

Empirical Network Structure of Malicious Programs

John Musgrave

*Department of Computer Science
University of Cincinnati
Cincinnati, OH, USA*

musgrajw@mail.uc.edu

Alina Campan

*Department of Computer Science
Northern Kentucky University
Highland Heights, KY, USA*

campana1@nku.edu

Temesguen Messay-Kebede

*Air Force Research Lab
Wright-Patterson Air Force Base
Dayton, OH, USA*

temesgen.kebede.1@us.af.mil

David Kapp

*Air Force Research Lab
Wright-Patterson Air Force Base
Dayton, OH, USA*

david.kapp@us.af.mil

Boyang Wang

*Department of Computer Science
University of Cincinnati
Cincinnati, OH, USA*

boyang.wang@uc.edu

Corresponding Author: John Musgrave

Copyright © 2024 John Musgrave, et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

A modern binary executable is a composition of various types of networks. Control flow graphs are a commonly used representation of an executable program used for classification tasks. Control flow and term frequency representations are widely adopted, but provide only a partial view of program semantics and present challenges to increases in resolution. By performing a quantitative analysis of program networks, we enable the identification of patterns within these features that are correlated to structure. This allows for increases in feature resolution and pattern recognition in classification tasks. These are necessary steps in order to obtain greater explainability in classification results. We demonstrate the presence of Scale-Free properties of network structure for program data dependency and control flow graphs, and show that data dependency graphs also have Small-World structural properties. We show that program data dependency graphs have a degree correlation that is structurally disassortative, and that control flow graphs have a neutral degree assortativity, indicating the use of random graphs to model the structural properties of program control flow graphs would show increased accuracy. An increase in feature resolution allows for the structural

properties of program classes to be analyzed for patterns as well as their component parts. By providing an increase in feature resolution within labeled datasets of executable programs we provide a quantitative basis to interpret the results of classifiers trained on CFG graph features. By capturing a complete picture of program networks we can enable future work in mapping a program's operational semantics to its structure.

Keywords: Malware analysis, Network Science.

1. INTRODUCTION

In this study we propose a quantitative analysis of program networks. Network features are inherently structured, and the networks present in binary executables are representations of the program's operational structure. When used for supervised learning, network features can yield accurate classification in machine learning models. Increases in resolution at the feature level can provide additional insights into interpreting the results of these models. We propose a fine grained analysis of program networks, which we detail in Section 2.

A fine grained analysis of network properties can lead to an increase in feature resolution, the identification of patterns present in the data, and a correlation between structure and operational semantics. Our study provides empirical measurements of network properties, and these are representative of program structure. Previous studies have successfully trained classifiers using Graph Neural Networks and Graph Convolutional Neural Networks. In some cases these studies have extracted program networks, but the identification of the structural properties of the networks was not the primary focus. The significance of network features is discussed further in section 1.2 Current Work and Motivations.

The use of machine learning methods in malware analysis propose to represent features of operational semantics through models and classification accuracy. The results and accuracy of classification techniques in machine learning are dependent on the feature representations used in these datasets. Features are extracted from datasets of malicious programs are collected at specific levels in an architectural hierarchy. Many useful features for classification can be extracted at multiple points in the architectural hierarchy as discussed in previous studies, e.g. instruction n-grams, sequences and patterns of bytecode or hex representations, as well as graphs, n-grams, and sequences of system API calls. The feature representation selected greatly determines the available granularity, resolution, and the degree to which classification accuracy is correlated with and representative of semantics. However, further analysis and increases in feature resolution are presented with several obstacles for an accurate classification of programs correlated with operational semantics. Class labels are often coarse grained, with one label representing the class of an entire program, without a clear method to provide increased resolution for supervised models which are dependent upon labeled data. Without an increased resolution of features that are descriptive of structure, explaining correlation between structure and semantic abstraction is very challenging. The degree to which a program's component parts contribute to a class cannot be determined without increased feature resolution in a labeled dataset. A program level of resolution is too low to provide meaningful information about the relationship between a class label and a program's operational semantics across abstraction layers. Therefore we view feature representations from two perspectives: as a description of program operational semantics, and the ability to describe structural properties. Struc-

tural properties enable the syntactic elements of a program to be interpreted. Semantic properties allow the correctness of a program to be verified across abstraction layers. The relationship between syntax and semantics in natural language has been successfully modeled by using topics in bi-partite networks [1–3].

Many previous studies using machine learning methods to classify malware focus on finding errors in high level languages. While this is useful for increased security in the automation of software development, it does not address the semantic interpretability of the classification results. Classification does not directly provide a fine grained description of the malicious program structure, and is dependent on the granularity of feature representation used for training. Features in labeled datasets are often at the most coarse grained level of the binary as a whole. Studies based on malware graph features often have a focus on methods of differentiation in control flow graphs through graph isomorphism. While this is a useful feature for the classification of malicious programs, it is only a partial view of the program’s operational semantics. The isomorphism of graphs and all subgraphs is used to determine class equivalence, and have not been directly measured for structural properties. Further, if the classification model was trained on a high level language such as C or Java, a correlation is required to be proven to the compiled artifact. The correlation of the structural properties identified through classification to an abstract semantic representation is an open question. We attempt to analyze structural properties of program networks that correlate directly to operational semantics [4, 5].

1.1 Related Work

Machine learning techniques have been applied in many contexts to successfully identify malicious programs based on a variety of features. Many classification methods have been used for supervised learning including deep neural networks and support vector machines. Several datasets have been collected with various kinds of features, including assembly instructions, n-gram sequences of instructions and system calls, and program metadata [2, 6–9].

A number of studies have explored the use of static features at the level of file format, and their impact on the classification of malicious programs. This shows that structural features can be correlated with program behavior such that a suitable classifier can be trained. Decision trees for the classification of Windows PE files have shown to be effective in classifying malicious programs. Subsequent studies have focused on malware classification using ensemble methods, which include random forest with support vector machines and principal component analysis that focused on features extracted from file headers in Trojan malware [10–12].

