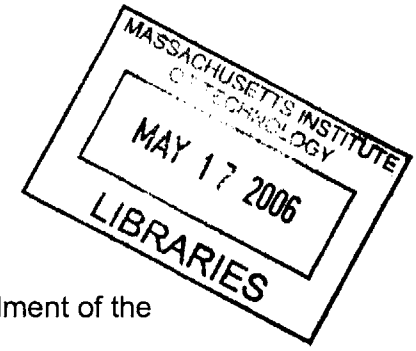
SCALE-FREE INFORMATION SYSTEMS NETWORKS

By

Wee Horng Ang

B. Sc. (Hons) Computer Engineering
University of Illinois at Urbana-Champaign 2004



Submitted to the Engineering System Division in Partial Fulfillment of the Requirements for the degree of

Masters of Science in Engineering Systems

At the

Massachusetts Institute of Technology

May 2006

Signature of Author.....
Engineering Systems
Engineering Systems Division
May 17, 2006

Certified by....
.....
Stuart Madnick
John Norris Maguire Professor of Information Technology
& Professor of Engineering Systems
Thesis Supervisor

Accepted by.....
.....
Richard de Neufville
Chair, Engineering Systems Division Education Committee
Professor of Engineering Systems

SCALE-FREE INFORMATION SYSTEM NETWORKS

By
Wee Horng Ang

Abstract

Many real, complex networks have been shown to be scale-free. Scale-free in networks mean that their degree distribution is independent of the network size, have short path lengths and are highly clustered. We identify the qualities of scale-free networks, and discuss the mathematical derivations and numerically simulated outcomes of various deterministic scale-free models. Information Systems networks are a set of individual Information Systems that exchange meaningful data among themselves. However, for various reasons, they do not naturally grow in a scale-free manner. In this topic, we will specifically examine a technique proposed by MITRE that allows information to be exchanged in an efficient manner between Information System nodes. With this technique, we will show that a scale-free Information System Network is sound in theory and practice, state the characteristics of such networks and demonstrate how such a system can be constructed.

Advisor: Stuart E. Madnick
Title: John Norris Maguire Professor of Information Technology & MIT Sloan School of Management & Professor Engineering Systems MIT School of Engineering

Table of Contents

Table of Figures

Acknowledgements

Firstly, I would like to thank Professor Stuart Madnick for the invaluable guidance he has shown me in the course of my research and graduate studies at MIT.

Dr. Michael Siegal, for the long hours of in-depth discussion we had, as well as suggesting possible research directions whenever I encounter any stumbling blocks.

MITRE, in particular Dr. Marwan Sabbouh, for providing financial support and initiating this study.

The Context Interchange (COIN) group, in particular Dr. Harry Zhu, Dr. Frank Manola and Allen Moulton, at MIT for providing me with a host of suggestions and improvements on this topic.

My family, for the emotional support they have given me for the course of my graduate studies.

My office mate, Lynn Wu, for providing me with a daily supply of freshly brewed coffee and keeping me sane.

My scholarship board, IDA, for sponsoring my overseas education.

Thank you!

Chapter 1: Introduction

Complex networks form an integral component of our daily lives. They have been extensively researched across several fields of study, including physics, sociology education, biology and the medical sciences. [1, 2] The growth of the Internet (see Figure 1) is an example of a vast, complex and constantly evolving network. Due to the scale and complicated nature of these networks, their seemingly unconstrained behavior and pattern-less growth, these complex evolving networks were initially viewed as completely random.

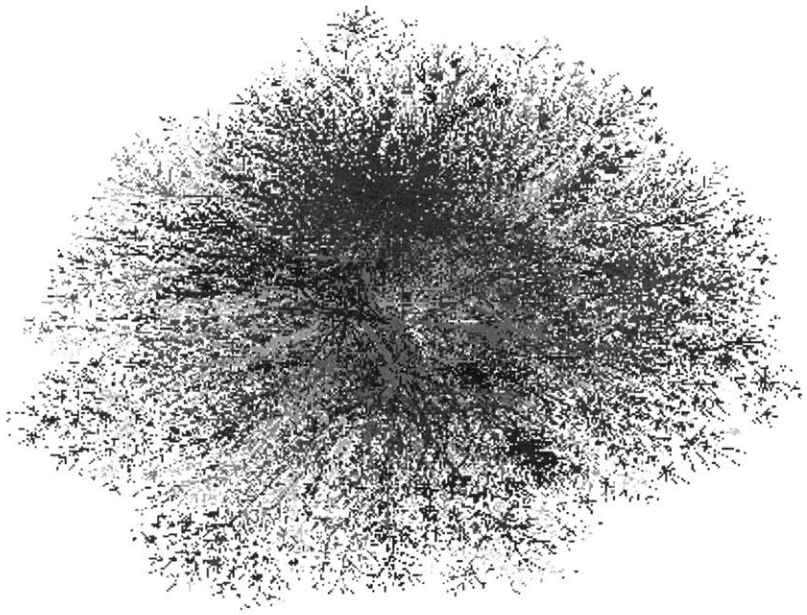


Figure 1: The Internet (Bill Cheswick, Lumeta Corp)

However, research has shown that these complex evolving networks share various similarities and many patterns can be distinguished within all these diverse networks. Exhaustive efforts have been made, in particular that of Barabási, Albert [3-5] and derivative works, to analyze how these complex networks are formed, as well as to

reproduce deterministically these networks in order to improve the understanding of how these complex networks grow.

An Information System Network (ISN) is a cluster of interconnected Information Systems (IS) or databases that allows information to be interchanged and be understood within the cluster. For Information Systems to work and share information in a network, semantic integration is needed. Semantics refer to the meaning of data as opposed to syntax, which only defines the structure of the information. Semantic heterogeneity indicates that there are similarities or differences in the meaning of local data between two or more data sources, such as when two schema elements in two separate data sources have the same intended meaning, but referencing different names [6]. For our purposes, we will define these equivalent fields as being semantically similar. Semantic integration is therefore the determination of these semantically similar fields that exist between different data sources or ISs. The process of creating the set of semantically similar attributes between two ISs is known as performing an interoperability mapping.

There has been little research in analyzing network growth in the field of Information System Networks. There are many features of these networks that do not lend themselves easily to form large complex networks, namely the huge production and maintenance costs associated with creating such a large interoperable ISN. Chief among these features is that semantic interoperability mappings are non-transitive. Non-transitive occurs when the Information System Network cannot determine that an object a is related to an object c, when it is separately known that object a is related to b, and b is related to c. This leads to the two poor ISN implementation choices: Either fully connect all the IS nodes within the networks, or suffer from a lack of interoperability.

In my thesis, I will show that, using a new semantic interoperability technique developed by MITRE, substantiated by network theories developed in the myriad field of network research, a third Information System Networks decision choice is available. I will show that this ISN will be scale-free, cost-efficient, robust, and maintain complete semantic interoperability without high implementation costs.

The thesis will be broken down into three major sections. Firstly, I will discuss some of the features unique to Information Systems and databases. I will examine the reasoning why Information Systems generally are unable to grow in an organic, self-sufficient growth, as well as why an n^2 network may ensue.

In the second portion of the thesis I will discuss some of the network theories and research that is relevant in our development of a robust Information System Network theory. In particular, I will discuss the evolution of network research, the formulation of the scale-free and competitive network models, mathematical derivations of the power-law degree distribution, as well as discuss deterministic methods that will lead to scale-free network growth.

Finally, I will discuss a methodology of Information System growth that will ensure that Information Systems Networks (ISN) will grow in an efficient, effective and robust network topology. In particular, I will discuss how the competitive network model can be applied directly to ISNs, derive the outcome behind a simple application of the competitive network algorithm under a given ISN setting. I will conclude by stating the potential limitations of this model and the conditions under which scale-free growth will be disrupted.

Throughout the paper, I will utilize two running examples of Information System Networks (ISNs), to illustrate certain characteristics of networks, the implications of adding the MITRE interoperability technique layer on top of ISNs, as well as susceptibility of ISNs to scale-free growth. The first example pertains to growth of an online shopping interoperability network. Through mergers and acquisitions, a single company presently has a set of online shopping information databases with diverse context, structures and data representations and wishes to achieve complete interoperability between its Information Systems. The second example revolves around a cluster of air flight mission systems, to illustrate how the MITRE interoperability technique is applied, and from which one can derive the benefits and drawbacks of the technique in the scale-free network growth context.

Chapter 2: Information Systems

Semantic integration allows information interchange between Information Systems. As each information system is created separately from each other, each system exist within its own unique context, based on the uses and needs of the user base, as well as the context of the database creator. Batini [7] stated that there are three main causes for semantic heterogeneity:

(1) Different Perspectives, where different IS designers adopt their own viewpoints when modeling the same information

(2) Equivalent Constructs, where a variety of combination of data constructs can model the same real-world information. An example is when a single attribute, *sales price*, models the association between tax and actual product cost in one IS, but is explicitly split into the two respective attributes in another IS.

(3) Incompatible design specifications, where different design specifications results in different schemas. One air mission IS design might allow for multiple missions for a single flight, while another IS might only allow a single mission profile during a specified flight.

Data Conflict Resolution

To allow for non-trivial information interoperability to exist, the following differences have to be resolved, which can be classified into two categories: (i) Structural or Syntactic Differences and (ii) Representational Differences [8].

Under Structural or Syntactic Differences, there are the following issues that need to be solved. The differences include: Naming Conflicts, Type Conflicts and Levels of Abstraction. Under Representational Differences, the potential incompatibilities include Scaling Variations and Domain Conflicts.

For naming conflicts, integration problems can be further categorized into homonyms and synonyms conflicts [9, 10]. Homonym inconsistencies occur when different concepts or properties in different information systems share the same name. On the other hand, synonyms are similar concepts but captured on different information databases through dissimilar names. Whereas homonyms can be detected by comparing concepts with the same name in different schemas, synonyms can only be detected after an external specification [7].

Type conflicts occur when the same concept is represented by different coding constructs in different schemas, such as when an object is represented as an entity in one construct and an attribute in another [11]. Levels of abstraction refers to when information is considered on a dissimilar scale between two information systems, such as when Total Costs in one system is segmented into Material Costs and Labor Costs in another.[12]

Scaling discrepancy is another factor, which is defined as when the same attribute is stored in disparate units in different IS [13]. For example, an Information System may utilize single units of US currency for its financial information, while another Information System may represent financial statements using Japanese Yen in thousand unit increments.

Another contextual variation exists when considering different data frames of reference, such as using different units of measurement like the metric and English system, or even having different Airport codes that would represent the same airport. These conflicts are also known as domain conflicts [14] Certain countries are denoted in two letter descriptors under a widely used airport naming system, while three letter descriptor also coexist.

Other information integration problems include accuracy variation, which occurs when the depth of information stored and structured varies among the databases [13], or when there is missing or conflicting information. Thus, it is often a difficult and tedious process to attempt to integrate information sources and systems together, and it is widely believed that these problems are non-trivial and will scale quadratically as more information systems are considered together. In all, Information Systems are sufficiently different from one another that it is necessary to resolve these non-trivial differences to allow semantic exchange of information.

Information System Networks also differ from other types of networks, in that there are very stringent requirements that govern whether edges between certain IS nodes can exist. Choosing an edge that is acceptable in the ISN is a tedious process, as well as the non-trivial work involved in edge creation. This is different from networks where edges can be added easily, such as when adding URL links in the World Wide Web, or networks with edges that exist naturally, such as when considering neural networks.

Transitivity

In an Information System Network, it is essential that disparate Information Systems are able to interoperate and share contextually meaningful data. For example, if

an IS node a_1 is mapped to an IS node a_2, it means that the two Information Systems can share meaningful data. But without network transitivity, IS node a_1 cannot share meaningful data with IS node a_3, if the two nodes are not directly connected. Figure 2 shows a situation when transitivity does not exist in the network.

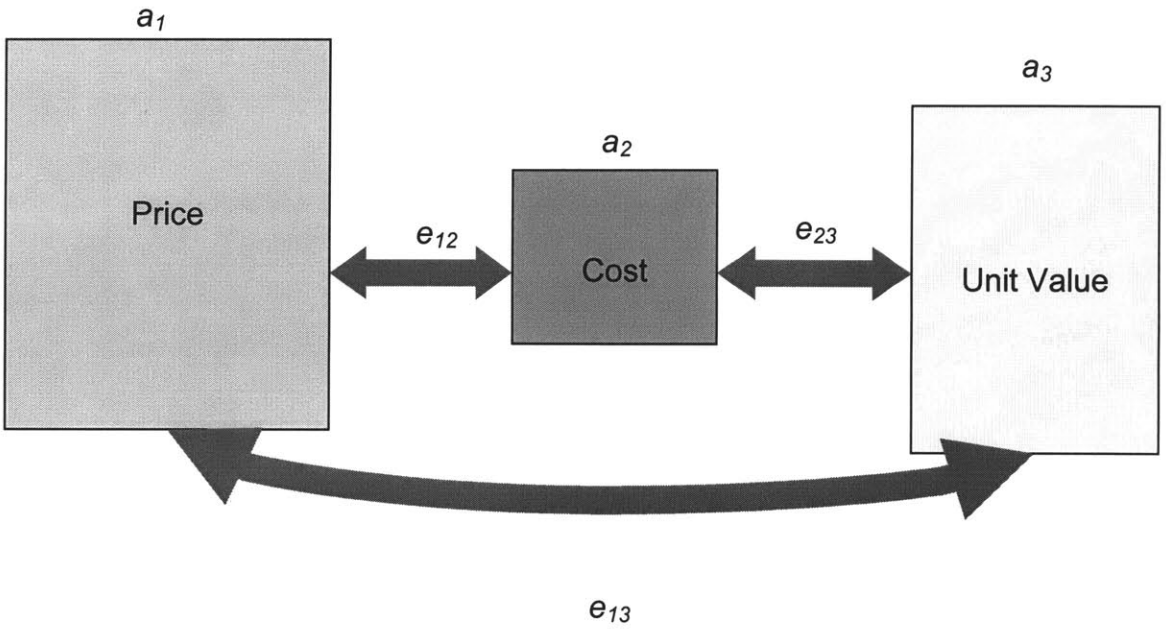


Figure 2: Transitivity example

In this example, consider the case with three nodes in the network, each modeling a single data attribute. Next, the attribute *"price"* in node a_1 is identified to be related to the attribute *"cost"* in node a_2, in the form of the interoperability mapping edge e_{12}. Separately, *"cost"* is also related to the attribute *"unit value"* in a third node a_3, through another mapping edge e_{23} that maps between a_2 and a_3. Transitivity dictates simply that if "*price*" is related to "*cost*", and "*cost*" is related to "*unit value*", the network would recognize that "*price*" is related to "*unit value*", without the need of an additional edge linking node a_1 and a_3. Without transitivity within the network, the network cannot establish any relation between "*price*" and "*unit value*" when only comparing the nodes

a_1 and a_3, and the edges e_{12} and e_{23}. To ensure no loss in interoperability, an additional edge, e_{13}, must be added into the Information System Network.

