1, \ldots, N$ samples. We consider the Pearson correlation

$$\mathrm{cor}(i, j) = \frac{\sum_{k=1}^{N}(a_{ik} - \mu_i)(a_{jk} - \mu_j)}{\sigma_i \sigma_j},$$

where μ_i and σ_i are respectively the mean and the standard deviation of gene i log-ratios, as a measure of similarity between the gene expression profiles. We define pairwise gene similarity weights $w_{ij} = |\mathrm{cor}(i, j)|$, for $1 \leq i, j \leq M$, with $w_{ij} = w_{ji} \in [0, 1]$ and $w_{ii} = 0$. A large weight w_{ij} indicates that genes i and j are highly co-expressed (or anti-expressed). In this representation $M \times M$ matrix W denotes the symmetric weight matrix encoding the strength of connection between pairs of genes.

Aware of the fact that different number of genes as well as samples in data sets can affect values of correlation and consequently distort comparisons of both weighted degrees and clustering coefficients,

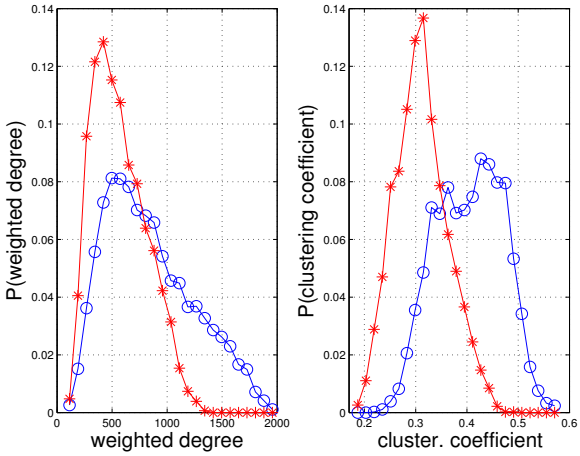


Figure 6: Probability of weighted degree (left) and clustering coefficient (right). P value 0.05. Lymphoma: normal (circles) and tumour (stars).

we looked for data consisting of the same number of normal and tumour samples for the same set of genes. In this experiment we used cDNA microarray data for normal and tumour tissues, taken from (Choi et al., 2005) and downloaded from http://centi.kribb.re.kr/MMA/index.html. Data processing performed by the authors included filtering of genes with more than 70% missing values or less than 4 observations, UniGene mapping, and imputation of missing values. The original data can be downloaded from the Stanford Microarray Database.

We selected data sets with more than ten samples in normal and tumour subsets. We present three of the results in this paper: breast cancer (5603 genes, 13 samples), liver cancer (12065 genes, 76 samples) lymphoma (4615 genes, 31 samples). Figure 2 shows the distribution of the weighted clustering coefficient (right), and also the distribution of the weighted degree (left) arising from breast cancer data. Circle-line and star-line represent the distributions of normal and tumour networks respectively. Figure 3 and Figure 4 show results for liver cancer and lymphoma.

We emphasize that our aim is to study the 'big-picture' issue of overall network topology, as opposed to the 'fine-detail' issue of clustering individual genes and/or samples (Kluger et al., 2003). The figures reveal global topological differences between the two networks. In general the tumour samples give rise to smaller and more peakily distributed clustering and degree. Degree ranges of normals and tumours start from a similar value but the degree range of tumours is narrower. Large numbers of genes in normal samples show a high degree of connection to other genes. Differences in clustering coefficient distributions are more striking. Distribution ranges of normal and tumour networks only partly overlap: most genes in normal networks have higher correlation than any gene in tumour networks.

Given that the weighted clustering coefficient produces interesting results, it is pertinent to ask whether careful thresholding to a discretised binary network (Rougemont and Hingamp, 2003) can also reproduce these findings. Clearly there is a whole parameterized family of such binary networks. In particular, high thresholding may exclude interesting features of the networks. For example, when weights above the threshold of 0.8 are re-set to 1 and the remaining weights are re-set to zero, the clustering coefficient and weighted degree distributions could not reveal the differences observed from original networks; see Figure 5.

For a more systematic approach, P values may be used to decide on significance of correlation. Even in this case, however, somewhat arbitrary thresholds must be imposed. For the lymphoma networks, suppose we take the view that correlations ≥ 0.355 are significant (corresponding to $P \leq 0.05$) and correlations ≥ 0.456 are highly significant (corresponding to $P \leq 0.01$). This would mean that only 18% (12%) of all possible edges are significant and 8% ($< 5\%$) are highly significant in the normal (tumour) lymphoma network, so that a large amount of data is being discarded. (Of course, there are computational benefits from introducing sparsity, but for the network sizes in these experiments this is not a significant issue.) In Figure 6 we plot data for the $P \leq 0.05$ binary networks. Comparing with Figure 4, we see that very similar topology is revealed. This allows us to conclude that the parameter-free weighted clustering coefficient approach is not affected by insignificant or "by chance" values, and automatically produces results consistent with the P value version.

5 Summary

Our aim here was to argue that out of the possible ways that have been proposed to generalise the clustering coefficient to the case of a weighted network, there is one very promising candidate; namely the Grindrod-Zhang-Horvath version (Grindrod, 2002; Zhang and Horvath, 2005). We gave a natural derivation and illustrated its behaviour on specific classes of network. Particular advantages of the definition are:

- It is a true generalisation, collapsing smoothly to

the binary case when edge weights tend to $\{0, 1\}$ values.

- It can provide meaningful results in cases where any type of binary thresholding produces breakdown.
- It reveals natural topological properties of real networks, and can do this without the need to specify parameters or discard potentially useful data.

Acknowledgements

Both authors are supported by EPSRC grant GR/S62383/01.

References

A. Barabasi and R. Albert. Emergence of scaling in random networks. *Science*, 286:509–512, 1999.

A. Barrat, M. Barthélemy, R. Pastor-Satorras, and A. Vespignani. The architecture of complex weighted networks. *PNAS*, 101:3747–3752, 2004.

X. Chen, S.T. Cheung, S. So, S.T. Fan, C. Barry, J. Higgins, K.M. Lai, J. Ji, S. Dudoit, I.O. Ng, M. Van De Rijn, and P.O. Botstein, D. Brown. Gene expression patterns in human liver cancers. *Molecular Biology of the Cell*, 13(6):1929–1939, 2002.

J.K. Choi, U. Yu, O.J. Yoo, and S. Kim. Differential coexpression analysis using microarray data and its application to human cancer. *Bioinformatics*, 21: 4348–4355, 2005.

T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caliguri, C.D. Bloomfield, and E.S. Lander. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286:531–537, 1999.

P. Grindrod. Range-dependent random graphs and their application to modeling large small-world proteome datasets. *Physical Review E*, 66:066702, 2002.

D. J. Higham. Spectral reordering of a range-dependent weighted random graph. *IMA Journal of Numerical Analysis*, 25:443–457, 2005.

R. Khanin and E. Wit. How scale-free are gene networks? *Computational Biology*, to appear.

Y. Kluger, R. Basri, J.T. Chang, and M. Gerstein. Spectral biclustering of microarray data: coclustering genes and conditions. *Genome Research*, 13 (4):703–716, 2003.

L. Lopez-Fernandez, G. Robles, and J.M. Gonzalez-Barahona. Applying social network analysis to the information in cvs repositories. *In Proc. of the 1st Intl. Workshop on Mining Software Repositories (MSR2004)*, pages 101–105, 2004.

M.E.J. Newman. The structure and function of complex networks. *SIAM Review*, 45:167–256, 2003.

J.-P. Onnela, J. Saramki, J. Kertsz, and K. Kaski. Intensity and coherence of motifs in weighted complex networks. *Phys. Rev. E*, 71(6):065103, 2005.

N. Pržulj, D.G. Corneil, and I. Jurisica. Modeling interactome: Scale-free or geometric? *Bioinformatics*, 20(18):3508–3515, 2004.

S. Ramaswamy, P. Tamayo, R. Rifkin, S. Mukherjee, C.-H. Yeang, M. Angelo, C. Ladd, M. Reich, E. Latulippe, J.P. Mesirov, T. Poggio, W. Gerald, M. Loda, E.S. Lander, and T.R. Golub. Multiclass cancer diagnosis using tumor gene expression signatures. *PNAS*, 98(26):15149–15154, 2001.

J. Rougemont and P. Hingamp. DNA microarray data and contextual analysis of correlation graphs. *BMC Bioinformatics*, 4(15), 2003.

T. Schank and D. Wagner. Approximating clustering coefficient and transitivity. *J. of Graph Algorithms and Applications*, 9(2):265–275, 2005.

M.R. Segal. Microarray gene expression data with linked survival phenotypes: Diffuse large-b-cell lymphoma revisited. *Biostatistics*, 2005.

J.M. Stuart, E. Segal, D. Koller, and S.K. Kim. A gene-coexpression network for global discovery of conserved genetic modules. *Science*, 302(5643): 249–255, 2003.

D. Watts and S. Strogatz. Collective dynamics of small-world networks. *Nature*, 393:440–442, 1998.

B. Zhang and S. Horvath. A general framework for weighted gene co-expression network analysis. *Statistical Applications in Genetics and Molecular Biology*, 4, 2005.

Scale-free Paradigm in Yeast Genetic Regulatory Network Inferred from Microarray Data

Kevin Kontos*
*ULB Machine Learning Group
Computer Science Department
Université Libre de Bruxelles
1050 Brussels – Belgium
http://www.ulb.ac.be/di/mlg/
kkontos@ulb.ac.be

Gianluca Bontempi†
†ULB Machine Learning Group
Computer Science Department
Université Libre de Bruxelles
1050 Brussels – Belgium
http://www.ulb.ac.be/di/mlg/
gbonte@ulb.ac.be

Abstract

A major challenge of computational biology is the inference of genetic regulatory networks and the identification of their topology from DNA microarray data. Recent results show that scale-free networks play an important role in this context. These networks are characterized by a very small number of highly connected and relevant nodes, and by numerous poorly connected ones. In this paper, we experimentally assess the predictive power of the scale-free paradigm in a supervised learning framework. The hypotheses we intend to test in this framework are: (i) regulatory genes are effective predictors of the expression of the genes they regulate; (ii) a subset of regulatory genes may explain most of the variability of the measures. More precisely, we use the expression levels of a subset of regulatory genes, returned by feature selection, as input of a learning machine which has to predict the expression levels of the target genes. We will show that (i) each gene can be predicted by a small subset of regulatory genes, and (ii) on a global scale, a small subset of regulatory genes, called the hubs, can have a non-negligible predictive power on all the target genes. Also, most of the regulatory genes returned by the application of this approach to Gasch et al. (2000) data were identified by Segal et al. (2003) and form a biologically coherent set of genes.

1 Introduction

The *inference* of genetic regulatory networks from DNA microarray data is one of the major challenges in systems biology. A critical issue in network inference is the identification of the *network topology* from noisy data. Recent results tend to show that *scale-free networks* play an important role in systems biology (Jeong et al., 2000; Barabási and Oltvai, 2004; Barabási et al., 2004), notably for the yeast *Saccharomyces cerevisiae* model organism (Farkas et al., 2003; van Noort et al., 2004). These networks are characterized by a very small number of highly connected and relevant nodes, called the *hubs*, and by numerous poorly connected ones.

In this paper, we experimentally assess the role of the scale-free paradigm in a *supervised learning* approach to network inference in the case of the *Saccharomyces cerevisiae* organism. The idea is that relevant genes should emerge as good *predictors* in a multi-input multi-output supervised learning approach where the inputs are the *regulatory genes*[1] and the outputs are the *target genes* (Zhou et al., 2004). In particular, we will show that (i) for each gene, few other genes have an important predictive power, and (ii) a few genes (the hubs) have an important predictive power on *all* the target genes.

The main contributions of the paper are described hereafter. First, a *supervised learning framework* for network inference is introduced in Sect. 2. This means that the dependency between genes is estimated by the predictive power that the *expression levels* of a set of regulator genes have on the expression levels of some target genes. Also, because of the large ratio between the number of genes and the number of experimental conditions, we propose a *feature selection* strategy based on the Gram-Schmidt (GS) orthogonalization procedure (Stoppiglia et al., 2003). This procedure generates an *individual variable rank-*

[1] Transcriptional regulatory genes, also known as *regulator genes* or simply *regulators*, produce transcription factors (TF), which are regulatory proteins that regulate non-coding DNA segments (so called TF binding motifs) of target genes and initiate the transcription process.

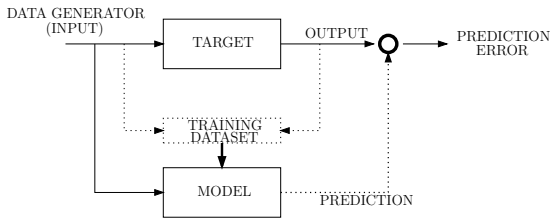


Figure 1: Supervised learning setting.

ing for each target gene. The first genes of each of these rankings constitute the small subsets of genes to be used as *predictors* in the supervised learning problem. According to the scale-free paradigm, only a small subset of genes plays a crucial role and these genes are thus expected to be well ranked in all the individual rankings. In order to assess this hypothesis, an *aggregated ranking* is built by "averaging" the individual rankings.

The experimental session relies on the dataset described in Gasch et al. (2000). The predictive power of the selected genes is assessed via a *cross-validation* strategy (Hastie et al., 2001) for a conventional *linear* model (D'haeseleer et al., 1999).

2 Supervised Learning Framework for Network Inference

Let us represent a DNA microarray dataset by a $N \times n$ matrix E, where N is the number of samples, n is the number of genes, and $E[C_j, G_i] = expr_{C_j}^{G_i}$ denotes the expression measure of gene G_i in mRNA sample C_j.

Let $\mathcal{T}$ be the set of target genes G_i, $i \in \{1, \ldots, |\mathcal{T}|\}$. Typically, this set is constituted by all genes, and thus $|\mathcal{T}| = n$. Also, let $\mathcal{R}$ be the set of regulatory genes RG_i, $i \in \{1, \ldots, |\mathcal{R}|\}$.

The issue of modeling the statistical dependencies between gene expression levels can be described as a supervised learning problem (see Fig. 1) characterized by the following elements: a data generator (the input), a target operator (the output), a training set and a learning machine (Vapnik, 1998).

The goal of a learning machine is to return a hypothesis with low *prediction error*, i.e. a hypothesis which computes an accurate estimate of the output of the target when the same value is an input to the target and the predictor. The prediction error is also usually called *generalization error* since it measures the capacity of the hypothesis to generalize, that is to return a good prediction of the output for input values not contained in the training set.

A typical way of representing the unknown input/output relation is the *regression plus noise form*[2]: $\mathbf{y} = f(x) + \mathbf{w}$, where $f(\cdot)$ is a deterministic function, also known as the *regression function*, and the term $\mathbf{w}$ represents the noise or random error. It is typically assumed that $\mathbf{w}$ is independent of $\mathbf{x}$ and $E[\mathbf{w}] = 0$.

Concerning the expression levels of the genes, the following dependency is assumed for each gene[3] $G_i \in \mathcal{T}$:

$$\mathbf{expr}^{G_i} = f_i(expr^{RG_1}, \ldots, expr^{RG_{|\mathcal{R}|}}) + \mathbf{w} \ .$$

The goal of the machine learning is to find a model $h(\cdot)$ which is able to give a good approximation of the unknown function $f(\cdot)$ by minimizing an estimation of the *mean integrated squared error* (MISE), which measures the generalization error in the case of a quadratic cost error.

In this paper, we will consider the *leave-one-out* (LOO) algorithm to return an estimate of the MISE prediction error:

$$\widehat{\text{MISE}}_{\text{LOO}} = \frac{1}{N} \sum_{i=1}^{N} (y_i - h(x_i, \alpha_N^{-i}))^2 \ ,$$

where α_N^{-i} is the set of parameters returned by the parametric identification performed on the training set with the ith sample set aside.

The parametric identification of the hypothesis is done according to the *empirical risk minimization* (ERM) principle (Vapnik, 1998).

In order to assess more easily the quality of a $\widehat{\text{MISE}}_{\text{LOO}}$ estimate, we will focus on the LOO estimate of the *normalized mean integrated squared error* (NMISE):

$$\widehat{\text{NMISE}}_{\text{LOO}} = \frac{\widehat{\text{MISE}}_{\text{LOO}}}{\text{Var}[\mathbf{y}]} \ .$$

The $\widehat{\text{NMISE}}_{\text{LOO}}$ of a predictor is by definition positive and the closer it is to zero, the better is the generalization accuracy of the predictor. Note that the simplest predictor of the output variable, i.e. the average:

$$\hat{y}_i = \frac{1}{N} \sum_{j=1}^{N} y_j \ , \qquad i = 1, \ldots, N,$$

has an $\widehat{\text{NMISE}}_{\text{LOO}}$ of one. It follows that a value of $\widehat{\text{NMISE}}_{\text{LOO}}$ close to one for a supervised learning predictor has to be interpreted as a sign of bad accuracy.

[2] Throughout this paper, boldface denotes random variables.

[3] If the target gene G_i to be predicted is a regulatory gene, then it will not appear among the inputs.

3 Feature selection

The supervised learning formulation of the dependency between expression levels leads to a prediction problem where the number of inputs is very large with respect to the number of samples. Due to the very high dimensional input space, conventional supervised learning techniques can perform badly (Guyon and Elisseeff, 2003). A preliminary feature selection step is then required.

Many feature selection algorithms include *variable*[4] *ranking*, i.e. ranking variables according to their individual predictive power, as a principal or auxiliary selection mechanism because of its simplicity, scalability, and good empirical success (Guyon and Elisseeff, 2003). Computationally, it is efficient since it requires only the computation of e scores, where e is the number of input variables, and the sorting of the scores. Statistically, it is robust against over-fitting because it introduces bias but it may have considerably less variance (Hastie et al., 2001). Therefore, although variable ranking is not optimal, it may be preferable to variable subset selection methods because of its computational and statistical scalability.

In this paper we adopt the Gram-Schmidt (GS) orthogonalization (Stoppiglia et al., 2003) ranking procedure. Given a set of e candidate features, there are 2^e possible models. Only e models for selection are considered in this paper: the model with the feature ranked first, the model with the first two features, and so on. The price paid for that complexity reduction is the fact that there is no guarantee that the best model is among the e models generated by the procedure.

4 Method

This section presents the algorithmic procedure adopted to assess the predictive power of a number of regulator genes on their targets and the effectiveness of a scale-free aggregation of the most relevant genes.

First, because of the very high dimensional input space, a subset $\mathcal{V}$ of the set of regulatory genes $\mathcal{R}$ is selected. For this purpose:

1. We use the GS orthogonalization algorithm to obtain a ranking of the regulatory genes for each gene G_i of set $\mathcal{T}$. These rankings are called the *individual rankings* $\mathcal{IR}^i$.

2. We build an *aggregate ranking* of the regulatory genes by averaging the positions of each regulatory gene in the individual rankings.

3. We build a random ranking $\mathcal{RR}$, where the regulatory genes are randomly ranked (this is done for comparison purposes – see Sect. 6).

4. We build the subset $\mathcal{V} = \{RG_{v_1}, \ldots, RG_{v_{|\mathcal{V}|}}\}$ of predictors (i) by fixing the number $|\mathcal{V}|$ of genes to be considered in the subset $\mathcal{V}$ of predictors, (ii) by choosing between the individual rankings $\mathcal{IR}^i$, $i \in \{1, \ldots, |\mathcal{T}|\}$, the aggregated ranking $\mathcal{AR}$ and the random ranking RR, and (iii) by taking the $|\mathcal{V}|$ first genes of the chosen ranking.

Second, the function $h_i(\cdot)$, corresponding to a conventional *linear* model (D'haeseleer et al., 1999), is estimated according to the ERM principle (Vapnik, 1998) for each[5] $G_i \in \mathcal{T}$, where

$$\widehat{expr}^{G_i} = h_i(expr^{RG_{v_1}}, \ldots, expr^{RG_{v_{|\mathcal{V}|}}}) \ .$$

Then, the generalization capacity of each model is assessed by estimating the generalization error with the $\widehat{\text{NMISE}}_{\text{LOO}}$. Finally, the average of the $\widehat{\text{NMISE}}_{\text{LOO}}$ obtained is computed.

For each target gene, we also computed the numbers of genes of its individual ranking to be added to $\mathcal{V}$ before the $\widehat{\text{NMISE}}_{\text{LOO}}$ starts to increase. In other words, for each target gene G_{tg}, we start by taking the first gene G of its individual ranking $\mathcal{IR}^{tg}$, i.e. $\mathcal{V} = \{G\}$. Second, if the $\widehat{\text{NMISE}}_{\text{LOO}}$ obtained with the set of genes $\mathcal{V} \cup \{G'\}$, where G' is the next gene of $\mathcal{IR}^{tg}$, is smaller than the one obtained with the genes of $\mathcal{V}$ solely, then G' is added to $\mathcal{V}$ and we continue by considering the following gene of $\mathcal{IR}^{tg}$. If this is not the case, we stop. We thus obtain sets of regulators of varying sizes. Finally, we count the number of genes "targeted" by each regulator, i.e. the number of occurrences of each regulator in these sets.

5 Materials

We now study the predictive power of the yeast *Saccharomyces cerevisiae* genes by applying our procedure on a DNA microarray data set described in Gasch et al. (2000). We used the list of know and *putative* regulatory genes of the yeast *Saccharomyces cerevisiae* used by Segal et al. (2003).

[4] The terms *variable* and *feature* are used interchangeably in this paper. Note, however, that a distinction is sometimes made in the literature (Guyon and Elisseeff, 2003).

[5] If the target gene G_i to be predicted is a regulatory gene, then it will not appear among the inputs.

Table 1: Averages of the $\widehat{\text{NMISE}}_{\text{LOO}}$ in the linear case for different number $|\mathcal{V}|$ of regulatory genes for the individual rankings, the aggregated ranking and the random ranking.

$\lvert\mathcal{V}\rvert$	$\mathcal{IR}$ (ind.)	$\mathcal{AR}$ (agg.)	$\mathcal{RR}$ (rand.)
1	0.606	0.781	1.003
2	0.550	0.768	0.854
3	0.527	0.758	0.839
4	0.515	0.741	0.829
5	0.508	0.740	0.787
6	0.504	0.732	0.785
7	0.501	0.725	0.779
8	0.499	0.710	0.763
9	0.497	0.710	0.762
10	0.496	0.698	0.713
15	0.494	0.665	0.678
20	0.497	0.662	0.666
25	0.502	0.652	0.645
30	0.510	0.644	0.649

6 Results

6.1 Predictive Power

The averages of the predictive $\widehat{\text{NMISE}}_{\text{LOO}}$ for different number $|\mathcal{V}|$ of regulatory genes are presented in Table 1. The table contains the results for the individual rankings, the aggregated ranking and the random ranking (average over 10 realizations). It is worthy to remark that all the differences between rankings reported in the table are statistically significant (p-value=0.01) according to a paired t-test on the error vectors and that the results obtained when permuting the outputs have a poor $\widehat{\text{NMISE}}_{\text{LOO}}$ (> 1.4) and are significantly worse than the non permuted results.

The leave-one-out assessment of the predictive power shows that a small subset of genes (about 4 to 8) can have a significantly better performance than the random case and, although inferior, not too far from the predictive power obtained with individual rankings. However, the improvement of the aggregated ranking with respect to a random selection vanishes for larger number of inputs. This suggests that the variability of most of the genes can be explained by a small subset of regulators composed of 4 to 8 genes.

Concerning the individual rankings, the leave-one-out assessment shows that small subsets of genes (about 4 to 8) can each have an important predictive power. Moreover, no significant improvements in terms of $\widehat{\text{NMISE}}_{\text{LOO}}$ occur when bigger subsets are considered. This suggests that, each gene can be predicted by a small number of regulatory genes.

Another outcome of the feature selection procedure is the histogram illustrating the distribution of the number of target genes regulated by a regulatory gene. As shown in Fig. 2(a), it appears that many regulatory genes regulate a small percentage (less than 1%) of genes, while few regulatory genes regulate around 10% of the genome (i.e. about 600 genes). The log-log plot (Fig. 2(b)) of the histogram suggests an underlying power-law distribution, as confirmed by (i) the fitting of a generalized Pareto distribution to the data (Fig. 2(c)) and (ii) the insufficient evidence to reject the null hypothesis that the underlying distribution is a generalized pareto distribution (p-value=0.49) as returned by the hypothesis testing method discussed in Goldstein et al. (2004).

6.2 Biological validation of results

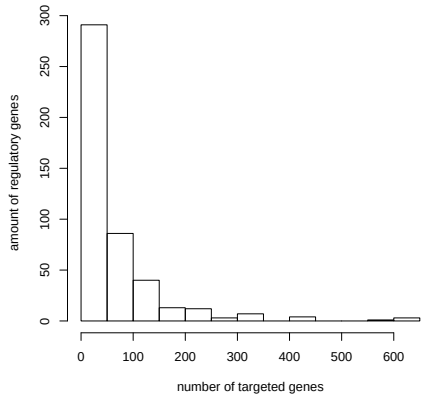
This section aims to show that most of the regulatory genes well ranked in our method's aggregated ranking $\mathcal{AR}$ (i) correspond to regulatory genes identified by Segal et al. (2003) and (ii) form a biologically coherent set of genes. Moreover, (iii) the predictive power of the regulatory genes found by Segal et al. (2003) is comparable to the one of the regulatory genes our procedure identified.

In order to compare with our method, let $\mathcal{AR}^i$ be the set composed of the i first genes of the aggregated ranking $\mathcal{AR}$, let $\mathcal{S}_{all}$ be the set of the 60 regulatory genes identified in Segal et al. (2003), and let $\mathcal{S}_{main}$ be the subset of 22 regulators considered as the "main" regulators of $\mathcal{S}_{all}$.

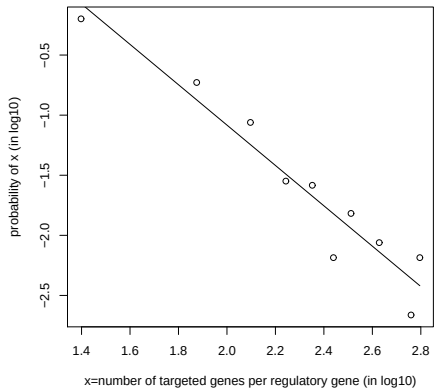
Table 2 shows that 32 to 50% and 60 to 90% of the regulatory genes identified by sets $\mathcal{AR}^i$, $i \in \{10, 15, 20, 25, 30\}$, are also in $\mathcal{S}_{main}$ and $\mathcal{S}_{all}$, respectively. Moreover, these results are highly significant. Indeed, the corresponding p-values of the hypergeometric distribution are all smaller than 1.0×10^{-4}.

The 10 first genes of the aggregated ranking $\mathcal{AR}$ obtained with our procedure are listed in Table 3 (genes in bold belong to set $\mathcal{S}_{main}$). We noted that 5 of these genes, identified by an asterisk (*), form a biologically coherent set of genes as they are involved in starvation, in nutrient limitation or in nutrient control.

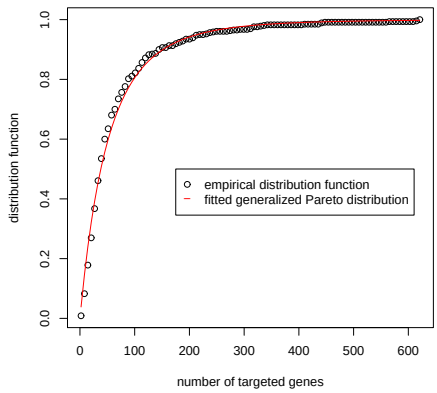
The predictive power of the regulators obtained by Segal et al. (2003) is similar, in terms of $\widehat{\text{NMISE}}_{\text{LOO}}$, to the one obtained with our method: averages of



(a) Histogram of the number of target genes regulated by a regulatory gene.



(b) Log-log plot of the histogram.



(c) Fitting of a generalized Pareto distribution to the data.

Figure 2: Number of target genes regulated by a regulatory gene.

Table 2: Number of regulatory genes of sets $\mathcal{AR}^i$, $i \in \{10, 15, 20, 25, 30\}$, that belong to $\mathcal{S}_{main}$ and $\mathcal{S}_{all}$, respectively.