The focus of many studies applying machine learning techniques to malware analysis is the task of classification for the purposes of identifying unknown programming errors. Zhou et al. used a graph neural network (GNN) model to classify various types of C functions in order to determine semantic errors in their abstract syntax tree (AST) and program flow. The model was trained on a dataset of functions which were drawn from several executable binaries including the Linux kernel. Wang et al. developed a synthetic dataset of 3 million Python programs with class labels, and extracted function call graphs from AST graphs generated from tokenization to be used for training with a novel GNN design. Park et al. have used sequence modeling to identify potential optimizations in

programs based on an intermediate representation. A program flow graph was extracted from this intermediate program representation and used in the sequence predictions [13–15].

Several studies have focused on function abstraction semantics through decompilation. LeDoux et al. represented a program as a graph of function abstractions obtained from reverse engineering and used semantic hashing as a measurement of similarity. However, this study did not take a bottom up approach, and basic block features were specifically not considered. There may be many equivalent programs for a given malware binary, and whether semantic function abstractions in a high level language are correlated to lower level binary representations is an open question. In a similar manner, Alrabaee et al. have used a *tf-idf* representation with Hidden Markov Models and graph kernels to obtain a graph of semantic function abstractions for a program. This was accomplished by constructing a Bayesian network for each of the features collected [16, 17].

Several studies have used control flow graphs as features in datasets used for malware classification tasks. Bruschi et al. have extracted control flow graphs from malware for the purposes of classification through comparing the graphs for isomorphism. Cesare et al. have presented several studies on the uses of control flow graphs in the classification of malware with efficient results [4, 18–20].

The use of control flow graphs as features for classification provides a model correlated to structural properties that is descriptive of semantics. Increasing the classifier accuracy requires increasing the resolution of features. Available methods to increase the resolution of these specific features have not been studied previously. Additionally, reasoning to larger classes of functionality are ambiguous due to context, and a lack of available resolution. A fine grained study of the structural properties of networks may show patterns in the features collected.

1.2 Current Work and Motivations

The utilization of methods of analysis and proof of program semantics require the use of a program specification to be checked for validity. Malicious programs in an adversarial environment do not have a specification available for verification prior to execution. Without the presence of a formal specification, proofs and verification for the operation of a program cannot easily be developed. Therefore, a semantic representation corresponding to a formal specification or other description of operational semantics must be constructed from structural elements to verify the semantic correctness. This method of construction represents a bottom up approach. The degree to which an abstract representation correlates to a program's semantics is an open question.

In order to explain how structural elements are correlated to their abstract semantics, the structural properties must first be formally defined. Without the empirical observation of structural properties, the patterns in structure cannot be identified, and therefore the semantic abstraction generating structural patterns cannot be derived. A definition of semantics must be specified, and in this context we refer to the operation of a program. Program semantics that are descriptive of the behavior of a program must also persist across architectural layers. While other semantic representations exist at various architectural levels, a structured feature representation and analysis of instructions composed of opcodes and their operands does not require proving a correlation between architectural levels. The instructions describe the operation of the machine at the most fundamental level. In the case of malware, a malicious binary is the initial artifact. While we are unable to make

assumptions about the values operands will have at runtime, we are able to derive the structure of their dependency. So when we discuss structure in a program, we are describing a structure of dependencies. This is explicitly specified in the program, and we cannot determine the values of the terms at runtime. By describing the structure of dependency we obtain a representation that is descriptive of a program's structural properties. By representing program structure at the lowest level of abstraction we directly describe the structural properties of program operation in terms of dependency. So the ability to recognize structural patterns that are tied to semantics would have wide applications.

When an executable program is viewed in automata theoretic terms as the operation of a Turing Machine, then the state of the Finite State Machine is subject to analysis. The potential state space of a program is not computationally feasible to analyze, and presents challenges for searching within this space. However, the set of state transitions and structure of their dependencies within the same program contain the program's structure and are subject to further analysis to gain insight. This is the structure that provides insight into the sequences present in the program. Network analysis methods provide meaningful features for this analysis. These features are needed for accurate classification tasks based on program behavior [19, 21].

The instruction sequence in the program is also not determined by the linear placement of the term in the document, but by the structure of the program's control flow network. This sequence is also segmented into blocks of potentially deterministic sequences of instructions [21].

Since programs are made up of sequences of instructions, by performing an analysis of the frequency of terms in the sequence of instructions using $tf-idf$ based methods, the result shows an overwhelming prevalence for a high frequency of data movement instructions. This drastic distribution of term frequencies shows the highest density in the body of the distribution, with the tail being extremely thin. In order to further analyze the body of the term distribution, a feature set with more structure is required for meaningful analysis. The underlying semantic structure is not captured by the term frequency distribution of the instructions alone. Networks of data dependencies however are descriptive of the structural relationships between the terms. This network structure is directly descriptive of the terms present in the body of the term frequency distribution represented by $tf-idf$, which is positively skewed. Our study measures network properties at a segment and program level for data movement and program execution respectively [22, 23].

At this time we are unaware of a study that directly measures the structural properties of the program networks. Previous studies have successfully trained classifiers using Graph Neural Networks and Graph Convolutional Neural Networks. In some cases these studies have extracted program control flow graphs and compiled labeled datasets for models trained on graph features, but the identification of the structural properties of the networks was not the focus, so the measurements are not known.

In order to correlate networks by their degree, we compare networks based on their assortativity. As we will see, networks' structural assortativity properties for the collected data appear to hold across samples. This means that predictions can be made as to the degree correlations for various networks. The structure of data dependency graphs can be predicted to not have a prevalence of links between nodes with high degree. The structure of control flow graphs can be predicted based on the network properties of random graphs. The values of the random graph properties will likely vary by sample, and this is an area of future research.

1.3 Outline

Section 2 covers the experiments performed. Section 3 contains the Results and Discussion. Section 4 is a Summary and Conclusion.

2. EXPERIMENTS

This section describes the data collection process as well as the metrics selected for our analysis.

2.1 Data Collection

The goal of this study is to provide a quantitative basis for analysis of the structure of malicious executable programs. The executable programs for our purposes are adversarial, and are provided in binary form.

Several feature representations exist in the binary, and additional data can be collected for more structured representations. For example, a program can be viewed as a document in a $tf-idf$ representation, with terms selected from a dictionary. Terms in a dictionary correspond to assembly instruction opcodes, which are explicitly specified in the binary. However, this ignores any data operands, and focuses on the term frequency distribution as the primary representation. Also, unless further structure is considered, this assumes a linear structure to the document. Executable programs are not structured linearly, but are divided into segments, which are structured in a network. This is one motivation for a structural analysis of the networks present within a program.