Consider a movie actors database such as Internet Movie Database, IMDB.com. A network analysis of thISN does not factor transitivity into its analysis, i.e. if actors a_1 and a_2 acted together in a movie, and actors a_2 and a_3 acted together in another movie, actors a_1 and a_3 have a two degrees of separation apart, but they do not have a direct relationship. Transitivity in this case would imply that because of the above relationships, actors a_1 and a_3 have acted together in a movie, which is untrue. Although such a loose relationship is sufficient to generate a movie actor network, it is not acceptable for Information Systems Networks.

There has been a lot of research to resolve data interoperability between databases. For example, the Context Interchange (COIN) project [15] solves contextual and representational differences through having a centralized knowledge representation and reasoning system that possesses the ability to resolve contextual variation for common concepts such as Time and Currency. The approach of COIN, however, addresses syntactic differences inherently in its implementation methodology by identifying and resolving semantically similar attributes in its coding structure. The MITRE Semantic Interoperability Technique [16] builds on the COIN approach by providing a systematic resolution for these syntactic differences as well.

Performance Metrics of Information System Networks

There can be many measures for IT artifacts, in both quantifiable and qualitative measures. Where analytical metrics are appropriate, one can perform a comparison assessment to determine which Information System Networks are better. Johansson et al.

[17] stated that previous cost-based approaches on Information Systems, which ascribe a unit cost to each computing resource and determined the total minimum costs, did not account sufficiently for response time. He proposed a model of response time that is interdependent on communication delays and parallelism. Other metrics for Information System Networks have been proposed by Salton [18] that assesses distributed information databases in terms of precision and recall.

The three metrics we will use to assess our Information System Networks are:

(1) Implementation Costs, denoted as the unit cost of creating and maintaining a network connection.

(2) Response Time, measured in how long it takes for an execution-time query to get a response. In our case, we will relate response time to the average number of edge traversals to answer a given query between two nodes in the network.

(3) Semantic Interoperability, denoted by how complete and lossless data can be exchanged within the network.

For a Complete Interoperability Information System Network

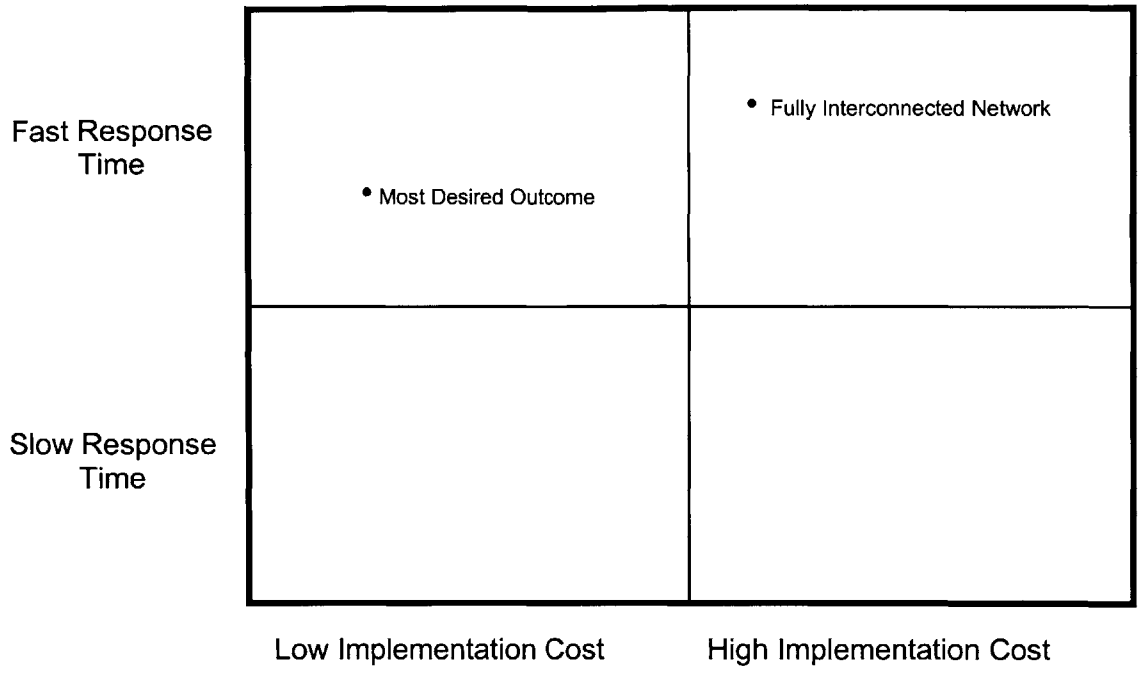


Figure 3: Performance Criterias for Information Systems

When we only consider an ISN that has complete information interoperability, the most desired ISN solution is one that has a fast response time, and requires a low implementation cost. We will first consider the fully interconnected ISN before establishing how one can achieve the most desired outcome ISN.

Fully Interconnected ISN

With the lack of transitivity, the way to ensure full information exchanges between all Information System nodes is for every node to be separately linked to every other nodes existing in the system. For a network system with n number of nodes, this would mean that there needs to be $n(n-1)/2$ connections to ensure full data interoperability. This leads to the problem of n^2 connections between nodes.

Having n^2 connections within a ISN entails different advantages and drawbacks as opposed to the usual network theories. The fully connected network is advantageous in certain ways. It is robust; with the loss of a single Information System in the network, information can still be exchanged between the remaining Information Systems. The shortest path distance between any two Information Systems is of unit length, which means that every node is directly connected to every other node. Information exchange between any two nodes can be performed with a single edge traversal, indicating fast query response time.

However, there are severe disadvantages of having a fully connected ISN. Firstly, mapping between IS is non-trivial work. Although there are many ways to automate the processes, a large amount of human intervention is still necessary to identify attributes that can be mapped to each other. Creating a fully connected network can be fairly simple initially, but the amount of work scales quadratically with the network size. A network with 100 nodes would require 4950 edges to be completely interconnected. When a new IS is added to this existing network, it will have to separately create 100 more edges, deterring the continued growth of the network. Also, as the purpose of an individual IS changes, its data schema and area of interest may change as well. A change in the data schema/ structure in an IS also signify that all connected mappings will also have to be updated. The creation costs and maintenance costs of the network are therefore prohibitively expensive in a large interconnected network.

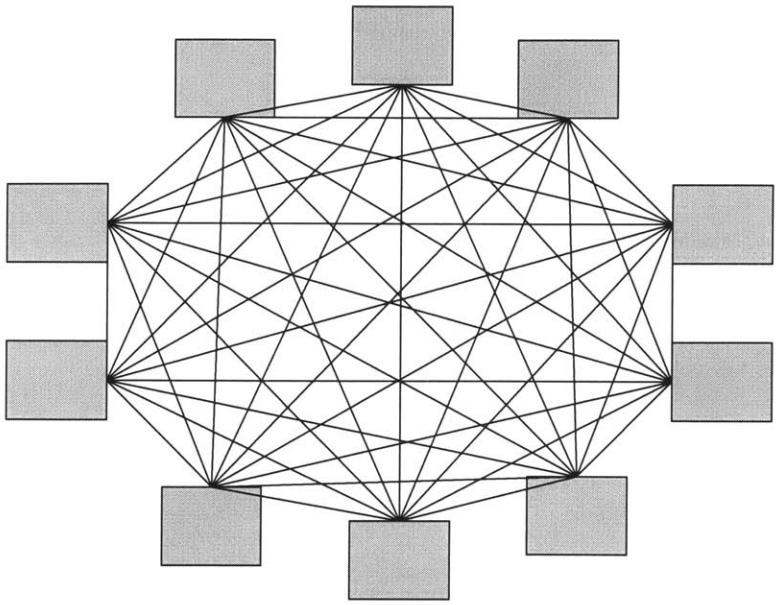


Figure 4: N-squared network

We thus need to examine the possible alternate scenarios in which we can achieve the most desired outcome for our Information System Network, one that can grow without having implementation costs scaling quadratically, while still maintaining a high level of interoperability.

Chapter 3: Information Systems as Networks

To relate Information Systems Interoperability Clusters to network theory, we need to conceptualize ISs on an appropriate abstract level. Networks are graphs consisting of nodes (or vertices) connected by edges (or links). Nodes are often used to denote independent, individual entities that are created and subsequently exist on a separate basis. Depending on the field of discussion, nodes can either be domain-level routers, when discussing physical web connections, people which is relevant to social networks or even human contagion networks. Under citation networks, nodes can be published papers which cite previous existing works and theories.

In the IS model, we will consider an individual internally consistent Information System as a individual node, consisting of its own data structure or schema, information context, and data assumptions. Figure 5 shows an example of a data schema that defines the data stored within an individual Information System that stores customer information. For our discussion, an IS node i will be denoted by the symbol a_i. A network with n nodes will have nodes labeled $a_1, a_2, \ldots, a_i, \ldots a_n$.

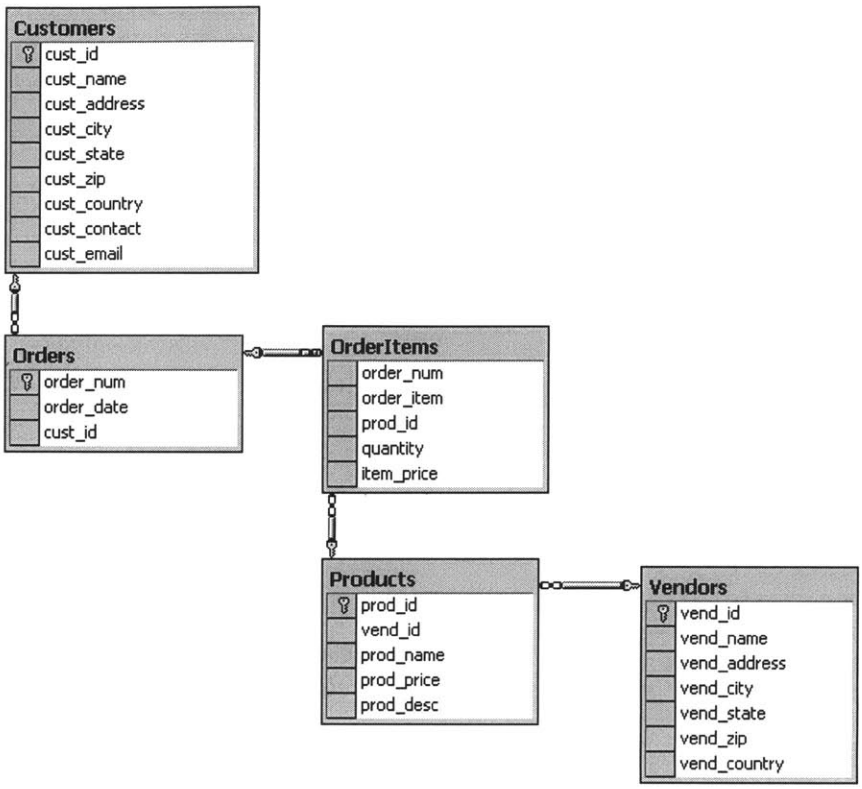


Figure 5: Example of a Data schema

Edges are the connections formed between two nodes. Edges can be directed, where traversals or flows can only occur in a single stated direction, or undirected, where bidirectional traversals are possible. For directed edges, they can flow "into" a node, or flow "out of" a node. Edges can be a physical connectivity between two nodes, such as a fiber optic connection between two routers, or more non-corporeal, such as friendship connections in a social network. In the Information System world, an edge will be defined as the creation and maintenance of the resolution of representational, contextual and structural difference between two information systems. Depending on the type of interoperability technique applied to the system, edges can be directed or undirected. In this discussion, an undirected interoperability technique will be examined. An edge formed between two nodes a_i and a_j will be denoted by e_{ij}.

Edges can also be weighted or non-weighted, depending on the network under discussion. In many fields, it is well known that the interaction strengths can vary widely, such variations being essential to the network's ability to carry on its basic functions. Researchers have repeatedly argued about the importance of assigning strengths to links, such as the link strength in neural or transportation networks[19],[20]. Although it is acknowledged that nodes and edges in Information Systems Networks are unique, the workload required to create an edge between two information systems is approximately equal, thus there is no requirement or basis to assign different weights to the nodes. Under the network model of Information Systems, only non-weighted edges will be discussed.

To apply the concept of a set of nodes and edges to an Information System Network, we need to address the concept of nodal depth. As stated earlier, each IS entity is not identical since the architecture and make-up depends on several factors such as the purpose of its creation and subsequent use. Also, the context of its origins plays an important factor. Finally, the IS characteristics is heavily influenced by the context and assumptions of its creator. All these factors serve to make each IS node, if not completely unique, then sufficiently different such that resolution between nodes is necessary. It is un-necessary to have to capture every nodal variation when relating Information Systems to network theory. Rather, a certain level of abstraction can be applied when modeling nodes of an ISN.

To understand the level of abstraction to be utilized, it is necessary to examine the actual task/work needed to achieve interoperability among ISs. Mapping between different IS or databases primarily depends on identifying concepts or attributes that

share the same semantic meaning between the entities that are being mapped. Semantically similar attributes is independent of naming variations, scaling, contextual or even referential differences.

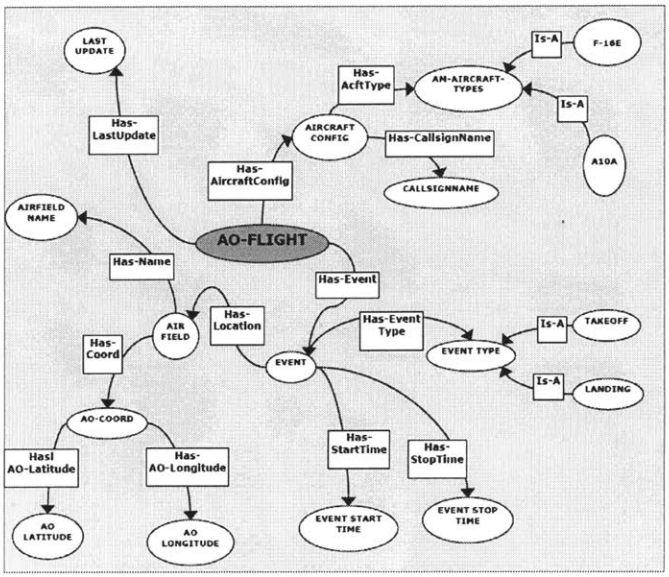


Figure 6: AO Flight IS schema [16]