	$\mathcal{AR}^{10}$	$\mathcal{AR}^{15}$	$\mathcal{AR}^{20}$	$\mathcal{AR}^{25}$	$\mathcal{AR}^{30}$
$\mathcal{S}_{all}$	9	11	12	15	18
$\mathcal{S}_{main}$	5	7	8	8	10

Table 3: The 10 first genes of the aggregated ranking obtained with our procedure. Genes in bold belong to $\mathcal{S}_{main}$. Genes identified by an asterisk (*) are involved in starvation, in nutrient limitation or in nutrient control.

Rank	Systematic name	Standard name
1	**YPL230W***	ORF Uncharacterized
2	**YJL164C**	TPK1
3	**YGL099W**	LSG1
4	**YDR096W***	GIS1
5	YER118C	SHO1
6	YLL019C	KNS1
7	YGL208W*	SIP2
8	YPL203W*	TPK2
9	**YIL101C***	XBP1
10	YJL103C	ORF Uncharacterized

the $\widehat{\text{NMISE}}_{\text{LOO}}$ by using both lists of main predictors $\mathcal{S}_{main}$ and of all predictors $\mathcal{S}_{all}$ are 0.688 and 0.620, respectively. The ranking produced by our method thus seems to yield better results compared to list $\mathcal{S}_{all}$ and less good results compared to $\mathcal{S}_{main}$ in terms of $\widehat{\text{NMISE}}_{\text{LOO}}$.

7 Conclusion

In this paper, we experimentally assessed the predictive power of the scale-free paradigm in a supervised learning framework. The results obtained, although preliminary, tend to validate this paradigm. Indeed, it appears that (i) each gene can be predicted by a small subset of regulatory genes, and (ii) on a global scale, a small subset of regulatory genes, called the hubs, can have a non-negligible predictive power on all the target genes.

Moreover, most of the regulatory genes well ranked in our procedure correspond to regulatory genes found by Segal et al. (2003) and form a bio-

logically coherent set of genes.

Future work will focus on testing other feature selection algorithms and using other learning algorithms. Another interesting direction consists in generating in silico DNA microarray data for given networks by using existing simulation techniques. The effect of network topology on the predictive power that the expression levels of a set of regulator genes have on the expression levels of some target genes could then be more accurately assessed by using different topologies (scale-free, random and small-world network topologies for example).

Acknowledgements

This research was supported by the "Communauté Française de Belgique" through an ARC project (no. 04/09–307) research grant. The project is entitled "Integrating experimental and theoretical approaches to decipher the molecular networks of nitrogen utilisation in yeast".

References

A.-L. Barabási and Z. N. Oltvai. Network biology: Understanding the cell's functional organization. *Nature Reviews*, 5:101–114, 2004.

A.-L. Barabási, Z. N. Oltvai, and S. Wuchty. Characteristics of biological networks. *Lecture Notes in Physics*, 650:443–457, 2004.

P. D'haeseleer, X. Wen, S. Fuhrman, and R. Somogyi. Linear modeling of mRNA expression levels during CNS development and injury. In *Proceedings of the Pacific Symposium on Biocomputing*, 4: 41–52, 1999.

I. Farkas, H. Jeong, T. Vicsek, A.-L. Barabási, and Z. N. Oltvai. The topology of the transcription regulatory network in the yeast, *Saccharomyces cerevisiae*. *Physica A*, 318:601–612, 2003.

A. P. Gasch, P. T. Spellman, C. M. Kao, O. Carmel-Harel, M. B. Eisen, G. Storz, D. Botstein, and P. O. Brown. Genomic expression programs in the response of yeast cells to environmental changes. *Molecular Biology of the Cell*, 11:4241–4257, 2000.

M. L. Goldstein, S. A. Morris, and G. G. Yen. Problems with fitting to the power-law distribution. *The European Physical Journal B*, 41:255–258, 2004.

I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics. Springer, 2001.

H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A.-L. Barabási. The large-scale organization of metabolic networks. *Nature*, 407(6804):651–654, 2000.

E. Segal, M. Shapira, A. Regev, D. Pe'er, D. Botstein, D. Koller, and N. Friedman. Module networks: Identifying regulatory modules and their condition-specific regulators from gene expression data. *Nature Genetics*, 34(2):166–176, 2003.

H. Stoppiglia, G. Dreyfus, R. Dubois, and Y. Oussar. Ranking a random feature for variable and feature selection. *Journal of Machine Learning Research*, 3:1399–1414, 2003.

V. van Noort, B. Snel, and M. A. Huynen. The yeast coexpression network has a small-world, scale-free architecture and can be explained by a simple model. *EMBO reports*, 5(3):280–284, 2004.

V. N. Vapnik. *Statistical Learning Theory*. Wiley Series on Adaptive and Learning Systems for Signal Processing, Communications, and Control. Wiley-Interscience, 1998.

X. Zhou, X. Wang, and E. R. Dougherty. Gene prediction using multinomial probit regression with Bayesian gene selection. *EURASIP Journal of Applied Signal Processing*, 1:115–124, 2004.

Generating networks with realistic properties: The topology of locally evolving random graphs

Andreas Krause*
*School of Management
University of Bath
Bath BA2 7AY
Great Britain mnsak@bath.ac.uk

Abstract

We present a model in which a random graph evolves locally by randomly changing edges in the immediate neighborhood of a node. We find that the emerging graph maintains its small diameter and obtains a high clustering coefficient as well as a power-law tail in the degree distribution. The scaling of the diameter and clustering coefficient of the model are also obtained.

Keywords: Evolving graphs, random graphs, power law, local search
PACS numbers: 89.75.Hc, 89.65.-s
Mathematics Subject Classification: 05C80, 91D30
JEL Classification: D85

1 Introduction

Many investigations into the properties of real networks have shown that they can mostly be characterized by three key properties: a small diameter, a high clustering coefficient and a power-law tail of the degree distribution, Albert and Barabási (2002) provide a comprehensive review. Since Watts and Strogatz (1998) introduced small world networks there has been an abundance of models with high clustering coefficients and small diameters as well as models of preferential attachment generating a power-law tail of the degree distribution following the introduction of scale-free networks by Barabási and Albert (1999).

Until the introduction of local edge formation, however, it had been proved difficult to generate graphs exhibiting all desired properties simultaneously. The vast majority of models proposed using local edge formation thus far are models of growing networks, either through adding new links and/or new nodes, (Davidsen et al., 2002; Jost and Joy, 2002; Vazquez, 2003; Csányi and Szendrői, 2003; Li and Chen, 2003; Blanchard and Krüger, 2004; Geng and Li, 2005). While growing networks might be appropriate for some applications, in many cases it would be more adequate to assume that the number of nodes and edges remains constant and a mechanism is required that relies only on re-connecting these edges and thereby generating realistic properties.

Blanchard et al. (2005) developed a model in which after an initial phase of a growing network, existing edges get re-connected without new nodes or edges being added. Their mechanism requires in each time period a single node to re-align one of his edges to another node which has a distance of two, what they call their "my friends are your friends" principle. After sufficient time periods this model shows a high clustering coefficient, small diameter and a fat-tailed degree distribution, but no evidence for a power-law tail as found in many real networks.

In this paper we employ a very similar rule for the evolution of a graph without adding or subtracting any nodes or edges. Our focus is on small networks having only a very limited number of nodes and we assume that the initial network is random rather than regular. As in Blanchard et al. (2005) we find evidence for a small diameter and a high clustering coefficient whose properties we compare with that of a random graph and in contrast to Blanchard et al. (2005) we find strong evidence for a power-law tail of the degree distribution.

2 The graph evolution

The starting point is a random undirected graph with $N > 2$ nodes in which a link exists between any two nodes with probability of $p \in (0; 1]$. We now let this graph evolve in discrete time steps using the following algorithm in each time step:

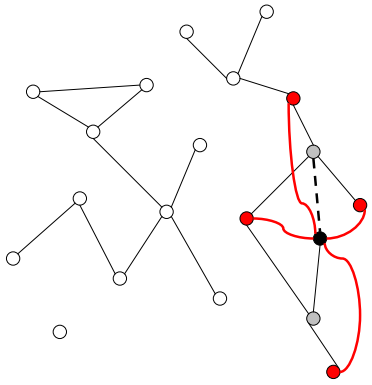


Figure 1: Local evolution of the graph. Suppose the black node is chosen for updating and intends to replace this thick dashed link. He will be able to form a link to any of his neighbor's neighbor, market red. Which link is actually chosen is randomly determined.

1. Select a node x randomly with equal probability for all nodes,

2. Select an edge i of this node randomly with equal probability for all edges,

3. Select randomly with equal probability for all nodes another node y which has a distance of 2 to the already selected node x,

4. If node x has no common edge with node y, remove edge i and replace it with an edge i' connecting nodes x and y,

5. If node x has no edges, there does not exist a node y which has a distance of 2 to node x or the selected node y already has a common edge with node x, no changes to the edges are made.

The number of time periods investigated in the analysis is denoted by T and we generally investigate the resulting properties after $T = 20N$ time periods. This algorithm is illustrated in figure 1.

This algorithm is in essence the same as used by Blanchard et al. (2005) with one important difference: their model starts with a simple circle to which they add links using a similar algorithm as above until they have obtained the desired number of links in the model and then from this point onwards let the graph evolve as described before.

3 The resulting network topology

Using the algorithm described in the previous section we conducted a number of simulations using a variety of parameter constellations, exploring any combination with $N \in \{50, 100, 150, 200, 250\}$ and $pN \in \{2, 4, 6, 8, 10\}$. For each parameter constellation we ran 100 simulations and use the average values for the diameter and clustering coefficient while we aggregate all nodes to obtain the degree distribution. Any analysis is conducted after $T = 20N$ time periods and we do not observe any significant changes when the number of time periods is extended further.

3.1 Network diameter

Most real networks have a diameter which is only slightly larger than that of a random graph. As we can establish from figure 2 this is also true in our model for $pN \geq 6$, i.e. if the average number of links of a node is at least 3. For random graphs the average path length ℓ scales as

$$\ell_{rand} \sim \frac{\ln N}{\ln pN}, \tag{1}$$

which we also find for our model, whose average path length is only about 10% higher than that of a random graph. For smaller values of pN, however, the scaling is approximately linear in N rather than logarithmic. A similar result has also been obtained by Blanchard et al. (2005) who found that an average of at least two links per node or more was required to obtain a small diameter.

3.2 Clustering coefficient

Another characteristic of real networks is a clustering coefficient which is significantly higher than that of a random graph with the same number of nodes and edges. For random graphs we know that

$$C_{rand} = p. \tag{2}$$

As can be seen from figure 3 the clustering coefficient in our model is significantly higher, with again the case of $pN = 2$ standing out slightly. For $pN \geq 4$ we find a different scaling of the clustering coefficient C as follows:

$$C = 20.6646 pN^{-0.2880}. \tag{3}$$

We thus observe a high clustering coefficient in our model which is slowly decreasing in the number of

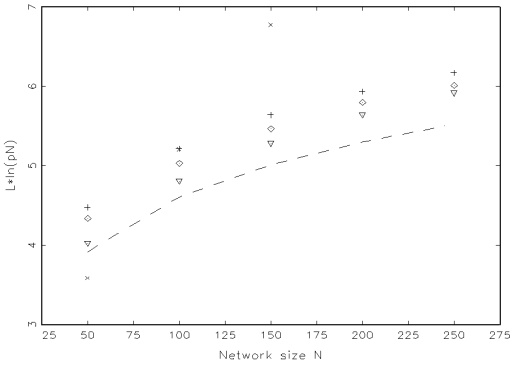


Figure 2: Average path length ℓ of the resulting graph for different sizes of the network N and probabilities of two nodes being connected p: $pN = 4$ ($\times$), 6 ($\bigtriangledown$), 8 ($\diamond$), 10 ($+$). Results are based on averaging over 100 runs for each of the 25 parameter constellations after $20N$ time steps. The dashed line represents the results from a random graph.

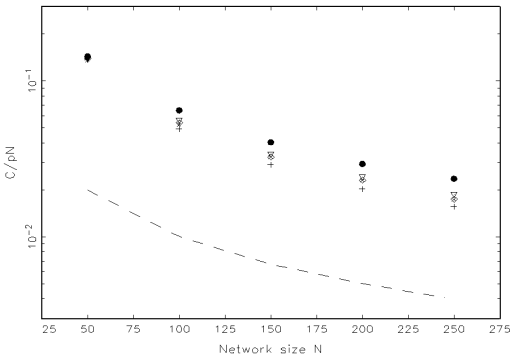


Figure 3: Clustering coefficient C of the resulting graph for different sizes of the network N and probabilities of two nodes being connected p: $pN = 2$ ($\bullet$), 4 ($\times$), 6 ($\bigtriangledown$), 8 ($\diamond$), 10 ($+$). Results are based on averaging over 100 runs for each of the 25 parameter constellations after $20N$ time steps. The dashed line represents the results from a random graph.

nodes, a property for which there is weak empirical evidence in the overview collated in Albert and Barabási (2002), where they mention the clustering to be nearly constant, although the graph they present suggests a small negative relationship.

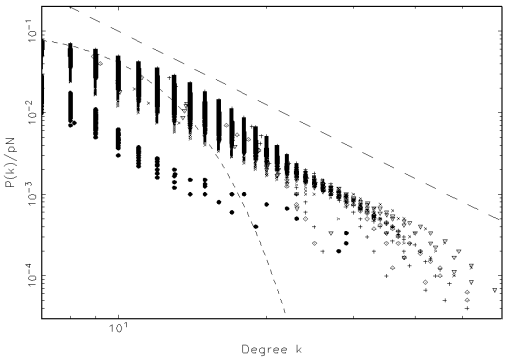


Figure 4: Degree distribution $P(k)$ of the tail from the resulting graph after $20N$ time steps for different sizes of the network N and probabilities of two nodes being connected p: $pN = 2$ ($\bullet$), 4 ($\times$), 6 ($\bigtriangledown$), 8 ($\diamond$), 10 ($+$). The short dashed line represents the results from a random graph with $pN = 10$ and the long dashed lines that of a power law distribution with an exponent of 3 (distribution shifted upwards for clarity). An exponential cut-off can be seen at approximately $k = 20$. The distribution is obtained from 100 simulations of each of the 25 parameter constellations using $pN = 2, 4, 6, 8, 10$ and $N = 50, 100, 150, 200, 250$.

3.3 Degree distribution

The degree distribution as illustrated in figure 4 shows clear evidence of a power-law tail with an exponent of approximately 3. This result is in clear contrast to the very similar model of Blanchard et al. (2005) who find evidence for fat tails but no sign of a power-law for the tail. Given that their algorithm is, apart from the initial phase, nearly identical to ours, this result is very surprising and merits further consideration of the relevance of the initial graph for these results which seems to have a significant influence on the results.

We observe from figure 4 that, again apart from the case of $pN = 2$, the degree distributions scale quite uniformly with a power-law tail. However, closer inspection of the distribution as illustrated in figure 5 shows that we do not observe a perfect power-law tail, but it rather appears to be the combination of two exponential tails. We generally observe an exponential cut-off at approximately $k = 20$. Given that we did not investigate larger graphs it has to be seen whether this observation can be explained with the finite size of the graph or is a more genuine feature of the algorithm used. Evidence from the graphs investigated

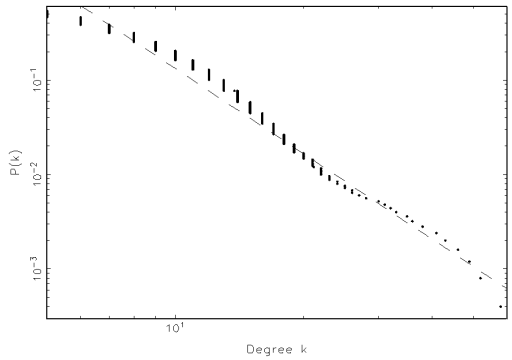


Figure 5: Degree distribution $P(k)$ from the resulting graph after 5,000 time steps for $N = 250$ nodes and $pN = 6$. The long dashed lines that of a power law distribution with an exponent of 3. An exponential cut-off can be seen at approximately $k = 20$. The distribution is obtained from 100 simulations with this parameter constellation.

here suggests that the finite size effect drives this result because the distribution moves closer to a power-law tail as we increase the number of nodes.

3.4 Graph structure

Apart from the properties discussed we see that one other important element of the resulting graph is that nodes with a small degree tend to reduce their degrees even further over time given that they are hardly attracting any new links but loose existing links at the same rate as any other node. This will inevitably result in such nodes becoming isolated or forming small subgraphs which are not connected with the large component or each other. Figure 6 shows an example of the initial random graph and the resulting graph after 2,000 time steps, clearly illustrating this property. We furthermore observe that the large component is usually very well connected and does not show clear evidence of any further distinguishable features, being quite homogeneous in its structure.

3.5 Explanation of findings

The algorithm used in this model generates a high clustering coefficient and a power-law tail of the degree distribution while maintaining the small diameter of the initial random graph. This result can easily be explained from the way new edges are formed. Any node with a large number of neighbors is also likely to have a large number of nodes with a distance

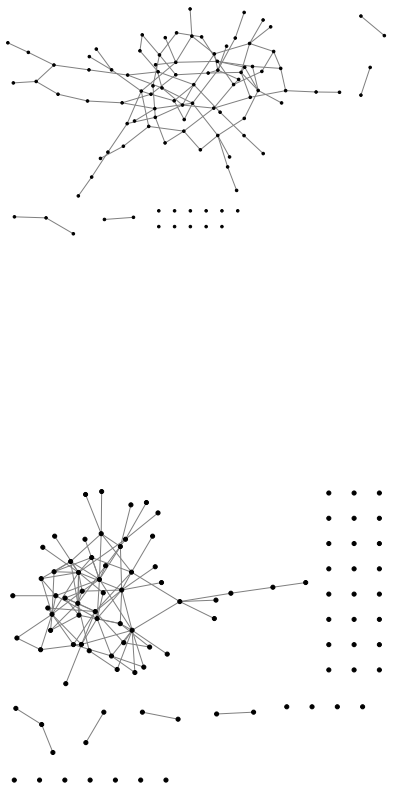


Figure 6: Single realization of an initial random graph with $N = 100$ and $Np = 2$ (top panel) and the evolved graph after $T = 2000$ time steps (bottom panel). It has to be noted that with a larger number of edges the results are qualitatively the same, but the large number of edges makes any representation very difficult to visualize.

of two, i.e. neighbor's neighbors, and thus is quite likely to be chosen as the destination of an edge by a randomly selected node. In contrast, a node with only few neighbors will also only have a small number of neighbor's neighbors and the probability of him being chosen as the destination of a link is relatively small, causing the number of edges connected to this node to fall and in some instances leading to a node becoming isolated. Thus nodes with a high degree tend to attract more links than those with a low degree, giving rise to a similar effect as the preferential attachment

of nodes in the Barabási and Albert (1999) model of scale-free graphs which exhibits a power-law tail of the degree distribution, also with an exponent of 3.

The high clustering coefficient emerges as the consequence of the local evolution of the graph where edges are established to a neighbor's neighbor, thus increasing the number of triangles in the graph and thereby increasing the clustering coefficient. The clustering coefficient will only be limited by the number of links available in the graph. The random nature of any connections is maintained, however, thus retaining the small diameter of the graph.

4 Discussion and conclusions

The model presented in this paper allows a graph to evolve randomly in its immediate neighborhood. The resulting graphs had properties that were largely consistent with those of real networks, namely the small diameter, high clustering coefficient and power-law tail of the degree distribution. The algorithm for the evolution of the graph needs to be adjusted, however, to avoid the appearance of a large number of isolated nodes or small components which are not realistic.

Nevertheless, the algorithm can be described as realistic for many social networks where new contacts are often made through existing contacts, the "friend's friend" or "neighbor's neighbor" principle, but it remains unclear at this stage how much the initial network structure affects the graph topology, especially in light of the results obtained by Blanchard et al. (2005) for the degree distribution. Future research needs to clarify the importance of the initial conditions for the results obtained here, in particular evaluating whether similar results can be obtained when starting with a regular graph.

It is furthermore of interest to evaluate the sensitivity of the way new links are determined by considering a wider range of rules, e.g. preferential attachment, as well as to include some random attachment outside the neighborhood to prevent nodes from becoming isolated as in our model.

References

Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Review of Modern Physics*, 74(1):47–97, 2002.

Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286:509–512, 1999.

Philippe Blanchard, Tyll Krüger, and Andreas Ruschhaupt. Small world graphs by iterated local edge formation. *Physical Review E*, 71(4):046139, 2005.

Phillippe Blanchard and Tyll Krüger. The "cameo principle" and the origin of scale-free graphs in social networks. *Journal of Statistical Physics*, 114 (516):1399–1416, 2004.

Gábor Csányi and Balász Szendrői. Structure of a large social network. cond-math/0305580, 2003.

Jörn Davidsen, Holger Ebel, and Stefan Bornholdt. Emergence of a small world from local interactions: Modelling acquaintance networks. *Physical Review Letters*, 88(12):128701, 2002.

Xianmin Geng and Qiang Li. Random models of scale-free networks. *Physica A*, 356:554–562, 2005.

Jürgen Jost and Maliackal P. Joy. Evolving networks with distance preferences. *Physical Review E*, 66 (3):036126, 2002.

Xiang Li and Guanrong Chen. A local-world evolving network model. *Physica A*, 328:274–286, 2003.

Alexei Vazquez. Growing network with local rules: Prefrential attachement, clustering hierarchy, and degree correlations. *Physical Review E*, 67(5): 056104, 2003.

Duncan J. Watts and Steven H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393: 440–442, 1998.

Evolving Genetic Regulatory Networks Performing as Stochastic Switches

André Leier* Kevin Burrage*
*Advanced Computational Modelling Centre
University of Queensland, Brisbane, QLD 4072 Australia
{leier,kb}@acmc.uq.edu.au

Abstract

Recent studies have shown that small genetic regulatory networks (GRNs) can be evolved *in silico* displaying certain dynamics in the underlying mathematical model. It is expected that evolutionary approaches can help to gain a better understanding of biological design principles and assist in the engineering of genetic networks. To take the stochastic nature of GRNs into account, our evolutionary approach models GRNs as biochemical reaction networks based on simple enzyme kinetics and simulates them by using Gillespie's stochastic simulation algorithm (SSA). We have already demonstrated the relevance of considering intrinsic stochasticity by evolving GRNs that show oscillatory dynamics in the SSA but not in the ODE regime. Here, we present and discuss first results in the evolution of GRNs performing as stochastic switches.

1 Introduction

Genetic regulatory networks (GRNs) are complex systems composed of molecular species including genes, RNAs, transcriptional factors and other proteins, that chemically interact by specific reactions, thereby controlling the expression levels of the genes. GRNs are the fundamental units governing cell activities. Understanding them is essential to gaining understanding of a cell's development, organization, function and, ultimately, control. Revealing the design principles of GRNs is an important step towards this goal.

Studies suggest that GRNs have a modular structure, that is, they are composed of small constituent subnetworks or "modules" representing basic building blocks (Hartwell et al., 1999). Breaking down a GRN into its modules and analysing their individual structure and dynamics as well as their interactions (reverse engineering) can facilitate the identification of the GRN's overall functionality. The inverse approach, synthesis of GRNs instead of decomposition, is an alternative way to acquire new insights. By trying to build GRNs showing specific behaviours we may recognize certain design principles. Engineering (synthetic) genetic circuits either by hand (Elowitz and Leibler, 2000; Gardner et al., 2000; Becskei et al., 2001; Kobayashi et al., 2004) or by using directed evolution in vivo (Yokobayashi et al., 2002) is both time-consuming and expensive. Alternatively, evolutionary approaches *in silico* have been applied to find genetic networks performing as bistable switches or oscillators in the underlying mathematical model (François and Hakim, 2004; Leier et al., 2006). Naturally, in terms of the *in vivo* performance, the resulting networks are only as accurate as the formalisms describing them. Thus, since the mathematical models are only simplifications of the biological processes, the evolved GRNs may not perform *in vivo* as they do *in silico*. Nevertheless, as first results show, it can still be insightful to study the structure and characteristics of the resulting networks, taking the mathematical model into account (François and Hakim, 2004; Leier et al., 2006).

The molecular character of GRNs makes them intrinsically stochastic and noisy (McAdams and Arkin, 1997; Arkin et al., 1998; Elowitz et al., 2002; Hasty and Collins, 2002). The uncertainty of knowing when a reaction occurs and which reaction it might be causes fluctuations that become increasingly noticeable with smaller numbers of interacting molecules. Although noise can adversely affect cell function, it is also considered a source of robustness and stability, signal amplification, and selection of signalling pathways. Intrinsic stochasticity can be modelled by using the stochastic simulation algorithm (SSA), introduced by Gillespie (1977, 2001). The SSA is a statistically exact simulation method assuming that the system is homogeneous and populations are well-mixed within a constant volume.

In Leier et al. (2006) we present a genetic programming (GP) approach for evolving biochemical reaction networks based on simple enzyme kinetics that demonstrate desired dynamics when simulated with the SSA. It is inspired by the work of François and Hakim (2004) that use a similar approach to evolve genetic networks that are modelled by ordinary differential equations (ODEs). However, the dynamic behaviour of biochemical systems in the ODE regime (deterministic, continuous) and the SSA regime (discrete, stochastic) can be very different (Elowitz and Leibler, 2000; Heuett and Qian, 2005). Therefore, it is important to see how certain dynamic behaviour can evolve in the presence of noise. Since one neglects the stochastic nature of molecular interactions when modelling GRNs with ODEs, stochastic simulation can give deeper insights into chemical dynamics when there are only small numbers of molecules in the system.

We used our GP system to evolve GRNs with noisy oscillatory dynamics (Leier et al., 2006). Evolutionary runs breed GRNs that clearly oscillate in the discrete, stochastic regime but not when modelled as ODEs. The outcomes also confirm results from François and Hakim (2004) stating the importance of post-translational modifications for the functioning of the networks.

Switching, in particular bistable switching, is another fundamental dynamics that can be observed in many biological systems. Genetic switches are known to be responsible for controlling developmental processes and responding to environmental and intercellular signals. Biological switches and their underlying mechanisms can be quite different (Wolf and Arkin, 2003): a switch can be mono-stable, also called memory-less (the system moves back into its single stable state once the switching stimulus ends) or multistable (the system switches between two or more stable states by a transient application of a stimulus). Switching can occur randomly or by induction. Mechanisms, identified to allow bistable switching include cross-repressive feedback loops with cooperativity ((Gardner et al., 2000) utilized this mechanism to synthesize a toggle switch in *E. coli*) and positive feedback with cooperativity (based on this mechanism (Becskei et al., 2001) construct a bistable switch in *S. cerevisiae*).

This work presents first results in the evolution of genetic switches under intrinsic noise using the GP approach. For a successful evolution identification of switching behaviour in GRNs is a crucial factor. Although switching is often associated with a stimulus or induction signal, we first evolved GRNs that show periodic, not externally stimulated switching behaviour. However, we also analysed the resulting networks in terms of induced switching by injection of certain key molecules. Since noise can bounce trajectories between quasi-equilibrium states SSA dynamics of switches can be very different from ODE dynamics when there are only small numbers of molecules involved. Therefore, we also compared the SSA trajectories with the solutions of the corresponding ODE models and tested the ODE models for induced switching. Here, we focus on GRNs that show switching behaviour in the SSA but not in the ODE regime.

2 Methods

In the following material we briefly describe the reaction model, the stochastic simulation algorithm and the genetic programming approach. For additional information on the reaction model and the GP system we refer to Leier et al. (2006).

2.1 Reaction Model

Our reaction model describes GRNs as a set of species (genes, mRNAs, proteins and complexes such as gene-protein bindings or protein complexes) and *master reactions* governing their interactions. Master reactions are small sets of elementary and irreversible chemical reactions that correspond to biologically meaningful processes. Each elementary reaction either is a first order reaction, a second order reaction or a homodimer formation determined by the reaction rate constant. The seven master reactions are: gene transcription and translation (including basal transcription), transcriptional regulation of genes provided with two regulatory binding sites operating in a mode of cooperativity (based on the model described in Goutsias (2005)), protein modification, dimerization and three types of degradation (partial, catalytic and partial catalytic). Master reactions and their associated elementary reactions are listed in Table 2.1.