The $tf-idf$ representation makes two assumptions prior to analysis, that data movement operations should be ignored, and that the distribution of term frequencies is representative of a program. However, the term distribution is heavily skewed towards the use of data movement instructions, and has a tail that is thin. The variance captured by data movement in the body is more than the variance of all other terms in the tail. So ignoring data movement operations discards a majority of the data, and this data is descriptive of the program's function. Further, we are unable to further analyze the body of the term distribution without additional quantitative information about the structure [2].

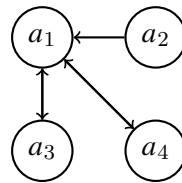
Since our use case focuses on adversarial examples, we have only selected malicious examples to be analyzed for their program structure. No specifications exist beforehand for verification for these samples. Additionally, each malware example must be able to be executed as a binary. The program samples were selected from the public malware repository *theZoo*, a collection of live malware samples. We plan to expand to other publicly available malware repositories in future studies [24].

Each sample was taken from a binary able to be executed on the target platform across several operating systems and architectures. Each binary was decompiled using GNU *objdump*, a tool which is able to reverse engineer a binary program to its assembly instruction set representation. Assembly instruction representations were collected for each program in the dataset. Our assembly artifacts were segmented into basic blocks, sequential segments of contiguous instructions separated

```

mov    ecx , rbp - 44
mov    eax , ecx
and    eax , 400
or     eax , 140
or     ecx , 1
cmp    rip + 170 , 0
cmovne ecx , eax
mov    rbp - 44 , ecx
mov    rip + 180 , 0
jmp    0x100000000

```



$$Addg = \{a_i \mid a_i \in A_{operand}\}$$

Figure 1: Basic block segment of assembly instructions and its data dependency graph. The data dependency graph shown is constructed from data movement instruction dependencies. *mov* instructions are the primary instructions with respect to term frequency.

by a jump instruction. Control flow graphs are obtained from static and dynamic analysis tools, and are represented in an adjacency matrix format. We used several tools for comparison purposes, but focus primarily on *radare2*, although some variation exists between the tools, and we do not offer a comparison of program static analysis tools in terms of their accuracy. The graphs recovered by static analysis tools are obtained by analyzing the structure of basic blocks as nodes in a program networks, and jump instructions as edges to these nodes. One control flow graph exists for each program in our dataset [25].

Since basic blocks determine the nodes in the program network, each node was analyzed for its data dependency structure. Each line in the segment is an assembly instruction composed of an opcode and a set of operands. Since these instructions are issued with a corresponding order, explicit and implicit data dependencies exist between the instructions in the sequence. Data dependency graphs were constructed by creating a network where nodes represent a data operand, and an edge between the nodes represents a *mov* instruction, or other opcode with implicit data movement between source and destination operands. This was done by using a program written to look for specific data operands of instructions in the same basic block being input, and output an adjacency list representing the data dependency graph. The graph of dependencies between operands in assembly instructions was constructed for each basic block in the program, and were represented in an adjacency list format for each graph. FIGURE 1 shows an example of a data dependency graph being constructed from a sequence of contiguous assembly instructions with

dependencies in a basic block. We constructed data dependency graphs for each block, and focused solely on dependencies between *mov* instructions, as the prevalence of data movement was the primary motivation for providing additional structure [26, 27].

The interaction of the two graphs provides a large amount of additional data that can be used for further analysis. This interaction is captured by a program dependence graph. The program dependence graph can be obtained by constructing a bi-partite graph from a tensor representation, where each cell in the tensor represents a data dependency graph, and the overall tensor structure is built from the adjacency matrix for the program's control flow graph. We have constructed a program dependence graph in a tensor representation, which we include as a note. A program dependence graph represents a composition of the networks analyzed, and does not differ in the structural properties of its components. A complete semantic analysis of the PDG composition is outside the scope of this work, but we present a structural analysis of its component parts [28].

Therefore three networks are available for analysis for each program sample, the control flow graph (CFG), the data dependency graphs (DDG) for each node, and the complete program dependence graph (PDG). The PDG represents the interaction of the CFG and DDG graphs. Each of the graphs collected are directed graphs which contain cycles, and can also be analyzed as undirected graphs [26, 28].

The quantitative network properties discussed in the results section were observed by using network libraries to measure the adjacency matrix and adjacency list representations of the networks collected. Additional analysis was performed in Matlab to obtain quantitative properties of the adjacency matrices and generate plots of the data [27, 29].

2.2 Network Metrics

Several metrics for measuring the properties of networks exist in previous work on network science and graph theory. These are used to quantify the properties of the network. We briefly introduce for background several metrics for our analysis that are commonly used in the relevant literature on networks. These metrics will provide us with a method of quantitative measurement for our analysis of the network data collected [30].

N represents the number of nodes in the network, or the size of the network.

L represents the number of edges or links in the network, directed edges in the context of a directed graph.

K represents the degree of a given node in the network, calculated by counting the number of links for a given node.

k_{max} represents the maximum degree for a node in the network.

$k_{max}/\ln(N)$ - the ratio of a network's maximum degree and the natural log of the network size in terms of nodes. This measurement is a predictor of network diameter, and also indicates cluster size in Small-World networks.

γ is a term that represents the power law exponent that the degree distribution follows, when present.

2.3 Scale-Free

Scale-Free networks have a number of interesting properties including the generative property of preferential attachment. For the purpose of this study, the most relevant feature is the presence of a degree distribution following a power law exponent that is sufficiently large to cause a hub and spoke pattern [30].

2.4 Small-World

A Small-World property of a network is characterized by a small network diameter and a high average clustering coefficient. In order to demonstrate the Small-World property, we use $k_{max}/\ln(N)$ as a measurement of network diameter over $\ln N / \ln < K >$, although we present both measurements for comparison. Since our degree distributions follow a power law distribution we use the former metric to measure the existence of the Small-World property as it is not dependent on mean degree. Both represent predictions of network diameter [31].