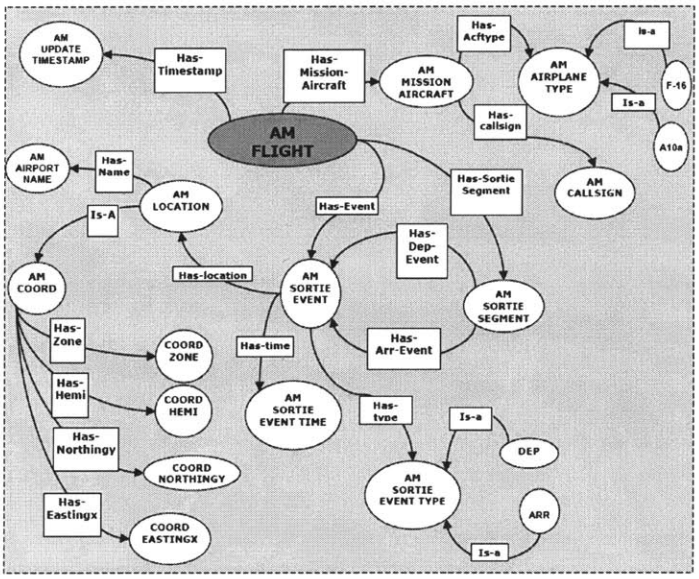
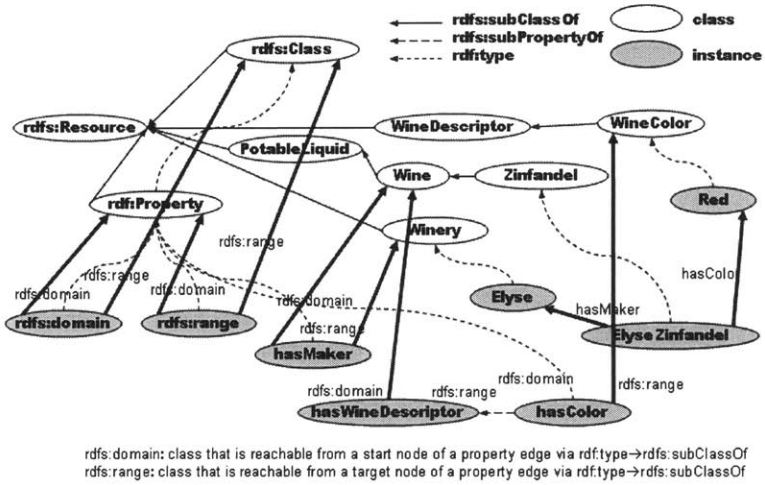


Figure 7: AM Flight IS schema [16]

For example, in Figure 6 and Figure 7, we have the data schemas of two Information Systems utilized for military air flight purpose tasked to handle different

types of air flight information. As with most Information Systems, they were created independently to serve separate functions. An interoperability agent will identify the various concepts that are similar between them. In this example, "AM Coord" in the AM IS and "AO-Coord" in the AO IS are identified to be sharing the same semantics. Similarly, "TakeOff" in the AM IS is the same as "DEP" in the AO IS.



rdfs:domain: class that is reachable from a start node of a property edge via rdf:type→rdfs:subClassOf
rdfs:range: class that is reachable from a target node of a property edge via rdf:type→rdfs:subClassOf

Figure 8: Example of a Wine schema [21]

Conversely, attempting to map an AM IS to a completely orthogonal Information System such as the one in Figure 8 which stores and disseminates wine data, there will be very little shared commonalities. A cursory glance will show that there is little to no semantically similar attributes between the two IS. As stated in Chapter 1, semantically similar attributes are two schema elements in two separate data sources that have the same intended meaning, but referencing different names. There is no additional value rendered from mapping between attributes of different semantics. Attempting to map a *"hasWineDescriptor"* attribute to an *"AMAirplaneType"* attribute will only produce trivial or erroneous data integration. Thus, one notes that when creating edges between IS, one is in fact mapping all the semantically similar attributes that exist in both ISs. Unless at least a semantically similar attribute exist between two Information Systems, a

connection or an edge cannot practically exist. This is a limitation that must be taken into consideration when extending the network model to the IS framework. As a means of mediating this difference, we would need to consider the inclusion of data schema attributes in our abstract model.

For an edge to exist there must be at least a set of semantically similar attributes present in both ISs. Since the attributes of an Information System plays an essential role in determining whether a mapping can exist between two ISs, it is essential that any network theory or algorithm recognize and compensate for this unique IS feature. As edges are thus defined, edges that loop onto the same nodes will not be considered, as well as the situation when multiple edges exist between any two ISs.

One factor to note is that certain ISs have unique attributes that do not exist elsewhere in the network. As these attributes have no semantically similar attributes, they need not be considered when mapping between nodes. Thus we will only consider attributes that can be mapped, and will denote γ_i as the semantically similar attributes that are captured in an Information System node.

To recap, in our abstract model of an Information System Network, a node a_i represents a single IS. Within it, there are any number of attributes, with each single attribute, denoted by γ_k, present within that IS. Between nodes a_i and a_j, there are semantic interoperability mappings that exist, denoted by edge e_{ij}. Edges are thus the sum of the mappings of semantically similar attributes that exist between any two IS. Edges will be undirected and of unit length, with at most a single edge existing between any two given nodes.

With that understanding, we will be able to extend our ISN model into network theory discussions, and be able to understand and utilize concepts and terms employed in network theory. We will now examine some of the network quality measurements used in network theory that is also pertinent to our discussion.

Network Quality Measurements

Over the course of the literature available regarding network research, there are several established measures of network quality that defines the behavior and attributes within the network. Those qualities are average path length, clustering coefficient and degree distribution.

Average Path Length

Path length, *l,* is the shortest distance necessary to traverse between two given nodes. *l* is also known as the diameter of a network, as it effectively establishes the linear size of a network, the average separation of pairs of nodes. In a fully-connected network, l=1. For non-weighted edges, when every edge between two different nodes is of uniform unit length, the path length is the shortest number of link traversals it takes to connect from one node to another. The average path length, $\bar{\ell}$, is the average of all the distance between nodes in the network.

Clustering Coefficient

Networks usually exhibit signs of clustering or cliques, whether social, neural or even citation networks. The inherent tendency to cluster is quantified by the clustering coefficient[22]. As a defining example, for a node i in the network, having k_i edges which connect to k_i other nodes. For a fully connected cluster, there would be $k_i(k_i\text{-}1)/2$ edges between the k_i nodes in that cluster. Thus, the clustering coefficient is defined as the ratio of the number E_i of edges that actually exist between these k_i nodes and the total possible number of edges that can exist.

$$C_i = \frac{2E_i}{k_i(k_i - 1)}$$

The clustering coefficient of the whole network is the average of all individual C_i's.

Degree Distribution

Degree of a node is the number of undirected edges that are connected to that node. Not all nodes in a network have the same number of edges. The spread in the node's degrees is characterized by a distribution function *P(k)*, which gives the probability that a randomly selected node has exactly k edges. Degree distributions are classified into in-degree or out-degree when referring to directed edge networks. The degree distribution is particularly relevant as recent network theories postulate that complex evolving networks grow in a manner independent of scale, but rather follows the network's degree distribution.

With the ability to extend Information System clusters into the network theory arena, we can now examine the various network theories that have been recently employed to explain complex, evolving networks.

Chapter 4: MITRE Semantic Interoperability Technique

Sabbouh [16] suggested a Semantic Interoperability Technique that utilizes Information System data models, context ontologies and a small number of simple OWL/RDF mappings to enable information originating in one part of the enterprise to be used in another in a way that is highly automated.

The technique solves the problem of semantic interoperability through a two-step process: 1) Resolving representational differences 2) Resolving structural and syntactic mismatches.

There are several ubiquitous enterprise concepts like types of Things, Time and Position. Representational differences would be having disparate levels of accuracy, scaling conflicts and dissimilar data context. Resolving representational differences is done by building or reusing a context ontology structure for each of the various concepts. For example the Position ontology can resolve differences between different grid reference systems, such as between UTM and WGE coordinate systems. The resolution mechanism is provided either through direct hard-coding, such as resolving unit-scale differences or through the use of appropriate Web Services, like GeoTrans for Position contextual differences.

Initially, a context ontology structure is constructed that captures common concepts across the enterprise space, while accounting for each IS's representation for a particular concept. When a new information system is added to the network, OWL/RDF is first utilized to construct the data schema. Next, context mediation is performed by mapping all the context relations from the information system to the context ontology.

This operation is only performed once, during the addition of a new Information System to the network, and the mapping occurs between the information system and the context ontology. Using the Position context, the mapping requires the Coordinate Systems, Coordinate Reference Frame and Datum to disambiguate any geo-Coordinate position. The corresponding attributes in the Information System schema will be mapped to the context ontology if it exists. The mapping occurs in the form of OWL/RDF encoding. Please see example [16] for the full documentation of the MITRE technique. It is important to note that this mapping occurs independently of other Information Systems, as there is no need to possess any knowledge of Geo-Coordinate context data from other Information Systems in the network during the mapping of a new Information System. This methodology is similar to the technique employed in Context Mediation [15].

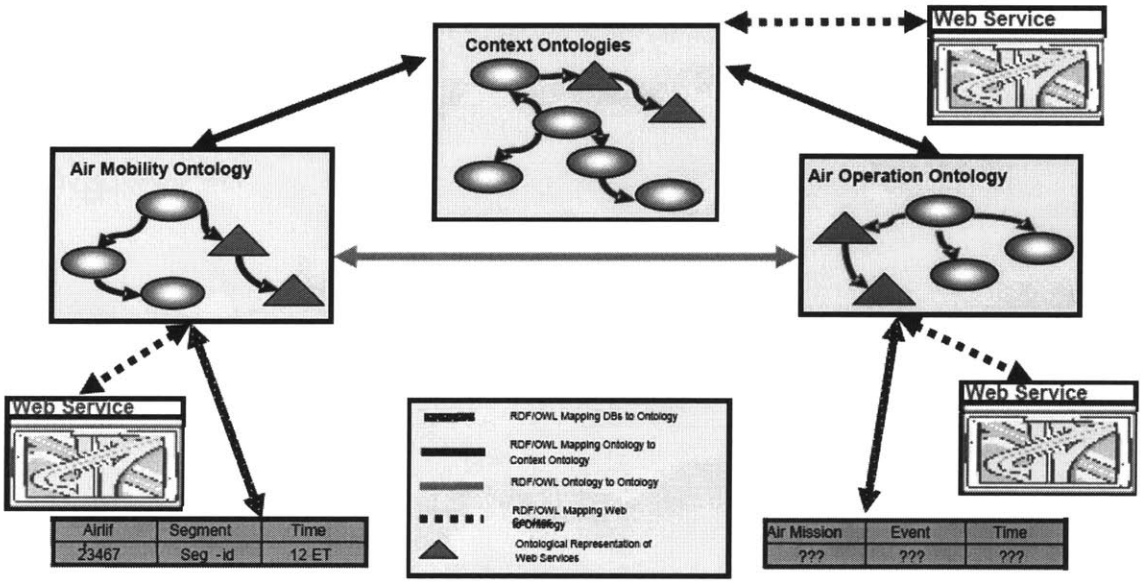


Figure 9: Methodology of MITRE Semantic Interoperability Technique[16]

Next is the resolution of structural and syntactic mismatches that would occur between information systems in the network. Structural differences include data behavioral conflicts, different levels of abstraction of data, and the identification of

related concepts. For example, TakeOff, Landing, Departure and Arrival are all extensions of the concept EventType, but they have inherently different semantic meaning. Mapping the attribute TakeOff in one Information System to the attribute Departure in another Information System would be an accurate mapping. Conversely, mapping the attribute TakeOff to the attribute Arrival will produce erroneous data interoperability assumptions. The example illustrates that the resolution of structural and syntactic differences occurs between Information Systems, and is dependent on the data that is to be shared between the two systems.

Features of MITRE Semantic Interoperability Technique

For data retrieval and semantic interoperability, the MITRE technique performs a series of constrained graph traversals that identify all connected conceptually similar data. Reasoning algorithms such as Directed Path Query (DPQ) or Incoming Intersection Query (IIQ) are used when there is a need to resolve instance data capture in one IS into the context and data structure of another. For a given list of inputs and a desired output, DPQ searches through the available paths between them. A direct path is the sequence of nodes and relations or mappings that connects them.

IIQ uses a two-pronged approach, first creating the set of direct paths that lead to the desired output, followed by creating another set of directed paths that lead to the given inputs. The intersection of the two sets will be the pathway between the given inputs and the desired output.

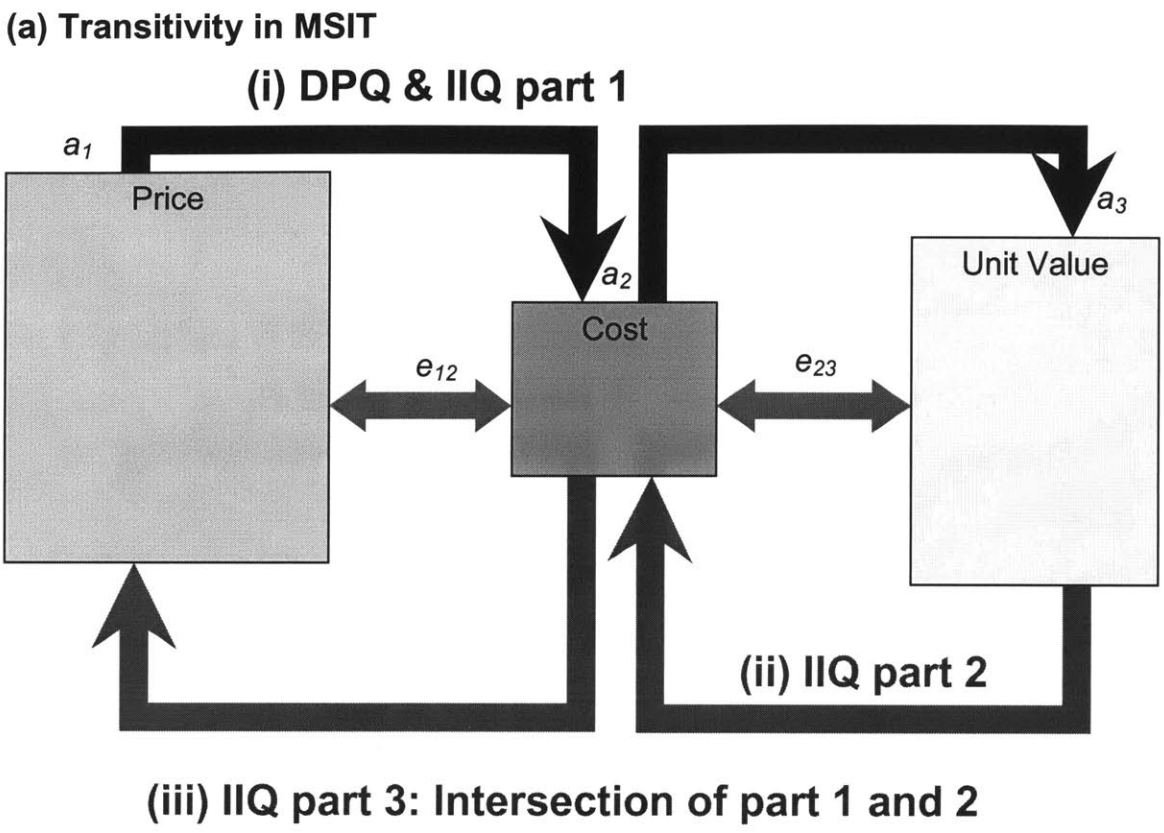


Figure 10: Direct Path Query and Incoming Intersection Query

Figure 10 shows how the two query algorithms layered on top of the Information System cluster enables transitivity within the network in relation to the earlier example of three Information Systems that have a single semantically similar data attribute. With DPQ and IIQ, the price attribute in System A will be logically linked to the cost attribute in System C, since a logical connection is made through the traversals on the edges e_{12} and e_{23}. This eliminates the need for an additional edge e_{13} that maps the price attribute in node a_1 to the unit value attribute in node a_3 for interoperability purposes.

Thus, with transitivity intelligence built on the MITRE layer on top of the Information System, through the utilization of path queries executed along the

connections between IS nodes, one can clearly see that full interoperability within Information System clusters can exist without requiring a fully interconnected network.