2.2 Stochastic Simulation Algorithm

Two issues led to Gillespie Gillespie (1977) introducing the SSA that exactly simulates the evolution of a discrete, stochastic chemical kinetic process in a well stirred mixture: (1) certain key molecules may be produced in quite low numbers (models based on continuous concentrations miss out on the discrete nature) and (2) the system is intrinsically noisy due to the uncertainty of knowing when a reaction and what

Master Reactions
Basal transcription and translation:
$a \rightarrow a + a_{\text{mRNA}}$
$a_{\text{mRNA}} \rightarrow a_{\text{mRNA}} + A$
$a_{\text{mRNA}} \rightarrow \emptyset$
$A \rightarrow \emptyset$
Transcriptional regulation:
$a + T \rightarrow aT$
$aT \rightarrow a + T$
$aT \rightarrow aT + a_{\text{mRNA}}$
$aT + R \rightarrow aTR$
$aTR \rightarrow aT + R$
Dimerization:
$A + B \rightarrow$ A:B
$A{:}B \rightarrow A + B$
Partial degradation:
$A{:}B \rightarrow A$
Catalytic degradation:
$A + B \rightarrow A$
Partial catalytic degradation:
$A{:}B + C \rightarrow A$

Table 1: The set of master reactions that are the building blocks of the genetic networks. Each gene has two regulatory binding sites (R_1 and R_2) that work in a cooperative manner: binding of a transcription factor at R_2 requires R_1 to be occupied by another factor. Lowercase letters (a,b, etc.) represent genes with unbound regulatory sites. When a transcription factor T is bound at R_1 of a the binding is denoted aT. The regulatory effect (positive or negative regulation) depends on the corresponding reaction rate constant. Binding at R_2 excludes any transcriptional activity and hence, represses transcription. The case of a repressor R bound to aT is denoted as aTR. mRNA is indicated such as in a_{mRNA}. Capitalized letters (A,B, etc.) represent the proteins translated from the associated mRNA. Protein complexes are represented using colons (i.e. a protein complex composed of proteins A and B is represented by $A{:}B$). Each reaction is specified by a reaction rate constant.

reaction takes place (deterministic models ignore stochasticity). Hence, in this context, the use of continuous differential equations methods is debatable.

In the following material we briefly describe the SSA. It is assumed that the biochemical system is well-mixed within a constant volume held at constant temperature. Let there be N molecular species $\{S_1, \ldots, S_N\}$ that chemically interact through M reactions $\{R_1, \ldots, R_M\}$. The system state at time t is described by a vector $X(t) \equiv (X_1(t), \ldots, X_N(t))^T$ where $X_i(t)$ is the number of molecules of species i at time t. Each reaction R_j can be uniquely defined by its *propensity function* a_j, where $a_j(X(t))dt$ is the probability that reaction R_j will occur somewhere in the system within the time interval $(t, t + dt)$, and its assigned *stoichiometric vector* ν_j that specifies the update of the system state when reaction R_j occurred. This is defined by ν_{ji} for $i = 1, \ldots, M$, which is the change in the number of S_i molecules produced by one R_j reaction. Our SSA implementation simulates the time evolution of a system according to the *direct method* Gillespie (2001): two independent samples r_1 and r_2 of the uniform random variable $\mathbf{U}(0, 1)$ are drawn consecutively. The length of the time interval $[t, t + \tau)$ is given by

$$\tau = \frac{1}{a_0(X(t))} \ln(\frac{1}{r_1}) \,,$$

where

$$a_0(X(t)) = \sum_{j=1}^{M} a_j(X(t))$$

is the sum of all propensities. The specific reaction R_j occurring in $[t, t + \tau)$ is determined by the index j satisfying

$$\sum_{j'=1}^{j-1} a_{j'}(X(t)) < r_2 a_0(X(t)) \leq \sum_{j'=1}^{j} a_{j'}(X(t)) \,.$$

The state vector is then updated as

$$X(t + \tau) = X(t) + \nu_j \,.$$

For the elementary reactions (first and second order reactions and homodimer formations) used in our reaction model, the corresponding propensity functions are shown in Table 2.2.

Since the SSA can become very computationally intensive when time steps become very small (due to large numbers of reactions, large reaction rates or large numbers of molecules), we limit our model to small numbers of species.

Reaction	Propensity Function
First order reaction $S_k \xrightarrow{c_j} S_l$	$a_j = c_j * X_k$
Second order reaction $S_k + S_l \xrightarrow{c_j} S_m$ with $S_k \neq S_l$	$a_j = c_j * X_k * X_l$
Homodimer formation $S_k + S_k \xrightarrow{c_j} S_l$	$a_j = c_j * X_k * \frac{X_k - 1}{2}$

Table 2: Propensity functions for three elementary types of reactions. c_j is the reaction rate constant of the respective reaction.

2.3 Genetic Programming System

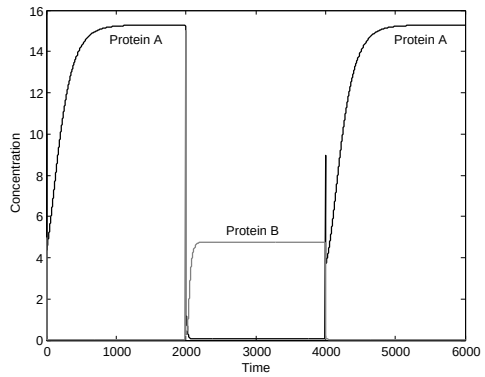
Individuals in the GP population are GRNs according to the reaction model. Each individual is assigned a fitness value describing how well its dynamics meets the prescribed requirements. This value is calculated by simulating the reaction system using the SSA over a predefined time and analysing the resulting trajectory. For our purposes, we define constraints the trajectory must satisfy in order not to be penalized. A penalty is equal to the amount by which the solution exceeds the constraint. Penalties are weighted according to how the constraints are met. Then, penalties are summed up to obtain the fitness value. That is, the lower the fitness the better the individual. To guide the evolution towards GRNs with a switching behaviour in the concentration of a particular protein (e. g. protein A), we define the following constraints (numerical values in parentheses exemplify the thresholds): (i) the molecular number n of the species has to be at "low level" ($n < 20$) for a minimum time period T_1 ($T_1 = 500$), (ii) the molecular number n of the species has to be at "high level" ($150 < n < 200$) for a minimum time period T_2 ($T_2 = 500$), (iii) the time period for a switch between low and high level (and vice versa) is limited by T_3 ($T_3 = 50$). There has to be at least one switch between "low" and "high" concentration levels in the trajectory of the corresponding protein, otherwise the individual's fitness is set to a maximum value.

Evolution is driven by repeated selection and mutation. The selection method is a simple (50+50) evolutionary strategy, that is, 50 individuals produce one offspring each and the best 50 out of 100 individuals build the new population. Offspring are produced by mutation of the parent individuals. The mutation operators involve random modifications of the reaction rate constants and additions and deletions of master reactions. That is, not only rate constants are evolved but also the structure of the GRN. To focus on small regulatory networks we fixed the number of genes at two. This reduced the search space and facilitated evolution. When reactions are added to the GRN the reaction rates are uniformly drawn from $[0, 1]$. A reaction rate is mutated by multiplication with a random number from $[0, 2]$. At the beginning of an evolutionary run, the initial concentrations are randomly chosen from $1, 2, \ldots, 10$ and remain fixed for the entire evolution. Evolution is terminated if the number of generations without fitness improvement exceeds a threshold (100 generations). The evolved networks are simulated several times to verify their dynamic behaviour.

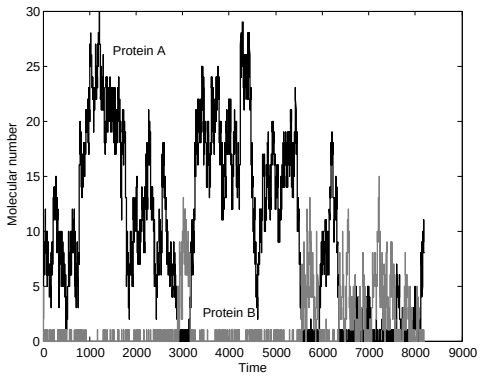
3 Results

We already mentioned in the introduction that ODE and SSA models of the same GRN can display very different dynamics. Figure 3 illustrates this for a bistable switch evolved in the ODE regime (François and Hakim, 2004) (Figure 3A). While the ODE model allows induced switching between two equilibrium states the SSA trajectories are very noisy with many irregular switches. However, protein A and B seem to be in opposite levels, that is, whenever A exists in higher numbers, B does not and vice versa. We note that François and Hakim (2004) also present the GRN with different reaction rates that works as a bistable switch in both regimes where the species in the high concentration level have several hundred molecules. For the rest of this section, we present and discuss evolved GRNs.

The resulting GRNs can be quite different in their dynamics and their structure, i. e. in the composition of the master reactions. Interestingly, although our fitness function is not geared to the evolution of bistable switches, a few GRNs show some sort of bistable switching behaviour, usually with different high level concentrations of protein A (as predefined) and B. Figure 2(a) shows such an evolved GRN where the periodic switching between the two states, either protein A in high and protein B in low numbers or vice versa, is driven by noise (Figure 2(b)). In this GRN, protein A is the transcription factor of gene b. It may bind to the gene's binding sites R'_1 and R'_2. Binding of A at R'_1 activates enhanced transcription of b, additional binding of A at R'_2 represses transcription. In this example the regulatory region of gene a is unused. As expected, the dynamic behaviour of the GRN in the ODE regime is quite different (cf. Figure 2(c)) from the SSA regime. The corresponding ODE model shows a single switch when



(a)



(b)

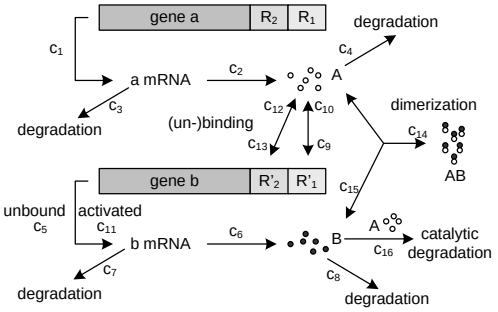
Figure 1: Dynamics of an evolved switch in the ODE regime by François and Hakim (2004). (a) ODE dynamics: the switching is induced by two pulses. At $t = 2000$ and $t = 4000$ we add protein concentrations $[A] = 15$ and $[B] = 9$, respectively. (b) SSA dynamics (without additional pulses): we observe very irregular, noisy behaviour. If one protein is expressed in high numbers the other is not and vice versa. Switching occurs irregularly without prior initiation. For SSA and ODE, the initial conditions are the same ($[A] = 10$, $[B] = 5$, $[A : B] = 0$).

starting with low concentrations $[A]$ and $[B]$ and remains stable at certain concentration levels over the monitored time whereas the SSA trajectories display periodical switches. When the state of the ODE solution is perturbed by injecting protein B the system immediately develops back into the stable state. Only for very large amounts of added protein concentrations (several hundreds or even thousends depending on the systems state at the time of injection), we can observe a short time period where protein B is present in a higher concentration than protein A, similar to the early dynamics shown in Figure 2(c). Thus, the GRN displays no bistable switching in the ODE regime. One-time injection of a sufficient number of protein A molecules during stochastic simulation let the system switch from low protein A level to high protein A level. However, injections of protein B did not necessarily lead to a switch back, irrespective of the number of molecules added to the system.

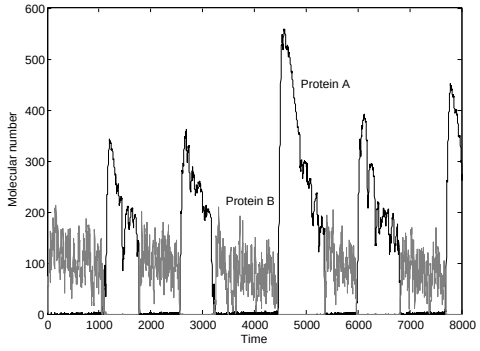
Figure 3(a) shows an evolved GRN that behaves in the stochastic simulation as a monostable switch with protein A at high level (about 150–250 molecules) and protein B at low level (0 molecules) as the only truely stable state. By injecting protein B molecules the state switches immediately into a state with very low numbers of molecules (< 10) for both A and B. Interestingly, the time the system remains in this state depends on the amount of protein B molecules added to stimulate the switch. Figure 3(b) and 3(c) demonstrate this for injections of 20 and 40 molecules, respectively. In this case the time is roughly twice as long for the second than for the first trajectory. From several simulations we got the impression that the larger the injection the longer the system stays in its state. However, this can only serve as a rule of thumb as large variations were observed as well.

4 Discussions

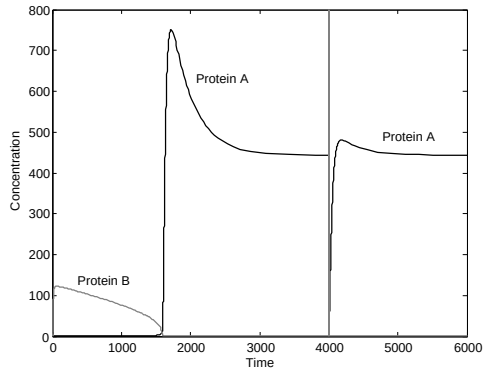
In this contribution we present two GRNs that our evolutionary approach produced. The resulting GRNs vary highly in their dynamics and not every solution exhibits a switching behaviour. According to the fitness function we search for GRNs where at least one protein dynamics displays a periodic but untriggered switching between a low and a high level (in terms of molecular numbers). The GRN in Figure 2(a) shows this form of switching behaviour for both proteins, that is, their molecular numbers mutually alternate between high and low levels. This does not occur because of any injection process or external control but because of the inherent noise which



(a)

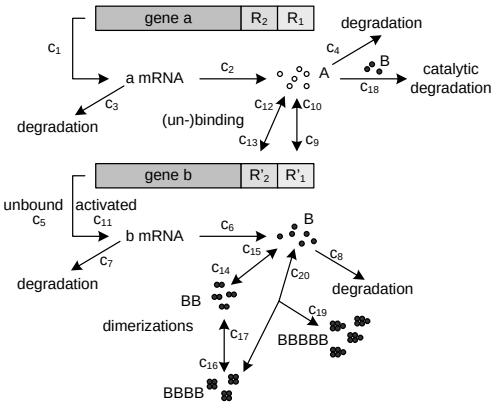


(b)

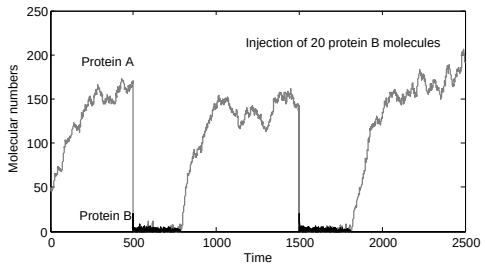


(c)

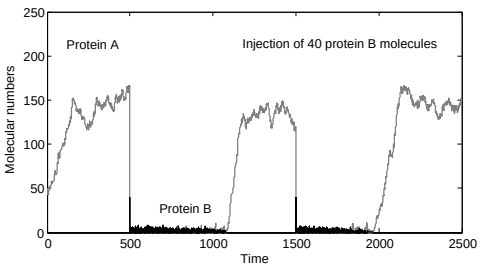
Figure 2: (a) Schematic representation of an evolved GRN. The evolved reaction constants are: $c_1 = 0.445, c_2 = 0.110, c_3 = 0.136, c_4 = 0.003, c_5 = 1.6, c_6 = 3.867, c_7 = 0.122, c_8 = 0.516, c_9 = 1.021, c_{10} = 0, c_{11} = 2.086, c_{12} = 0.013, c_{13} = 0.092, c_{14} = 0.92, c_{15} = 0.089$ and $c_{16} = 0.446$. (b) Simulation results showing the concentration dynamics of protein A and B. The (non-triggered) periodic switching behaviour is evident. (c) Solution of the corresponding ODEs with initial concentrations $[A] = [B] = 5$ and $[AB] = 0$. After injecting a strong concentration of protein B at time $t = 4000$ ($[B] = 800$) the system quickly evolves towards its previous stable state.



(a)



(b)



(c)

Figure 3: (a) Schematic representation of an evolved GRN performing as a stochastic monostable switch. It differs from the GRN in Figure 2(a) in the post-translational modifications and in the reaction rate constants ($c_1 = 0.939, c_2 = 0.208, c_3 = 0.155, c_4 = 0.008, c_5 = 0.047, c_6 = 0.022, c_7 = 0.312, c_8 = 0.044, c_9 = 0.852, c_{10} = 0.169, c_{11} = 0.251, c_{12} = 0.523, c_{13} = 0.013, c_{14} = 0.783, c_{15} = 0.222, c_{16} = 0.063, c_{17} = 0.494, c_{18} = 1.364, c_{19} = 0.121, c_{20} = 1.363$). (b) System dynamics when injecting 20 protein A molecules at time $t_1 = 1000$ and $t_2 = 2000$. Initial concentrations are: $[A] = 50, [B] = 0$ (c) This time, 40 protein A molecules are added to the system at t_1 and t_2. The system remains roughly twice as long in the state with $[A]$ at low level.

drives the dynamics backwards and forwards. Indeed, the corresponding ODE solution has only one equilibrium state and after transient stimulation the system develops back into this state.

The monostable GRN in Figure 3(a) shows interesting properties as well. Here, switches from high level (stable state) to low level (instable) molecule numbers of protein A do not occur randomly. Instead, they have to be induced by protein B injection. The switch back into the stable state is driven by noise, though. Apparently, the time period for which the system is not in its equilibrium depends on the injection.

Experiments with other monostable GRNs showed that often injections need to exceed a certain threshold to become effective. This threshold can depend on the molecular numbers of key proteins in the system. Also, the same dosage of injection can be more efficient, in that it keeps the monostable switch for a longer time in the nonstable state if it is spread over a certain time period. This might be interesting from an experimental perspective and needs to be analysed in more detail.

So far, we did not evolve real bistable switches. This is mainly because of the fitness function which does not put selection pressure on the evolution of such switches. Nevertheless, evolutionary runs with modified fitness functions suggest that it is difficult to find switches which are bistable but work with key molecules in low numbers. When dealing with low numbers of molecules there is a natural fluctuation due to stochasticity and the lower the numbers the larger the possibility of accidental switching (cf. Figure 3).

Whether the evolved GRNs in Figure 2(a) and 3(a) have a structure that occurs in nature remains to be seen. A comparison of this and other evolved solutions with known biological switches might lead to further insights.

This work underpins the necessity of stochastic simulation by evolving GRNs that perform as switches in the SSA regime (but not necessarily in the ODE regime) and might contribute to the finding of functional design principles and a better understanding of regulation in cells.

Acknowledgements

KB would like to thank the Australian Research Council for funding via the Federation Fellowship program.

References

A. Arkin, J. Ross, and H. H. McAdams. Stochastic kinetic analysis of developmental pathway bifurcation in phage λ–infected Escherichia coli cells. *Genetics*, 149(4):1633–1648, Aug 1998.

A. Becskei, B. Séraphin, and L. Serrano. Positive feedback in eukaryotic gene networks: cell differentiation by graded to binary response conversion. *EMBO J*, 20(10):2528–2535, May 2001.

M. B. Elowitz and S. Leibler. A synthetic oscillatory network of transcriptional regulators. *Nature*, 403 (6767):335–338, Jan 2000.

M. B. Elowitz, A. J. Levine, E. D. Siggia, and P. S. Swain. Stochastic gene expression in a single cell. *Science*, 297(5584):1183–1186, Aug 2002.

P. François and V. Hakim. Design of genetic networks with specified functions by evolution *in silico*. *Proc. Natl. Acad. Sci.*, 101(2):580–585, Jan 2004.

T. S. Gardner, C. R. Cantor, and J. J. Collins. Construction of genetic toggle switch in *escherichia coli*. *Nature*, 403(6767):339–342, 2000.

D. T. Gillespie. Approximate accelerated stochastic simulation of chemically reacting systems. *J. Chem. Phys.*, 115(4):1716–1733, 2001.

D. T. Gillespie. Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.*, 81(25): 2340–2361, 1977.

J. Goutsias. Quasiequilibrium approximation of fast reaction kinetics in stochastic biochemical systems. *J. Chem. Phys.*, 122(184102):1–15, may 2005.

L. H. Hartwell, J. J. Hopfield, S. Leibler, and A. W. Murray. From molecular to modular cell biology. *Nature*, 402:C47–C52, 1999.

J. Hasty and J. J. Collins. Translating the noise. *Nature Genetics*, 31(1):13–14, May 2002.

W. J. Heuett and H. Qian. A stochastic model of oscillatory blood testosterone levels. To appear in Bulletin for Mathematical Biology, 2005.

H. Kobayashi, M. Kaern, K. Chung, T. S. Gardner, C. R. Cantor, and J. J. Collins. Programmable cells: Interfacing natural and engineered gene networks. *Proc. Natl. Acad. Sci.*, 101(22):8414–8419, 2004.

A. Leier, P. D. Kuo, W. Banzhaf, and K. Burrage. Evolving noisy oscillatory dynamics in genetic regulatory networks. In P. Collet, M. Tomassini, M. Ebner, A. Ekart, and S. Gustafson, editors, *Proceedings of the 9th European Conference on Genetic Programming (EuroGP 2006)*. Springer, 2006. In press.

H. H. McAdams and A. Arkin. Stochastic mechanisms in gene expression. *Proc. Natl. Acad. Sci.*, 94(3):814–819, Feb 1997.

D. M. Wolf and A. P. Arkin. Motifs, modules and games in bacteria. *Curr. Opin. Microbiol.*, 6:125–134, 2003.

Y. Yokobayashi, R. Weiss, and F. H. Arnold. Directed evolution of a genetic circuit. *Proc. Natl. Acad. Sci.*, 99(26):16587–16591, 2002. doi: 10.1073/pnas.252535999.

Network Entropy and Cellular Robustness

Thomas Manke*
*Max Planck Institute for Molecular Genetics
Ihnestr. 73, 14195 Berlin, Germany
manke@molgen.mpg.de

Lloyd Demetrius†
†Dept. of Organismic and Evolutionary Biology
Harvard University, Cambridge MA02138, USA
Second@else.where

Martin Vingron*
‡Max Planck Institute for Molecular Genetics
Ihnestr. 73, 14195 Berlin, Germany
vingron@molgen.mpg.de

Abstract

Here we present a novel approach to the study of networks and their resilience properties. Our work is based on a fundamental relation from dynamical systems theory which states that the macroscopic resilience of a steady state is correlated with the uncertainty in the underlying microscopic processes, a property which can be measured by entropy. Here we apply these ideas to the analysis of biological networks as obtained from large-scale protein interaction screens in yeast and *C. elegans*. In this context we characterize the diversity of possible pathways in terms of network entropy. Our analysis shows that knockouts of proteins with large contribution to network entropy are preferentially lethal. This observation is robust with respect to several possible errors and biases in the experimental data. Our analytical approach goes beyond the phenomenological studies based on local network observables, such as connectivity. It provides a rationale to study proxies of cellular resilience and to rank network elements (proteins) according to their importance within the global network context.

1 Introduction

Recent experimental efforts have highlighted the pervasiveness of molecular networks in biological sciences (Alm and Arkin, 2003). While a large number of molecular interactions and associations have been mapped qualitatively, we have yet to understand the relation between the structure and the function of biological networks which control the information flow and regulation of cellular signals. One particularly important functional characterisation is the resilience of an organism against external and internal changes (Stelling et al., 2004; Kitano, 2004), which, at the molecular level, amounts to perturbations in the network parameters. In recent experiments this resilience has been studied in direct response to gene deletions or RNA interference (Giaever et al., 2002; Kamath et al., 2003). It has been demonstrated that a large number of such network perturbations do not result in any phenotypic variation under a given experimental condition. This observation has led to a simple classification into 'viable' and 'lethal' proteins, according to whether the organism survives the removal of this component or not. In the following we also refer to the latter as "essential" proteins.

If network topology gives rise to behavioral complexity, one may ask if there is any topological correlate for lethality. In this work we present a natural framework to derive macroscopic parameters that characterise the topological and structural resilience of a network against random perturbations. Our analytical framework goes beyond the seminal studies of Albert et al. (2000) and Jeong et al. (2001), which addressed the same problem in terms of phenomenological parameters such as degree. The key idea and underlying assumption of our work is that many biological systems operate at steady state, where characteristic macroscopic observables (the "phenotype") remain constant for relatively long times. This, however, does not imply that the underlying microscopic variables (such as protein activities and concentrations) are static, but rather that their complex and continuous interplay results in a stable phenotype which can be experimentally observed. Indeed, it is the diversity and uncertainty of microscopic processes which determines the resilience of macroscopic steady states against random perturbations. In the context of dynamical systems this uncertainty is quantified by the dynamical entropy. The relationship between entropy and the robustness of macroscopic

observables (their rate of return to steady state values) is the content of a fluctuation theorem (Demetrius et al., 2004), which states that changes in entropy are positively correlated with changes in robustness. As a great simplification, and in recognition of our ignorance about the actual molecular processes, we assume that the microscopic processes on the network are Markovian. This leads to the notion of *network entropy* as a global measure of pathway diversity and as a correlate of cellular resilience. In this paper we will demonstrate that this global description leads to a natural ranking of network elements (proteins) according to their contribution to network entropy. In terms of functional perturbation experiments, we will test the hypothesis that proteins with higher entropic contribution to the cellular network more frequently have a lethal phenotype when they are impaired (knock-out/knock-down). Previously this question has been addressed in terms of various notions of network centrality: degree Jeong et al. (2001), shortest path length Yu et al. (2004) and more recently betweenness Hahn and Kern (2005). Here we also provide a rationale for why these ad-hoc measures are sometimes convenient proxies for network resilience and how they could be extended.

2 Network Entropy and Entropic Ranking

First recall the definition of the dynamical entropy for a Markov process, $P = (p_{ij})$, which is given by (Billingsley, 1965)

$$H = -\sum_{ij} \pi_i p_{ij} \log p_{ij} \quad . \tag{1}$$

Here p_{ij} denotes the transition probabilities and the π_i are the components of the stationary distribution. There are many other ways to investigate complex dynamical systems through microscopic modelling, such as differential equations, but our simple stochastic description of dynamical uncertainty is based on random walks on the network and has a long tradition in the analysis of diffusive systems Berg (1993).

In our context the dynamical entropy of a Markov process characterizes the diversity of possible pathways and is related (through the fluctuation theorem) to the systems response to perturbations. If only the network topology is known, we associate the following process with a given adjacency matrix $A = (a_{ij})$:

$$p_{ij} = \frac{a_{ij} v_j}{\lambda v_i} \quad . \tag{2}$$

It has been shown (Arnold et al., 1994) that this process is the unique solution to a variational principle for the leading eigenvalue, λ, of the adjacency matrix. For irreducible matrices the components of the corresponding leading eigenvector, v_i, are all strictly positive. For Boolean matrices the process matrix of Eq.2 maximizes the entropy and provides the most parsimonious choice of p_{ij}. See (Demetrius and Manke, 2004) for a more detailed discussion.

In the following we utilize the decomposition of network entropy into contributions from all individual proteins

$$H = \sum_i \pi_i H_i \quad , \tag{3}$$

where H_i is the Shannon entropy associated with protein i. This decomposition suggests that network elements with a higher contribution to the overall entropy have a larger effect on the network's resilience and functionality when removed.

3 Biological Networks and Functional Studies

Here we analyse biological networks of protein-protein interactions for a single-cellular organism (budding yeast) and the multi-cellular worm (C.elegans) which we retrieved from public databases (Mewes et al., 2002; Chen et al., 2005). For both organisms this information is supplemented by functional studies of large-scale gene disruption experiments (Giaever et al., 2002; Kamath et al., 2003).

For *S.cerevisiae* we retrieved a bidirected interaction network of 3854 proteins with 11912 yeast-two hybrid interactions and 1170 essential proteins among the total set of 6203 proteins. The protein-protein interaction network of *C.elegans* consists of 2800 proteins and 8740 interactions. Of all the proteins with a recorded interaction 322 are classified as essential, because their inhibition resulted in a lethal phenotype. It should be noted that both interaction data and functional screens have a number of associated errors, resulting from experimental insufficiencies of large-scale studies and their limitation to certain environmental conditions.