2.5 Degree Assortativity

Degree assortativity is the correlation of nodes with high degree. This is representative of the existence of a trend for nodes with high degree to have links between them, or not. Networks with nodes that have high degree that show a preference to link together over low degree nodes are assortative. Networks with high degree nodes that show a preference not to link together, and prefer to link to low degree nodes are disassortative. Networks that do not show a preference among high degree nodes are neutrally assortative.

3. RESULTS

As discussed in the experiments, FIGURE 1 outlines the construction of data dependency graphs from basic block segments. These graphs were constructed for dependencies of data operands between data movement instructions, which were the primary motivation for the analysis of graphs. Data movement instructions make up the most significant portion of the term variance with respect to frequency.

FIGURE 2 shows various measurements taken from program's control flow graph. The network's degree histogram is shown in the bottom right, which is positively skewed with the largest number of nodes in the network having a low degree. The topology of connected components is shown in the top of the figure, which shows several nodes with few connections, and a small number of nodes with high degree. The network's degree rank plot is shown in the bottom left. The control flow graph for a program represents a network of contiguous program instructions as nodes with transitions between them as edges. The degree histogram shows that the degree distribution follows

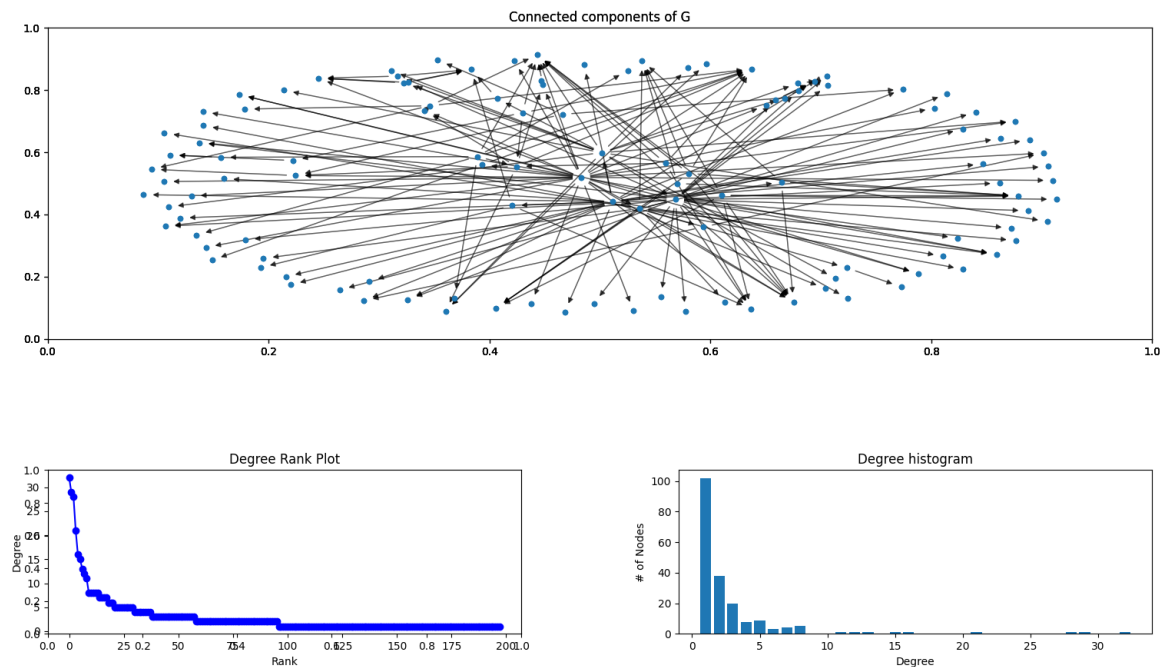


Figure 2: *The program's Control Flow Graph has a power law degree distribution* - Figure showing a network with Degree Histogram, and Degree Rank Plot of a program's Control Flow Graph (CFG), which shows a power law degree distribution with a positive skew, in that most nodes have very few connections and a small number of nodes have a high degree.

a power law, a small number of nodes have a very high node degree. FIGURE 4 and FIGURE 7 both show the comparison of $k_{max}/\ln(N)$ with $\ln(N)/\ln \langle K \rangle$ for comparison purposes. While degree distributions of control flow graphs follow a power law, this distribution is not stable within a program. The degree-rank plot in FIGURE 2, shows a power law trend, but when this power law distribution is plotted on a log scale, the cumulative distribution function shows that the power law exponent decays from the linear trend. This shows that the power law exponent of the degree distribution is not stable, and decays as the rank increases. Methods of matching a program to specific degree distribution would need to take this decay into account. Both control flow graphs and data dependency graphs have a degree distribution which follows a power law exponent, and that is higher than that of $\gamma = 3$, a measurement which holds across the data collected. From this we can draw the conclusion that control flow and data dependency networks have Scale-Free properties [32].

FIGURE 3 shows a comparison of data dependency graphs. This figure shows a sample of data dependency graphs taken from a malicious program with the largest number of segments. A program was segmented into basic blocks, and a data dependency network was constructed for each basic block. When analyzing data dependency graphs for a program, the result is a large number of small networks, one per each block segment. The graphs have been ordered by number of nodes

Malware Block Index	N	L	k_{max}	$k_{max}/\ln(N)$	$\ln(N)/\ln \langle K \rangle$	Pearson	γ
390	48	32	3	0.774	2.904	-0.153	18.264
527	40	30	9	2.466	2.397	-0.416	4.538
263	29	26	4	1.187	1.878	0.105	3.281
358	32	25	5	1.442	2.218	-0.577	4.279
526	21	13	4	1.313	2.459	-0.326	8.574

Figure 3: Comparison of network structure between data dependency networks of operands for largest 5 DDG networks.