(b) Attribute Independence

In general, a **binary relation** consists of the following terms: a **key** and a **value**, which refer to entities; a predicate, which is an **access function** connecting the terms together in a relationship; cardinality, which states the number of elements in the relationship. An example of a binary relation statement is:

relation(access function, key, value, cardinality)

Binary relationships are associations that are frequently utilized in databases and information systems, specifically the relational models. Quite often, information systems utilize ternary and higher-degree to describe relationships between entities, as they are indicative of the natural understanding of the relationship that exists between several objects. Methods to reduce N-ary relations to binary relations has been an intense subject of research [23-30]. One method for resolving N-ary relation is Reification, which uses an Entity-Relationship model to resolve the N-ary relation into several binary relations while preserving most of an N-ary relationship semantic integrity. [31] The formula for reification works as this:

relation(t1,...,tn) ---> (exists e)(relation(e) & first(e, t1) & second(e, t2) &...& nth(e, tn))

The Semantic Web OWL/RDF language represents properties as a set of binary relations. A W3C Working Paper Draft, dated 21 July 2004, states how OWL/RDF is able to Define N-ary relations in terms of binary relations. The solution utilizes reification, and models a complicated N-ary relationship into a set of binary relations by

the introduction of a new complex object. All previous objects that shared the original N-ary relations now share a binary relation with the new complex object. [32]

MSIT uses the OWL/RDF language to describe the ontology of the various information systems. Each IS is represented by a set of binary relations that completely illustrates the features and context of that particular Information System. For example, in our AO system ontology, there are several binary relationships to describe the entire schema.

relation(Has-AO-Longitude, AO-Coord, AO Longitude, 1-1)

relation(Has-AO-Latitude, AO-Coord, AO Latitude, 1-1)

relation(Is-A, Event Type, TakeOff, 1-n)

...

Similarly for our AM system ontology, there are several similar binary relationships.

relation(Has-Northingy, AM Coord, Coord Northingy, 1-1)

relation(Has-Eastingx, AM Coord, Coord Eastingx, 1-1)

relation(Is-A, AM Sortie Event Type, Dep, 1-n)

...

Under the modeling technique employed by MITRE, all information models, data ontologies and schemas are expressed as a set of binary relationships. Within boundary of similar concept mapping, binary relationships are distinct and non-interfering. For the purposes of mapping similar concept, each binary relationship is distinct and does not affect the performance and relationship of other binary relationships present within the same information model. This binary relation expression of an Information System's data ontology means that the assortment of data attributes captured within the data schema can be expressed into distinct quantized units that can be considered individual element,

rather than a combined set of information qualifiers. From Figure 7, using the AM Information System as an example, we can see clearly that the AM IS has the following data attributes: *"AM-Sortie-Event"*, *"AM-Coord"*, *"Coord Northingy"*, *"Dep"*, *"Arr"*, *"AM Location"*, etc. All these attributes, though related to each other as defined by the data schema, are distinct data elements that can be considered on their own.

When performing mappings between information systems, one seeks semantically similar concepts or elements present within the two systems. For example, the *"AM-Coord"* attribute in the AM IS can be mapped to the attribute *"AO-Coord"* in the AO IS. Conversely, there is no conceivable value obtained from mapping the *"AM-Coord"* attribute in one IS to an Event Type attribute in another. One cannot map divergent attributes and hope to produce useful, consistent information interchange. Only attributes with the same semantics can be mutually mapped. Divergent attributes present in the Information Systems that are to be mapped thus **do not** play a role in determining the mapping. **We can therefore set each Information Systems' attributes for consideration separately from other attributes.**

(c) Need to Map all Attributes to Ensure Interoperability

Having established that transitivity exists for attributes under the MITRE Semantic Interoperability technique, as well as that attributes within a single Information System are independent of one another with regards to interoperability mapping, it leads to the conclusion that to ensure complete interoperability between all the nodes within the network, one has to first consider every attribute independently. Next, every semantic

type attribute within the system must be separately and completed connected to all semantically similar nodes.

Using the shopping Information System cluster, as shown in Figure 11, consider three Information System nodes that separate holds data pertaining to online shopping. Nodes a_1 and a_3 both describes shopping data in terms of five distinct semantic information types: 1) Price/Unit Value, 2) Quantity/Number of units, 3) Tax/Sales Tax, 4) Shipping/ Ship/Handle 5) Item Description/Item Review.

Now, if a_1 and a_3 are not directly mapped to each other, but rather through a third information system a_2 instead. a_2 describes shopping data held within its data schema in terms of 4 attributes: Cost, Quantity, Sales Tax, Description. When the edge e_{12} is created, information about shipping costs, captured in the a_1's data schema under the attribute *Shipping*, is not translated/mapped to a_2's data structure.

This also holds true for the edge e_{23}, where there is no semantically similar attribute in a_2 to account for the attribute "*Ship/Handle Charges*" present in a_3. This indicates that the attributes "*Ship/Handle*", as well as its semantically similar attribute "*Shipping*", are not mapped to each other.

Loss of information quality occurs if one assumes that if transitivity exists on the nodal level (between nodes a_1 and a_3), it will similarly exist on the attribute level (between all the attributes in a_1 and a_3). To retain complete interoperability, an additional edge e_{13} is required, that will provide the mappings between the *Shipping* attribute in a_1 and the *Ship/Handle Charges* attribute in a_3. This example shows that every single attribute must be attached to every other attribute that shares the same semantic type in the network within the network to enjoy complete semantic interoperability.

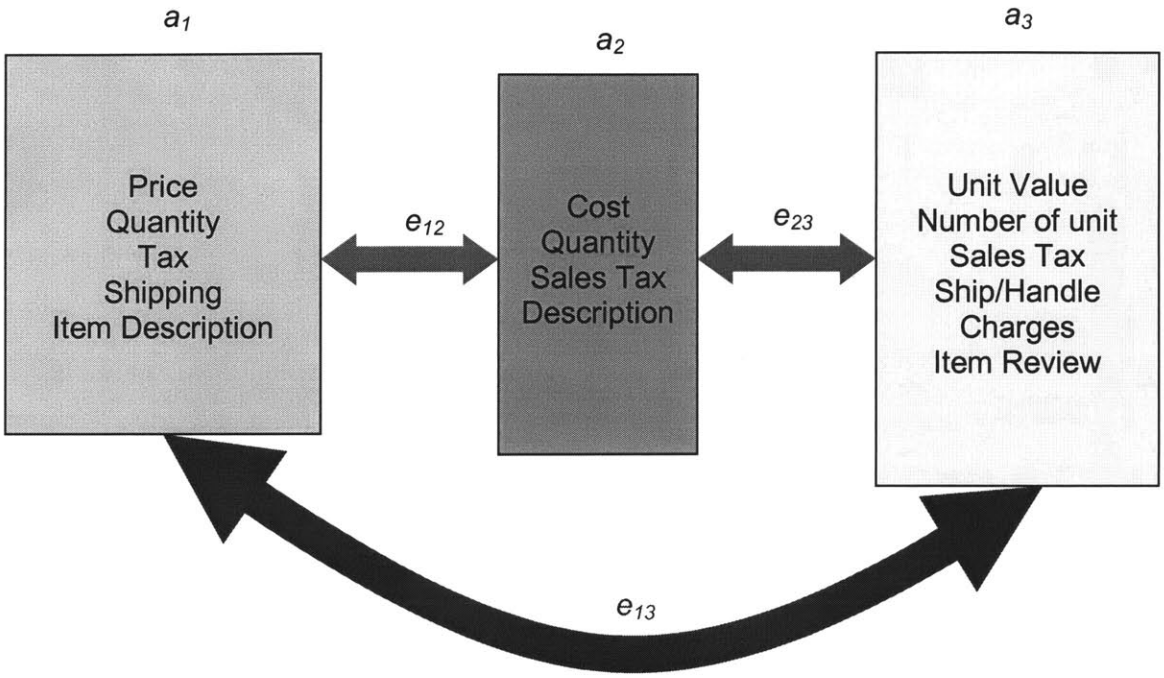


Figure 11: Transitivity Insufficiency Example

Interoperability Criterion: When a new IS node is introduced into the existing ISN, all of new IS node's attributes must be semantically mapped to at least one semantically similar attribute already existing in the network.

This is the case for a network that performs interoperability between existing nodes in the network. The conditions are different in the case of a complex, evolving network. Assume initially that the nodes in the Information System Network are all interconnected on the attribute level. When a new IS node is introduced into the network, if each attribute in the new IS is mapped to at least one semantically similar attribute existing in the network, full interoperability is still ensured.

In Figure 12, when a new node a_4 is added to the network, by fulfilling the interoperability criterion, we will maintain a completely semantic interoperability. For

example, when only a single edge is added to the network , if the edge added is either e_{14} or e_{34}, all 5 attributes present in a_4 is mapped, which means that only a single additional edge is required to ensure complete interoperability. Conversely, the interoperability criterion is not fulfilled if edge e_{24} is the only edge added between the new IS node and the existing network, since there is one attribute, Shipping, not mapped into the system. Thus another edge, e_{14} or e_{34} must be added to fulfill the interoperability criterion.

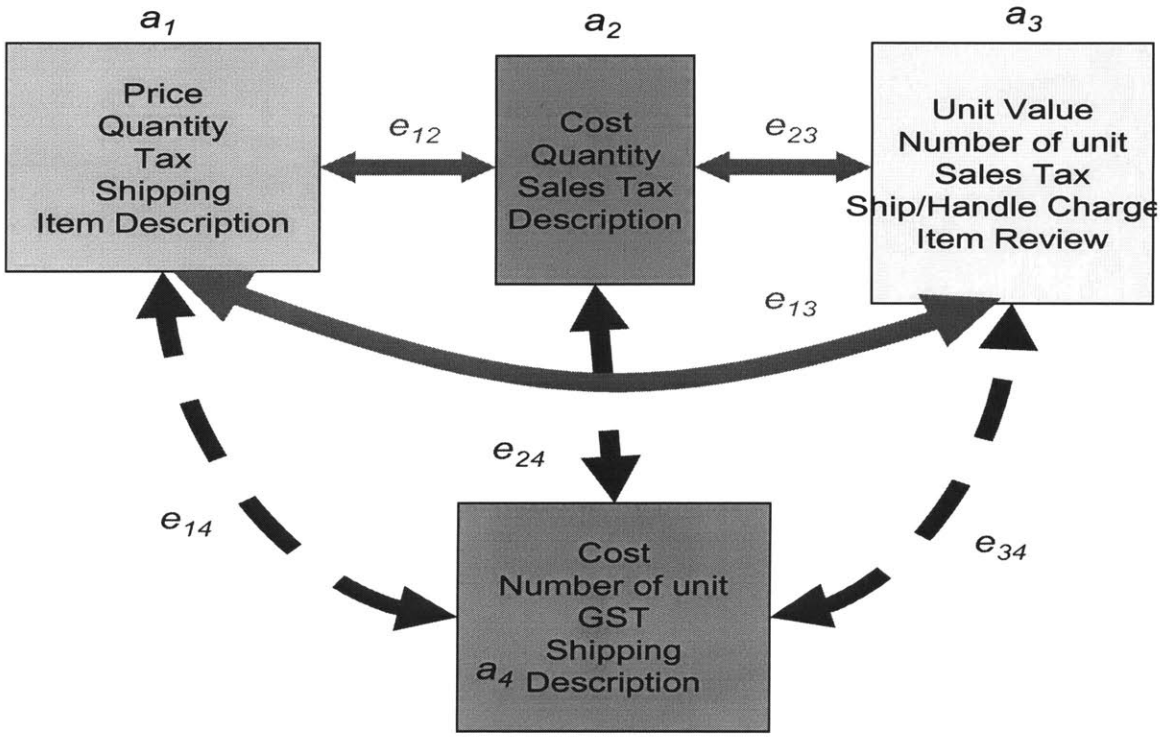


Figure 12: When adding new IS node to existing interoperable ISN

Thus, for Information Systems interoperability, all corresponding attributes within a network must be fully connected to possess a fully interoperable network system. We will now examine various network theories and its applicability to Information System Networks.

Chapter 5: Scale-Free Networks and the Barabási-Albert Model

It was originally perceived that such complex real evolving networks could not possibly arise out of any pre-determined sets of patterns. Networks such as the neural network, collaboration networks, public relations nets, citation of scientific papers, transportation networks, biological networks, food and ecological networks, social networks, the Internet and the World Wide Web are all examples of complex evolving networks that have only recently been shown to have general similar properties and structures that are a natural consequence of the principles underlying their growth.

The simplest and most common network initially used to explain the growth of networks was the Erdös-Rényi classical random network model (ER model). Their model states that the total number of nodes N in a network is fixed, and that the probability of two arbitrary nodes being connected together equals p. Conclusions drawn from the ER model state that the network would contain $pN(N\text{-}1)/2$ edges and that the degree distribution would be binomial, which means that

$$P(k) = \binom{N-1}{k} p^k (1-p)^{N-1-k}$$

And that the average degree is $\bar{k} = p(N-1)$. For large N, the degree distribution takes the Poisson form

$$P(k) = e^{-\bar{k}} \bar{k}^k / k!$$

This implied that the distribution rapidly decreases at large degrees. Also, the estimate for an average shortest-path length of such networks is $\bar{\ell} \sim \ln N / \ln[pN]$. Networks that followed such a Poisson degree distribution and statistically uncorrelated nodes are known as classical random graphs.

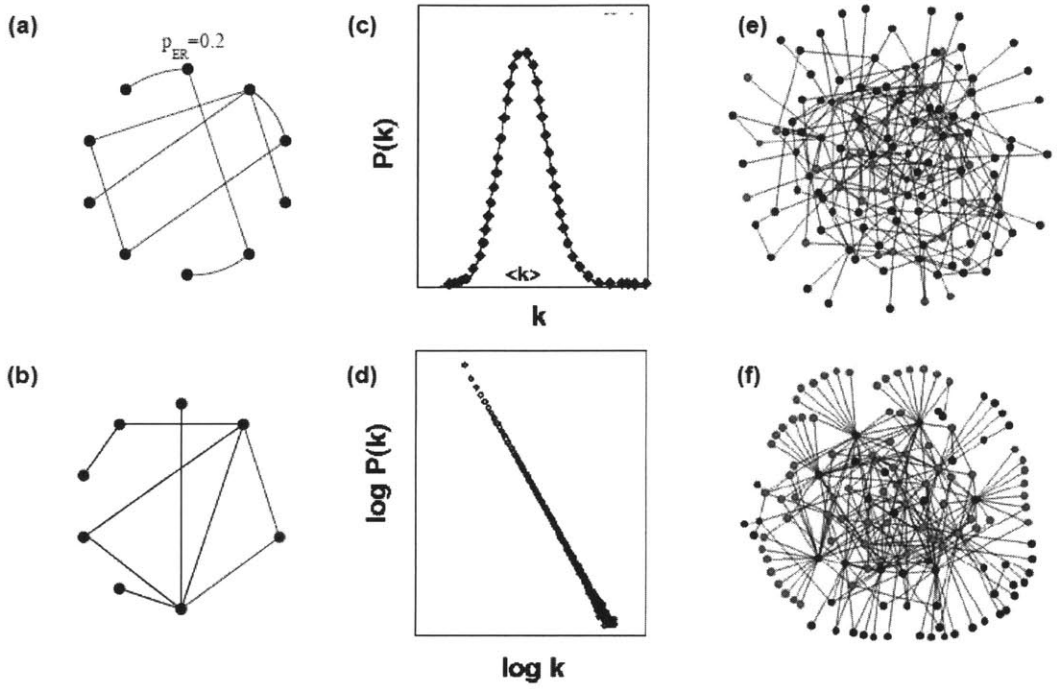


Figure 13: Comparison the Erdös-Rényi Network and the Scale-Free Model [4]

As the ability to garner information regarding complex evolving networks increased, it has allowed the realization that such networks, although complex and different in nature, often share several similar characteristics and properties. Research literature on several key characteristics of real networks indicated that there is a clear set of similarities shared by the complex real evolving networks.