4 Proteins with high entropic contribtion tend to be essential

The network data described in the previous section lends itself to a structural analysis, which has conventionally been done in terms of various connectivity

measures Jeong et al. (2001); Yu et al. (2004); Hahn and Kern (2005). Here we utilize network entropy as a global characteristic measure and its decomposition according to Eq.3, which provides an alternative measure to rank the importance of proteins within the network. Figure 1 shows that proteins with high rank are more often essential than expected by chance.

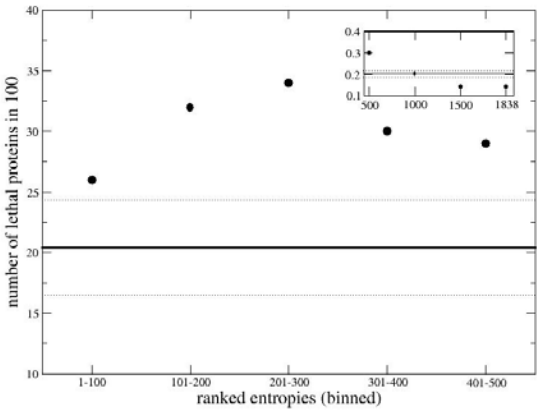


Figure 1: In the main figure we define 5 classes of *C.elegans* proteins according to their rank with respect to entropic contribution: 1-100, 101-200 .... In all these high ranking cases the fraction of essential proteins is significant. The expectation from 100 random proteins is shown as horizontal lines ($\pm$ one standard deviation). The inset shows the same analysis for taking larger bins of 500 proteins. Again we can see an enrichment for high-ranking proteins, while there is an under-representation of essential proteins for proteins with small entropic contributions (for ranks > 1000)

In figure 2 we plot this probability against the number of top-ranking proteins for several setups. Our initial analysis (full circles) shows a steep decline of hypergeometic p-values and a systematic deviation from p-values obtained from randomized list of proteins (solid line). This reiterates the observation from the previous section. Next we tested the effect of false positive interactions by randomly deleting 50% of all edges (interactions) from the protein interaction network. This gave rise to a new ranking of proteins. Figure 2 illustrates that, despite this drastic change, the correlation is only moderately affected (triangles down). Missing interactions, on the other hand, can be expected to have a more pronounced effect. If, in a similar spirit, we increase the number of all interaction by 50% (random link addition), the entropic ranking will also change and the corresponding lethality assignment will become less and less predictive (triangles up). Notice though, that even with such large assumed error rates, the lethality assignment based on entropy is still significantly better than random (solid line).

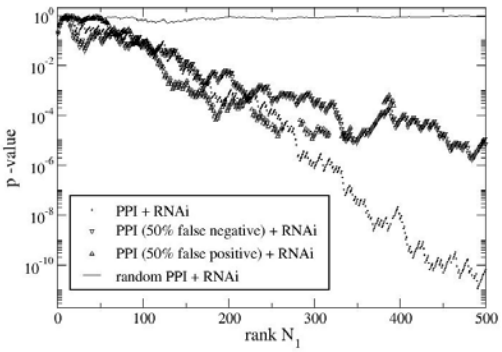


Figure 2: Randomized Networks. Here we analyse the correlation of entropic contribution and lethality assignment in the light of possible experimental errors for *C.elegans*. To this end we have added and removed a sizeable fraction of random interactions to the original data (full circles). While a large fraction of false positive errors (triangles down) does not significantly change the observed correlation, a large number of false negatives would reduce the significance of the observed correlation (triangles up). As expected, completely randomised interaction networks do not show any correlations as signified by the flat behaviour of the solid line.

5 Systematic Errors

As was mentioned above, the current large-scale data has sizable errors. Therefore we now investigate, whether the observed enrichment of essential proteins in top-ranking lists is robust against known sources of systematic errors.

First we extended the analysis of the previous section and evaluate the prevelance of essential proteins more systematically. For a given number, N_1 of top-ranking proteins we observe a certain number, N_{12} of essential proteins. This fraction can be translated into a probability (hypergeometric score) to observe such an overlap by chance, given a total of N proteins of which N_2 are essential.

Different cellular locations are known to influence

the results of protein interaction screens. Therefore we have also tested our result against this possible bias by selecting randomized groups of "top-ranking" proteins, while maintaining their distribution with respect to cellular components. We find that the observed correlation is robust against this experimental artefact (data not shown).

Given the predominance of degree-based methods for network analysis, we also compared our novel importance measure, entropic contribution, to protein connectivity. While the two measures show a correlation for large degree, there are also clear differences, see Fig.3. Since entropy is a global measure, entropic contribution also takes into account the overall position of a protein within the network. This has the effect that proteins with highly connected interaction partners make a higher contribution to network entropy than proteins (with the same connectivity), but less connected neighbours.

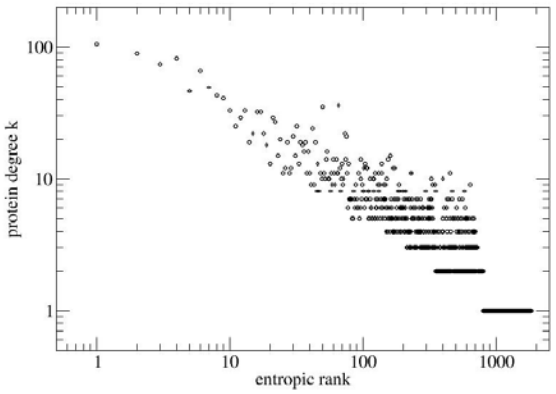


Figure 3: Here we show (for *C.elegans*) that connectivity and entropic contribution are correlated, but distinct from one another. For the process defined in Equation 2, proteins with high degree tend to have high entropic contribution. On the other hand there are also lowly connected proteins with high contribution to network entropy and hence robustness.

6 Conclusions

In summary, we have shown that the entropic characterisation of protein interaction networks can account for a significant fraction of proteins whose removal results in a lethal phenotype.

In our framework proteins are ranked according to their contribution to network entropy, which is a measure of microscopic uncertainty (pathway diversity) and is correlated with the macroscopic robustness of a dynamical system defined on the network.

We introduced a systematic method to assess the correlations between the entropic ranking scheme and phenotypic lethality data, and we have carefully tested the observed correlations against a number of possible errors. Our new conceptual framework provides a rationale to understand macroscopic resilience in the light of microscopic uncertainty, as characterized by entropy, rather than structural network observables. From this perspective, the observed enrichment of essential proteins in ranked lists of proteins has a natural and clear interpretation: proteins with higher contribution to cellular resilience are more often essential. Heuristic constructs, such as node degree, emerge as effective descriptors of dynamical properties, but our work also illustrates where one can go beyond such structural measures. Moreover, and in contrast to degree based-methods, our approach is extendable to networks where more quantitative data is available.

In the following we want to point to possible limitations of our approach. First, the phenotypic assessment of a gene disruption is usually done for one given condition and the observed correlation is strictly with respect to this single condition. It has been remarked that so-called viable proteins may actually play a significant role in untested environments and their disruption could cause lethal phenotypes. An exhaustive study of all possible conditions is clearly beyond experimental capabilities. Therefore we take the present lethality data as representative for other conditions and implicitly assume that the classification of lethal and viable proteins is at least robust against environmental changes.

A related problem concerns the static representation of interaction data which discards all dynamical dependencies. Just as many genes are expressed only under specific conditions, we also should think of different network realisations of an underlying blueprint which experimental interaction screens try to establish. Since the concept of entropy is based on the notion of dynamical diversity of the microscopic processes underlying the cellular states, we believe that this approach will ultimately be more fruitful than network characterisations which are solely based on topology. We should, however, stress that in the present application we relied exclusively on structural information of only a part of the complete cellular network - namely protein-protein interactions. Furthermore, we characterized the microscopic diversity through a Markov process that maximizes the entropy based on a Boolean adjacency matrix, rather than

quantitative information about transition rates. Needless to say that actual processes may be different from this representative one.

To the extent that real processes resemble the one defined in this work, we can now better understand the importance of structural network observables as correlates of dynamical properties. We expect that structural properties will become less useful concepts for processes that deviate from the one with maximal entropy. Our approach is a first attempt to bridge these two domains and to address structural and dynamical questions in a single framework.

This situation can be likened to thermodynamics, where some properties of large systems can be effectively described by a number of macroscopic parameters, regardless of our ignorance about the microscopic processes. For equilibrium systems, this simplification is made explicit through relations between the Gibbs distribution over microstates and various macroscopic properties that can be derived from it Gibbs (1901). Formally, our work builds on an extension of the Gibbs formalism, which also applies to non-equilibrium systems at steady state Ruelle (2004). We implicitly assumed that the cellular processes on protein interaction networks fall into this larger class. If these assumptions hold, our approach should also apply to other complex networks, and there is hope that some systemic properties can be elucidated without having to resort to microscopic details.

Acknowledgements

This work has been supported by a grant from the German National Genome Research Network (NGFN).

References

Albert, Jeong, and Barabasi. Error and attack tolerance of complex networks. *Nature*, 406(6794): 378–82, Jul 2000.

Eric Alm and Adam P Arkin. Biological networks. *Curr Opin Struct Biol*, 13(2):193–202, Apr 2003.

L. Arnold, V. Gundlach, and L. Demetrius. Evolutionary formalism for products of positive random matrices. *Annals of Probab.*, 4(3):859–901, 1994.

Howard C. Berg. *Random Walks in Biology*. Princeton University Press, 1993.

P. Billingsley. *Ergodic Theory and Information*. Wiley, New York, 1965.

Chen *et al.* WormBase: a comprehensive data resource for Caenorhabditis biology and genomics. *Nucleic Acids Res*, 33(Database issue):D383–9, Jan 2005.

Lloyd Demetrius, Volker Matthias Gundlach, and Gunter Ochs. Complexity and demographic stability in population models. *Theor Popul Biol*, 65 (3):211–25, May 2004.

Lloyd Demetrius and Thomas Manke. Robustness and network evolution – an entropic principle. *Physica A*, 346(3-4):682–696, 2004.

Giaever *et al.* Functional profiling of the Saccharomyces cerevisiae genome. *Nature*, 418(6896): 387–91, Jul 2002.

J Willard Gibbs. *Elementary Principles in Statistical Mechanics*. Dover, New York, 1901.

Matthew W Hahn and Andrew D Kern. Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. *Mol Biol Evol*, 22(4):803–6, Apr 2005.

H Jeong *et al.* Lethality and centrality in protein networks. *Nature*, 411(6833):41–2, May 2001.

Kamath *et al* Systematic functional analysis of the Caenorhabditis elegans genome using RNAi. *Nature*, 421(6920):231–7, Jan 2003.

Hiroaki Kitano. Biological robustness. *Nat Rev Genet*, 5(11):826–37, Nov 2004.

Mewes *et al.* MIPS: a database for genomes and protein sequences. *Nucleic Acids Res*, 30(1):31–4, Jan 2002.

David Ruelle. *Thermodynamic Formalism*. Cambridge Mathematical Library, 2004.

Stelling *et al.* Robustness of cellular functions. *Cell*, 118(6):675–85, Sep 2004.

Yu *et al.* Genomic analysis of essentiality within protein networks. *Trends Genet*, 20(6):227–31, Jun 2004.

Effects of Dimensionality over Cooperation Dynamics

Ivette C. Martínez*
*Grupo de Inteligencia Artificial
Universidad Simón Bolívar
martinez@ldc.usb.ve

Klaus Jaffe†
†Laboratorio de Comportamiento
Universidad Simón Bolívar
kjaffe@usb.ve

Abstract

An Agent Based Model was used to explore the effects of simulating one and two dimensional grids over the dynamics of cooperation, under scenarios of biological evolution (BE) and cultural evolution (CE). Our results show that the way space is simulated does affect the dynamics of evolution. Interestingly, biological evolution was more susceptible to this effect that cultural evolution.

1 Introduction

Biologists, economists, computer scientists and physicists have all worked to further our understanding of human and animal cooperation. Yet different premises underlay these efforts. The main difference among them is the assumption that social behavior arrived through biological evolution among animals, and that culture and rational decision making is a principal driver of the evolution of cooperation and sociality among humans (Richardson et al., 2004). Human cooperation seems to be molded by both, cultural and biological forces (Kurzban and Houser, 2005).

Important differences between the dynamics of cultural evolution (Richardson et al., 2004; Ehrlich and Levin, 2005) and biological evolution (Nowak and Sigmund, 2004) exist. One important feature differentiating systems driven by biological and cultural evolution, is that transmission of information in BE is vertical (heredity), and that in CE is horizontal (imitation of the behaviour of the majority). This feature affects the speed information is transmitted, and is sufficient to explain important differences in the dynamics between both types of evolution (Jaffe and Cipriani, 2006). Here we want to explore the effect of the dimensionality of space on the different dynamics reported. To do so, we modify a one-dimensional spatial model (Cipriani and Jaffe, 2005) to study the differences between the dynamics of cooperative, group-forming individuals subject to a selective pressure (in this case predation). We based our model on the well known 'selfish herd' concept (Hamilton, 1971) and assume that cultural and biological dynamics is driven by natural selection on the phenotypes (i.e., roles) of individuals: cooperators and non-cooperators.

2 The Model and Experiments

Based on a cellular automata model that represents a population of interacting individuals with different social roles proposed by (Cipriani and Jaffe, 2005), we construct an agent-based model that incorporated environments with different spatial structures. Our initial implementation, made in Python, is for one-dimensional and two-dimensional toroidal grids environments of size 10000.

The majority rule implemented here to simulate cultural evolution (CE) assumed that individuals had a given probability of imitating the behavior of their neighbours. Behavioural traits were transmitted 'horizontally' via learning by imitation. We contrast this mechanism with biological evolution (BE) where learning does not take place and information was transmitted to offspring via hereditary rules. The summation of both mechanisms is a metaphor of species driven by both cultural and biological evolution.

We studied five scenarios characterized in our model by the way agents were allowed to interact. In CE, the rate T determined the probability an agent would imitate the behaviour (cooperate or not) of the majority of its neighbors. Production of new agents for CE was uniform $(50/50)$ and T took three values: $\{0, 0.5, 1\}$. In the BE scenario, empty cells were replenish in proportional to the current number of individuals of each kind (cooperators and noncooperators) and $T = 0$. In the BE+CE scenario $T = 1$ and proportional replenish were applied..

Our experiments consisted in 5 series of simulations, corresponding to the described scenarios. In

each series (101 simulations) we varied the cost of cooperation in steps of 0.01. Each simulation was run for 400 time-steps. For all simulations the "fitness differential" was 0.6 (the difference between the predation rate of isolated individuals (0.8) and that for cooperators being part of a group of cooperators (0.2)).

3 Results

The results of the different experiments are shown in figures 1 and 2 where each figure summarizes the results from simulations with environments of either a 1-dimensional grid 1 or a two-dimensional grid 2. Interestingly, the differences between the dynamics of the various scenarious explored were larger when we simulated a 1-dimentional grid than when using the 2-dimentional grid. The basic morphology of the resulting dynamics was not affected by the dimension of the simulated space. That is, BE has very sharp thresholds compared to CE and CE+BE was the strategy favouring most cooperation.

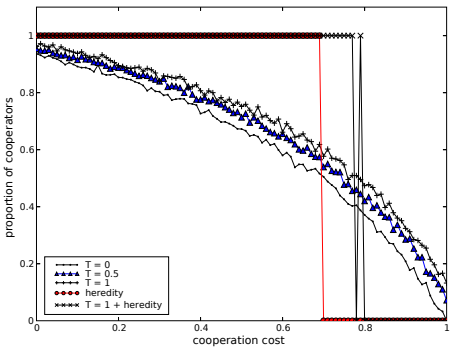


Figure 1: Influence of the cooperation's cost over the proportion of cooperators at the end of simulations. 1-dimensional grid.

4 Discussion

Our experiments confirmed the result obtained by (Jaffe and Cipriani, 2006). That is, we showed that the dynamics of CE and BE differed in very basic aspects. The fact that we used our own model implementation, confirms that this observed effects is not an artefact of the specific model implemented.

We also showed that the dimensionality of the simulated space, and thus also probably also the topology of the space (i.e. grids, networks, small world, scale

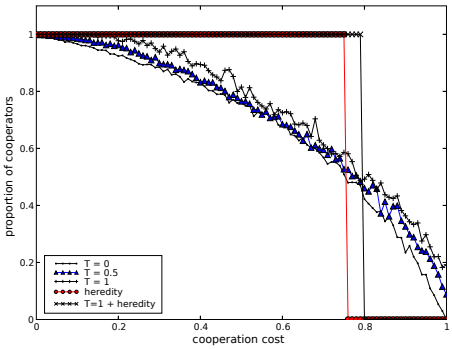


Figure 2: Influence of the cooperation's cost over the proportion of cooperators at the end of simulations. 2-dimensional grid.

free worlds, etc) will affect the dynamics under study, making this a complex subject to study.

Our results strongly suggest that simulations reporting on the evolution and/or dynamics of cooperation (or probably of anything) should specify if it is simulating BE or CE, avoiding future confusions when comparing results of simulations from different authors.

References

R. Cipriani and K. Jaffe. On the dynamics of grouping. In *Proceedings of the Fifth IASTED International Conference on Modelling, Simulation and Optimization*, pages 56–60, 2005.

P.R. Ehrlich and S.A. Levin. The evolution of norms. *PLoS Biology*, 3(6):e194, 2005.

W.D. Hamilton. Geometry for the selfish herd. *Journal of Theoretical Biology*, 31:295–311, 1971.

K. Jaffe and R. Cipriani. Nature and nurture, two different routes leading to sustainable cooperation. Submited, 2006.

R. Kurzban and D. Houser. Experiments investigating cooperative types in humans: A complement to evolutionary theory and simulations. *PNAS*, 102: 1803–1807, 2005.

M.A. Nowak and K. Sigmund. Evolutionary dynamics of biological games. *Science*, 303:793–799, 2004.

P.J. Richardson, J.E. Strassmann, and C.R. Hughes. *Not by Genes Alone: How Culture Transformed Human Evolution*. Chicago Univ. Press, 2004.

Community structure in group living animals

David Mawdsley*
*Department of Physics
University of Bath, BA2 7AY
d.mawdsley@bath.ac.uk

Richard James†
†Department of Physics
University of Bath, BA2 7AY
r.james@bath.ac.uk

Abstract

We present a technique using networks to detect intermediate level community structure within animal fission-fusion societies. The technique uses simulated annealing to optimise the quality of a proposed division of the network into communities. We also present a method that allows the statistical significance of the communities to be determined. We illustrate this technique by the detection of communities in systems of wild guppies and Galápagos sea lions. In each case, we show that this technique allows new levels of statistically significant structure to be revealed. In both cases, this allows new insights into the structure of the system under investigation.

1 Introduction

The social structure of animals in group living populations is likely to exert a profound effect of many aspects of each animal's life, as well as that of the whole population. For example, an animal's breeding and reproductive success, its foraging behaviour, the spread of disease and of information through a population will all be affected by with whom, and how frequently the individuals interact (e.g. Barnard (2004)). In many group living animals such contacts are made and broken frequently (fission-fusion groups). Despite the high frequency with which these individual associations occur, there is increasing evidence that robust non-random structures can exist (Whitehead et al., 2005). Detecting such structures give us new insights into the social and structural organisation of such systems, as well as underlying assortative tendencies that may drive them.

We present a method using networks which allows us to both detect structures at an intermediate level between that of the dyad and that of the population within systems of group living animals, and evaluate the statistical significance of such structures.

The results for wild populations of guppies and Galápagos sea lions are presented. Both systems are fission-fusion systems, where individuals are exchanged between groups with a timescale much less than the frequency with which group populations are observed.

2 Method

Traditionally, intermediate level structure in group living animals has been studied using techniques such as cluster analysis, which agglomerate animals based on some given similarity measure (e.g. Kaufman and Rousseeuw (1990)). We adopt an alternative, but complementary, approach and simply consider the structure of (repeated) interactions over an extended period of time. From this data, we construct a network. The presence of an edge in this network (representing an interaction between two animals) is determined *solely* from the observations of individual animals, without appealing to any external parameters.

Such a network is constructed from repeated censuses of the population. Animals observed in the same group on a given census are considered socially connected, and an edge is made between them. We combine many such censuses, each sufficiently separated in time from the others that they may be considered independent. We apply filters to the network to keep only strong interactions between dyads, and to remove animals that are only weakly associated with the population.

We use the technique of community detection to look for intermediate level structure. The idea of a community within a network is simply put; it is a region of the network that has a greater density of connections within it than to other parts of the network. The detection of communities within a network is, however, non-trivial, though several techniques have been proposed in recent years to do so (see Newman

(2004) for a review).

The size of systems that we have studied contain several hundred nodes. They are thus (by modern network analysis standards) quite small. This does, however, allow more sensitive, but computationally intensive techniques to be used than would be possible with larger networks. This is especially important in fission-fusion systems, where the structures that we seek may be subtle and difficult to detect. This, in turn, motivates the need to test for *significance* in the resulting structures that we find.

We maximise a measure, Q, that was originally proposed as a stopping criterion for earlier community detection algorithms (Newman and Girvan, 2004) that quantifies the quality of a set of communities.

Q is maximised via the well known optimisation technique of simulated annealing (Kirkpatrick et al., 1983). We impose an initially random division of community upon the network, and, via a series of trial moves attempt to find the global maximum of the function. The partition of communities that result thus represent our best effort at their detection. We have performed tests which show the accuracy and sensitivity of this technique outperforms existing community detection algorithms.

Significance testing

Almost all community detection algorithms, including the simulated annealing technique, will find some division of community, regardless of whether such a division is meaningful. Although large values of Q suggest a strong community structure, as noted by Guimerà et al. (2004), this isn't a sufficient condition for a *meaningful* community structure. We thus need to test the significance of the communities we find, to determine whether they are real, or simply artefacts of the process. We do this via a randomisation test (Manly, 1997).

We perform a simulation of the census rounds, "observing" the animals that were seen in each round at random, but preserving the group sizes that were observed. We thus assume random interaction between the animals. By comparing the true value of Q with the ones that result from repeated runs with the randomised network, we are able to determine the significance of the community structure we find.

The communities that we find in the guppy system are significantly assorted by body-length and the median depth of water that each animal in the community was observed. In previous studies, phenotypic assortment has been found at the level of the shoal, but our result suggests that this may also occur at a higher level of organisation.

Analysis of the sea lion data shows that geographically separated communities exist. If we apply our algorithm to each community as a separate entity, we can seek a further layer of sub-community structure, whose presence can not be explained by simple category assortment or space use, suggesting that there is evidence of genuine sociality in this system, revealed by our method.

Acknowledgements

We thank J. Krause and D. P. Croft of the University of Leeds for providing us with the guppy data, and J. B. W. Wolf and F. Trillmich from Universität Bielefeld for the sea lion data. We thank them all for stimulating and fruitful discussions.

References

C J Barnard. *Animal behaviour: mechanism, development, function, and evolution*. Prentice Hall, 2004.

R Guimerà, M Sales-Pardo, and L A Nunes Amaral. Modularity from fluctuations in random graphs and complex networks. *Phys. Rev. E*, 70:025101, 2004.

L Kaufman and P Rousseeuw. *Finding groups in data: an introduction to cluster analysis*. Wiley Interscience, 1990.

S Kirkpatrick, C D Gelatt, and M P Vecchi. Optimization by simulated annealing. *Science*, 220(4598): 671–680, 1983.

B F J Manly. *Randomization, Bootstrap and Monte Carlo Methods in Biology*. Chapman and Hall, London, 2nd edition, 1997.

M E J Newman. Detecting community structure in networks. *Eur. Phys. J. B*, 38:321–330, 2004.

M E J Newman and M Girvan. Finding and evaluating community structure in networks. *Phys. Rev. E*, 69:026113, 2004.

H Whitehead, L Bejder, and C A Ottensmeyer. Testing association patterns: issues arising and extensions. *Anim. Behav.*, 69:e1–e6, 2005.

Noise R Us: From Gene Regulatory Networks to WWW

Margaritis Voliotis*
School of Computing
University of Leeds, UK
mar@comp.leeds.ac.uk

Carmen Molina Parìs
Department of Applied Mathematics
University of Leeds, UK
carmen@maths.leeds.ac.uk

Liverpool B. Tanniemola
Department of Applied Mathematics
University of Leeds, UK
tannie@maths.leeds.ac.uk

Netta Cohen
School of Computing
University of Leeds, UK
netta@comp.leeds.ac.uk

Abstract

Gene regulatory systems are complex biological systems accounting for gene expression. Like many complex systems, they are subject to inherent noise as well as external perturbations posing a threat to their robust functionality. We review different types and sources of noise in the context of gene regulatory systems as well as different mechanisms that such systems have adopted to effectively deal with this deficiency. In so doing we consider whether there might be some link between how certain structural and architectural properties might have evolved due to a system's requirement to function under noise.

1 Introduction

Today's world is dominated by complex engineered systems of interacting components, ranging from electrical power grids to the World Wide Web (WWW) and the internet. Many of these complex systems are prone to random failures, noise or even deliberate attacks that can have devastating and far reaching implications. Indeed, two massive power failures in the summer of 2003 (in Italy and across North America) left many tens of millions of people without electricity and caused damages estimated in many billions of dollars. Such large scale events are typically due to cascading series of failures, reflecting the vulnerability of many of these systems even to local failure. Other relevant examples could be denial-of-service attacks that render specific parts of the WWW unreachable or even the spread of computer viruses through the internet.

In principle, our world is also in abundance of naturally occurring (as distinct form traditional engineered) complex systems. Such systems appear across the whole spectrum of life. A cell can be thought as a complex system of interacting biomolecules. Cells, in turn, combine to form tissues, organs and neural networks. Moving up one level, multicellular organisms can be regarded as systems consisting of a multitude of complex subsystems such as the metabolic and nervous systems. Finally, living organisms can be regarded as the fundamental components of ecosystems, forming prey-predator relationships, complex food webs and interactions with the environment.

An interesting property of biological or naturally occurring systems, that might be used as a basis for distinction from classical engineered systems, is that of adaptability. Many such systems have the ability to respond (or adapt) to environmental or internal perturbations and therefore can achieve greater robustness with regard to failures, attacks and noise.

In order to better understand the ability of biological systems to withstand or even exploit noise, it is important to pin down the sources and types of noise and their effects within specific biological contexts. This paper focuses on gene regulation networks. In particular, we present a review of a selection of recent work which provides an overview of noise in gene regulatory systems, manifestations of noise therein and implications for gene expression dynamics and long-term evolutionary processes. In doing so, this paper does not offer a comprehensive coverage of the above but rather attempts to demonstrate how certain design and organisational principles can effectively be used as noise barricades and may have evolved to do so

The remainder of the paper is organised as follows. Firstly, a brief introduction to the process of gene regulation is given. We then set out to review different

*Corresponding author

sources and types of noise that dominate gene regulatory systems as well as some of the mechanisms that are related to the ability of such systems to preserve robust and adaptive functionality. From an evolutionary perspective, such a discussion can serve as a basis for interesting questions concerning the role of noise in the evolutionary process. In this context, we propose a simple toy model, based on the work of (Kashtan and Alon, 2005), that may provide insight into the evolution of gene regulatory networks. Finally, we look at some examples of classical engineered systems and discuss how noise might be affecting their structure and topology.

2 Gene Regulatory Systems

Progress in molecular biology and experimental methods has paved the way for a more comprehensive, system-level understanding of cell function (Kitano, 2001). Indeed, many basic biological processes occurring within a cell have been extensively studied over the past decades and one can state that their basic functionality has been mastered, at least to some degree. However, addressing how biological processes are managed and synchronised, so as to produce robust functionality requires a more integrative, system-level approach.

One major and rather complex cellular system is that which regulates gene expression and determines the protein profile of the cell. Based upon interactions between DNA, RNA and protein molecules, the gene regulatory system effectively switches on and off the expression of genes to accommodate various intra-cellular needs and changing environmental conditions.

2.1 Gene Expression

Gene expression refers to the set of biomolecular processes that result in the production of proteins from their corresponding genes. Despite its rather complex nature, gene expression is a relatively well understood process (Orphanides and Reinberg, 2002) and can effectively be divided into two main steps: transcription and translation (Fig. 1).