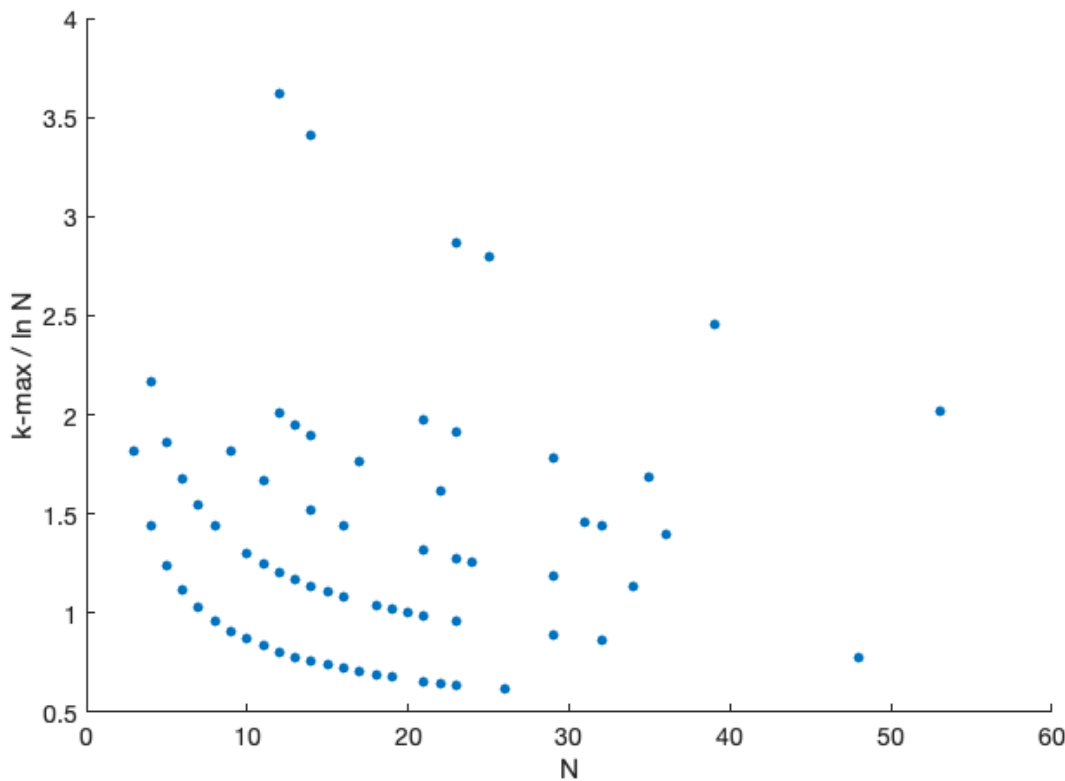


Figure 4: Scatter plot of network size and ratio of natural log of maximum node degree and natural log of network size for data dependency networks collected for a single program sample. Network diameters are small, and decrease as network size increases, additionally indicating the existence of network hubs.

for the networks with the largest size. It is important to note that degree distributions for the data dependency graphs also follow a power law, and therefore we have not computed the mean degree. Shown is the network size with respect to number of nodes N , the number of edges between nodes or links L , the maximum degree k_{max} , the ratio of k_{max} to $\ln(N)$, the Pearson degree correlation between nodes in the network, and the power law exponent of the degree distribution γ . The variable

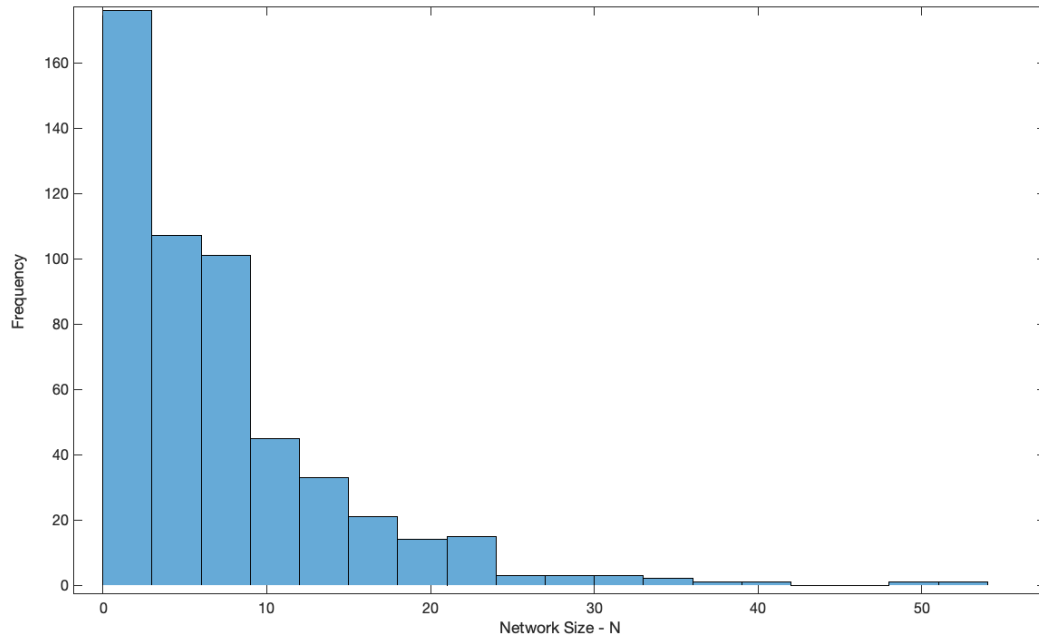


Figure 5: *Data Dependency Graph sizes for this program are skewed positively and follow a power law distribution* - This histogram shows network sizes N of data dependency networks (DDG) extracted from *mov* instructions per block segment in a single program. This shows a power law distribution where most data dependency networks for *mov* instructions are very small, N less than 5.

γ was calculated by finding a function approximating the cumulative distribution function for the data collected of node degrees. This distribution of node degrees was used to approximate the cumulative distribution function, and is used for the degree distribution exponent. In this figure we can see that each network has a cumulative distribution that follows a power law exponent. From this we can conclude the existence of the Scale-Free property. We can see that each network has a negative Pearson correlation value, from which we can make an inference about the degree correlation and assortativity. We can see that the value of $k_{max} / \ln(N)$ is low, and this indicates a small network diameter. This in combination with a high clustering coefficient demonstrates the existence of the Small-World property [33].