Table 1: Characteristics of Real Networks Studied by Albert, Barabási [33]

Network	Size, *N*	*Average Degree <k>*	Average Path Length, *l*	Clustering Coefficient, C	Reference
WWW, site level	153,127	35.21	3.1	0.1078	Adamic, 1999
Internet, domain	3015	3.52	3.76	0.18-0.3	Yook et al, 2001
Movie actors	225,226	61	3.65	0.79	Watts and Strogatz, 1998
Co-authors, neuro	209,293	11.5	6	0.76	Barabási et al, 2001
Words, Synonyms	22,311	13.48	4.5	0.7	Yook et al, 2001
Power Grid	4941	2.67	18.7	0.08	Watts and Strogatz, 1998
Silwood Park food web	154	4.75	3.40	0.15	Montoya and Sole, 2000
C. Elegans	282	2.65	2.65	0.28	Watts and Strogatz, 1998

For example, the path length of networks does not scale with the network size. For the Movie actors network, even when considering a network size of 225,226 nodes, the average path length stays low at 3.65, which means that, on average, every actor is less than 4 degrees away from one another. Though the average path length varies with the type of network analyzed, it remains low compared to the network system size.

Clustering is also shown here, with the average cluster coefficient C high, higher than would be predicted under a classical random network. The various literature studied different co-authorship networks, and though the C varies from 0.066 to 0.726, most of the C values are high, indicating that there is a high tendency of nodes to cluster in real evolving networks as well.

New theories were established, namely that of the small world effect, which explains that average shortest path length is unusually small. Watts and Strogatz [34] noted that the average shortest-path length between nodes is small and of the order of the logarithm of their size, the clustering coefficient is much greater than allowed for under classical random graphs. This would theoretically explain the shorter path length of most

real networks. The WS model, also known as the small-world model, was proposed to demonstrate such a possibility, and the model is in the class of networks displaying a crossover from ordered to random structures and are those with 'small' average shortest-path lengths and 'large' clustering coefficients. The networks introduced by Watts and Strogatz are generally constructed from ordered lattices by random rewiring of edges or by addition of connections between random nodes.

One common feature of real networks is that often there are a few nodes that have an unusually high degree, while most other nodes are of low-degree, typically characterized by the concept of hubs and spokes. Examples can be seen when considering viral networks, where a highly connected hub, once infected, becomes an effective disease vector, and spreads the disease to a high percentage of other nodes [35].

A-L. Barabási and R. Albert, through their empirical studies on many large networks, demonstrate that these networks often display scale-free characteristics. They studied real networks in terms of their degree distribution, and noted they follow a power-law distribution, up to a very large degree. For example, for the World Wide Web, analyzing a network size of 325,729 nodes, they noted that the network follows a power-law distribution up to nodes with degrees greater than 900 ($k > 900$). The analyzed in-degree γ_{in} and out-degree γ_{out} is 2.45 and 2.1 respectively. In simpler terms, it means that for nodes with a lower degree than 900, the probability of the nodes in the network follows a simple distribution, $P(k) \sim k^{-2.45}$ (for in-degree networks)

As shown in Table 2, such power-law distribution is common in several networks, from the World Wide Web, the Internet (domain or router level), Co-authorship networks, citation networks, protein networks etc.

Table 2: Real Networks Analysis [33]

Network	Size, *N*	*Average Degree <k>*	Degree of scale-free cutoff, *K*	*indegree exponent, yin*	Outdegree [33]exponent, *yout*	Reference
WWW	325,729	4.51	900	2.1	2.45	Albert, Jeong, and Barabási, 1999
WWW	4x10^7	7		2.1	2.38	Kumar et al, 1999
WWW	2x10^8	7.5	4000	2.1	2.72	Broder et al, 2000
WWW,site	260,000			1.94		Huberman and Adamic, 2000
Internet, domain	3015-4389	3.42-3.76	30-40	2.1	2.1	Faloutsos, 1999
Internet, router	3888	2.57	30	2.48	2.48	Faloutsos, 1999
Movie actors	212,250	28.78	900	2.3	2.3	Barabási and Albert, 1999
Co-authors, neuro	209,293	11.54	400	2.1	2.1	Barabási et al, 2001
Sexual contacts	2810			3.4	3.4	Liljeros et al, 2001
Citation	783,339	8.57		3		Redner, 1998
Words, Synonyms	22,311			2.8	2.8	Yook et al, 2001
Metabolic E. Coli	778	7.4	110	2.2	2.2	Jeong et al, 2000

The scale-free network growth provided a theoretical basis for the growth and evolution of complex networks that more closely display the characteristics of real networks than previous network theories. Scale free networks are created from the observation that most networks have several common features and dynamics. An example of a scale-free network topology would be like Figure 13(f), where there are several nodes (denoted in red) that are highly connected, while most of the remaining nodes have a low degree (denoted in green and black)

Scale-free networks are significantly different from random connectivity networks in the presence of failure. If nodes fail randomly, scale-free networks behave even better than random connectivity networks, because random failures are unlikely to harm an important hub. Scale-free networks can be a disaster if the failure of nodes is not random. For instance an intelligent attacker can destroy the whole network by attacking key hubs.

Also, as mentioned earlier, the average path length of scale-free networks is short relative to its system scale.

The Barabási-Albert Model

Barabási and Albert argued that the scale-free nature of real networks is due to two generic mechanisms shared by many real networks. Unlike previous models which assumes a time-invariant fixed number N of nodes in the network. These nodes are then connected by edges according to the model used. Barabási postulated that real networks are open systems that grow by the continuous addition of new nodes. Real networks would usually start with a small nucleus of nodes, where nodes will increases throughout the lifetime of the network. Examples are the citation networks, where for a given topic, there would be a seminal set of initial papers from which additional research and papers will build on and cite respectively.

Next, most network models assume that the probability of two nodes being connected is independent of the nodes' degree. Real networks, however, often exhibit preferential attachment, such that the likelihood of connecting to a node depends on the node's degree. For example, in the citation network, a new research publication is more likely to cite well-known, highly cited previous research literature in the same field of study, rather papers that are less-cited and consequently less-known. Similarly, in an ISN, there can be certain Information Systems that, for reasons of length of period of existence, ease of interoperability, the importance or the universality of the data captured, are more likely to be semantically mapped than others.

The Barabási-Albert model therefore replicates these two factors and produces a network with a power-law degree distribution. The algorithm of the model is the following:

(1) Growth: Starting with a small number (m_0) of nodes, at every time step, we add a new node with m(m_0) edges that link the new node to m different nodes already present in the system.

(2) Preferential attachment: When choosing the nodes to which the new node connects, we assume that the probability that a new node will be connected to node i depends on the degree k_i of node i, such that

$$\prod_i = \frac{k_i}{\sum_j k_j}$$

After t time steps, there are $N=t+m_0$ nodes and mt edges.

Numerical simulations by Barabási and Albert indicate that thISN evolves into a scale-invariant state with the probability that a node has k edges follows a power law with an exponent $\gamma=3$, with the scale exponent independent of m.

The dynamic properties of the scale-free model can be addressed using various analytical approaches. The continuum theory proposed by Barabási [4] focuses on the dynamics of node degrees. Other approaches include the master equation approach of Dorogotsev, Mendes and Samukhin [36], and the rate equation approach of Krapivsky, Redner and Leyvraz [37] In this thesis, we will only examine the mathematical derivations of scale-free behavior using the continuum theory.

Continuum theory: The continuum approach introduced by Barabási, Albert and Jeong [3, 38] calculates the time dependence of the degree k_i of a given node i. This degree will increase every time a new node enters the system and links to node i, the

probability of this process being $P(k_i)$. Assuming that k_i is a continuous real variable, the rate at which k_i changes is expected to be proportional to $P(k_i)$. Consequently k_i satisfies the dynamical equation

$$\frac{\partial k_i}{\partial t} = m\pi(k_t) = \frac{mk_i}{\sum_{j=1}^{N-1} k_j} \tag{1}$$

The sum in the denominator goes over all nodes in the system except the newly introduced one; thus its value is $\sum_j k_j = 2mt - m$, leading to

$$\frac{\partial k_i}{\partial t} = \frac{k_i}{2t}$$

The solution of this equation, with the initial condition that every node i at its introduction has $k_t(t_i)=m$, is

$$k_t(t) = m\left(\frac{t}{t_i}\right)^{\beta} \text{ with } \beta = \frac{1}{2} \tag{2}$$

Equation (2) indicates that the degree of all nodes evolves the same way, following a power law, the only difference being the intercept of the power law. Using Eq. (2), one can write the probability that a node has a degree $k_i(t)$ smaller than k, $P[(k_t(t)<k]$, as

$$P[(k_t(t) < k] = P\left(t_t > \frac{m^{1/\beta} t}{k^{1/\beta}}\right) \tag{3}$$

Assuming that we add the nodes at equal time intervals to the network, the t_i values have a constant probability density $P(t_i) = \frac{1}{mt_0 + t}$

Substituting this into Eq. (3) we obtain

$$P\left(t_t > \frac{m^{1/\beta}t}{k^{1/\beta}}\right) = 1 - \frac{m^{1/\beta}t}{k^{1/\beta}(m_0 + t)} \tag{4}$$

The degree distribution *P(k)* can be obtained using

$$P(k) = \frac{\partial P[k_t(t) < k]}{\partial k} = \frac{2m^{1/\beta}t}{(m_0 + t)} \frac{1}{k^{1/\beta+1}} \tag{5}$$

predicting that asymptotically (t → ∞)

$$P(k) = 2m^{1/\beta}k^{-\gamma} \text{ with } \gamma = \frac{1}{\beta} + 1 = 3 \tag{6}$$

being independent of m, in agreement with the numerical results.

Average Path Length:

R. Albert and A.-L.-Barabási [39] performed a comparison study of average path lengths of two networks with an average degree $\langle k \rangle$=4 and a similar network size. As shown in Figure 14, the average path length of a random network, as shown by the solid line along the □ symbols, is numerically contrasted to that of a Barabási-Albert network, denoted here by the dashed line drawn along the ○ symbols. There is a shorter average path-length in B-A networks than in random networks, which would be favored in the implementation of Information System Network interoperability techniques such as the one employed by MITRE.

A lower average path length translates into lower search cost by the network when performing the DPQ and IIQ queries to identify semantically similar attributes within the interoperable network. This characteristic of scale-free networks ensure that fewer nodal and edge traversals are required to confirm or deny any relationship that may

exist between attributes in two separate information systems. A shorter path length also reduces the possibility of transitivity losses, as stated earlier in Chapter 4.

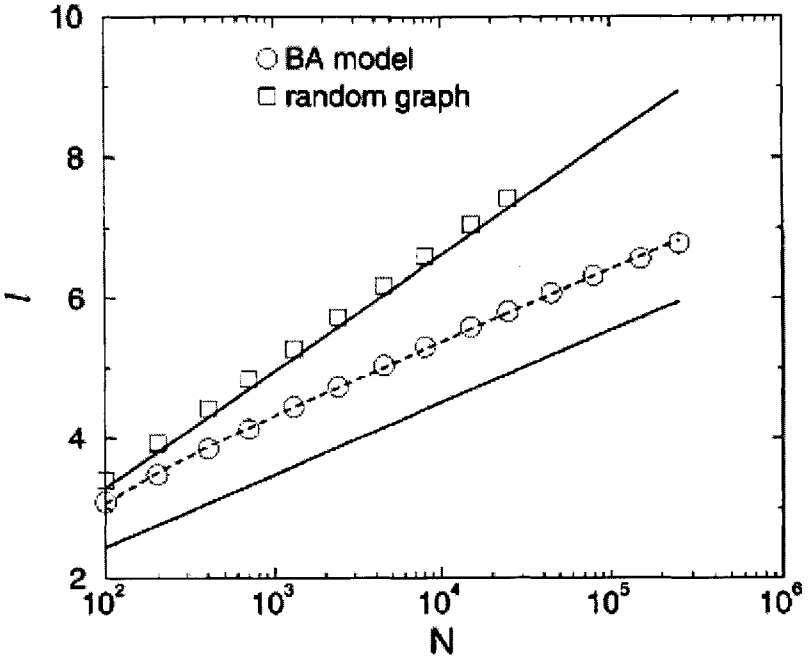


Figure 14: Characteristic average path length of B-A network vs random network of comparable size and degree [33]

Clustering Coefficient:

The clustering coefficient of a Barabási-Albert model also differs significantly from that of a random network of comparable size and average degree, as shown in Figure 15. Comparatively, the scale-free Barabási-Albert network has a clustering coefficient that is five times higher than that of the random network, and this factor increases with the number of nodes in the system. Once again, the scale-free model is of strong relevance to the Information System Network, since clustering often occurs in groups known as communities of interests (COI), such as a cluster of Information Systems referencing a particular subject topic. This clustering often occurs due to the

higher levels of similarities that exist between these nodes, and is not duly reflected in the characteristics exhibited in random networks.

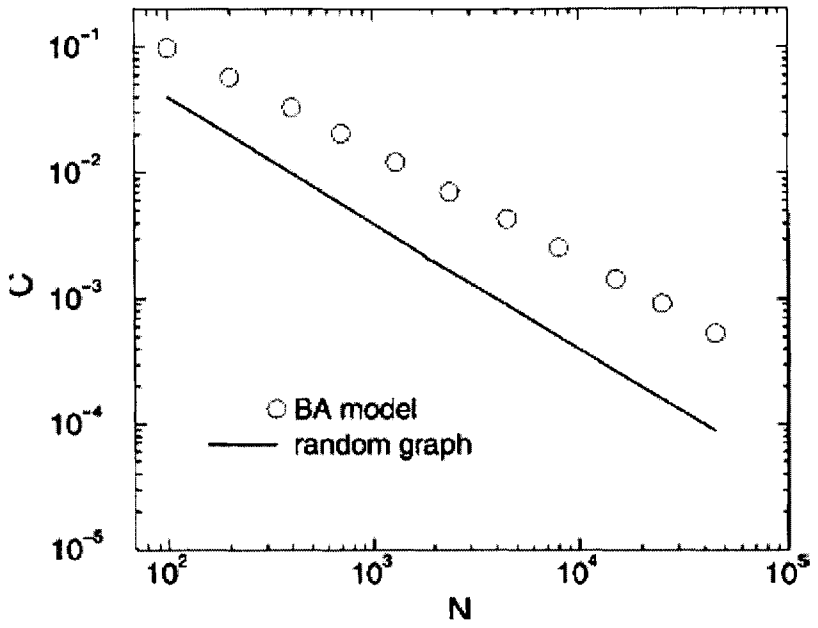


Figure 15: Clustering comparison of B-A network vs random network [33]

As shown, scale-free networks exhibit much of the similar growth patterns that exist in Information Systems, and would thus provide a good predictor for the kind of growth that would arise in ISNs. We will now examine the main points of conflicts between the Barabási-Albert method of scale-free growth and the growth patterns of an interoperable Information Systems network.