During transcription genes (or sets of genes) are copied into intermediary (mRNA) molecules. This step usually involves the utilisation of specific regulatory proteins, known as transcription factors (TF), which bind to the DNA, either inducing (activating) or restraining (inhibiting) transcription. On the subsequent step of translation, the mRNA molecules are used as templates for the synthesis of proteins or other

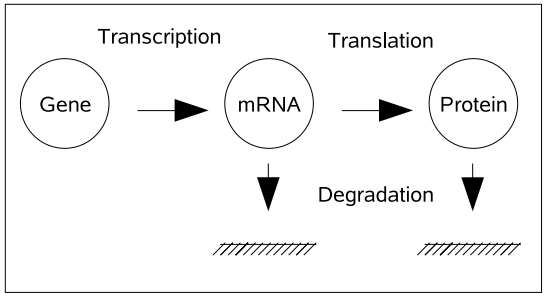


Figure 1: Simple model of the gene expression process. Intermediary (mRNA) molecules are produced from genes (transcription) and are then used as templates for the production of proteins (translation). Proteins and mRNA molecules are subject to degradation.

amino-acid chains. It should be noted that gene expression comprises of multiple interacting processes, of which transcription and translation form the basic universal core. Hence, modelling transcription and translation, simplifies the overall picture of gene expression but nonetheless captures its essence

2.2 Transcriptional Regulation

It is known that the process of gene expression can be regulated at different levels by different means (Orphanides and Reinberg, 2002). Nonetheless at the simplest level, one often focuses only on regulation accomplished via TFs. This type of regulation, dubbed as transcriptional regulation, gives rise to transcriptional regulatory systems, which essentially constitute a part of gene regulatory systems[1]. This simplification is based on our notion that transcriptional regulation is the dominant type of gene regulation, at least in prokaryotes that are widely used as model organisms.

Nevertheless, transcriptional regulatory systems demonstrate a high degree of complexity. This inherent complexity partially arises from the relatively large size of such systems, which usually account for the regulation of a few hundred to a few thousand genes. Additional complexity stems from the fact that interactions between genes are not trivial. In particular, a single TF can regulate a number of different genes and not necessarily all in the same manner or with the same strength. On the other hand, it is also possible that the same gene is regulated by multiple

[1] In the remainder of this paper, the terms gene regulation and transcriptional regulation will be used interchangeably.

TFs, which can act either cooperatively or antagonistically.

2.3 From Network Representation to Dynamics

A network (or graph, consisting of nodes and edges) is often a useful abstraction for modelling complex systems. In the case of transcriptional regulatory systems, a network can be constructed by representing distinct genes as nodes, and denoting transcriptional regulation by directed edges between nodes. Examined under this perspective, it was shown that the transcriptional regulatory network of *Saccharomyces cerevisiae* (yeast) and *Escherichia coli* revealed certain architectural (macroscopic) properties, such as small world compactness and modularity (Shen-Orr et al., 2002; Maslov and Sneppen, 2002). On a finer (mesoscopic) level, it was also shown that the these networks were in significant abundance of simple building blocks, sometimes referred to as network motifs (Shen-Orr et al., 2002; Lee et al., 2002).

In addition to mere structural and topological observations, the network abstraction can effectively be used to examine how specific interconnectivity influences the dynamics of the system. In the simplest instance, one may interpret nodes as basic dynamical systems and edges as coupling between nodes. Thus, in the context of regulatory networks, nodes shall denote the actual gene expression process of different genes while edges encode how and with what strength the protein product (TF) of one gene regulates the expression of another.

3 Noise in Gene Regulatory Systems

In the example of the power failure in North America and Italy, reports suggested that possible causes could be local fluctuations in the demand load, as well as external factors such as storms damaging the power lines. In a similar way, robustness of transcriptional regulatory systems is threatened by external perturbations as well as inherent noise. Internal and external sources of noise may sometimes act on different timescales, however they both can have significant effects on the overall functionality of the system.

3.1 Noise in Gene Expression

Perhaps the most intuitive explanation behind noise at the microscopic level, is the fact that gene expression is essentially driven by biochemical reactions that are inherently stochastic processes. In our simplistic model of gene expression (Fig. 1) the rates at which transcription, translation and degradation proceed are not fixed, mainly due to stochastic fluctuations. However when the reactants, in our case DNA, RNA and protein molecules, are in great abundance, fluctuations are insignificant and the behaviour of individual steps can be modelled, with a great degree of accuracy, in a deterministic manner. Unfortunately, this is not the case in a typical intra-cellular environment where DNA, RNA, and protein molecules are usually present only in relatively small numbers. In such an environment, the effect of stochastic fluctuations becomes a significant factor and thus a stochastic framework is needed to fully capture the dynamics of gene expression.

Over the past years, various probabilistic models of gene expression have been proposed in literature (McAdams and Arkin, 1997; Thattai and van Oudernaarden, 2001). Summing up this theoretical work, origins of noise in gene expression can be effectively modelled by the following probabilistic events:

- gene activation and deactivation
- transcription initiation
- translation initiation
- decay of the mRNA and protein molecules.

The random activation and deactivation of a gene is usually attributed to random TF binding on DNA as well as to other events (e.g. chromatin remodelling). This switching in gene activity between *on* and *off* states leads to production of mRNA molecules in bursts of random sizes. Transcriptional bursting in turn leads to considerable fluctuations in the protein product, especially when transition rates between *on* and *off* states are slow (Blake et al., 2003; Raser and O'Shea, 2004).

Similarly to transcription, translation also occurs in random size bursts. This is due to the random lifetime of mRNA molecules during which several protein copies can be produced. According to the translational bursting mechanism a gene with high translational efficiency (number of proteins produced per mRNA molecule) is predicted to show a wide distribution of protein abundance, especially when mRNA molecules exist in low numbers (Thattai and van Oudernaarden, 2001). The above theoretical speculation for the translation mechanism was indeed verified by experimental work on *Bacillus subtilis* strains (Ozbudak et al., 2002).

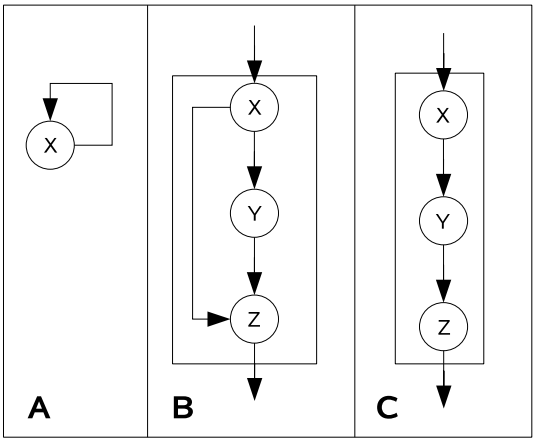


Figure 2: Examples of network motifs in gene regulatory networks (Lee et al., 2002; Shen-Orr et al., 2002). **A** Autoregulation. **B** Feed-forward Loop. **C** Regulatory Chain.

3.2 Noise in Network motifs

It has been proposed that transcriptional regulatory networks, both in prokaryotes and eukaryotes, are essentially assembled from basic structural units known as network motifs (Lee et al., 2002; Shen-Orr et al., 2002). From a dynamical point of view, network motifs can be thought of as simple signal transducing and/or controlling mechanisms. They are usually subject to some input regulatory signals (TF) and produce a corresponding output in the form of a protein product. Quite similar to logic gates in digital circuits, the functionality and dynamical behaviour of network motifs is dictated by their internal structure.

At this level of organisation, notions of intrinsic and extrinsic noise are of particular relevance. Focusing on a simple structural unit of the system, such as a network motif, the former type reflects noise produced internally by its components (as discussed above). This inherent noise propagates through the network motif leading to a noisy output. On the other hand, the term extrinsic noise accounts fluctuations in the output signal that originate from a noisy regulatory cue. Of particular importance is the notion of extrinsic noise and how such noise is propagated through the network motif. As discussed in section 4.2 the strong nonlinearity of the control architecture implies that in some cases external noise will be suppressed, whereas in other cases, it will be significantly amplified.

3.3 Network noise

Gene regulation networks attempt to capture the entire set of regulatory interactions within a cell. As such, they comprise an intricate web of network motifs, that are organised into larger structures (or modules) and form the global gene network of the cell. These cell networks can have characteristic topologies and statistics. At the network level, one often focuses on noise produced by fluctuating environmental and intra-cellular conditions that affect the overall stability and robustness of the regulatory system.

Environment is a basic factor compromising the stability of gene regulatory systems. Environmental conditions (e.g. temperature, pH) provide cues that can trigger the system, which in turn responds by modifying its expression pattern, or switching between alternate gene expression profiles. Therefore, environmental fluctuations are effectively transformed into noise in the system. Intra-cellular sources of noise can be ascribed to a wide variety of factors directly or indirectly affecting the process of gene expression. Such factors can be fluctuations in metabolite concentrations and variability in the activity of utility macromolecules (e.g. ribosomes and polymerases). It has also been known that cell specific characteristics such as cell size, cell age and the stage of the cell cycle can alter the gene expression profile (Kaern et al., 2005).

4 Noise Related Mechanisms

From the above discussion, it is perhaps striking that transcriptional regulatory systems are not only functioning under fluctuating environments, but are also comprised of unreliable, inherently noisy components. However, these systems demonstrate remarkably robust and adaptive functionality that sustains life. To accomplish that, gene regulatory systems utilise certain mechanisms, at their different organisational levels. Such mechanisms not only barricade the system against detrimental effects of noise but also have the ability to exploit noise in advantageous ways when this is possible.

4.1 Gene Level

At the component level, various distinct strategies can be adopted by a gene so as to achieve the same levels of protein expression (Fraser et al., 2004). These strategies essentially differ in the average rates in which transcription and translation are proceeding. One such strategy, for example, might yield high tran-

scription rates while imposing low translation rates. Simply put, this strategy produces high numbers of mRNA molecules each one producing in turn low proteins numbers. On the other extreme a gene can accomplish the same protein numbers by producing limited mRNA molecules (low transcription rate) but each one yielding a high number of proteins (high translation rate). Finally, intermediate strategies can also be realised where both transcription and translation proceed at intermediate rates.

Following our discussion in section 3.1, the least noisy strategy is the one that maximises transcriptional efficiency while minimising the rate of translation, since such a combination minimises the effects of translational bursting. Indeed, Fraser et al. (2004) in their bioinformatics study discovered that the most essential genes of yeast showed a strong bias towards utilising this most uniform expression pattern. Moreover, the fact that not all genes follow the same strategy can be reasoned under the perspective that such a noise reducing mechanism is energetically expensive (more mRNA molecules have to be produced) and it should thus be adopted only by vital genes whose fluctuations might lead to deleterious effects (Fraser et al., 2004).

4.2 Network Motifs

Among the network motifs identified in the transcriptional regulatory network of yeast and *E. coli*, are those of autoregulation, the feedforward loop and the regulator chain (Lee et al., 2002; Shen-Orr et al., 2002). The properties of such structures have been recently studied both in theory and experimentally, using synthetically engineered gene circuits, providing us with deeper understanding of how noise is managed in gene regulatory systems.

Autoregulation can be thought of as an elementary form of control mechanism where the output of the gene expression process is fed back as a regulatory input (Fig. 2A). Autoregulation can either be negative or positive, depending on how the protein-product regulates its corresponding gene. Becskei and Serrano (2000) engineered regulatory circuits in *E. coli* cells to assess the importance of negative autoregulation with regard to noise. The results showed that the amount of protein produced from autoregulation-free circuits showed great variability among the cell population as opposed to the protein produced by the autoregulated circuit, which demonstrated significant stability. A similar study by Isaacs et al. (2003) focused on positive autoregulation, and demonstrated that the amounts of protein expressed under such a mechanism follow a bimodal distribution as a result of the inherent noise. In other words positive autoregulation amplifies noises to the point that two distinct phenotypes arise.

Feed-forward loops consist of a gene regulating another in a both direct and indirect manner, through a third gene (Fig. 2B). There are basically two types of Feedforward Loops: coherent ones where the sign of both regulation paths is the same, and incoherent ones where the signs of regulation are opposite. While both types were found in the studied networks the coherent type appears to be far more abundant (Shen-Orr et al., 2002). Summarising the theoretical work of Mangan and Alon (2003) coherent feed-forward loops may act as low pass filters for extrinsic noise, responding only to persistent input stimuli.

In the regulator chain motif, a gene regulates a second gene which in turn regulates a third one and so forth (Fig. 2C). The regulatory cascades can be of varying size and it has been suggested that they account for series of transcriptional events that happen sequentially (Lee et al., 2002). A recent experimental study by Hoosangi et al. (2005) dealt with noise in transcriptional regulatory cascades as a function of their length. In this study transcriptional cascades of several repressing steps were engineered in *E. coli* cells. The experimental results were consistent with the theoretical predictions that long cascades essentially act as extrinsic-noise filters, just as in the case of coherent feed-forward. However, intrinsic noise accumulates as the cascade length is increased.

In contrast to the noise filtering properties that some of the network motifs demonstrate, the noise amplification accomplished by others seems rather counter intuitive. Remarkably enough, however, the latter mechanisms can effectively be used to produce phenotypic diversification out of noise. This is particularly beneficial, especially in the case of bacterial populations, that can exploit diversity to survive and adapt under fluctuating environments. It has also be speculated that in a similar way cell differentiation is accomplished, in the developmental stages in multicellular organisms, form initially homogeneous cell populations (Kaern et al., 2005). It therefore appears that proper interaction between different genes is essential not only for the system to be shielded against noise but also for stochasticity to be exploited.

4.3 Architecture

Little is known about the actual architectural design of gene regulatory systems and much research is still in progress. Such enormous biological systems, con-

sisting in general of thousands of components require vast experimental and theoretical work to be explored in their entirety. Even in the case of well studied model organisms, such as *E. coli* and yeast, their fully detailed regulatory networks are yet to be completed[2]. Circumventing this limitation, studies, focusing on subsets of the actual networks, reveal a rather modular design (Shen-Orr et al., 2002; Lee et al., 2002). However, for one to generalise these findings to the statement that gene regulatory systems are indeed modular, an implicit assumption is often made posing modularity as a design principle underpinning such biological systems.

Nonetheless, modularity is an interesting property with regard to noise. In particular, it can effectively be used by gene regulatory networks to isolate inherent noise into constrained subnetworks, thus minimising the risk of overall failures. Finally, even in extreme cases of deleterious failures, the functional decoupling that modularity provides, might under some conditions, give rise to graceful degradation.

Another interesting architectural property that was observed in the model regulatory network of yeast was that of small-world compactness (Maslov and Sneppen, 2002). This property, reflects the shortness of regulatory pathways, in the sense that only a few regulation steps are usually involved in the expression of a given gene. This property has perhaps an intuitive role if one takes into account the fact that noise propagates through the network. Therefore, minimal number of regulatory steps could prove to be an essential way of controlling the accumulation of noise.

5 Evolution

From an evolutionary perspective, the above discussion can serve as a basis for interesting questions concerning whether and how the notions of adaptation and evolution are linked to noise at every level of organisation. For example at the gene level, Fraser et al. (2004) provided strong support that certain vital genes have evolved towards utilising certain expression strategies that effectively reduce inherent noise. One might also consider alternative ways of gene regulation, other than transcriptional, and examine their noise properties. In doing so, deeper insight can be gained on whether mechanisms controlling inherent noise are subject to evolutionary pressure.

Focusing on the mesoscopic and macroscopic organisational levels one can go even further seeking

[2]RegulonDB (http://regulondb.ccg.unam.mx), an online database of transcriptional regulation in *E. coli*, currently includes 139 experimentally verified TFs.

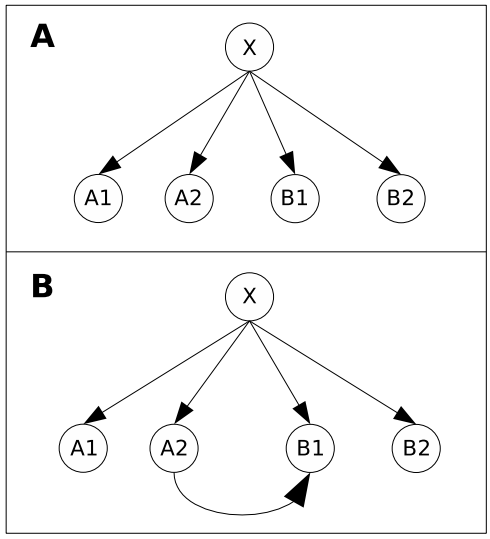


Figure 3: Simple toy models of regulatory networks with one TF. **A** Gene X activates the four genes. **B** Regulated genes can also interact with each other.

ways in which evolution may have affected the topology of gene regulatory networks. Although such an issue is still open to discussion, one can examine the different viewpoints and draw some general conclusions. For instance, one may ask whether modularity and network motifs could evolve purely by mutational drift under neutral evolution. This would imply that such traits provide no selective advantages for the organism. Alternatively, such traits could have been subject to selection if, indeed, they offer some evolutionary advantage. In the latter case noise might prove quite significant, since as discussed above, there are strong indications that modularity and internal structures, such as network motifs, provide a framework that can effectively deal with and/or reduce noise.

In an attempt to model and gain insight into such evolutionary processes, it has been proposed that modularity and network motifs might have spontaneously evolved as a result of an ever-changing environment (Lipson et al., 2002; Kashtan and Alon, 2005). In the example of Kashtan and Alon (2005), the environment defines modular goals, consisting of basic subgoals. As these goals are varied it was shown that modularity was evolved to make the system more adaptable to these changes. One may ask whether such a proposition holds for the case of transcriptional regulatory networks.

Simple toy models of a gene regulatory network and *in silico* evolutionary simulations might be used to demonstrate the effectiveness of different network

structures in performing modular tasks. Let us consider a simple regulatory network where gene X encodes for a TF, which is actively regulating four genes, namely A1, A2, B1 and B2 (Fig. 3A). The biological function of these genes is not independent but assumed to be coupled in some way. For example, genes A1 and A2 could cooperatively metabolise substance A whereas genes B1 and B2 could metabolise substance B. On a more complex network, regulatory interaction can also exist between the metabolising genes (Fig. 3B) and indirect feedback can reach the TF X (not shown). The initial regulatory network does not demonstrate high modularity and one can readily think of a more modular version, where for example genes B1 and B2 are regulated separately from genes A1 and A2, through a second TF (Fig. 4A-B). Not all network structures achieve the same efficiency under similar conditions, since they cannot reproduce exactly the same expression profiles. For example, in the case where the metabolites A and B act independently, the network in Fig. 4A can be more efficient than that in Fig. 3A, since it can independently regulate the production of different subsets of genes.

Following the example of Kashtan and Alon (2005), mutations in our simple model will not account for changes in the biological functionality of genes (i.e. metabolism) but affect the structure of the network by removing or adding regulatory relationships between genes. Taking the model one step further, one can also assume that such mutations might as well affect the dynamics of the system by altering the rates which govern gene expression. In addition, rare events of gene duplication can also be incorporated. Such events affect the topology of the network by duplicating nodes while preserving the relationships between them.

Now suppose that under a typical steady environment, substances A and B are present at fixed (possibly different) concentrations and the evolutionary process can optimise the parameters of the system (rates of gene expression, and interactions) for the cell to be energetically satisfied. However, if our environment is not constant, in the sense that concentrations of substances can fluctuate, a different network topology may be better suited. This notion of a changing environment can be thought of as a form of noise acting at different timescales, even at an evolutionary one. Under a noisy environment, the mere optimisation of expression rates may be insufficient, and evolution may act on the topology of the network, possibly giving rise to more modular topologies

More generally, genes A1, A2, B1 and B2 of our

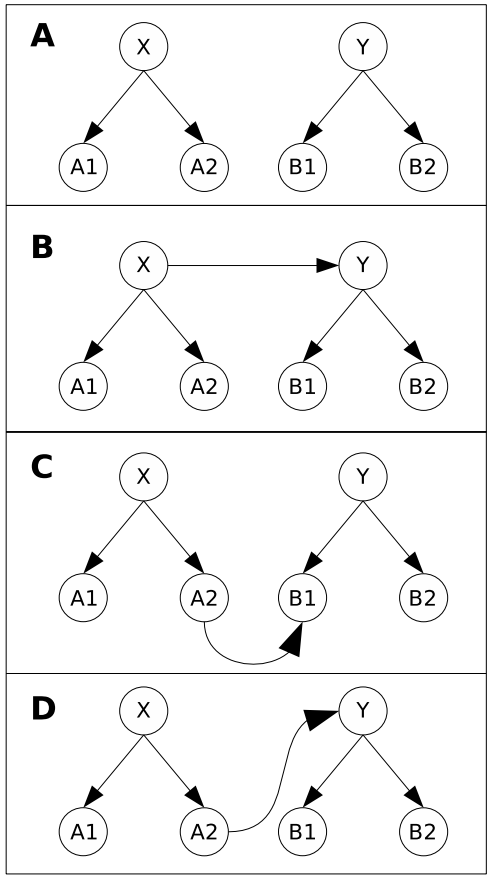


Figure 4: Simple toy models of regulatory networks with two TFs (A-B) and up to 6 TFs (C-D). **A** Two independent subnetworks where TFs X and Y activate genes (A1/A2) and (B1/B2) respectively. **B** Example demonstrating a possible regulatory interaction between the TFs. **C** Example demonstrating a possible transcriptional interaction between regulated genes. **C** Example demonstrating a possible feedback interaction between the regulated genes and their regulators. Combinations of the possible kinds of interaction **B-D** are also possible.

model may take on any role, and may even be TFs themselves. In that case, we may also envision regulatory interactions as illulstrated in Fig. 4C-D. In all these cases, the same principles apply: changes in the environment may provide selective pressures that eventually manifest themselves in the topology of the network.

Comparing evolutionary simulation with and without environmental noise can lead to some general insights on whether the evolutionary design of such toy models is affected by noise. In fact, one can go even further examining a more complicated toy model that also accounts for inherent noise and determining how this source of noise might affect the topology of the network in the long run.

6 The WWW and Power Grids

Similar to gene regulatory systems, classical engineered systems are also prone to noise. Surprisingly enough, some of these systems also demonstrate basic structural and organisational commonalities with biological systems, such as scale-free architectures, small world interconnectivity, modularity and abundance of network motifs (Dorogovtsev and Mendes, 2003; Milo et al., 2002). Therefore, it is perhaps interesting to try and generalise the above discussion on gene regulatory networks and examine whether noise, across biological and engineered systems, might have any relevance to the emergence of such structural properties. We pick the WWW and electric grids as two examples.

The WWW is essentially a vast information platform that enables efficient storage, retrieval and exchange of information across a network of nodes (web pages) and edges (hyper links). Fluctuations in the demand load of specific information can compromise its accessibility, as they might result in information retrieval latency or even unavailability due to limitations of the underlying communication network (the internet). To secure the WWW against this type of noise, certain measures have been adopted including mirror links and cached information on different (independent) web servers[3]. These measures are effectively altering the network topology since they involve the creation of new nodes and links and thus may give rise to certain structural properties. In fact, the creation of mirror sites can be likened to gene duplications in evolution. However, perhaps unlike gene regulatory networks, the WWW is also under a rapid evolutionary process, with information being published and/or withdrawn and links being created and/or removed on at least as fast a time scale. This process is has a dramatic effect on network topology and is likely to mask the topological effects due to mirroring and caching. Thus in general the question of whether noise might be related to the emergence of certain structural network properties seems somewhat obscure in the case of the WWW.

A more intuitive and pronounced example might be that of power grids, where growth takes place on a much slower time scale, so as to meet the needs of growing demand, population centres and industrial development. Even for a relatively static network, power grids are usually subject to load demand fluctuations as a consequence of various external factors (e.g. seasonal temperature fluctuations). Since noise in this case cannot be effectively controlled one intuitive way of minimising its effects is by adopting a modular grid design so that probable blackouts can be localised. Indeed, studies of power grids have revealed a highly clustered design (Dorogovtsev and Mendes, 2003), which has emerged during their evolution, in part due to our need to secure these systems to intrinsic and extrinsic sources of noise, as well as to direct insults.

7 Conclusions

Understanding how complex biological systems, such as gene regulatory systems, control noise, and how they evolved to accomplish that, is a vital step towards their structural and functional understanding. However, it can also provide valuable lessons for the design of complex engineered systems. Most engineered systems were specifically designed to utilise relatively reliable components. Nonetheless, even rare events may, under certain circumstances, lead to cascading failures, whether such failures are due to an internal component or external effects. In order to shield a system against such rare events, lessons learnt and design principles gleaned from biological systems may prove useful. In particular, the prevalence of modularity and small scale control structures in a range of engineered complex systems, could be suggestive of possible relevance of biological networks. It is hoped, therefore, that further research into the structure and function of biological systems, as compared with artificial or engineered ones, and specifically such systems' ability to effectively handle noise, will lead to practical applications in the design and regulation of complex systems.

[3]Note, of course, that the creation of mirror sites cached pages etc. is motivated by a combination of facts of which network noise is only one.

Acknowledgements

NC was funded by an EPSRC grant (EP/C011953/1).

References

A. Becskei and L. Serrano. Engineering stability in gene networks by autoregulation. *Nature*, 405: 590–593, 2000.

W. J. Blake, M. Kaern, C. R. Cantor, and J. J. Collins. Noise in eukaryotic gene expression. *Nature*, 422: 633–637, 2003.

S. N. Dorogovtsev and J. F. F. Mendes, editors. *Evolution of Networks. From Biological to the Internet and WWW*. Oxford University Press, 2003.

H. B. Fraser, A. E. Hirsh, G. Giaver, J. Kumm, and M. B. Eisen. Noise minimization in eukaryotic gene expression. *PLoS Biology*, 2(6):834–838, 2004.

S. Hoosangi, S. Thiberge, and R. Weiss. Noise characteristics of feed forward loops. *Proc. Natl. Acad. Sci. USA*, 102(10):3581–3686, 2005.

F. J. Isaacs, J. Hasty, C. R. Cantor, and J. J. Collins. Prediction and measurment of an autoregulatory genetic module. *Proc. Natl. Acad. Sci. USA*, 100 (13):7714–7719, 2003.

M. Kaern, T. C. Elston, W. J. Blake, and J. J. Collins. Stochasticity in gene expression: from theories to phenotypes. *Nature Reviews Genetics*, 6:451–464, 2005.

N. Kashtan and U. Alon. Spontaneous evolution of modularity and network motifs. *Proc. Natl. Acad. Sci. USA*, 102(39):13773–13778, 2005.

H. Kitano, editor. *Foundations of Systems Biology*. The MIT Press, 2001.

T. I. Lee, N. J. Rinaldi, F. Robert, D. T. Odom, Z. Bar-Joseph, G. K. Gerber, N. M. Hannett, C. T. Harbison, C. M. Thompson, I. Simon, J. Zeitlinger, E. G. Jennings, H. L. Murray, D. B. Gordon, B. Ren, J. J. Wyrick, J. Tagne, T. L. Volkert, E. Fraenkel, D. K. Gifford, and R. A. Young. Transcriptional regulatory networks in saccharomyces cerevisiae. *Science*, 298:799–804, 2002.

H. B. Lipson, J. B. Pollack, and N. P. Suh. On the origins of modular variation. *Evolution*, 56(8):1549–1556, 2002.

S. Mangan and U. Alon. Structure and function of the feed-forward loop network motif. *Proc. Natl. Acad. Sci. USA*, 100(21):11980–11985, 2003.

S. Maslov and K. Sneppen. Specificity and stability in topology of protein nwtworks. *Science*, 296:910–913, 2002.

H. H. McAdams and A. Arkin. Stochastic mechanisms in gene expression. *Proc. Natl. Acad. Sci. USA*, 94:814–819, 1997.

R. Milo, S. Shen-Orr, N. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network motifs: Simple building blocks of complex networks. *Science*, 298:824–827, 2002.

G. Orphanides and D. Reinberg. A unified theory of gene expression. *Cell*, 108:439–451, 2002.

E. M. Ozbudak, M. Thattai, I. Kurtser, A. D. Grossman, and A. van Ourdanaarden. Regulation of noise in the expression of a single gene. *Nature genetics*, pages 69–73, 2002.

J. M. Raser and E. K. O'Shea. Control of stochasticity in eukaryotic gene expression. *Science*, 304:1811–1814, 2004.

S. S. Shen-Orr, R. Milo, S. Mangan, and U. Alon. Network motifs in the transcriptional regulation network of escherichia coli. *Nature genetics*, 31: 65–68, 2002.