In FIGURE 4, $k_{max} / \ln(N)$ is a predicted network diameter. If we assume that a network is Scale-Free, and not a random network, then we would expect the existence of network hubs. Random networks do not follow a hub-and-spoke pattern, so the existence of hubs would also demonstrate a Scale-Free network. So the ratio of k_{max} to $\ln(N)$ is one metric of a Scale-Free network. We would expect the network diameters to be very small. We would also expect the network diameters to decrease as the size of the network grows. If hubs were not present, then as the size of the network grows, the diameter of the network would grow as well. The network's diameter is logarithmically dependent upon the network size. $\ln \langle k \rangle$ is typically used as a metric of network density to

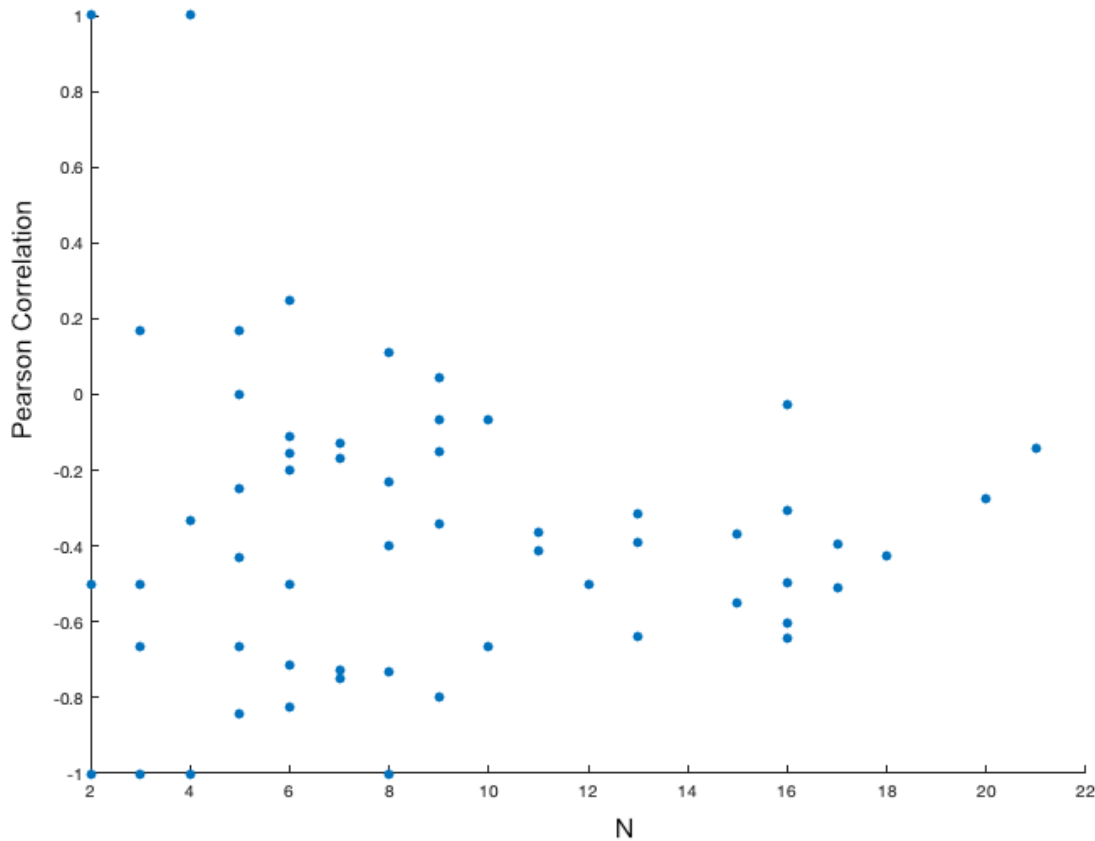


Figure 6: *Data Dependency Networks' Degree Correlations are Disassortative and are not random for the sample* - Scatter plot of network size N on the x-axis and Pearson correlation of the network in terms of degree on the y-axis for data dependency networks of *mov* instructions, *DDG*, per block segment. This shows that the degree correlation coefficient for a majority of networks is below 0, meaning that a majority of the data dependency networks in this program are Degree Disassortative, and do not link to nodes with high degree.

demonstrate both Scale-Free and Small-World properties. Since our networks follow a power law, measured by the exponent γ , we do not base our measurements on mean degree here. Instead, if we plot the network diameters in FIGURE 4, we see that the predicted diameters of the network with respect to k_{max} and $\ln(N)$ decrease as the network size N increases. $k_{max}/\ln(N)$ also gives us a prediction of the size of the hub in the network, a result which would not be present in random networks. We can also see the effect of the hub's presence on the network diameter as N increases [30].

Since a diameter of path length greater than or equal to 4 does not exist, we can derive the conclusion that any two randomly selected data dependencies will be less than a path distance of 3 nodes away.

Malware Sample CFG	N	L	k_{max}	$k1$	$k2$	Pearson	γ
Win32_APT28_SekoiaRootkit	1,495	2,779	246	33.653	5.566	-0.098	6.204
Win32_AgentTesla	21,732	18,394	578	57.877	18.971	-0.057	11.539
Win32_Avatar	928	1,669	23	3.366	5.337	-0.012	5.999
Win32_BigBangA	57,344	120,007	2,308	210.644	7.653	-0.042	2.211
Win32_BigBangB	46,937	97,470	2,288	212.707	7.554	-0.041	2.281
Win32_BigBangC	71,109	155,022	1,153	103.204	7.587	-0.054	2.250
Win32_Boaxxe.BB	2,507	5,129	118	15.076	5.555	-0.073	3.618
Win32_Caphaw_ShylockA	1,929	3,450	76	10.046	5.935	-0.046	5.934
Win32_Caphaw_ShylockB	1,713	3,336	45	6.043	5.476	0.038	8.291
Win32_Cridex	1,155	1,386	58	8.224	8.054	-0.040	6.713
Zeus_Gameover_2014_partA	22,169	42,845	712	71.154	7.409	-0.033	2.595
Zeus_Gameover_2014_partB	20,488	39,836	599	60.336	7.310	-0.039	2.544

Figure 7: Control Flow Graph metrics. $k1$ and $k2$ represent $k_{max}/\ln(N)$ and $\ln(N)/\ln < K >$ respectively. $k_{max}/\ln(N)$ is large because k_{max} is large. $\ln(N)/\ln < K >$ indicates a small world property for large N . γ shows the presence of the Scale-Free property. Control Flow Graphs can have a high number of nodes N , and has a high number of links L , but still have low Pearson correlation, indicating that their degree assortativity is neutral.