Chapter 6: Variations between Barabási-Albert and the Information Systems Model

As stated in Chapter 5, at every time step, a new node and m number of edges are added to the network, with an uneven or preferential attachment. The probability of attachment of a new node to an existing node a_i is described by: $\prod_i = \frac{k_i}{\sum_j k_j}$

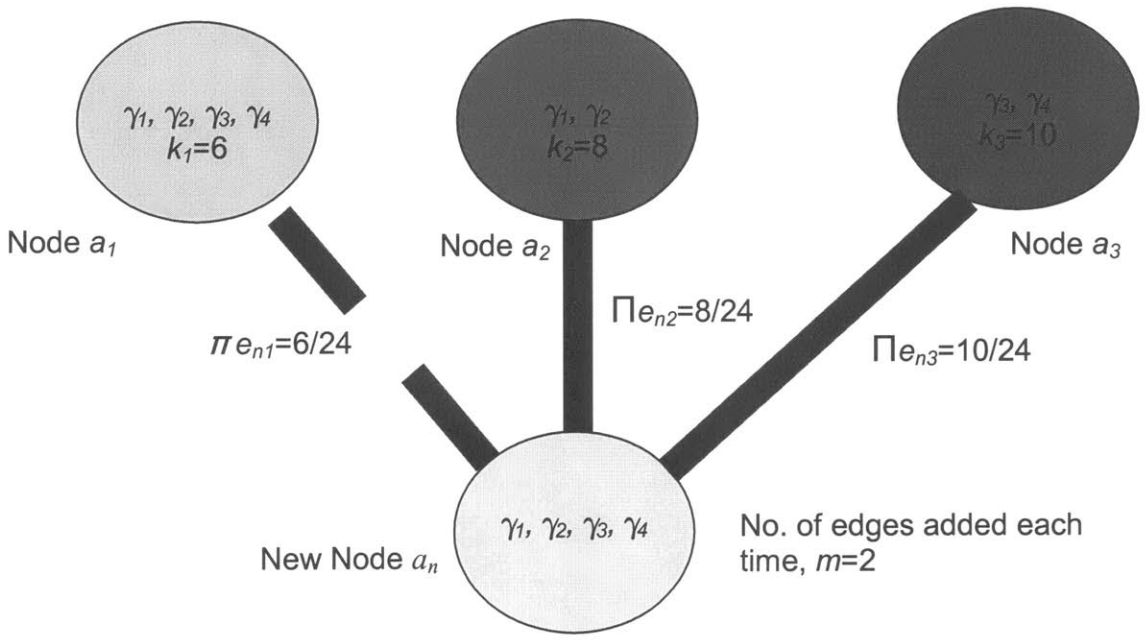


Figure 16: Example of a Barabási-Albert model of growth

Figure 16 shows an example of how the Barabási-Albert might be employed in the Information System context. Consider that an Information System Network exists with n number of nodes, of which we will examine 3 separate nodes a_1, a_2 and a_3, with degrees of 6, 8 and 10 respectively. Node a_1 has four attributes, γ_1, γ_2, γ_3, and γ_4, while nodes a_2 and a_3 have attributes γ_1, γ_2 and γ_3, γ_4 respectively.

When a new Information System node, a_n, is added, with all four attributes γ_1, γ_2, γ_3, and γ_4 as in node a_1, following the preferential attachment probability, where the

probability of a new node attaching to an existing node is dependent on the degree of the existing node over the total degree of the network, one could see that the edge e_{n3} between the new node a_n and the existing node a_3 is favored over the edge e_{n1}. It is important to note that even though nodes a_1 and a_n can be very similar, where a connecting edge would present the best possible interoperability improvement to the network, under this algorithm, such a connection would be disfavored. Rather, a less valuable connection, in this case, e_{n2} and e_{n3}, with only two out of four possible attributes, will be picked instead of e_{n1}.

Although, as stated in this simple example, a theoretical application of the B-A model of growth is possible in the Information System Network, for a practical application within the ISN context, it is necessary to understand the underlying differences between the two models and analyze if these differences can be bridged. The Barabási-Albert model makes several assumptions of the network nodes, before enforcing a growth algorithm on the network. We will now examine these assumptions and determine whether they are applicable to the IS model. Of those differences, we will examine the extent of the incompatibilities and appropriately disregard superficial inconsistencies, and focus only on resolving the main issues.

Problem 1: Nodal Similarity

The biggest conflicts resolve around the concept of nodal depth. In the Barabási-Albert model, nodes have zero depth, i.e. they are completely similar from each other and share the same exact features from each other. Nodes therefore have basically the same basis to compete with one another for edge attachments; the only permitted distinction

between the nodes is their present connectivity to other nodes, or degree. Each node has a non-fixed number of edges attached to it, following the connectivity probability of connecting a new node to an old node.

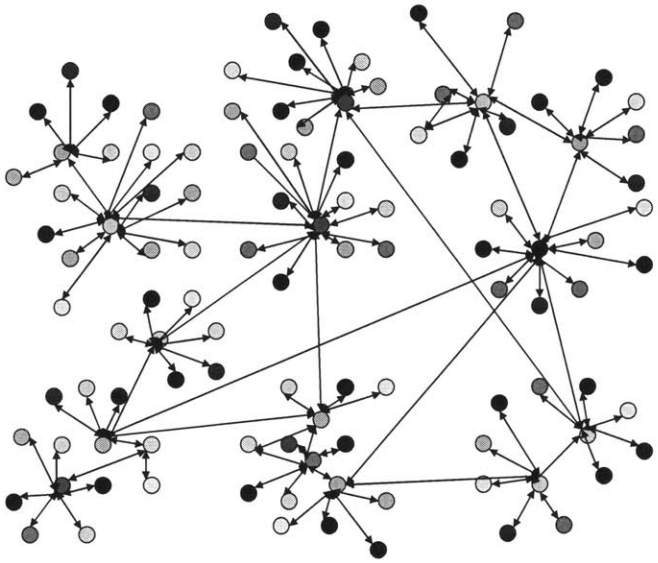


Figure 17: B-A Growth on Heterogeneous Information Systems

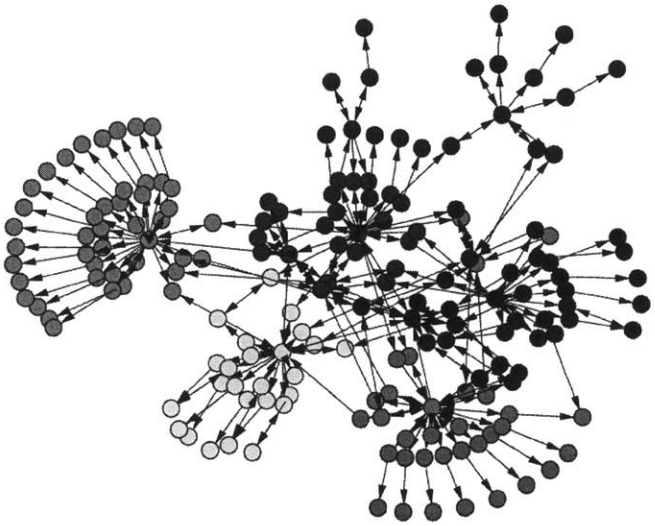


Figure 18: Desired Heterogeneous Information Systems Network

Information systems tend to grow as a set of independently developed data sources. As shown in Chapter 2, these independently developed data sources have their own schema, with contextual, structural and representational individuality. Thus, heterogeneous information systems are predominant.

Forcing a model of growth on an Information System Network that disregards the individuality of Information Systems will result in an unsustainable, impractical and unfeasible solution. Figure 17 shows the expected results of a Barabási-Albert growth in Information System Networks. Nodes of the same color represent Information Systems with many data attributes that can be mapped to one another. In this topology, the ISNs do not take advantage of this nodal variation, and so correlated nodes are not mapped to

each other. This results in reduced information flow across the network. A more desired outcome would be Figure 18, where the network recognizes that closely correlated IS nodes exist, and clusters them together. This clustering allows nodes with a similar set of data attributes to be closely connected, allowing easier interoperability mappings and better overall information exchange.

One possible remedy would be to examine extensions of the original scale-free model proposed by Barabási and Albert, and determine which alternative theories best relate to the Information System Network. This important issue will be addressed in Chapter 7, when we examine the Competitive and Multi-Scaling Model proposed by Bianconi and Barabási as an extension of the original scale-free model.

Problem 2: Edges Similarity

The Barabási-Albert algorithm specified that any two nodes can be connected, and the probability of connectivity is only attributed to the degree of that existing node. However, as elaborated in Chapter 4, mappings can only exist between Information Systems when there are at least one semantically similar attribute shared between them. Attempting to replicate the B-A model in Information System Networks will create several problems. Firstly, un-necessary, zero-value added edges will be added to the ISN that would place additional burdens to network maintenance. Secondly, by the pure emphasis of preferential attachment due to the degree distribution, nodes with high degree will be sought after by the new node for edge attachment. This buries the benefits added to the network through the attachment to essential but lower degree nodes.

Essential nodes in this case will refer to nodes that have a high percentage of semantically similar attributes with the new node introduced into the system.

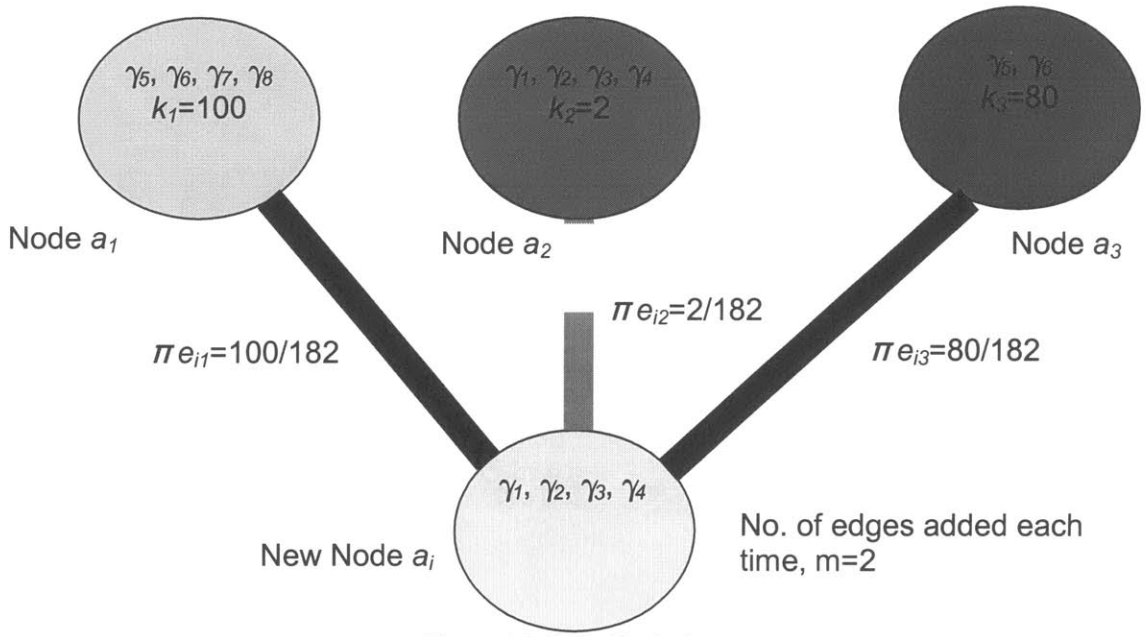


Figure 19: Edge Variation

As an example of how Information System Networks differ from the B-A model, consider the following situation as shown in Figure 19. When a new node a_i is added to the network, preferential attachment would dictate that edges will be formed with nodes a_1 and a_3, rather than a_2. However, when noting the attributes present within each of the Information Systems, nodes a_1 and a_3 have no common attributes with a_i, thus edges cannot exist between them.

In the B-A model, as all nodes and all edges are identical, this allows any node to be connected to any other node, i.e. there is no prohibition that a certain node cannot be connected to another particular node. Though this implication is obvious, it represents a significant difference between the Barabási-Albert model and the Information Systems model. Thus an accurate network model of Information Systems must account for the

condition that edges can only be formed between nodes that share conceptually similar attributes.

Another reason is that, over time, certain links degrade while others strength, depending on the usefulness and strengths of the existing relationships. In social networks, as time passes, people don't remain in contact with loosely affiliated acquaintances, and thus these former links wither and disappear, while stronger links remain. All edges in the Barabási-Albert model have the same strength and longevity, and the usefulness and costs of ISN interoperability mappings are not accounted for. In real Information System Networks, maintaining connections between nodes are costly, especially if its utility has been superseded by other newer edges. Utility is also reduced when the Information System node changes.

One way to compensate for this inadequacy is to establish a measure of quality for links between the nodes in the ISN that accounts for the strength of the fitness between the two nodes. A strong link is a connection between two highly correlated IS nodes. A criterion can be added to the network construction, stating that only links over an acceptable threshold of acceptability will be created. This will be further elaborated in Chapter 8, when an aggregated Information System Network solution is proposed.

Problem 3: Non varying number of edges added per time-step

Under the Barabási-Albert mode, a new node is introduced into the existing network at each time-step. Alongside the new node, a fixed number of new edges, m, are created between the new node and the existing nodes within the network. The number of nodes added each time step is arbitrary, since it does not affect the overall degree distribution of the network. The only criterion is that the number of edges added each

time step must be less than m_0, the initial number of nodes in the network. Thus at time t, there are a maximum of tm edges present in the network.

For an Information System Networks, there are more restrictions imposed on the number of edges added each time-step. Firstly, the number of edges added each time step is dependent on the type of node added. Recall that to ensure the full interoperability between IS in a network that has transitivity of attributes, all attributes must be fully connected to their semantic counterparts. For a node that has very few common attributes, to ensure that the node's attributes are fully mapped, that would most likely encompass just the mapping between a few nodes. However, should a Information System supernode appear, which has attributes that are covered separately by all the existing nodes, it is necessary to map this supernode to all the existing nodes to ensure full interoperability, thus conflicting with the non-variant edge addition. This would occur in real networks when the intention of the newly created node is to aggregate all the present data and meta-data into a central node to facilitate easy information access and/or data manipulation. This could potentially pose a problem when relating ISNs to the Barabási-Albert model. We will examine the likelihood of such an event and analyze the effects of this feature.

We will now propose an alternative model to the Barabási-Albert model. This new model will address the failures of the original B-A model and will be more applicable in the context of Information System Networks.

Chapter 7: Competitive and Multi-Scaling in Evolving Networks

Nodes have an inherently different ability to compete for links. In real networks, often a new node introduced into the network tends to gain an uncharacteristically large number of links than can be purely predicted by degree distribution alone. On the World Wide Web, some URLs acquire a large number of links within a short timeframe, due to the content or marketing of the website. Seminal research papers also acquire a large number of citations over a very short duration.