M. Thattai and A. van Oudernaarden. Intrinsic noise in gene regulatory networks. *Proc. Natl. Acad. Sci. USA*, 98(15):8614–8619, 2001.

Statistical Analysis of Dynamic Graphs

Xiaomeng Wan
[⋆]Faculty of Computer Science
Dalhousie University
Halifax, Canada
xwan@cs.dal.ca

Jeannette Janssen
[†]Dept. of Mathematics and Statistics
Dalhousie University
Halifax, Canada
janssen@mathstat.dal.ca

Nauzer Kalyaniwalla
[‡]Faculty of Computer Science
Dalhousie University
Halifax, Canada
nauzerk@cs.dal.ca

Evangelos Milios
[§]Faculty of Computer Science
Dalhousie University
Halifax, Canada
eem@cs.dal.ca

Abstract

Communications between large numbers of individuals can be modeled as a dynamic graph. The graph is the integrated effect of the individuals acting autonomously. To identify and analyze communication patterns, we study dynamic graphs by examining the global behaviour of local, vertex-specific measures. In this paper, we introduce novel vertex-specific measures, and apply scan statistics to examine the global extremes of these measures. We apply our methods to a set of email data, and show that the different measures offer complementary views of the data.

1 Introduction

A dynamic graph is a graph whose edges and vertices may appear and disappear over time. Examples are phone call graphs, email communication graphs and graphs representing visits of web pages by users. In these graphs, vertices represent entities and edges represent communication transactions. The variation of dynamic graphs over time can be used to profile what normal behaviour is, which is the basis for detecting anomaly and predicting future behaviours. Cortes et al. (2003) study phone call graphs, and present ideas on how to predict behaviour based on historic data. Priebe et al. (2005) introduce the use of scan statistics to study dynamic graphs. They study an e-mail communication graph using a density-based scan statistic, with the aim of detecting anomalies.

The dynamic graphs that are of interest to us can be considered as social networks. In most studies of such graphs, the dynamic nature of the graphs is summarized into the formation or weighting of a graph representing a general notion of "connectedness" or "contact" between vertices over the time period considered. This summarized graph can then be studied with the tools of social networks.

Our approach is, instead, to consider the dynamic graph as a time series of graphs, and to study it by focusing our attention on a number of locality measures derived from the links present in the neighbourhood of a specific vertex. We present a number of promising measures targetting different features of communication patterns. Local statistics are generated for these measures and their extremes are identified and analysed with scan statistics. Those nodes generating extremes which deviate from the general trend can be considered as anomalies and are worth detailed investigation. We applied our methods to a large collection of email data; results are presented and discussed in Section 4.

2 Locality measures

The dynamic graphs we consider are all derived from communications between individuals. Hence the graph is the global result of a large number of individual actions. It therefore stands to reason that we can model global behaviour by analyzing and integrating the local behaviour of each vertex. For each vertex, a variety of time-dependent locality statistics can be defined. A statistical summary of the behaviour of these statistics over time can be used to create a vertex-specific signal. The set of all signals can be used to model normal behaviour, and thus to classify vertices and filter out noise.

Specifically, a dynamic graph is considered as a time series of static graphs. Typically, the time in-

terval in which communications have been observed is divided into shorter intervals. The static graph corresponding to each interval will have a link between vertices if an interaction took place between those vertices in that interval. The locality measures considered are all defined with respect to the edges present in the neighbourhood of a vertex. The *k-th order neighbourhood* of a vertex v consists of all vertices that are at most k "hops" away from v. We distinguish the *dynamic neighbourhood*, which consists of all vertices that received a link from v in a specific time interval for $k = 1$, and the *permanent neighbourhood*, which consists of all vertices that received at least one link from v during the whole time period under consideration for $k = 1$.

A time dependent locality measure based on the density of links in the dynamic neighbourhood of a vertex was introduced by Priebe et al. (2005). For $k = 1, 2$, their measure is the number of edges present in the k-th order dynamic neighbourhood of a vertex v. For $k = 0$, the measure is defined as the number of edges originating from v. Interesting results can be obtained if this measure is compared with a density measure derived from the permanent neighbourhood. For $k = 0$, permanent and dynamic neighbourhood density measures are equal, but for $k \geq 1$ they can catch different aspects of a vertex's behaviour.

Sometimes, a shift in communication patterns may occur without a change in activity level. To catch such shifts, we introduce a *novelty measure*. This measures the number of "new" links in a neighbourhood, i.e. links that have not been observed in a fixed number of previous intervals. Note that the novelty measure must be defined with respect to the past τ-week neighbourhood.

To capture the behaviour of locality measures over time, the running mean and standard deviation can be computed for each measure, and used to standardize the signal. More specifically, let $\Psi_{k,t}(v)$ be a time-dependent locality measure defined with respect to the k-th neighbourhood. Then, the mean of locality statistic based on the recent history of the previous τ time intervals is defined by (Priebe et al., 2005):

$$\hat{\mu}_{k,t,\tau}(v) = \frac{1}{\tau} \sum_{t'=t-\tau}^{t-1} \Psi_{k,t'}(v) \tag{1}$$

The variance of locality statistic based on the recent history of the previous τ time intervals is defined by:

$$\hat{\sigma}^2_{k,t,\tau}(v) = \frac{1}{\tau - 1} \sum_{t'=t-\tau}^{t-1} (\Psi_{k,t'}(v) - \hat{\mu}_{t,\tau}(v))^2 \tag{2}$$

The vertex-standardized locality statistic is defined by (Priebe et al., 2005):

$$\hat{\Psi}_{k,t}(v) = \frac{\Psi_{k,t}(v) - \hat{\mu}_{k,t,\tau}(v)}{\max(\hat{\sigma}_{k,t,\tau}(v), 1)} \tag{3}$$

3 Scan statistics

A common technique to detect local anomalies in behaviour by a global analysis is the use of scan statistics. Scan statistics are commonly used in signal analysis, and in the detection of anomalies in localized health data (Glaz et al., 2001). Priebe et al. (2005) first applied scan statistics techniques to the local density measures described in the previous sections. The idea behind a scan statistics approach is to study a large number of local measures by studying the extremities of its values over all localities.

In our case, it makes sense to study both the maximum and the minimum of the standardized locality measures. More precisely, the statistic studied is the maximum (minimum) of the vertex-standardized locality measure which is defined by (Priebe et al., 2005):

$$\tilde{M}_{k,t} = \max_v \tilde{\Psi}_{k,t}(v) \tag{4}$$

The minimum of vertex-standardized locality measure is defined by:

$$\tilde{M}'_{k,t} = \min_v \tilde{\Psi}_{k,t}(v) \tag{5}$$

Note that the maximum of the standardized locality measure represents a sudden increase in activity, while a minimum represents a sudden drop. To determine whether or not certain values of this maximum (minimum) represent an anomaly, this statistic is itself temporally normalized as follows (Priebe et al., 2005).

$$S_{k,t} = \frac{\tilde{M}_{k,t} - \hat{\mu}_{k,t,\ell}}{\max_{k,t,\ell}(\hat{\sigma}_{k,t,\ell}, 1)} \tag{6}$$

Where $\hat{\mu}_{k,t,\ell}$ and $\hat{\sigma}_{k,t,\ell}$ are the running mean and variance of $\tilde{M}_{k,t}$ defined by:

$$\hat{\mu}_{k,t,\ell} = \frac{1}{\ell} \sum_{t'=t-\ell}^{t-1} \tilde{M}_{k,t'} \tag{7}$$

$$\hat{\sigma}^2_{k,t,\ell} = \frac{1}{\ell - 1} \sum_{t'=t-\ell}^{t-1} (\tilde{M}_{k,t'} - \hat{\mu}_{k,t,\ell})^2 \tag{8}$$

4 Experimental setup and results

We apply the methods described above to a large collection of email data from the Faculty of Computer Science at Dalhousie University. The data are derived from the log files of the email server of Faculty of Computer Science covering the period from May 2004 to September 2005. There are 16,580 email addresses involved with 1,500 active accounts (defined as sending emails to more than five distinct users during the period under study). The email addresses are anonymized before being used in the study, but the categories of email addresses (faculty, student, staff, mailing lists, ...) are preserved.

We divide the data into disjoint, one-week intervals. The locality statistics are calculated for the vertices in each window. Using the number of edges for each vertex as the locality statistic (for k = 0, 1, 2), we compute scan statistics based on *(i)* the original number of edges and *(ii)* the novelty measure, as described in Section 2. Figure 1 below shows the vertex-standardized maximum results for the novelty measure.

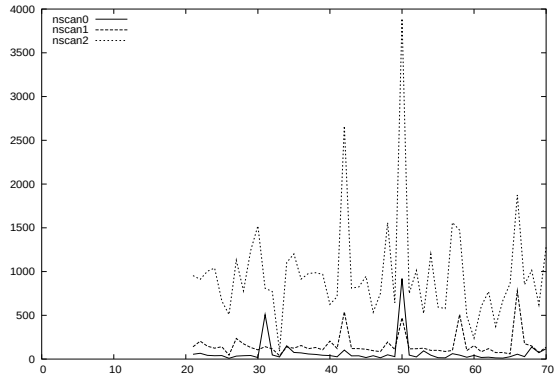


Figure 1: Time series of standardized scan statistics and max degrees for k = 0,1,2 with locality statistics as new links (Novelty) on CS data

The peaks in Figure 1 correspond to dormant or low activity vertices that suddenly come alive. The peaks in scan0 (degree of the vertex), are due to mailing lists that suddenly start transmitting after not sending any messages in the previous τ weeks. On the other hand, the peaks in scan2 are, in general, caused by low activity vertices communicating with a very high volume vertex *viz.* an un-moderated mailing list. The situation for scan1 is more complex, since the first neighbourhood of a very active node can mask the signal for a node that ultimately shows

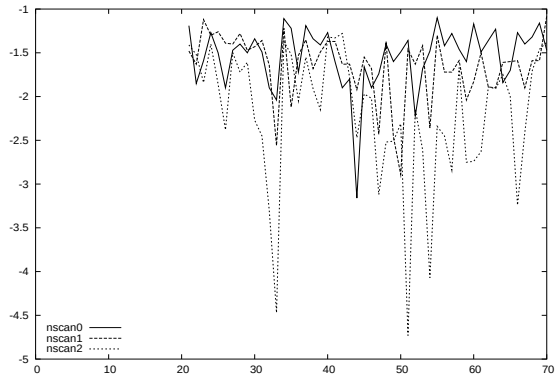


Figure 2: Time series of standardized scan statistics and min degrees for k = 0,1,2 with locality statistics as new links on CS data

a prominent peak in scan2.

The minima of the novelty measure are shown in Figure 2. In this plot, minima in scan0 simply show the sudden drop in activity of a normally active node, the effect is purely individual, but may be of interest. Since scan2 captures a large group of nodes, the minima in scan2 correspond to a general drop in activity. The pronounced minima in scan2 correspond to Christmas (week 32), the first week of May (week 52) which falls in the interval between the winter and summer terms, and the end of August (week 67) all of which correspond to general drop in activity around the university.

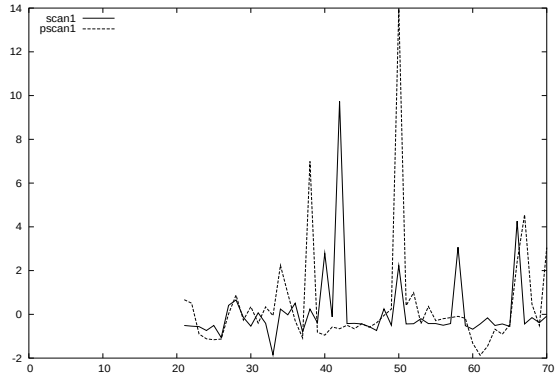


Figure 3: Time series of temporally-normalized standardized scan statistics and max degrees for k = 1 with locality statistics as links for dynamic and permanent windows on CS data

Figure 3 compares the temporally-normalized stan-

dardized maximum results of the dynamic and permanent windows. The two measures show different peaks. On closer examination of the data we notice that every peak for the moving window (the scan statistic) captures mailing lists, which are principally bursty vertices. The peaks of the permanent window however, all correspond to individual users. The effect is clearly due to the scan statistic's susceptibility to picking up on bursty behaviour. Consequently, bursty nodes mask the underlying changes in communication patterns of individual users which might be of interest. This effect points to separating the nodes into homogeneous groups and then using the machinery of scan statistics to look for anomalies within the groups.

These are preliminary results, which, however, indicate that locality measures are indeed a promising tool to analyze dynamic graphs. Currently, more experiments are being performed to study the effect of dynamic versus permanent neighbourhoods, and to compare different varieties of the locality measures. In the future, we hope to aggregate all locality measures with the aim of defining a vertex-specific signal, which can be used to categorize vertices according to their behaviour.

5 Conclusions and future work

Our experiments show that the application of scan statistics to local, vertex-specific measures is successful in identifying anomalous behaviour. Different vertex-specific measures appear to identify different types of anomalies. Moreover, the relative behaviour of scan statistics derived from different vertex measures, or from different levels of locality can indicate abnormal patterns of special interest. To interpret the anomalies identified by scan statistics, the vertices and time periods where peaks occur should be studied in detail by other methods.

The analysis of our results shows that the peaks of almost all scan statistics are achieved by vertices corresponding to mailing lists. This is no surprise, since scan statistics are especially sensitive to their bursty behaviour (long periods of inactivity followed by mass mailings). Changes in communication patterns between individual email correspondents are easily masked by the presence of such bursty nodes. In future work, we will develop methods to filter out "noisy" nodes, by characterizing nodes into groups that exhibit fairly homogeneous behaviour. Application of scan statistics to such groups should lead to the discovery of a wider variety of anomalies.

Acknowledgements

The authors gratefully acknowledge the financial support from the MITACS Network of Centres of Excellence, IT Interactive Services Inc., and the Natural Sciences and Engineering Research Council of Canada.

References

C. Cortes, D. Pregibon, and C. Volinsky. Computational Methods for Dynamic Graphs. *Journal of Computational and Graphical Statistics*, 12(4): 950–970, 2003.

J. Glaz, J. Naus, and S. Wallenstein. *Scan Statistics*. Springer-Verlag, New York, 2001.

C. Priebe, J. Conroy, D. Marchette, and Y. Park. Scan Statistics on Enron Graphs. *Computational and Mathematical Organization Theory*, 11(3): 229–247, 2005.

Reconstruction of Flexible Gene-Protein Interaction Networks using Piecewise Linear Modeling and Robust Regression

Ronald L. Westra[⋆]
[⋆]Dept. Mathematics, Maastricht University
Maastricht, The Netherlands
westra@math.unimaas.nl

Ralf L.M. Peeters[†]
[†]Dept. Mathematics, Maastricht University
Maastricht, The Netherlands
ralf.peeters@math.unimaas.nl

Goele Hollanders[‡]
[‡]Dept. Computer Science, Hasselt University
Hasselt, Belgium
goele.hollanders@uhasselt.ac.be

Karl Tuyls[§]
[§]Dept. Computer Science, Maastricht University
Maastricht, The Netherlands
karl.tuyls@cs.unimaas.nl

Abstract

In this study we will focus on piece-wise linear state space models for gene-protein interaction networks. We will follow the dynamical systems approach with special interest for partitioned state spaces. From the observation that the dynamics in natural systems tends to punctuated equilibria, we will focus on piecewise linear models and sparse and hierarchic interactions, as for instance described by Glass, Kauffman, and de Jong. Next, the paper is concerned with the identification (a.k.a. reverse engineering and reconstruction) of dynamic genetic networks from microarray data. We will describe exact and robust methods for computing the interaction matrix in the special case of piecewise linear models with sparse and hierarchic interactions from partial observations. Finally, we will analyze and evaluate this approach with regard to its performance and robustness towards intrinsic and extrinsic noise.

Keywords: piecewise linear, robust identification, hierarchical networks, gene expression data, gene regulatory networks.

1 Introduction

This paper is concerned with the identification of dynamic gene-protein interaction networks with intrinsic and extrinsic noise from empirical data, such as a set of microarray time series.

Prerequisite for the successful reconstruction of these networks is the way in which the dynamics of their interactions is modeled. The formal mathematical modeling of these interactions is an emerging field where an array of approaches are being attempted, all with their own problems and short-comings. The underlying physical and chemical processes involved are multifarious and hugely complex. This condition contrasts sharply with the modeling of inanimate Nature by physics. While in physics huge quantities of but a small amount of basic types of elementary particles interact in a uniform and deterministic way provided by the fundamental laws of nature, the situation in gene-protein interactions deals with tens of thousands of genes and possibly some million proteins. The quantities thereby involved in the actual interactions are normally very small, as one single protein may be able to (in)activate a specific gene, and thereby change the global state of the system. For this reason, gene regulatory systems are much more prone to stochastic fluctuations than the interactions involved in normal anorganic reactions. Moreover, each of these interactions is different and involves its own peculiar geometrical and electrostatic details. There are different processes involved like transcription, translation and subsequent folding. Therefore, the emergent complexity resulting from gene regulatory networks is much more difficult to comprehend.

In the past few decades a number of different formalisms for modeling the interactions amongst genes and proteins have been presented. Some authors focus on specific detailed processes such as the cir-

cadian rhythms in *Drosophila* and *Neurospora* (10), (11), or the cell cycle in *Schizosaccharomyces* (Fission yeast) (14). Others try to provide a general platform for modeling the interactions between genes and proteins. For a thorough overview consult de Jong (2002) in (2), Bower (2001) in (1), and others (6), (13).

We will focus on dynamical models, and not discuss static models where the relations between genes are considered fixed in time. In discrete event simulation models the detailed biochemical interactions are studied. Considering a large number of constituents, the approach aims to derive macroscopic quantities. More information on discrete event modeling can be found in (1).

2 Modeling gene-protein interactions as a piecewise linear system

The traditional approach to modeling the dynamical interactions amongst genes and proteins is by considering them as biochemical reactions, and thus representing them as 'rate equations'. The concept of chemical rate equations consists of a set of differential equations, expressing the time derivative of the concentration of each constituent of the reaction as some rational function of the concentrations of all the constituents involved. Though the truth of the underlying biochemical interactions between the constituents is generally accepted, a rate equation is not a fundamental law of Nature, but a statistical average over the entire ensemble of molecular collisions that contribute to an actual chemical reaction (22). So, rate equations are statistical approximations that – under certain conditions – predict the average number of reactive collisions. The actual observed number will fluctuate around this number, depending on the details of the microscopic processes involved. In case of biochemical interactions between genes and proteins the applicability of the concept of rate equations is valid only for genes with sufficient high transcription rates. This is confirmed by recent experimental findings by Swain and Elowitz (5), (16), (18), (19).

From the above, we may conclude that modeling can only be successful for genes with sufficiently high transcription rates. Even in the optimal case, we would obtain a high-dimensional (reflecting the number of genes, RNAs, and proteins involved – so tens of thousands), non-linear, differential equation, that is subject to substantial stochastic fluctuations. Much more problematic is the fact that the precise details of most reactions are unknown, and therefore cannot be modeled as rate equation. This could be compensated by a well-defined parametrized generic form of the interactions, such that the parameters could be estimated from sufficient empirical data. A generic form based on rational positive functions is proposed by J. van Schuppen (23). However, in the few cases where parts of such interaction networks have been described from experimental analysis, like the circadian rhythms in certain amoeba (10), or the cell cycle in fission yeast (14), it is clear that such forms have a too extensive syntax to be of any practical use.

Let us for the moment forsake these problems, and consider the dynamics of gene-RNA-protein networks. When we assume a stochastic differential equation as model for the dynamics of the interaction network, the relation can be expressed as:

$$\dot{x} = f(x, u|\theta) + \xi(t) \tag{1}$$

Here $x(t)$, called the state-vector, denotes the N gene expressions and M RNA/protein densities at time t – possibly involving higher order time derivatives. $u(t)$ denotes the P controlled inputs to the system, such as the timing and concentrations of toxic agents administered to the system observed. $\xi(t)$ denotes a stochastic Gaussian white noise term. This expression involves a parameter vector θ, that contains the coupling constants between gene expressions and protein densities. We can consider this system as being represented by the state vector $x(t)$ that wanders through the (at least) $(N + M)$-dimensional space of all possible configurations. In the formalism of dynamic systems theory, eventually x will enter an area of attraction, and become subject to the influence of an attractor. An attractor here can be an uniform convergent attractor, a limit cycle, or a 'strange attractor'. We can understand the entire space as being partitioned into cells, where such attractors – or their antagonists so-called repellers – reign. Thus, the behavior of x can be described by motion through this collection of cells, swiftly moving through cells of repellers, until they enter the basin of attraction of an attractor. Under the effects of external agents via the vector $u(t)$ or by stochastic fluctuations via $\xi(t)$ they can leave this cell, and start wandering again, thereby repeating the process. Now, a vital assumption is that in each cell the behavior is governed by specific (un)stable equilibrium points, and therefore it is possible to make a linear approximation of equation 1 in the cell with index l as:

$$\dot{x}(t) = F_l x(t) + G_l u(t) \tag{2}$$

In case of a uniform attractor the largest eigen-value

of F_l will be negative, and in case of a uniform repeller the smallest eigen-value will be positive. We can now formalize the qualitative behavioral dynamics of gene-protein interactions as predominantly linear behavior near the stable equilibria – called the steady states, interrupted by abrupt transitions where the system quickly relaxes to a new steady state, either externally induced or by process noise.

In biology such behavior is frequently observed, as for instance in embryonic growth where the organism develops by transitions through a number of well-defined 'check points'. Within each such checkpoint the system is in relative equilibrium. There is an ongoing debate on mathematical modeling of cell division as *checkpoint mechanisms* versus *limit-cycle oscillators*, see (20). We will follow the view of *piecewise linear behavior* (PWL, also known more appropriately as piecewise *affine* behavior). This approach corresponds to the piecewise linear models introduced by Glass and Kauffman (9), and the qualitative piecewise linear models described by de Jong et al. (2), (3).

3 The identification of *piecewise linear networks* by L_1-minimization

Next, we will be concerned with the identification (a.k.a. *reverse engineering* or *reconstruction*) of piecewise linear gene regulatory systems from microarray data. The nature of our problem – few microarray experiments and lots of genes – implies that we are dealing with *poor data* (as opposed to *rich data*), where the number of measurements is *a priori* insufficient to identify all parameters of the system. One standard approach to circumvent this problem is by dimension reduction through the clustering of related genes. We consider the case where time series of genome-wide expression data is available. The case of the identification of a *simple* linear system is discussed in Peeters and Westra (15), (26), and Yeung et al. in (27). In the following, we will be concerned with the identification of *piecewise* linear systems. Our aim is to obtain the gene-gene interaction matrix. This matrix can be interpreted as a connectivity matrix, and so directly relates to the graph of the gene regulatory network. With this network we are able to make statements like: 'the expression of this gene causes that and that cluster of genes to alter their expression in this and this way'.

Let us in the following assume a dynamical input-output system Σ that switches irregularly between K linear time-invariant subsystems $\{\Sigma_1, \Sigma_2, \ldots, \Sigma_K\}$. Let $S = \{s_1, s_2, \ldots, s_{K-1}\}$ denote the set of – possibly unknown – switching times, i.e. the time instants $t = s_l$ that the system switches from subsystem Σ_l to Σ_{l+1}. Similarly as with the simple linear networks, we assume *Hankel matrices* $X = (x[1], x[2], \ldots, x[M])$, and $U = (u[1], u[2], \ldots, u[M])$ at M sampling times $T = \{t_1, t_2, \ldots, t_M\}$, representing full observations of the N states and P inputs. The interval between two sample instants is denoted as $\tau_k = t_{k+1} - t_k$. In first instance we assume that the system is sampled on regular time intervals, i.e. that the sample intervals are equal to τ. Within one subsystem Σ_l the relation between the inputs $u(t)$ and outputs $y(t)$ is represented as a state-space system of first-order differential (for continuous time systems) or difference equations (for discrete time systems), using an auxiliary vector $x(t)$ spanning the so-called subspace.

Continuous time:

$$\dot{x}(t) = F_l x(t) + G_l u(t), \tag{3}$$
$$y(t) = H_l x(t) + J_l u(t). \tag{4}$$

Discrete time:

$$x[k+1] = A_l x[k] + B_l u[k], \tag{5}$$
$$y[k] = C_l x[k] + D_l u[k]. \tag{6}$$

The relation between these is given by:

$$A_l = e^{\tau F_l}, \tag{7}$$
$$B_l = e^{\tau F_l} G_l. \tag{8}$$

with $x[k] = x(t_k)$.

3.1 Determination of the new state equilibrium points

Moreover, in each new state the new equilibrium point $\mu_l \in \mathbb{R}^N$ has also to be established. The linearization near μ_l can be written as:

$$\frac{\partial}{\partial t}(\mu_l + (x - \mu_l)) = F_l(x - \mu_l) + G_l u + \mathcal{O}(\|x - \mu_l\|^2) \tag{9}$$

which can be rewritten as: $\dot{x} = F_l x + \tilde{G}_l \tilde{u}$, with:

$$\tilde{G}_l = (G_l | - F_l \mu_l), \tag{10}$$
$$\tilde{u} = \begin{pmatrix} u \\ 1 \end{pmatrix}. \tag{11}$$

The reasoning is similar in the discrete case, and we obtain: $x[k+1] = A_l x[k] + \tilde{B}_l \tilde{u}[k]$. Therefore,

we can follow the original formulation and, using $\tilde{u}$ rather than u as input, estimate A_l and $\tilde{B}_l$, and using:

$$\tilde{B}_l = (B_l| - A_l\mu_l)\,, \tag{12}$$

to compute μ_l and B. We will follow this approach, and from here on drop the *tilde*, and simple write B_l for $(B_l| - A_l\mu_l)$, and $u[k]$ for $\begin{pmatrix} u[k] \\ 1 \end{pmatrix}$.

3.2 General dynamics of switching subsystems

In the context of piecewise linear systems of gene regulatory systems, the dynamics is slightly different to the case of simple linear systems as in (15). In our context we assume that we observe *all* N genes, and that there is no direct through-put. This means that $C_l = I$ and $D_l = 0$ for all l. Therefore, we can suffice with equation 5 corrected for the equilibrium point:

$$x[k+1] = A_l x[k] + B_l u[k]. \tag{13}$$

We furthermore assume that the system matrices in these equations are constant during intervals $[s_l, s_{l+1} >$, and abruptly change at the transition between the intervals at $t = s_{l+1}$. We assume that on the time scale τ the system has relaxed to its new state. This means that we do not observe *mixed states*, which would severely complicate the problem of identification.

Finally, we define the *weights* w_{kl}, as the membership functions of observation k to subsystem Σ_l; if observation $\{x[k], u[k]\}$ belongs to system Σ_l then $w_{kl} = 1$, if $\{x[k], u[k]\}$ does not belong to Σ_l then $w_{kl} = 0$. This definition allows for a *fuzzy* definition of weight, such that $w_{kl} \in [0, 1]$. *A priori*, we thus can state two constraints on w:

$$\forall_{k,l} w_{kl} \in [0, 1], \tag{14}$$

$$\forall_l \sum_l w_{kl} = 1. \tag{15}$$

The challenge in system identification is to estimate the relevant model parameters in piecewise linear dynamics from empirical observations. The success of this approach depends on the amounts of empirical data available – *rich* or *poor*, the validity of the mathematical model, the levels of process noise and measuring noise, and the nature of the sampling process. In case of regular sampling the discrete model 5 can be applied which leads to more straightforward techniques than the continuous model 3 that should be used in case of irregular sampling. In the following sections we will study a number of these conditions in more detail.

3.3 Identification of PWL models with *unknown* switching and *regular* sampling from *poor* data

The assumption that the switching times between the linear subsystems are completely known suits various experimental conditions, as for instance when toxic agents are administered. In many biological situations, however, the exact timing between subsystems is not known, as during embryonic growth and in many metabolical processes.