Interestingly this appears to hold only for data dependency networks, a property that does not hold for program control flow graphs with large N . From this figure we can also see that small networks have higher density, with the highest degree node taking up a larger proportion of the total network size. The overall density decreases as the size of the data dependency networks increases.

FIGURE 5 shows the frequency of network sizes in a frequency histogram of data dependency graphs. This figure shows the number of nodes in a network, N , for data dependency graphs in a program sample. This shows a power law distribution where most networks are very small, N less than 5. A small number of networks have a very high number of nodes in the network. Since this is the distribution, it is not suitable to take the arithmetic mean of the node sizes for the dataset. The networks for control flow graphs and data dependency graphs both follow power law distributions for their network size.

3.1 Data Dependency Networks are Degree Disassortative

FIGURE 6 shows a scatter plot of data dependency networks. The network size N is shown on the x-axis, and the Pearson degree correlation is shown on the y-axis. This graph shows that regardless of network size, the degree correlation coefficient is negative for a large majority of the networks. While graphs exist with positive Pearson correlation values, no graph with a network size greater than 5 nodes has a Pearson correlation over 0.2. This indicates the complete absence of graphs of significant size with high Pearson correlation values. A low or negative correlation coefficient of network degrees indicates that these networks are degree disassortative, and show a preference for not connecting to nodes with high degrees. This is the case for the sample being analyzed. This

appears to hold only for data dependency networks, and does not hold for control flow graphs with large N .

Through this finding we are able to make the prediction that nodes with high degree will not connect to other similar nodes with high degrees, but show a preference to connect to nodes with low degrees, as shown by the degree correlation value. This is a structural feature that is less than a purely random network with a degree correlation between hubs based on pure probability. Since k_{max} can still be a very high degree value, the network topology resulting from this structure is one of a *hub – and – spoke* pattern, where many nodes with small degree must connect to one of a small set of high degree nodes acting as hubs, and hubs will have fewer connections between each other [30].

FIGURE 7 shows a comparison of program control flow graph properties. These include the degree distribution power law exponent. The structural properties measured here provide additional structure to the representation of control flow graphs that can be used for further analysis. FIGURE 7 shows $k_{max} / \ln(N)$ as a measurement of network density and diameter for control flow graphs, and these values are typically very high. This metric of network diameter is one indication that hubs are not present, and that random networks are accurate models. We can expect the number of links between nodes in the network to be high, and for individual nodes to have high degrees. FIGURE 7 also shows the Pearson correlation values with the network sizes for data dependency graphs. This shows that a majority of the networks have a negative correlation, and this holds as the network size increases. The Pearson correlation values for node degrees are very low. This shows that nodes in a control flow graph do not show a preference for connecting to nodes with high degree, nor do they show a preference for connecting to nodes with low degree. From this we can draw the conclusion that control flow graphs have Neutral Degree Assortativity. This assortativity property appears to hold across samples of program control flow graphs. Since this is the case, random graph models are likely to display the behavior of nodes in this network well through growth and preferential attachment properties.

An analysis of adjacent k-cliques would provide a measurement of the hubs and components that are fully connected, and does not include node communities with less than maximum density. The distribution of adjacent k-cliques in various malware samples show the prevalence of several small communities. For example, the ZeusGameover_Feb2014 control flow graph shows the number of 3-cliques to be 3113, the number of 4-cliques to be 31, the number of 6-cliques to be 1 [30, 34].

Other types of networks such as semantic networks have been analyzed and shown to have the Scale-Free and Small-World properties. It is of note that the networks we have studied have these properties in addition to being degree disassortative [3, 35].

4. CONCLUSION

In this study we have shown the measurement and quantitative analysis of several networks which compose structural features of malicious programs. In this study we have found through empirical observations that data dependency graphs and control flow graphs in programs are Scale-Free. DDG and CFG networks have a power law degree distributions, and the degree distribution of control flow graphs is not stable on a log scale and decays with network size. CFG networks have high

diameters, which indicates the absence of hubs. DDG networks have low diameter, and follow a hub-and-spoke pattern. CFG network assortativity is neutral and nodes are connected based on a probability distribution. The distribution of network sizes is positively skewed and follows a power law distribution. DDG networks correspond to the Small-World property outlined by Watts and Strogatz. Degree correlations of DDG networks are structurally disassortative. While DDG nodes are connected to hubs, hubs show a preference for not connecting to similar high degree nodes, and most path lengths between nodes in a data dependency graph are very small. Since control flow graphs show low correlation and high diameter, random graph models are likely to be better models through modeling growth and preferential attachment of program control flow. These network properties show that while measurements are skewed, the networks have identifiable structural properties based on degree assortativity and probability. This serves to provide quantitative analysis and additional structure to the representation of malicious programs for static analysis that can be used for further insights. The structural properties outlined provide increased feature resolution. In future studies we intend to use the network features discussed for supervised learning to train models for classification tasks and to more accurately identify patterns of malicious programs correlated to operational semantics.

4.1 Future Work

In future studies we hope to explore the implications of networks as feature representations and their impact on classification accuracy. Since these features are tied to both structure and operational semantics, we believe that this may lead to increases in classifier accuracy for supervised learning, and the identification of classes of functionality in unsupervised learning. The use of random graphs to model the structural properties of program control flow graphs would show increased accuracy. This is due to control flow graphs having a neutral degree assortativity. Additionally we hope to answer questions of overlap in classes of malicious programs categorized by functionality with features tied to structural properties.

5. ACKNOWLEDGEMENTS

This research was supported in part by Air Force Research Lab grant #FA8650 to the University of Cincinnati.

References

- [1] Sebesta RW. Programming languages. Addison-Wesley. 1999.
- [2] Souri A, Hosseini R. A State-Of-The-Art Survey of Malware Detection Approaches Using Data Mining Techniques. *Hum Centric Comput Inf Sci*. 2018;8:1-22.
- [3] Griffiths TL, Steyvers M, Tenenbaum JB. Topics in Semantic Representation. *Psychol Rev*. 2007;114:211-244.