This is the main criticism of the original Barabási-Albert model: The model does not possess the capability to provide a proper assessment and subsequent network growth of systems where not all nodes are equally successful in acquiring links. Furthermore, one consequence of the B-A model is that the oldest nodes in the network tend to have the most number of links, due to the simple growth mechanism of attributing edges to nodes of higher degrees. In the consideration of ISNs, this might be true, as evidenced by the importance of decades-old legacy databases that are still relevant in today's context. But there is no compensating effect under the original B-A model for IS supernodes, when a node that accumulates all present data schemas appear to form a centralized hub between existing information systems.

To that end, a new model, the competitive and multi-scaling model proposed by Bianconi and Barabási [5], was formulated to address these shortcomings. To acknowledge that the nodes are no longer identical, a new parameter for fitness, η_i, is

assigned to each node. This new fitness parameter is assumed to be unchanged over time. η_i is chosen from the distribution $\rho(\eta)$ and is used to account for the inherent quality present within each node that determines how well that node competes for links.

At each time step, a new node a_i with fitness η_i is added to the network. Also, a fixed number of links, m, are connected from the new node a_i to the nodes already present in the system. The probability that that a new node will connect to a node i already present in the network is: $\prod_i = \frac{\eta_i k_i}{\sum_j n_j k_j}$

Thus the characteristics of the network is dependent on the distribution of η, or $\rho(\eta)$ as well as the degree distribution.

Mathematical Derivation of Competitive Network Model Outcome

Using the continuum theory, as explained in Chapter 5, we can see that a node a_i increases its degree k_i at a rate proportional to the probability that a node will attach to it, giving $\frac{\partial k_i}{\partial t} = m\frac{\eta_i k_i}{\sum_j \eta_j k_j}$ (7)

From this, assume that similarly to the scale-free model the time evolution of k_i follows a power law, but with multi-scaling incorporated. Multi-scaling implies that time dependence of a node depends on the fitness of the node. Mathematically, it states that the dynamic exponent depends on the fitness η_i, $k_{\eta_i}(t,t_0) = m(\frac{t}{t_0})^{\beta(\eta_i)}$ where t_0 is the time when the node is introduced to the network. We can observe that $0 < \beta(\eta) < 1$, since a

node always increases in degree over time (>0), but cannot increase more than the number of edges added per time (<1).

We calculate the mean of the sum $\sum_j \eta_j k_j$ over all possible quenched noise $\{\eta\}$.

Since each node is born at different times t_0, the sum over j can be written as an integral over t_0:

$$\left\langle \sum_j \eta_j k_j \right\rangle = \int d\eta \rho(\eta) \eta \int dt_0 k_n(t,t_0) \\ = \int d\eta \eta \rho(\eta) m\left(\frac{t - t^{\beta(\eta)}}{1-\beta(\eta)}\right) \tag{8}$$

With $\beta(\eta) < 1$ and with t→ ∞, $t^{\beta(\eta)}$ can be neglected compared to t, giving

$$\left\langle \sum_j \eta_j k_j \right\rangle^{t\to\infty} = Cmt(1 + O(t^{-\xi}) \tag{9}$$

$$\varepsilon = (1 - \max(\beta(\eta)) > 0$$

where $C = \int d n \rho(\eta) \frac{\eta}{1-\beta(\eta)}$

Using this and the notation $k_\eta = k_{\eta_i}(t,t_0)$, (1) can be rewritten as

$\frac{\partial k_\eta}{\partial t} = \frac{\eta k_\eta}{Ct}$, which can be solved, giving $\beta(\eta) = \frac{\eta}{C}$ (10)

Substituting $\beta(\eta) = \frac{\eta}{C}$ in (3), we have:

$$1 = \int_0^{\eta_{\max}} \partial\eta \rho(\eta) \frac{1}{\frac{C}{\eta} - 1} \tag{11}$$

With this we will discuss three fitness distributions that affect the scale-free characteristics of the network. The three fitness distributions discussed pertain to three

different scenarios: (1) When all nodes are identical, (2) When nodes follow a finite uniform fitness distribution, and (3) When there is infinite support for the fitness distribution.

(1) Identical Nodes

For $\rho(\eta) = \delta(\eta - 1)$, when all fitness is equal, it reduced exactly to the scale-free model, the results can be seen in Chapter 5. The probability connectivity distribution follows perfect scale-free behavior such that $P(k) \sim k^{-3}$

(2) Finite Uniformly Distributed Fitness

Competitive networks become more interesting when we consider a uniform fitness distribution, where nodes with different fitness compete for edges attachment. Consider when $\rho(\eta)$ is uniformly distributed over the interval [0,1]. The constant C in (11) can be determined, where exp[-2/C]=1-C, whose solution is C*=1.255. Since $\beta(\eta) = \frac{\eta}{C}$, each node will have a different dynamic exponent.

If $\rho(\eta)$ is chosen uniformly from the interval [0,1], the probability connectivity distribution of the network will be:

$$P(k) \propto \int_0^1 d\eta \frac{C^*}{\eta} \frac{1}{k^{1+C^*/\eta}} \sim \frac{k^{-(1+C^*)}}{\log(k)} \tag{12}$$

This states that the connectivity distribution follows a generalized power law, albeit with an inverse logarithmic correction. C in this case is 1.255, so following Equation (12) the degree distribution can be generalized as thus: $P(k) \sim k^{-2.255}$

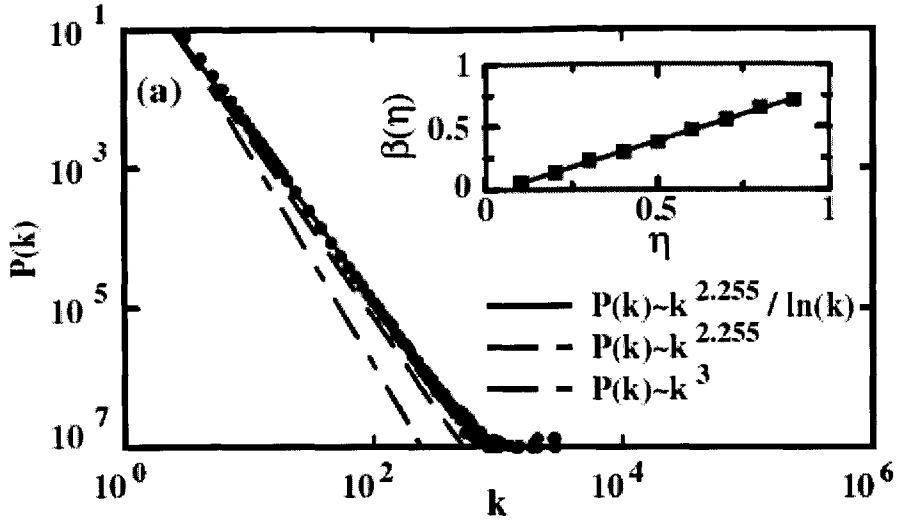


Figure 20: Degree Distribution of Finite Uniform Fitness Distribution [5]

Figure 20 shows the outcome of numerical simulations of the degree distribution in the competitive network model. In thISN growth analysis, $m = 2$, $N = 106$ nodes and nodal fitness is uniformly distributed. The top solid line that is lined with dots corresponds to the model predictions, with the exponent γ of the scale-invariant probability equal to 2.25. The dashed line corresponds to a simple fit $P(k) \sim k^{-2.255}$ without consideration of a logarithmic correction. The long-dashed curve correspond to $P(k) \sim k^{-3}$, as predicted by the scale-free B-A model in the first scenario, when all nodes are identical.

(3) Infinite Support Fitness Distribution

Infinite support in fitness distributions indicate that at any time there will be a finite probability that a new node will have a fitness $\eta >$ maximum fitness η_{max} will be added to the system. This implies that the fitness scale keeps growing without bound. As

η_{max} keeps growing in the network over time, this is indicative that the fitness distribution function has a time dependent aspect. As stated earlier, a time step occurs when a new node is added to the system, i.e. time and system scale are associated with each other. The fitness distributions in infinite support systems are thus scale-varying or scale-dependent.

As an example of an infinite support fitness distribution, consider that fitness distribution $\rho(\eta)$ follows an exponential curve. For a $\rho(\eta)$ following an exponential distribution, $\rho(\eta)=e^{-\eta}$, $k(t)$ starts to scale as a power of $ln(t)$, indicating that it is no longer scale-free.

$$k(t) = k(t_0)\left(\frac{\ln(t)}{\ln(t_o)}\right)^{\xi(\eta)} \tag{13}$$

Thus, not all $\rho(\eta)$ distributions will result in a power law time dependence and connectivity distribution. This result is important to our discussion of scale-free networks in the information systems model, as the fitness distribution of the nodes is now a determinant of the characteristics of the information systems network. Depending on the type of fitness distribution, the competitive network model retains the characteristics of a power-law distribution, indicative that scale-free behavior is still retained.

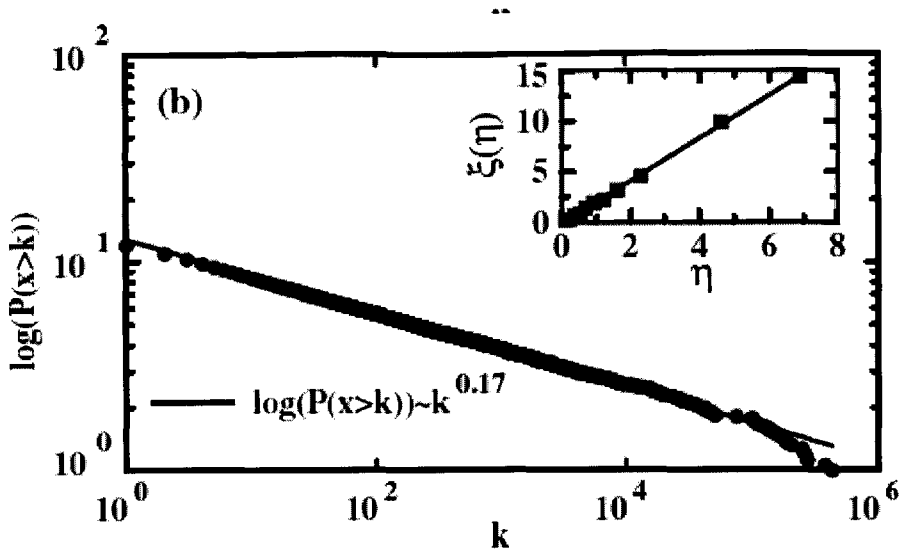


Figure 21: Degree Distribution of Infinite Support Fitness Distribution Networks [5]

Figure 21 shows the outcome of numerical simulations when modeling a network of similar size and number of edges as the model in Figure 20, but with an exponential fitness distribution. One can see that there is a more even spread of nodes of varying degrees, as compared to a scale-free network, where there is a significant drop-off in the number of nodes as the degree increases.

With the competitive network model, we can now address the biggest flaw that was inherent in the original B-A model; that not all nodes are equally competitive in attracting new nodes. We will now examine how the competitive network model can be applied to our Information System Network context.

Chapter 8: Solution approach

Instead of the original B-A model, we used the competitive, multi-scaling network model that designates a fitness quality to each IS node. Starting with an initial number of nodes, at every time step a fixed number of edges, *m,* are added from the new node to the existing nodes, where *m* is less than m_0, the initial number of nodes in the network. The preferential attachment is determined by the composite probability $\prod_i$ of connecting a new node a_n to an existing node a_i. $\prod_i = \frac{\eta_i k_i}{\sum_j n_j k_j}$

When an IS node is introduced into the network, it brings into the Information System Network a unique set of data attributes that models the data it stores and disseminates. When performing semantic interoperability, the edges between nodes are actually conversion mappings between semantically similar attributes that exist in both nodes. This means that an edge can only exist between nodes if there is an overlap between the attributes in the first node and the attributes in the second node, i.e. e_{in} exists if and only if $\sum \gamma_{a_i} \cap \sum \gamma_{a_n} \neq 0$

Attributes therefore forms the basis for an Information System Network to have a fitness comparison measure. The fitness quality of a node has a dependent relationship with the number of attributes present in the node's schema. It is apparent, however, that a direct relationship between the number of attributes within an IS node and its fitness for connectivity does not adequately model the complexity of fitness determination, i.e. $\eta_i = Count(\gamma_{a_i})$ is not an accurate modeling of fitness. A wine IS, with thousands of

attributes modeling wine aspects, has little in common with an air mission IS, and thus would not be a very good fit with the air mission IS.

A better unit of measurement of an existing node a_i to have links with a newly introduced node a_n, would therefore involve the number of semantically similar attributes that are shared between a_i and a_n. Mathematically, η_i, the fitness of an existing node a_i in relation to the new node a_n, is:

$$\eta_i = \frac{Count[\gamma_{a_n} \cap \gamma_{a_i}]}{Count[\gamma_{a_n}]} \tag{14}$$

Inadequacies of Barabási-Albert model addressed by the Competitive Network Model

This modeling of fitness will account for all the inadequacies of the Barabási-Albert model implementation on an Information System Network.

(1) Removal of attachment of incompatible IS nodes

This accounts for the scenario when incompatible nodes (with zero information in common) attempts to form an interoperability mapping. As equation (14) stipulates, $\eta_i = 0$ when there are no semantically similar attributes between the nodes. The composite probability of having that connection under the competitive network model, $\prod_i$,is also equal to zero, indicating that such a connection will not be formed. Thus it is a much more accurate depiction of IS nodes than would be possible under the B-A model.

(2) Infinite Support for Fitness Distribution

The bounds of the fitness η_i is [0,1] indicating that there is no infinite support in this system. Infinite support would have led to the eventual degradation of the scale-free aspect of this growth. However, due to the definition of the fitness level, infinite support can still occur when the number of unique attributes in the system continues to grow with time.

(3) Uneven number of edges added each time step

Barabási's model sets a fixed number of edges added per time step, since the calculations show that degree distribution is independent of the number of edges added each time. However, the interoperability criterion determined in Chapter 4 states that to ensure full interoperability in the network, the attributes of each new IS node introduced into the system must be mapped to at least one existing node's attributes. Clashes between these two networks features will occur.