3.3.1 As an extension to the simple linear systems in case state derivatives are available

When a sufficiently accurate record of estimates of the state derivatives $\dot{X} = \{\dot{x}[1], \dot{x}[2], \ldots, \dot{x}[M]\}$ is available, we can simply rewrite this problem as a special case of the method described in the case of a simple linear problem as in (15). In fact, by exploiting the data $\mathcal{D} = \{X, U, \dot{X}\}$, the problem can be stated as a linear equation in terms of new matrices H_1 and H_2 as:

$$\dot{X} = H_1 X + H_2 U. \tag{16}$$

In this equation the matrices H_1 and H_2 relate to the – unknown – system matrices $\{A_1, B_1, \ldots, A_K, B_K\}$ and ditto unknown weights $\{w_{kl}\}$ as:

$$\mathsf{vec}\ (H_1) = W \cdot \mathsf{vec}\ (A), \tag{17}$$

$$\mathsf{vec}\ (H_2) = W \cdot \mathsf{vec}\ (B). \tag{18}$$

The matrices A, B, and W are composed as follows:

$$A = \begin{pmatrix} A_1 \\ \ldots \\ A_K \end{pmatrix}, \quad B = \begin{pmatrix} B_1 \\ \ldots \\ B_K \end{pmatrix}, \tag{19}$$

$$W = w \otimes I_{N^2} = \begin{pmatrix} w_{1,1} I_{N^2} & \ldots & w_{1,K} I_{N^2} \\ \ldots & \ldots & \ldots \\ w_{M,1} I_{N^2} & \ldots & w_{M,K} I_{N^2} \end{pmatrix} \tag{20}$$

where $\otimes$ is the Kronecker-product, and I_{N^2} is the $N^2 \times N^2$ identity matrix. Note that equation 16 is not anymore a linear problem, as the unknown matrices A, B, and W appear in a non-linear way in the equation. This equation is exactly of the type of simple linear networks as in (15). Therefore, its solution method is fully applicable, so that an efficient and accurate algorithm is available for solving this problem in terms of H_1 and H_2. However, now the problem has shifted to solving two additional non-linear equations:

$$W \diamond A = H_1, \tag{21}$$

$$W \diamond B = H_2. \tag{22}$$

where A, B, and W have to be solved from the known – i.e. computed – matrices H_1 and H_2. The operation $\diamond$ makes the relations in equations 21 and 18 explicit. This is an underdetermined system that can only be solved by additional information, such as assuming sparsity for A, and a block structure for W, such as the two constraints in equations 14 and 15.

This non-linear problem can thus be solved in terms of H_1 and H_2, but not in terms of A, B, and W. It is a bilinear problem in terms of A and B for fixed W, otherwise it is a quadratic problem. As a quadratic programming problem this is not a a well-posed problem, i.e. it has a nonsingular Jacobian at optimality and is ill-conditioned as the iterates approach optimality. Therefore, we follow a different approach and split the problem in two LP-problems that are well-posed. The approach is as follows: (i) initialize A, B, and W, (ii) perform the iteration:

1. Compute H_1 and H_2, using the approach from Peeters and Westra (15) on equation 16,
2. Using fixed values for the weights W, compute A and B using equations 21, and 22,
3. Using fixed values for matrices A and B, compute the weights W using equations 14, 15, 21, and 22,

until: (iii) a cumulative weighted error criterion $\mathcal{E}$ has converged sufficiently – or a maximum number of iterations has passed. A proper choice for the criterion function is:

$$\mathcal{E}(A, B, W|\mathcal{D}) = \sum_{k,l} w_{kl} \|A_l x[k] + B_l u[k] - \dot{x}[k]\|_2^2 \tag{23}$$

This problem can be solved by minimizing the quadratic L_2-criterion subject to mentioned constraints, for instance by a gradient descent method. We can, however, formulate a different approach for solving this problem by defining an alternative criterion function $\mathcal{E}$, namely as a linear L_1-criterion:

$$\mathcal{E}_1(A, B, W|\mathcal{D}) = \sum_{k,l} w_{kl} \|A_l x[k] + B_l u[k] - \dot{x}[k]\|_1 \tag{24}$$

This expression allows for an LP-formulation of the problem, in which $\mathcal{E}_1$ serves as the objective function. Thus, we can split the non-linear optimization problem as two separate LP-formulations that are successively applied in the iteration; (i) an LP-problem LP_1 for obtaining the system matrices A and B from minimizing objective function $\mathcal{E}_1$ with given weights w, subject to the constraints in equations 21 and 22; and (ii) an LP-problem LP_2 for obtaining the weights w from minimizing objective function $\mathcal{E}_1$ with given system matrices A and B , subject to the constraints in equations 14, 15, 21, and 22.

We will revisit this philosophy in the next Section, when reviewing the more realistic case when the state derivatives of the gene expressions are *not* available.

4 Numerical experiments and performance of the approach.

This approach resulted in an efficient and fast algorithm that is able to accurately estimate the gene-gene coupling matrix for tens of thousands of genes based on only several hundred genome wide measurements, and that is robust towards measurement noise. With increasing measurement noise or decreasing number of measurements the approach retains the strongest gene-gene coupling links - i.e. the largest modal value of the coupling matrix A - longest, see Figure 1. A basic assumption in the approach is the sparsity of the underlying gene-gene coupling matrix, represented by the number of non-zero entries per row. If this number grows above a certain threshold the performance of the approach is severely affected, see Figure 2*b*. A number of numerical experiments were performed with this approach. These controlled experiments consist of the comparison of reconstructed network with the - known - original network structure. They were all performed on a PC with an PIV dual XEON processor of 3.2 GHz and 4096 MB RAM memory under Linux fedora core 3, using Matlab 6.5 release 13 including the optimization toolbox. The Matlab routine `linprog` was used to solve LP problems; its default solution method is a primal-dual interior point method, but an active set method can optionally be used too. *For larger problems it turned out to be essential for obtaining reasonable computation times, that the LP problems were solved by application of the active set method on the dual problem formulation. Therefore this method was adopted throughout all the experiments.* In line with the definitions above, we use the parameters N, M, K to quantify the size and complexity of the input. In addition, the sparsity of the interaction matrix A is measured by the number of nonzero entries per row and denoted by k (which should be much less than N). To quantify the quality of the resulting approximation A_{est} of A^* two performance measures are introduced: the number of errors N_e and the CPU-time T_c as clocked on the same platform.

1. *The number of errors N_e.*
   Errors in the reconstruction are generated by the

failure of the algorithm to identify the true non-zero elements of the original sparse vector x_0. These errors stem from false positives and false negatives in the reconstructed vector x_d. Their numbers are added up to produce the total number of errors N_e.

2. *The CPU-time T_c.*
   Using internal clocking, the time T_c required to perform the full computation was measured. As all numerical experiments are executed on the same platform under similar conditions, this provides a measure to compare problem instances.

The numerical experiments clearly demonstrate the range where the approach is effective. For relatively moderate noise levels and a high degree of sparsity i.e., a small number k of nonzero elements in the rows of matrix A - and not too many external stimuli p and switching times K, the approach allows one to reconstruct a sparse matrix with great accuracy from a relative small number of observations $M \ll N$. For example, a row of A with 30,000 components of which all but 10 are equal to zero, can be efficiently reconstructed from just 150 independent measurements, see Figure 4a. The sparsity property of A fits in nicely with the technique of L_1-minimization, which automatically will always set many entries of the solution A^* to zero, whereas L_2-regression would spread out the error over all components, thus creating many small components. Reconstruction of large networks from this approach is straightforward: each of the rows of the gene-gene interaction matrix can be computed independently from the same set of microarray experiments.

Acknowledgements

The authors express their gratitude to Gert-Jan Bex and Marc Gyssens from the University of Hasselt for their valuable contribution.

References

[1] Bower J.M., Bolouri H.(Editors), Computational Modeling of Genetic and Biochemical Networks, *MIT Press*, 2001. bibitemDavidson1999Davidson E.H. (1999), A View from the Genome: Spatial Control of Transcription in Sea Urchin Development, *Current Opinions in Genetics and Development*, **9**, pp. 530 – 541.

[2] de Jong H., Modeling and Simulation of Genetic RegulatorySystems: A Literature Review, Journal of Computational Biology, 2002, Volume 9, Number 1, pp. 67–103

[3] de Jong H., Gouze J.L., Hernandez C., Page M., Sari T., Geiselmann J., Qualitative simulation of genetic regulatory networks usingpiecewise-linear models, Bull Math Biol. 2004 Mar;66(2): pp 301–40.

[4] D'haeseleer P., Liang S., Somogyi R., Genetic Network Inference: From Co-Expression Clustering to Reverse Engineering, *Bioinformatics*, vol. **16**, no. 8, 2000, pp. 707–726.

[5] Elowitz M.B., Levine A.J., Siggia E.D., Swain P.S., Stochastic gene expression in a single cell, *Science*, vol.**297**, August 16, 2002, pp.1183–1186.

[6] Endy, D, Brent, R. (2001) Modeling Cellular Behavior, *Nature* 2001 Jan 18; 409(6818):391-5.

[7] Fuchs J.J. (2003), More on sparse representations in arbitrary bases, in: Proc. 13th IFAC Symp. on System Identification, Sysid 2003, Rotterdam, The Netherlands, August 27-29, 2003, pp. 1357–1362.

[8] Fuchs J.J. (2004), On sparse representations in arbitrary redundant bases, IEEE Trans. on IT, June 2004.

[9] Glass L., Kauffman S.A. (1973), The Logical Analysis of Continuous Non-linear Biochemical Control Networks, *J.Theor.Biol.*, 1973 Vol. 39(1), pp. 103–129

[10] Goldbeter A (2002) Computational approaches to cellular rhythms. Nature 420, 238-45

[11] Gonze D, Halloy J, and Goldbeter A (2004) Stochastic models for circadian oscillations : Emergence of a biological rhythm. *Int J Quantum Chem* **98**, pp 228–238.

[12] Guthke R., Möller U., Hoffmann M., Thies F., Töpfer S., 2004, Dynamic network reconstruction from gene expression data applied to immune response, *Bioinformatics*, 2004, pp 2261

[13] Hasty J., McMillen D., Isaacs F., Collins J. J., (2001), Computational studies of gene regulatory networks: in numero molecular biology,*Nature Reviews Genetics*, vol. 2, no. 4, pp. 268– 279, 2001.

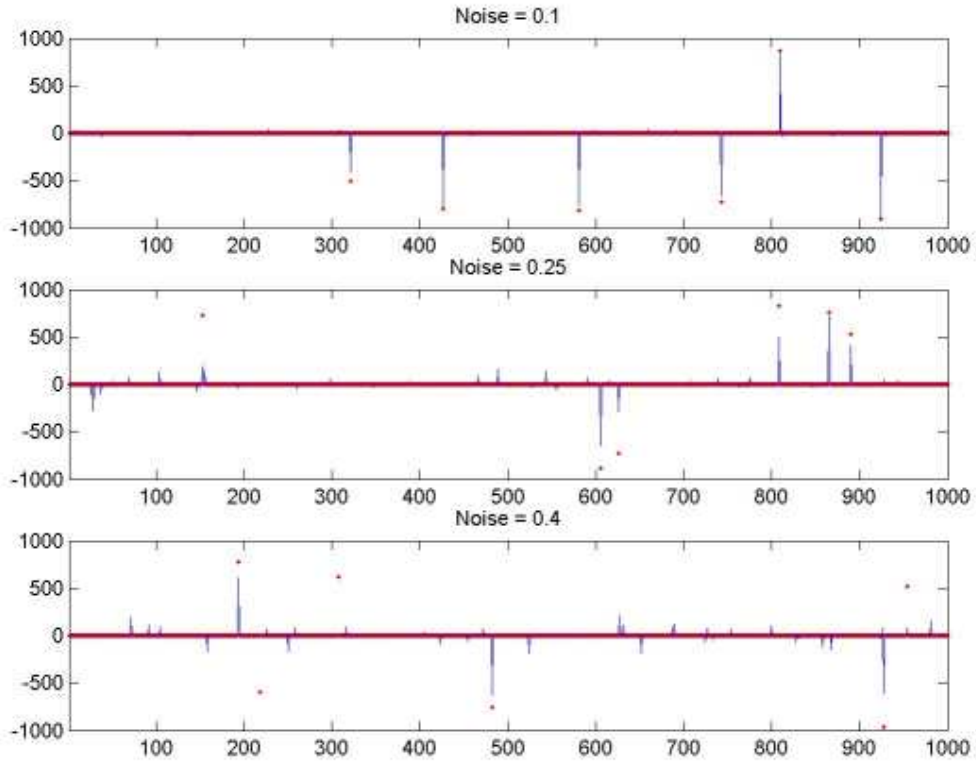


Figure 1: The influence of increasing intrinsic noise on the identifiability. The plot shows the corresponding values of the gene-gene matrix $a \equiv vec(A)$, and increasing zero-mean Gaussian noise added to A. The red dots indicate the true value of a, and the blue line the reconstructed values a^*. For low noise levels, like 0.1, the non-zero values of a are recovered without exception. At noise level 0.4 only the largest modulus maxima values have a chance to be found.

[14] Novak B, Tyson JJ (1997) Modeling the control of DNA replication in fission yeast, PNAS, USA, Vol. 94, pp. 9147-9152, August 1997.

[15] Peeters R.L.M., Westra R.L., On the identification of sparse gene regulatory networks, *Proc. of the 16th Intern. Symp. on Mathematical Theory of Networks and Systems* (MTNS2004) Leuven, Belgium July 5-9, 2004

[16] Rosenfeld N, Young JW, Alon U, Swain PS, Elowitz MB, Gene regulation at the single-cell level, *Science* 307 (2005) pp 1962.

[17] Somogyi R., Fuhrman S., Askenazi M., Wuensche A. (1997). The Gene Expression Matrix: Towards the Extraction of Genetic Network Architectures. Nonlinear Analysis, *Proc. of Second World Cong. of Nonlinear Analysis* (WCNA96) 30(3) pp 1815–1824.

[18] Swain P.S., Efficient attenuation of stochasticity in gene expression through post-transcriptional control, J Mol Biol 344 (2004) pp 965.

[19] Swain P.S., Elowitz MB, Siggia ED, Intrinsic and extrinsic contributions to stochasticity in gene expression, *PNAS* 99 (2002) pp 12795.

[20] Steuer R. (2004), Effects of stochasticity in models of the cell cycle:from quantized cycle times to noise-induced oscillations, Journal of Theoretical Biology 228 (2004) 293-301.

[21] Tegnér J., Yeung M.K.S., Hasty J., Collins J.J., Reverse engineering gene networks: Integrating genetic perturbations with dynamical modeling, *Proc. Nat. Acad. Science*, vol. **100**, no. 10, 2003, pp. 5944–5949.

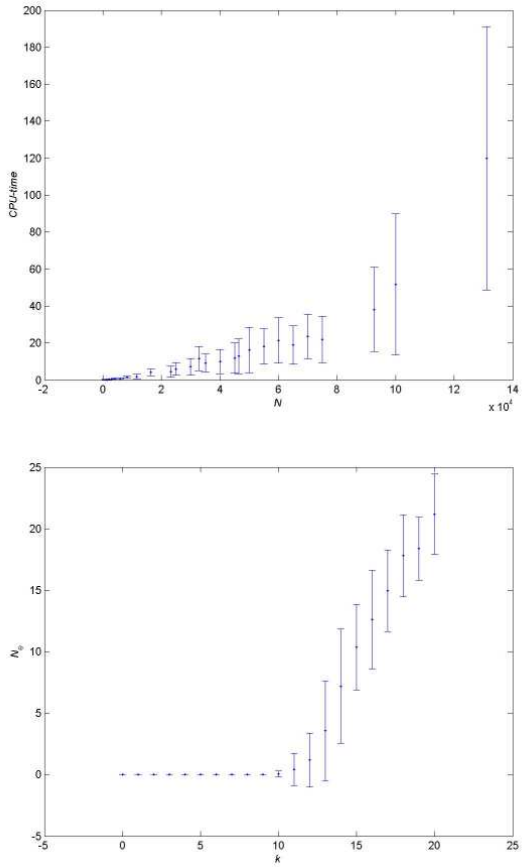


Figure 2: *a*: CPU-time T_c as a function of the problem size N, *b*: Number of errors as a function of the number of nonzero entries k in x_0, for $M = 150$, $m = 5$, $N = 50000$.

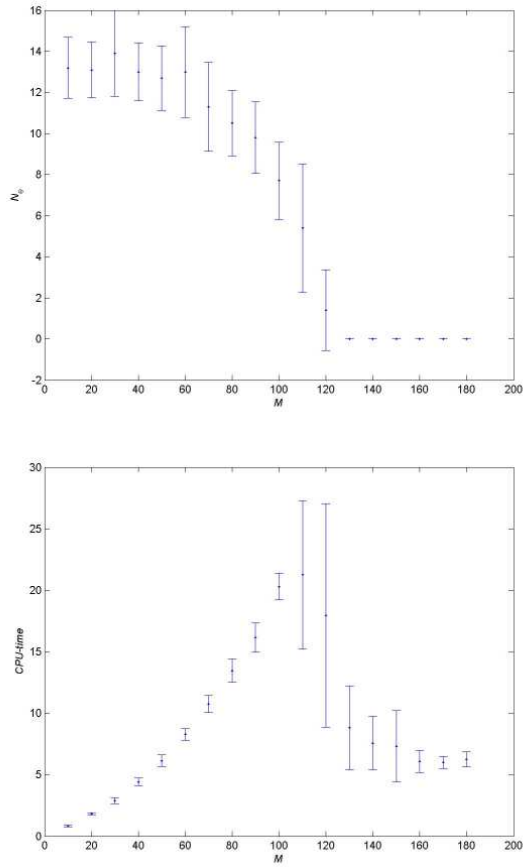


Figure 3: *a*: Number of errors as a function of M for $N = 50000$, $k = 10$, $m = 0$, *b*: Computation time as a function of M, for $N = 50000$, $k = 10$, $m = 0$.

[22] van Kampen N. G. (1992), Stochastic Processes in Physics and Chemistry, Elsevier ScienceB. V., Amsterdam, (1992).

[23] van Schuppen J.H. (2004), System theory of rational positive systems for cell reaction networks, CWI Report MAS-E0421, December 2004, ISSN 1386-3703

[24] Verdult V., Verhaegen M., Subspace Identification of Piecewise Linear Systems, In *Proc. 43rd IEEE Conference on Decision and Control (CDC)*, pp 3838–3843, Atlantis, Paradise Island, Bahamas, December 2004.

[25] Westra R.L., Peeters R.L.M. (2004), Modelling and identification of dynamical gene interactions: presentation, Workshop Intelligent Technologies for Gene Expression Based Individualized Medicine, 14th May 2004, Jena/Germany

[26] Westra R.L.,(2005*a*), Piecewise Linear Dynamic Modeling and Identification of Gene-Protein Interaction Networks, Nisis/JCB Workshop reverse engineering, Jena, June 10, 2005.

[27] Yeung M.K.S., Tegnér J., Collins J.J., Reverse engineering gene networks using singular value decomposition and robust regression, *Proc. Nat. Acad. Science*, vol. **99**, no. 9, 2002, pp. 6163–6168.

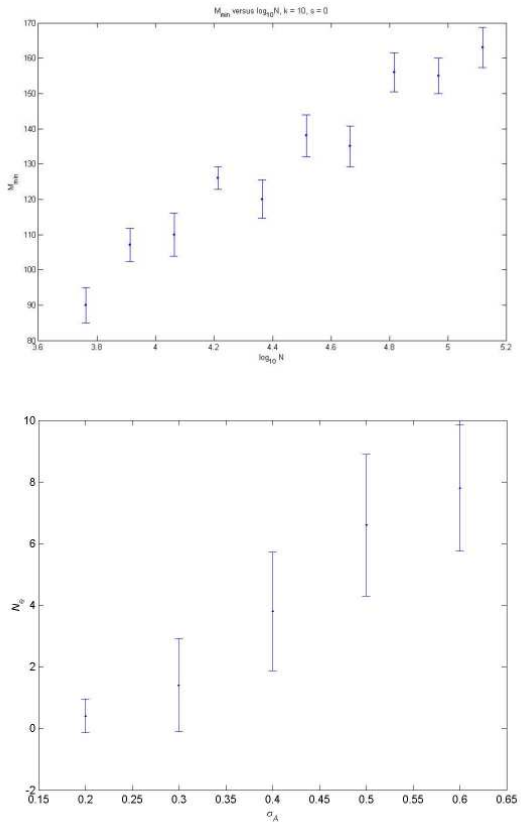


Figure 4: *a*: Dependency of the critical value M_{min} required to compute the matrix free of error versus the problem size N, *b*: Number of errors as a function of the intrinsic noise level σ_A, for $N = 10000$, $k = 10$, $m = 5$, with $M = 150$ and measuring noise $\sigma_B = 0$.

General Classification of Networks

Thomas Wilhelm*
*Theoretical Systems Biology
Leibniz Institute for Age Research - Fritz Lipmann Institute
Beutenbergstr.11
D-07745 Jena, Germany
wilhelm@fli-leibniz.de

Jens Hollunder*
hollund@fli-leibniz.de

Jong-Kwang Kim*
gensdei@fli-leibniz.de

Andreas Beyer*
beyer@fli-leibniz.de

Abstract

We present a new general classification scheme of networks, which is appropriate for any type: weighted or unweighted, and directed or undirected networks. We show that all networks can be grouped into one of two general classes: democracy or dictatorship networks. In democracy networks nodes tend to play equal roles in the network, whereas in dictatorship networks some nodes are more prominent for the network function. In other words, in democracy networks there is more cycling of information (or mass, or energy), while dictatorship networks are characterized by a straight through-flow from sources to sinks. The classification is based on information theoretic measures. If the *redundancy* of a given network is smaller than in a randomized version ($R < R_r$), we call it democracy network. In dictatorship networks it is $R > R_r$. We use the measure *Medium Articulation* to quantify the complexity of networks. Complex networks ($MA > MA_r$) are always between the pure democracy and the pure dictatorship networks. Taken together, we distinguish four different network types: pure democracy networks ($R < R_r, MA < MA_r$), complex democracy networks ($R < R_r, MA > MA_r$), complex dictatorship networks ($R > R_r, MA > MA_r$), and pure dictatorship networks ($R > R_r, MA < MA_r$). Finally, different real networks (weighted and unweighted, directed and undirected) are classified according to our proposed scheme.

1 Introduction

Starting with two pioneering works (2; 18), the last years have seen a surge of papers dealing with networks on fields as diverse as social networks (9; 10; 19), food webs (6; 21; 22), communication networks (4; 23), transportation networks (7), and sub-cellular networks, such as metabolic (13), protein interaction (5; 24), and genetic networks (14). Many interesting network properties have been found, providing us insights into both, the dynamics on networks (17) and (on another time scale) the evolutionary processes leading to such networks (1; 2). However, nearly all of these papers deal with simple unweighted networks, where links are either present or absent. Here we go one step further and present a characterization of the most general form of networks: directed weighted (non-binary) networks. Here n nodes are connected by maximally n^2 directed weighted edges (links) t_{ij} from node i to node j. Note that both, undirected and unweighted networks can also be analysed in this framework: in unweighted networks the edges have only two different weights (0-no edge, 1-edge), and undirected networks can be understood as directed with all links pointing in both directions (with the appropriate weight).

Note that many networks could better be described by their weighted form. For instance, in acquaintance networks one can quantify the acquaintances, or in coauthorship networks one can count the number of jointly written papers. Especially for food webs a weighted network description seems to be inevitable. Very small fluxes of mass or energy between nodes (usually trophic species) are probably not observed in reality, or may simply be neglected if deemed irrelevant. However, the decision whether to take small fluxes into account affects the statistics of the corresponding unweigted network (22). Up to now there is much ambiguousness in food web theory. More than ten years ago the leading "students of food webs" urgently demanded a common food web standard (3).

However, up to now such a standard has not been established, but it seems to be generally accepted that food webs should contain weighted links (3; 22).

It was shown that weighted networks are substantially more informative than unweighted networks (21; 22). The latter can simply be deduced from the former (21; 22): a special *cut*-value is defined, if the weight is larger *cut* the link is set, if not it is neglected. We have shown that such a deduced unweighted network may not be distinguishable from a random (Erdős-Renyi-) network, although the original weighted network was clearly structured (21). A simple *Gedankenexperiment* shows the potential ambiguity of the characterization of weighted networks with measures developed to describe unweighted networks: the highly weighted links can show a scale-free property (2) (power-law distribution of node degrees, taking into account only strong links), but a different degree distribution may be obtained if also links with small weights are considered. This problem is of importance especially for food webs where no standard exists: it depends on the personal choice of the ecologist if weak links are also counted. For instance, lions are usually considered as top predators, but gnats also bit them, so there is a small material flux also from lions to gnats which is usually not considered in corresponding food webs.

In either case, it is important to develop a deeper understanding of weighted networks. Appropriate statistical measures to characterise such networks are therefore needed. Theoretical ecologists developed different information theoretic measures for the description of directed and weighted food webs (11; 12; 16; 20; 21; 22), which can serve as a starting point for a corresponding general network theory. Recently, we proposed *Medium Articulation* as the first complexity measure for networks (20; 21).

In the first part of this paper we shortly review different appropriate measures to characterize weighted and unweighted, directed and undirected networks in a unifying manner. In the second part we show that an analysis of these measures allows to classify a given network into one of the four classes: pure democracy networks, complex democracy networks, complex dictatorship networks, and pure dictatorship networks. In the last part we discuss the corresponding class-membership of different real networks.

2 Information theoretic measures for the characterisation of networks

In the following T_{ij} exclusively denotes the normalized link from i to j: $T_{ij} = t_{ij}/T$ with the total sum of links $T = \sum_i \sum_j t_{ij}$ (t_{ij} is the non-normalized value). The most important measures we need are the *joint entropy* H, the *redundancy* of the network R and the *mutual information* I which are defined as follows (cf.(11)):

$$H = -\sum_i \sum_j T_{ij} \log T_{ij}, \quad (1)$$

$$R = -\sum_i \sum_j T_{ij} \log \frac{T_{ij}^2}{\sum_k T_{kj} \sum_k T_{ik}}, \quad (2)$$

$$I = H - R = \sum_i \sum_j T_{ij} \log \frac{T_{ij}}{\sum_k T_{kj} \sum_k T_{ik}} \quad (3)$$

From (3) one directly recognizes $I = I_{min} = 0$ if $T_{ij} = \sum_k T_{kj} \sum_k T_{ik}\ \forall t_{i,j} \neq 0$. Such a network is shown in Fig.1a. Furthermore: $H = H_{max} = 2 \log n$ if $T_{ij} = 1/n^2\ \forall i,j$ (Fig.1a); $H = H_{min} = 0$ if $T_{ij} = 1$ for any i,j and the remaining links equal zero (eq. (1)); $R = R_{max} = 2 \log n$ if $T_{ij} = 1/n^2\ \forall i,j$ (Fig.1a) and $R = R_{min} = 0$ if $T_{ij}^2 = \sum_k T_{kj} \sum_k T_{ik}\ \forall T_{ij} \neq 0$ (eq. (2)) (Fig.1c). $I = I_{max} = \log n$ if H is as large as possible ($H_{I_{max}} = \log n$) under the condition $R = R_{min} = 0$ (Fig.1c). Note that all networks in the extreme cases $H = H_{max} = R_{max}$ and $I = I_{max}$ belong to the class of Kirchhoff-networks where $\sum_i T_{ij} = \sum_i T_{ji}$ $\forall j$. Summarizing, highly connected networks are characterized by high H and high R, but low I-values. Sparsely connected networks, i.e. highly "articulated" ones, have low H- and R-values, but a high mutual information I.

Recently, we introduced *Medium Articulation* as the first measure for the complexity of networks(20):

$$MA = I \cdot R. \quad (4)$$

MA is a typical complexity measure in the sense that it is zero in the extreme cases (here: if either $I = 0$ or $R = 0$), but maximum in between (8; 15). Thus, $MA = 0$ for the networks given in Fig.1a,c. We have shown previously (20) that $MA = MA_{max} = (\log n)^2/2$ for the network in Fig.1b (link weights should all be equal in the extreme cases).