- [4] Bruschi D, Martignoni L, Monga M. Detecting Self-Mutating Malware Using Controlflow Graph Matching. In: International conference on detection of intrusions and malware, and vulnerability assessment. Springer. 2006:129-143.
- [5] Arora V, Kumar Bhatia RK, Singh M. Evaluation of Flow Graph and Dependence Graphs for Program Representation. *Int J Comput Appl*. 2012;56:18-23.
- [6] Rawashdeh O, Ralescu A, Kapp D, Kebede T. Single Property Feature Selection Applied to Malware Detection. In: NAECON IEEE National Aerospace and Electronics Conference. IEEE Publications. 2021:98-105.
- [7] Kebede TM, Djaneye-Boundjou O, Narayanan BN, Ralescu A, Kapp D, et al. Classification of Malware Programs Using Autoencoders Based Deep Learning Architecture and Its Application to the Microsoft Malware Classification Challenge (Big 2015) Dataset. In: 2017 IEEE National Aerospace and Electronics Conference (NAECON). IEEE Publications. 2017:70-75.
- [8] Djaneye-Boundjou O, Messay-Kebede T, Kapp D, Greer J, Ralescu A, et al. Static Analysis Through Topic Modeling and Its Application to Malware Programs Classification. In: 2019 IEEE National Aerospace and Electronics Conference (NAECON). IEEE Publications. 2019:226-231.
- [9] Chandrasekaran M, Ralescu A, Kapp D, Kebede TM. Context for API Calls in Malware vs Benign Programs. In: International Conference on Modelling and Development of Intelligent Systems. Springer. 2020:222-234.
- [10] Shafiq MZ, Tabish SM, Mirza F, Farooq M. Pe-Miner: Mining Structural Information to Detect Malicious Executables in Realtime. In: International Workshop on Recent Advances in Intrusion Detection. Springer. 2009:121-141.
- [11] Siddiqui M, Wang MC, Lee J. Detecting Trojans Using Data Mining Techniques. In: International Multi Topic Conference. Springer. 2008:400-411.
- [12] <https://researchcommons.waikato.ac.nz/handle/10289/1040>
- [13] Zhou Y, Liu S, Siow J, Du X, Liu Y, et al. Devign: Effective Vulnerability Identification by Learning Comprehensive Program Semantics via Graph Neural Networks. *Adv Neural Inf Process Syst*. 2019;32.
- [14] Wang Y, Wang K, Gao F, Wang L. Learning Semantic Program Embeddings With Graph Interval Neural Network. *Proc ACM Program Lang*. 2020;4:1-27.
- [15] Park E, Cavazos J, Alvarez MA. Using Graph-Based Program Characterization for Predictive Modeling. In: Proceedings of the tenth international symposium on code generation and optimization. 2012:196-206.
- [16] LeDoux C, Lakhota A, Miles C, Notani V, Pfeffer A. FuncTracker: Discovering Shared Code to Aid Malware Forensics. In: 6th USENIX Workshop on Large-Scale Exploits and Emergent Threats (LEET 13) 2013.
- [17] S, Shirani P, Wang L, Debbabi M. Fossil: A Resilient and Efficient System for Identifying Fossil Functions in Malware Binaries. *ACM Transactions on Privacy and Security (TOPS)*. 2018 Jan 31;21(2):1-34.

- [18] Cesare S, Xiang Y. A Fast Flowgraph Based Classification System for Packed and Polymorphic Malware on the Endhost. In: 2010 24th IEEE International Conference on Advanced Information Networking and Applications. IEEE Publications. 2010:721-728.
- [19] Cesare S, Xiang Y, Zhou W. Control Flow-Based Malware Variant Detection. IEEE Trans Depend Sec Comput. 2013;11:307-317.
- [20] Cesare S, Xiang Y. Classification of Malware Using Structured Control Flow. In: Proceedings of the eighth Australasian symposium on parallel and distributed computing-volume. 2010;107:61-70.
- [21] Hopcroft JE, Motwani R, Ullman JD. Introduction to Automata Theory, Languages, and Computation, 2nd Edition. SIGACT News. 2001;32:60-65.
- [22] Musgrave J, Purdy C, Ralescu AL, Kapp D, Kebede T, et al. Semantic Feature Discovery of Trojan Malware Using Vector Space Kernels. In: 2020 IEEE 63rd International Midwest Symp Circuits Syst (MWSCAS). 2020:494-499.
- [23] Musgrave J, Messay-Kebede T, Kapp D, Ralescu A. Latent Semantic Structure in Malicious Programs. In: International Conference on Modelling and Development of Intelligent Systems. Springer. 2022:234-246.
- [24] <https://thezoo.morirt.com/>
- [25] Nar M, Kakisim AG, Yavuz MN, Soğukpınar İ. Analysis and Comparison of Disassemblers for Opcode Based Malware Analysis. In: 2019 4th International Conference on Computer Science and Engineering (UBMK). IEEE Publications. 2019:17-22.
- [26] Hennessy JL, Patterson DA. Computer Architecture: A Quantitative Approach. Elsevier. 2011.
- [27] Hagberg A, Swart P, Chult DS. Exploring Network Structure, Dynamics, and Function Using Networkx, Los Alamos National Lab. (LANL). SCIPY. 2008.
- [28] Ferrante J, Ottenstein KJ, Warren JD. The Program Dependence Graph and Its Use in Optimization. ACM Trans Program Languages Syst. 1987;9:319-349.
- [29] SM Toolbox, Matlab. Mathworks Inc, 1993.
- [30] Barabási A-L, Network science, Philosophical. Transactions of the Royal Society à Mathematical, Physical and Engineering Sciences. 2013;371:20120375.
- [31] Watts DJ, Strogatz SH. Collective Dynamics of 'Small-World' Networks. Nature. 1998;393:440-442.
- [32] Li L, Alderson D, Doyle JC, Willinger W. Towards a Theory of Scale Free Graphs: Definition, Properties, and Implications. Internet Mathematics. 2005;2:431-523.
- [33] Alstott J, Bullmore E, Plenz D. Powerlaw: A Python Package for Analysis of Heavy Tailed Distributions. PloS one. 2014;9:e85777.
- [34] Palla G, Derényi I, Farkas I, Vicsek T. Uncovering the Overlapping Community Structure of Complex Networks in Nature and Society. Nature. 2005;435:814-818.
- [35] Steyvers M, Tenenbaum JB. The Large-Scale Structure of Semantic Networks: Statistical Analyses and a Model of Semantic Growth. Cogn Sci. 2005;29:41-78.