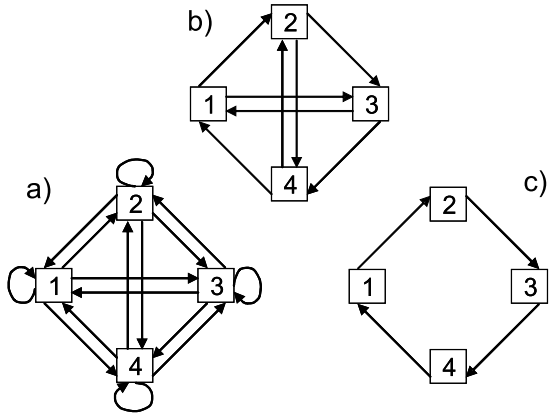


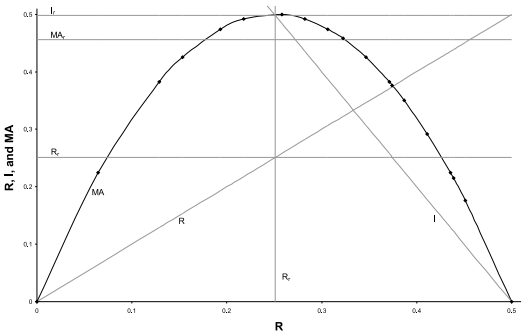
Figure 1: Three different 4-node-networks. a) maximally connected, $H = H_{max}$, $R = R_{max}$, $I = 0$, $MA = 0$; b) moderately connected, i.e. moderately articulated, $R = R_{max}/2$, $I = I_{max}/2$, $MA = MA_{max}$; and c) minimally connected, i.e. maximally articulated, $R = 0$, $I = I_{max}$, $MA = 0$.

3 Complex and non-complex democracy and dictatorship networks

Here we show that the measures described above can be used to classify all networks into one of four different classes. For that purpose, the redundancy R and the medium articulation MA of a given network are compared with the mean R_r and MA_r of correspondingly randomized networks (edges are randomly rewired). Note that the joint entropy H of a given network does not depend on the network's topology, but only on the number and weights of the edges. It follows $H = H_r = R + I = R_r + I_r$. If $R < R_r$ it follows $I > I_r$ and vice versa, thus I does not carry any additional information. Fig.1 shows that $R = 0$ for minimally connected networks with a ring structure. A corresponding analysis shows that minimally connected networks can also have a vanishing mutual information (i.e. maximum redundancy for the given edges), namely star-shaped networks: $I = 0$ if all links are going out from one single node or if all links are pointing to it. Obviously, ring-networks with $R = 0$ have a lower redundancy than their random counterparts $R < R_r$. Because in such networks the nodes play equal roles we call them *democracy networks*. Dictatorship networks, in contrast, have $R > R_r$ (i.e. $I < I_r$). The network complexity measure medium articulation serves to subdivide the two major groups: complex networks with $MA > MA_r$ lie between pure democracy and pure dictatorship networks. Summarizing, pure democracy networks have $R < R_r, MA < MA_r$, complex democracy networks $R < R_r, MA > MA_r$, complex dictatorship networks $R > R_r, MA > MA_r$, and pure dictatorship networks $R > R_r, MA < MA_r$.

In the following part of this section we exclusively deal with directed unweighted networks, i.e. $T_{ij} = 1/L\,(\forall\, T_{ij} > 0)$, where L denotes the number of links (directed edges). Each directed unweighted network with L links has a joint entropy $H(n, L) = log(L)$, for any number of nodes.

a



b

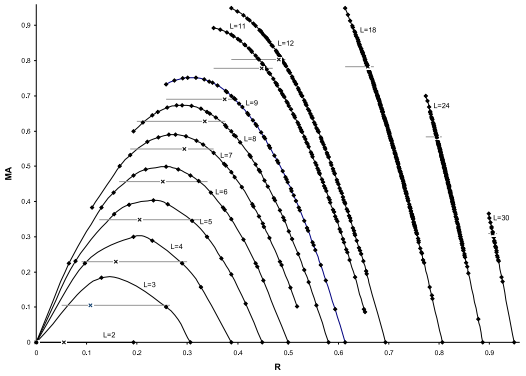


Figure 2: Classification of all directed unweighted 6-node-networks (normalized R, R_r, I, I_r, MA, MA_r). All networks above the horizontal MA_r line are complex networks (below are non-complex (pure) networks), all networks left of R_r are democracy networks (right are dictatorship networks). a) all networks with $L = 6$ edges b) all networks with $L = 2, 3, \ldots$ edges. x indicates the corresponding exact arithmetic mean values (R_r, MA_r) of the randomized networks.

Fig.2a shows the redundancy R, mutual information I, and medium articulation MA for all networks with $n = 6$ nodes and $L = 6$ directed unweighted edges. It can be seen that all four different network types can be found within the class of directed $n = 6, L = 6$ networks. Fig.2b shows for some selected L the corresponding R and MA for all $n = 6$ networks, as well as the corresponding random network values R_r and MA_r. It can be seen that complex dictatorship networks only exist for small L. Analysis shows that for a given n there are many more complex democracy, than complex dictatorship networks. Pure democracy networks only exist for even smaller L. In other words, most democracy networks are complex, whereas most dictatorship networks are non-complex.

Our classification scheme also allows to extract a special information about a given network: if $R < R_r$ (democracy), the information (or mass, or energy) tends to cycle in the network, if $R > R_r$ (dictatorship) there is a tendency for straight information through-flow from sources to sinks. Thus, democracy networks are cycling networks, and dictatorship networks could also be named source-sink networks.

4 Classification of real networks

In contrast to the well-known network classifications "small-world"(18) and "scale-free"(2) our classification scheme is of maximum generality. It is applicable to all four network types: directed and undirected, and weighted and unweighted networks. Table 1 shows the classification for some real networks of each of these types. The analysed food webs are always dictatorship networks. This seems plausible, because of the underlying trophic hierarchy. It is well-known that predators are mostly controlling different prey one trophic level below themselves. Four of the five directed unweighted food webs are non-complex, whereas 11 of 12 directed weighted food webs are complex. For a first corresponding comparison we have taken the largest weighted network ($n = 66$) as unweighted (i.e. all fluxes above the $cut = 0$ are 1 ($T_{ij} = 1/66$), the others are 0) and obtained also a pure dictatorship network ($R = 0.68 > R_r = 0.60, MA = 0.65 < MA_r = 0.94$). With other cut-values we again obtained complex dictatorship networks ($cut = 1/10000, 1/100, 1/10 \cdot t_{ij,max}$). In future we will study the dependency of the classification on cut-values more in detail.

In the analysed two neural networks the nodes are neurons and the weights correspond to the synaptic strength between the neurons (male adult worm (jsh), hermaphrodite worm (n2u)). Interestingly, both networks are pure dictatorship networks, that means there is a tendency to straight information throughflow from sources to sinks. This feature is even more pronounced in the neural network of the male adult worm.

In the undirected weighted railway network of the German federal state Brandenburg nodes are stations and weights correspond to spatial distances. It is a pure democracy network. The cycling property can easily be understood, because the whole railway network has the form of a cycle: it is circled around Berlin (Fig.3). It will be interesting to compare this result to other transportation networks.

Figure 3: The Brandenburg railway network.

The analysed protein-protein interaction network is of the pure democracy type, which indicates that, on average, proteins play similar roles in the corresponding networks and the information is cycling. It seems possible to extract biologically important information from a classification analysis of different protein networks. A lower complexity could, for instance, indicate perturbation or disease, but future studies are needed.

5 Conclusion

Using information theoretic measures to characterize networks, we have introduced the four network classes pure democracy networks, complex democracy networks, complex dictatorship networks, and pure dictatorship networks. This general classification scheme holds for all types of networks, weighted and unweighted, as well as directed and undirected

Table 1: Classification of real networks. Data from: the directed unweighted networks (www.cosin.org/extra/data/foodwebs/web.html), the directed weighted food webs (www.cbl.umces.edu/ ulan/ntwk/network.html), the directed weighted neural networks (elegans.swmed.edu/parts/neurodata.txt), the Brandenburg railway network (www.bahnstrecken.de/strecken.htm), and the *E.coli* protein-protein interaction network (www.cosin.org/extra/data/proteins).

Networks	n	L	R	R_r	MA	MA_r
directed and unweighted (food webs):						
Grassland	88	137	0.232	0.18	0.589	0.532
Little Rock Lake	183	2494	0.613	0.508	0.675	0.986
Silwood Park	154	370	0.396	0.219	0.605	0.645
St. Martin Island	45	224	0.524	0.446	0.784	0.945
Ythan Estuary	135	601	0.452	0.328	0.724	0.851
directed and weighted: (food webs)						
fw1	21	82	0.11	0.037	0.09	0.044
fw2	21	61	0.124	0.036	0.097	0.047
fw3	36	122	0.186	0.104	0.319	0.243
fw4	36	172	0.295	0.084	0.121	0.173
fw5	21	55	0.303	0.249	0.598	0.595
fw6	66	791	0.159	0.022	0.132	0.041
fw7	43	348	0.233	0.057	0.107	0.104
fw8	32	158	0.263	0.099	0.247	0.22
fw9	51	270	0.269	0.133	0.365	0.323
fw10	34	158	0.259	0.164	0.451	0.407
fw11	34	149	0.245	0.139	0.385	0.333
fw12	34	115	0.284	0.177	0.462	0.438
(neural networks)						
C. elegans (jsh)	190	4336	0.451	0.399	0.868	0.933
C. elegans (n2u)	202	3963	0.446	0.403	0.883	0.936
undirected and weighted:						
Railway network	213	332	0.108	0.115	0.339	0.352
undirected and unweighted:						
E. coli prot-prot interaction	270	1432	0.403	0.43	0.793	0.966

networks. Our first analyses show that special types of real networks belong to special classes. Unweighted food webs, for instance, are pure dictatorship networks, whereas the analysed weighted food webs are complex dictatorship networks.

R. Albert, A.-L. Barabasi, Statistical mechanics of complex networks. Rev. Mod. Phys. 74,47-97, 2002.

A.-L. Barabasi, R. Albert, Emergence of scaling in random networks. Science 286,509-512, 1999.

J.E. Cohen, et al., Improving food webs. Ecology 74, 252-258, 1993.

R. Cohen, K. Erez, D. Ben-Avraham, S. Havlin, Resilience of the internet to random breakdowns Phys.Rev.Lett. 85,4626-4628, 2000.

H. Jeong, S.P. Mason, A.-L. Barabasi, Z.N. Oltvai, Lethality and centrality in protein networks. Nature 411,41-42, 2001.

A.E. Krause, K.A. Frank, D.M. Mason, R.E. Ulanowicz, W.W. Taylor, Compartments revealed in food-web structure. Nature 426,282-285, 2003.

V. Latora, M. Marchiori, Is the Boston subway a small-world network? Physica A 314,109-113, 2002.

R. Lopez-Ruiz, H.L. Mancini, X. Calbet, A statistical measure of complexity. Phys.Lett.A 209,321-326, 1995.

M.E.J. Newman, The structure of scientific collaboration networks. Proc.Natl.Acad.Sci.USA 98,404-409, 2001.

M.E.J. Newman, D.J. Watts, S.H. Strogatz, Random graph models of social networks. Proc.Natl.Acad.Sci.USA 99,2566-2572, 2002.

C. Pahl-Wostl, The Dynamic Nature of Ecosystems. Wiley, New York, 1995.

R.W. Rutledge, B.L. Basore, R.J. Mulholland, Ecological stability: an information theory viewpoint. J.Theor.Biol. 57,355-371, 1976.

S. Schuster, D.A. Fell, T. Dandekar, A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. Nature Biotech. 18,326-332, 2000.

S. Shen-Orr, R. Milo, S. Mangan, U. Alon, Network motifs in the transkriptional regulation network of E. coli. Nature Genetics 31,64-68, 2002.

J.S. Shiner, M. Davison, P.T. Landsberg, Simple measure for complexity. Phys.Rev.E 59,1459-1464, 1999.

R.E. Ulanowicz, Ecology, the Ascendent Perspective. Columbia Univ. Press, 1997.

D.J. Watts, Small Worlds. Princeton Univ. Press, Princeton, NJ, 1999.

D.J. Watts, S.H. Strogatz, Collective dynamics of 'small-world' networks. Nature 393,440-442, 1998.

D.J. Watts, P.S. Dodds, M.E.J. Newman, Identity and Search in Social Networks. Science 296,1302-1305, 2002.

T. Wilhelm, R. Brüggemann, Information theoretic measures for the maturity of ecosystems. in: Integrative Systems Approaches to Natural and Social Sciences - Systems Science 2000, M. Matthies, H. Malchow and J. Kriz (eds.) (Springer, Berlin), 2001, pp. 263-273.

T. Wilhelm, An elementary dynamic model for non-binary food-webs. Ecol.Model. 168,145-152, 2003.

T. Wilhelm, A study of equal node models for food webs. Ecol.Model. 181,215-227, 2005.

T. Wilhelm, P. Hänggi, Power-law distributions resulting from finite resources. Physica A, 329(3-4),499-508, 2003.

T. Wilhelm, H.-P. Nasheuer, S. Huang, Physical and functional modularity of the protein network in yeast. Mol.Cell.Prot. 2.5,292-298, 2003.

Network Analyses to Understand the Structure of Wikipedia

Takeshi Yamada* Kazumi Saito* Kazuhiro Kazama
*NTT Communication Science Laboratories
Kyoto, JAPAN
yamada,saito @cslab.kecl.ntt.co.jp
NTT Network Innovation Laboratories
Tokyo, JAPAN
kazama@ingrid.org

Abstract

We investigated the structure of Wikipedia, the well known Web-based encyclopedia, as a very large-scale hyperlink network by using network analysis methods. We analyzed the time evolution of the number of articles and links, applied core-extraction methods to identify highly interconnected subnetworks, and compared different centrality measures to understand its characteristics.

1 Introduction

According to its website[1], Wikipedia is a multilingual Web-based, free-content encyclopedia. It is written collaboratively by volunteers, allowing articles to be changed by anyone with access to a Web browser. The project began on January 15, 2001 as a complement to the expert-written Nupedia, and is now operated by the non-profit Wikimedia Foundation. The English-language version of Wikipedia has more than 880,000 articles as of November 2005.

It is an interesting challenge to apply network analyses to such a large-scale network like Wikipedia to unveil its basic structure and extract useful information. There are a number of reasons why the Wikipedia network is so attractive. First, the size of the network is enormous: there are more than 880,000 articles and more than 27,000,000 links as of November 2005. Sophisticated but time-consuming methods, such as clique-based community extraction approaches, are not applicable: even calculating the graph diameter of the entire network takes a lot of time. We, therefore, need to invent fast and efficient algorithms for existing methods or develop alternative approximation methods. Second, it is a rapidly growing network. The revision records show that there were only about 1,000 articles and 9,000 links in April 2001. We can study growth patterns by using the network as a typical example of asynchronously updated and collaboratively built network. Third, it is a relatively well-organized and comprehensible network. The semantic information inherent in the Wikipedia network as an encyclopedia, such as article titles, article contents, and category hierarchies can be used to verify and evaluate the correctness and usefulness of a method for network analysis.

[1]http://en.wikipedia.org/wiki/Wikipedia

This paper is organized as follows. In Section 2, we analyze the growth pattern of the Wikipedia network, and show that it follows the densification power law. In Section 3, we apply community extraction methods to an undirected network generated from Wikipedia and investigate its community structure. In Section 4, we calculate and compare centrality measures for the directed Wikipedia category network and discuss their different features. Section 5 concludes the paper.

2 Time evolution

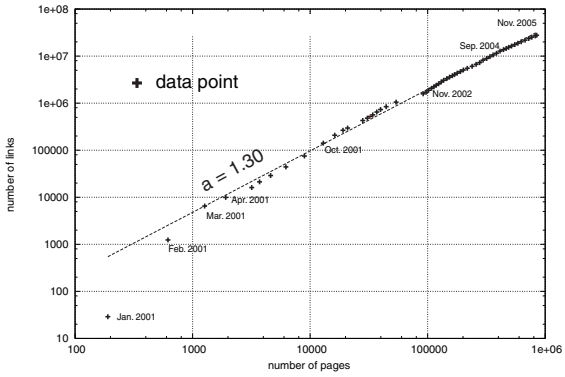


Figure 1: Number of articles versus number of links

The articles in Wikipedia are divided into several namespaces including the main namespace, consisting of all regular articles, and the category namespace, consisting of articles that index regular articles. We first focus on the main namespace. Each Wikipedia article has number of references to other articles, expressed as hyperlinks. We treat a reference

from an article to another as a link.

Figure 1 shows the time evolution of the number of articles (x-axis) and the number of links (y-axis) on a logarithmic scale. Each point in the graph corresponds to a monthly snapshot of the Wikipedia network in the period from January 2001 to November 2005. This graph illustrates the power law relationship (R^2=0 98), and is called a densification power law plot (Leskovec et al., 2005). The fact that the slope of the plot is a=1 30 > 1 indicates that the network is becoming denser over time, with the number of links growing super-linearly with the number of articles. The value is lower than that of the citation network reported in (Leskovec et al., 2005), but still clearly exhibits super-linear growth. It is clear, then, how rapidly the Wikipedia network has been growing and densifiying the relationship between articles in a scale-free fashion.

3 Community extractions

In this section, we focus on an undirected network by considering that two articles are linked if both articles refer to each other, and taking the maximal connected component, resulting in an undirected network with 536,724 nodes and 1,337,902 links This undirected network, as well as the directed one discussed in the next section, is obtained from the Wikipedia snapshot created on 24th September 2005.

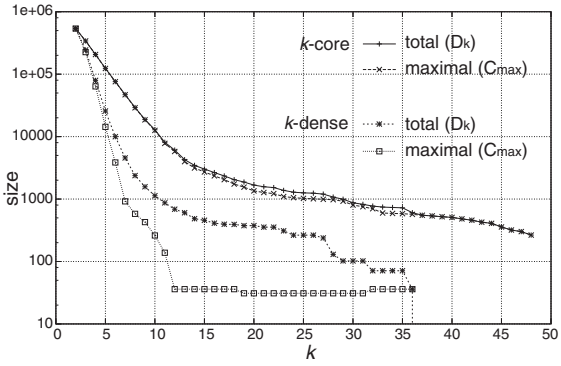


Figure 2: Maximal and total community sizes for k-core and k-dens methods for each k

To understand the structural and functional properties of a large-scale network, it is crucial to identify subnetworks (communities) in which the nodes are more highly interconnected than to the rest of the network. There are several such community extraction methods published in the literature. For example, Palla et al. (2005) proposed a method called "Clique Finder" based on the notion of k-clique (complete subgraphs of size k) [2]. They define a k-clique community as a union of all k-cliques that can be reached from each other through a series of adjacent k-clique (where two k-cliques are called adjacent when they share $k-1$ nodes). Unfortunately, their method is not suitable for a large-scale network, because finding all k-cliques is NP-hard and algorithmically intractable.

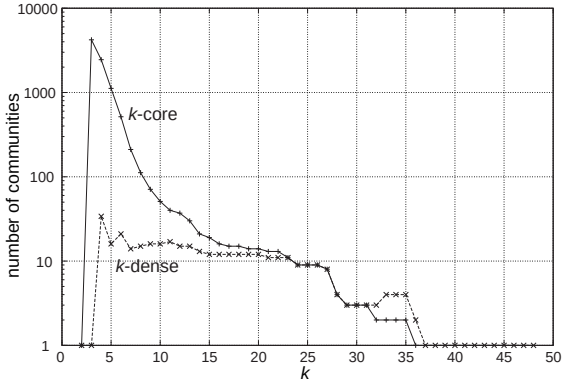


Figure 3: Number of communities obtained by the k-core and k-dens method for each k

Another well known method is called k-core community extraction or k-core decomposition. The notion of k-core was first introduced by Seidman (1983). A k-core community is defined as a maximal subgraph in which each node is adjacent to at least $k-1$ nodes in the subgraph.

Saito et al. (2006) proposed the k-dense community method that extends the concept of k-core and approximates the k-clique method. The k-dense community is defined as a maximal subgraph in which each two-clique (i.e., pair of adjacent nodes that are connected by a link) has at least $k-2$ adjacent nodes in common that connect to both of the nodes in the clique, in the subgraph. The k-dense method is more computationally efficient than the k-clique method and as simple as the k-core method. It is obvious that a k-clique is included in a k-dense component, which is included in a k-core component.

We applied k-core and k-dense extraction methods to the Wikipedia network for all possibly k values. For each k, the network is divided and pruned, and a set of N communities $\{C_i^k\}_{1\leq i\leq N}$ is obtained. Let C_{max}^k be the largest community in $\{C_i^k\}$ and D^k be the total number of nodes (articles) in $\{C_i^k\}$. Figure 2 illustrates how D^k and C_{max}^k change as k changes, for both k-core and k-dense methods. Figure 3 shows the number of communities for each k.

[2] We use the term k-clique as a clique consisting of k nodes.

From these figures, we can see that the k-core results consist of one dominant community and possibly other much smaller ones. The dominant community is either too large as a single community (for small ks) or is the only community extracted (for large ks), which makes the extraction results not sufficiently informative. On the other hand, the k-dense results consist of an appropriate number of smaller communities which are comparable in sizes for appropriate choices of ks and thus they are more informative.

For example, the k-dense method for $k = 12$ extracts total 37 communities. They include a community of baseball events, airplanes, airports, authors, and German cities. Twenty-two communities, including these five communities, belong to one same community if the k-core method for $k = 12$ is used.

Figure 4 shows a two-dimensional layout of the articles, calculated by using the spring method (Kamada and Kawai, 1989) (here, links are not shown). Because it is hard to plot all the 536,724 pages at once, articles not included even in low-degree communities are omitted from the spring-model calculation for computational efficiency. Thus, 9,773 pages are plotted in total including gray dots corresponding to articles in low-degree (6-dense) communities, and black dots (amount to 687 pages) in high-degree (12-dense) communities. The graph distance between articles are calculated from the original network. We can observe that the high-degree communities exist across the base network as clusters and form a characteristic structure.

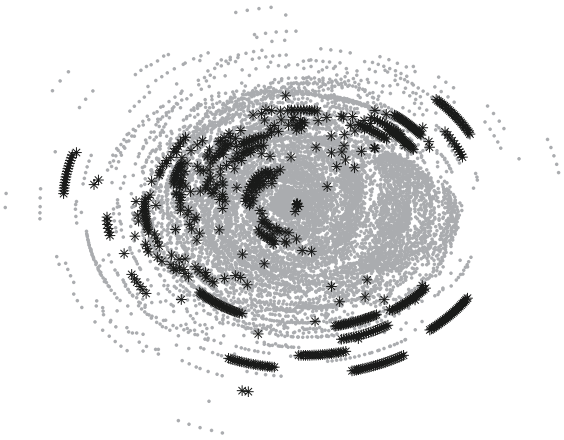


Figure 4: The Wikipedia network structure

4 Centralities

Each regular article in Wikipedia belongs to one or more categories, and each category has its own article (in the category namespace) to index corresponding regular articles. They form category hierarchies, but ones that are loose and loopy. In this section, we consider a directed network with 71,993 articles and 117,426 links in which a directed link corresponds to a reference, with its direction from a sub-category to a super-category.

We applied well known centrality measures such as PageRank (Brin and Page, 1998) and HITS (Kleinberg, 1999) to the Wikipedia category network. Because of the inherent nature of the category structure, we would expect that some fundamental concepts in the human knowledge should appear at the top of the rankings. Table 4 shows a portion of the category article rankings. PageRank ranking and HITS authority ranking, as well as the number of links directed to the article (in-link degree) and the number of articles that are reachable by iteratively following the in-links backward, are respectively shown in the columns labeled PRNK, HITS, INDEG and NRCH.

The top five rows correspond to the top five articles ranked by PageRank. As expected, very fundamental concepts such as *Categories*, *Fundamental*, *Humans*, and *Cultures* as well as *Wikiportals*, which is Wikipedia specific, are listed. This suggests that PageRank most effectively reveals the basic Wikipedia category structure. PageRank importance is determined by "votes" in the form of links from other articles, and the importance of a vote from any source should be tempered (or scaled) by the number of articles for which the source is voting (Langville and Meyer, 2005). In the Wikipedia category network, the votes from the articles located in the lower level of the category hierarchy are accumulated to the articles in a higher level.

If we look for a ranking measure that gives those fundamental concepts high rankings, the number of reachable articles given in the NRCH column (the NRCH value) would be a straightforward alternative. Articles corresponding to fundamental concepts located near the top of the category hierarchy must have a large NRCH value. This is because iteratively following in-links backward corresponds to moving all the way down the hierarchy. In fact, the top five articles have all large NRCH values. However, it should be noted that the converse is not always true. For example, the article titled *Albert Einstein* has a large NRCH value but a lower PageRank ranking.

This gap can be attributed by the distribution of the number of steps to reach the rest of the articles. Figure 5 illustrates the relationship between the number

Table 1: Category article rankings

No.	Title	PRNK	HITS	INDG	NRCH
1	Categories	1	11662	5	62018
2	Fundamental	2	6586	9	62015
3	Humans	3	5468	20	61966
4	Culture	4	1749	62	61966
5	Wikiportals	5	5472	8	61966
6	Geography	13	1970	48	41991
7	Information	744	3806	4	61968
8	Albert Einstein	1834	12735	1	61967
9	Albums by artist	76	1	1387	1436
10	American albums	885	2	373	158
11	Canadian albums	1276	3	313	99
12	Alternative rock albums	923	4	255	135
13	British albums	1991	5	235	55

of steps to reach other articles (x-axis) and the number of articles at those steps' reach (y-axis). The number in parentheses for each article title is the PageRank ranking. The article *Categories* has a more concentrated and steeper distribution than *Albert Einstein*, which would explain the fact that the former has a higher PageRank ranking than the latter even though their NRCH values are almost the same. Likewise, the PageRank ranking of *Geography* is higher than *Information* and *Albert Einstein*, although the NRCH value for *Geography* is smaller. The fact that *Geography* has a more concentrated and steeper distribution than *Information* and *Albert Einstein* would explain this.

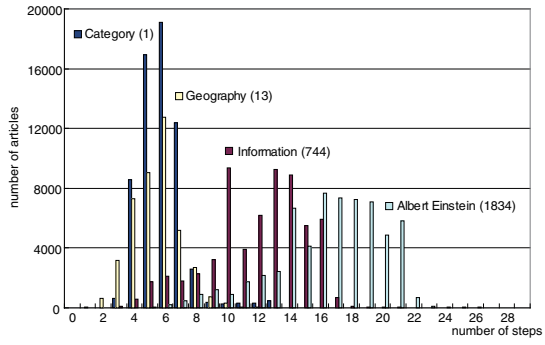


Figure 5: Distributions of the number of steps to reach other articles

The last five rows correspond to the top five articles ranked by HITS authority rankings. The set of articles with a high HITS ranking is quite different from that of a high PageRank ranking, and the fundamental concepts do not appear at all in the former. We know that HITS should be applied to a small network retrieved by a query and not to the entire network, but it is still worth investigating. In short, HITS rankings are affected strongly by in-link degrees: the top article *Albums by artist* has the highest in-link degree. In fact, the top fifty articles are all album-related articles that have strong connections with the top article. The top fifty hub articles are also all album-related ones.

5 Conclusions

We have investigated the structural properties of Wikipedia. The network grows by following the densification power law. The k-dense method gives us the most informative view of the community structure. In fact, we can observe that the high-degree communities exist across the base network as clusters and form a characteristic structure. PageRank most effectively reveals the basic category structure.

References

S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. In *7th int. conf. on WWW*, pages 107–117. Elsevier, 1998.

T. Kamada and S. Kawai. An algorithm for drawing general undirected graph. *Information Processing Letters*, 32:7–15, 1989.

J. Kleinberg. Authoritative sources in a hyperlinked environment. *J. ACM*, 46(5):604–632, 1999.

A. Langville and C. Meyer. A survey of eigenvector methods for Web information retrieval. *SIAM Review*, 47(1):135–161, 2005.

J. Leskovec, J. Kleinberg, and C. Faloutsos. Graphs over time: Densification laws, shrinking diameters and possible explanations. In *KDD '05*, pages 177–187. ACM Press, 2005.

G. Palla, I. Derenyi, I. Farkas, and T. Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(9):814–818, 2005.

K. Saito, T. Yamada, and K. Kazama. Extracting network communities based on the k-dense method. Technical report, NTT Communication Science Laboratories, 2006.

S.B. Seidman. Network structure and minimum degree. *Social Networks*, 5:269–287, 1983.