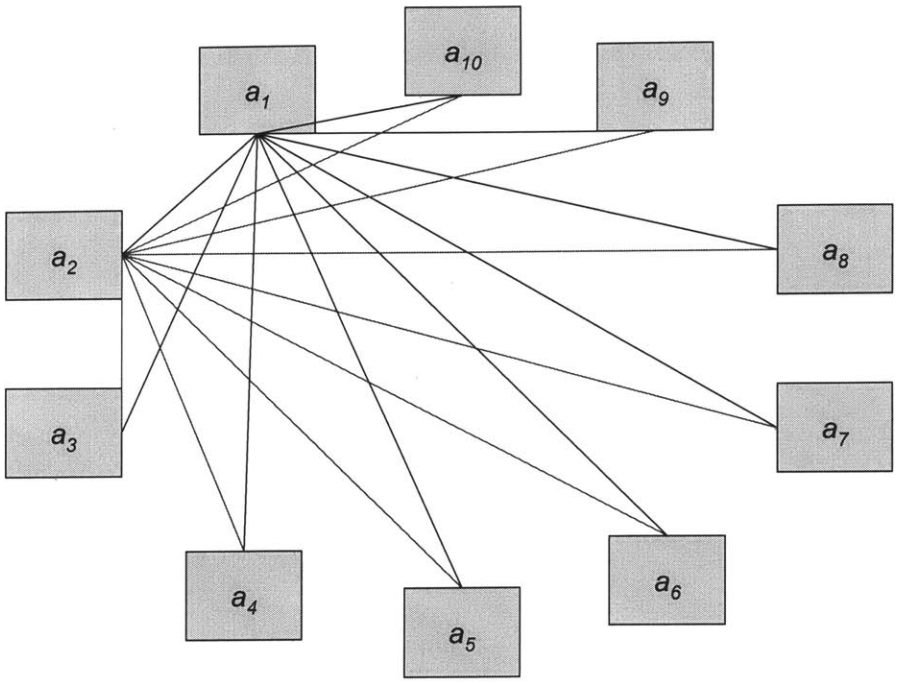


Figure 22: N-squared behavior in transitive ISN

Consider a node a_1 that exists in a network of 10 nodes in Figure 22. For this node to attach itself to every other node in the network to ensure interoperability, it will need to have at least nine, or *n-1,* conceptually distinct attributes. Each of the *n-1* distinct attributes (γ_{a1_1}, γ_{a1_1}, ... , γ_{a1_1}) in a_1 will need to have a corresponding similar attribute in all the other nodes in the system to perform an attachment. Thus for a_1 and a_2, there exists an edge e_{12} that attaches the two nodes together for that specific similar attribute. With the attachment of the *n-1* attributes in a_1, we now have *n-1* attachments, connecting a_1 to all the other nodes.

Now consider a_2. a_2 has an existing edge to a_1, e_{12}, and a_2's degree is currently 1. To achieve a degree of *n-1*, it needs to have an additional *n-2* distinct attributes that are also distinct from a_1's *n-1* distinct attributes. This distinction is necessary, since if the attributes in a_1 and a_2 have similarities in addition to the single conceptually similar

attribute that produced e_{12}, transitivity over e_{12} would mean there would be less than n-2 additional edges needed that connects a_2 to the rest of the edges.

As this continues for the entire network, to have n^2 connections between all the nodes in the network, the system would require at least $n(n-1)/2$ conceptually unique attributes within the networks. The distribution of the attributes would be such that each node a_i has at least n attributes, and that between nodes, for all nodes, they only share a single conceptually similar attribute.

As an example, for networks of two nodes, there is only one link between them, and they require at least one conceptually unique attributes. For networks of three nodes, there are three edges to fully connect the nodes, and it will require at least three conceptually unique attributes between the three nodes. Achieving two or three conceptually unique attributes, uniformly distributed between two or three nodes is fairly common. Thus initially for any network, it will resemble a fully connected network. However, as the number of nodes goes up, the probability of having $n(n-1)/2$ conceptually diverse attributes uniformly distributed between the n nodes markedly decreases.

Though this situation is unlikely to occur, one can take steps to prevent a n^2 network scenario. One is to restrict the number of semantically similar attributes that would be covered across the entire network. This is based on the fact that, as network size increases, the number of common attributes that are common across the entire network of IS nodes decreases, until the most general set of attribute types is obtained. These set of attributes varies with the ISN under construction. For an ISN built for military purposes, the set would be attributes relating to *Time*, *Geographical Location* and *Event Type*. As

such, infinitely increasing the number of attributes to be made interoperable reaches a diminishing point of value after it passes these general attributes, while the complexity increases exponentially (n^2 edges).

The best solution is to establish a fixed set of attributes, $\sum\gamma^*$, that an interoperability agent considers when performing interoperability mappings between IS nodes. These attributes will derive from those in common use within the context of the Information System Network in construction. Initially when the network size is small, $\sum\gamma^*$ will be larger than n. However, as network size increases, $n >> \sum\gamma^*$.

The determination of the number of edges added, m, can be set to the average number of mappings needed to cover $\sum\gamma^*$. In Figure 12, there is a network with $\sum\gamma^*=5$. Every node contains between four to five semantically similar attributes. When a new node a_n, with all five attributes, enters the network, an interoperability agent requires at most two mappings to cover the $\sum\gamma^*$ attributes in the system. So in this case, m will be set to two.

Overall Conditions for Information System Network growth

As a recap, we will state out all the conditions necessary that will ensure a scale-free interoperable Information System Network:

(1) Establish a set of attributes in common use within the space of Information System Network, $\sum\gamma^*$. This set of attributes will be the attributes that is targeted for interoperability in the network. $\sum\gamma^*$ can grow over time, so long as

the criteria $\sum \gamma^* << n$ is maintained. However, as shown in the infinite support example in Chapter 7, if $\sum \gamma^*$ continues to grow with respect to time, the network will deviate more and more from ideal scale-free behavior. This is as time is related to network size, thus time dependency indicates scale-dependency in preferential attachment probability, indicating that the network is no longer scale-free.

(2) Establish a fitness measure for acceptability of a semantic interoperable connection between two IS nodes, η_i. The fitness measure is to be based on the number of semantically common attributes shared between the two nodes in question.

(3) Ensure that all the attributes of a new IS node within the set of attributes targeted for interoperability purposes, $\sum \gamma^*$, are mapped to at least one semantically similar attribute that is currently existing in the network.

(4) Establish a fixed number of edges added to the network at each new node's inclusion into the network. This set number can be related to $\sum \gamma^*$, such as setting the fixed number of edges added to the network, m, equal to the average number of edges needed to map $\sum \gamma^*$.

In the best case, if every new IS node only required at most m edges to map $\sum \gamma^*$, the number of edges in the entire network will be $n \times m$.

In the worst case, if every new IS node needed $\sum \gamma^*$ edges to interoperate the attributes in its schema, the number of edges will be $n \times \sum \gamma^*$, where $\sum y^* << n$.

(5) Adhere to the competitive network model's composite probability of preferential attachment, $\prod_i = \frac{\eta_i k_i}{\sum_j n_j k_j}$, when considering which existing IS nodes should form an attachment with a new IS node.

Example of Information System Network Growth through the Use of Competitive Network Model

As a simple example, consider a network that has altogether seven distinct semantically-similar attributes that are to be mapped within the system. This occurs in communities of interests, where the cluster of nodes share similar interests and capture the same set of data attributes, though with different contextual information. Consider that every attribute has the same level of fitness as every other attribute, which means that nodes with three similar attributes have the same level of competitiveness for edge attachment. Assume a uniform discrete fitness distribution across the ISN, indicating that the probability of having a node with three semantically similar attributes is the same as having a node with seven semantically attributes. The fitness distribution will therefore resemble this:

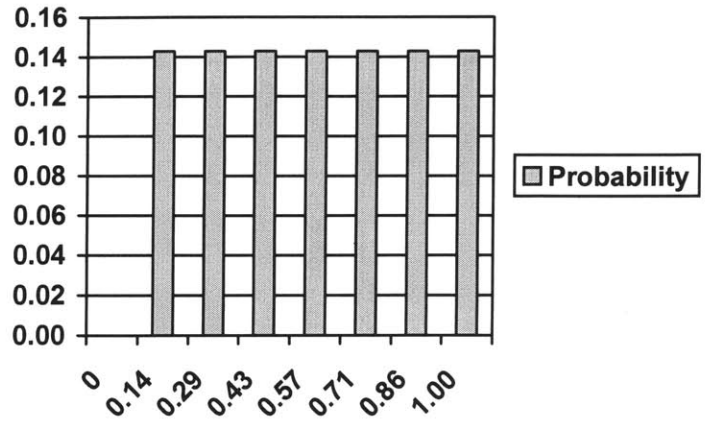


Figure 23: Theoretical Fitness Distribution of 7-Attribute ISN

Next, we state that the median number of edges that need to be established between a new node and existing nodes to ensure interoperability of these seven nodes is two. At every time step, two edges will be extended from a new node to the existing nodes in the network.

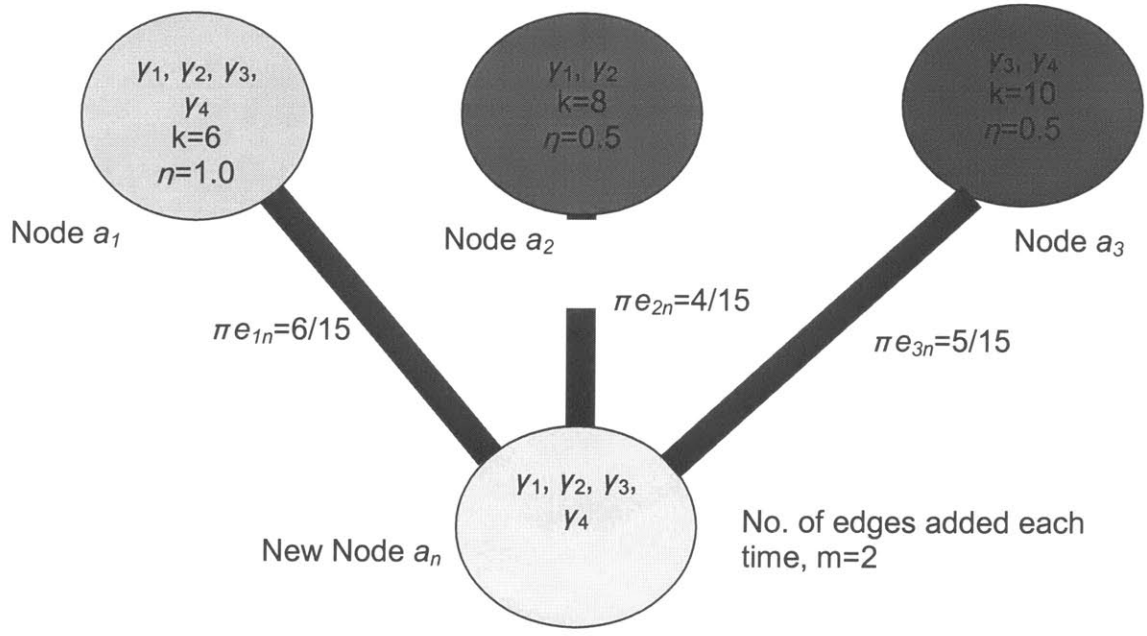


Figure 24: Information System Fitness Model

In Figure 24, when the new node a_n is added, the probability of a link is dependent on both the degree as well as the fitness of an existing node. Note that as the node a_1 has a sufficiently high fitness that even with a lower degree than nodes a_2 and a_3, the composite for both the degree and the nodal fitness means that the probability of attachment e_{1n} is higher than e_{2n} and e_{3n}.

Next, as m has been determined *a priori* to be two for this particular system, the two most likely edges to be added in this network are e_{1n} and e_{3n}. As can be seen, all the attributes in the new node a_n have been interoperability mapped to the existing attributes in the system, indicating complete interoperability still exist in the network.

Using the above fitness distribution, we can calculation from equation (11) that C=1.37101, and $P(k)$, while depending on the nodes of different fitness levels, maintains a generalized power law distribution where $P(k) \sim k^{-2.37101}$. Average path lengths and clustering coefficients will also extend from this scale-free growth. Apart from achieving a scale-free growth in this Information System Network, we have also maintained the contextual relevancy of our proposed solution, by adequately addressing all the unique features and issues arising from an Information System Network.

Chapter 9: Conclusions and Future Discussions

Information Systems, by their inherent nature, do not form networks easily. These features that inhibit networks include: 1) Lack of data transitivity between ISs, resulting in little to no value of having large scale ISNs, 2) Distinctiveness of individual ISs, with their unique set of data attributes and structure, making any interoperability mapping between two ISs difficult to create and maintain, and 3) Scaling problems associated with ISNs.

Using the MITRE Semantic Interoperability Technique, most of the features that inhibit the organic growth of Information Systems in a network are addressed or diminished in significance, by adding a layer of intelligence on top of the network. Central to the features addressed is the allowance of transitivity of data attributes between Information Systems. Data attributes transitivity is of the utmost importance in Information System Networks as they are the primary enablers of information exchange between disparate Information Systems. Transitivity on the data attribute level also implied that complete semantic interoperability of Information Systems did not require a fully connected network (or n^2 connections). Rather, so long as the criteria that all relevant attributes in every IS node are mapped to their semantically similar counterparts, complete interoperability is still maintained.

We next analyzed the various theories relating to complex evolving networks. Random classical theories proposed by Erdös and Rényi is no longer viewed as adequate in addressing the complexities of growth in evolving networks as well as the characteristics that are inherent in real complex networks. These characteristics are

namely: Degree Distribution of network is independent of system size (Scale-Free), the average path lengths of large networks remain small (Small-World) and large clustering coefficients. Instead we examined the scale-free network theory proposed by Barabási and Albert that examines the growth mechanism of real networks and postulates a preferential growth algorithm based on the degree distribution of the network. Using the continuum theory, we derived that scale-free behavior is a direct result of such preferential growth. Numerical simulations using the preferential attachment probability reinforced the fact that such growth and preferential attachment in networks resulted in a scale-free network.

However, the Barabási-Albert model failed to model one of the most important aspects in Information Systems: IS node uniqueness. This uniqueness also affects how well each IS node is able to compete for edge attachments. An extension of the original B-A model, the competitive and multi-scaling network model, addresses this issue, through the use of a fitness quality measure for nodes, η_i. The fitness quality measure in the competitive network model can be correlated to the number and type of attributes present within the data ontology of each Information System node. With the preferential attachment probability now a composite function of an existing node's degree and fitness, additional restrictions must be placed on the fitness distribution range, so as to avoid the problems of infinite fitness support. This restriction is in the form of the number of distinct attribute types that will be mapped within the given Information System Network. This minimally restricts the interoperability of an Information System Network, since the larger the network of Information Systems, the smaller the set of common attributes that will exist within most of the IS nodes.

The set of conditions are:

(1) Establish a set of attributes in common use within the space of Information System Network, $\sum \gamma^*$.

(2) Establish a fitness measure for acceptability of a semantic interoperable connection between two IS nodes, η_i.

(3) Ensure that attributes of a new IS node are mapped to at least one semantically similar attribute existing in the network.

(4) Establish a fixed number of edges added to the network at each new node's inclusion into the network.

(5) Utilize the competitive network model's composite probability of preferential attachment, $\prod_i = \frac{\eta_i k_i}{\sum_j n_j k_j}$

By following the set of conditions, one can ensure that a generalized, scale-free growth will ensue in Information System Networks. The ease of implementation is shown in an example of a set of Information Systems with only seven common attributes modeled in the network, specifically how the competitive network model's preferential attachment can be applied. As determined under these conditions, the network will maintain a generalized power law distribution where $P(k) \sim k^{-2.37101}$, therefore presenting scale-free growth.

Future Work

As the ease of implementation improves, large-scale Information System Networks will become a common feature in the future. As such, many more work will be devoted to produce an efficient network of information nodes that is robust and easy to implement and maintain. Performance metrics for different implementations of ISNs should be proposed as they become increasingly popular. These efficiencies exist in the form of shorter average path-lengths between nodes and therefore faster computation time, reduction in costs associated with creating edges, or even the balance of traffic distribution along the various interconnections.

Extrapolations of how large-scale Information System Networks can grow can also be extended from various other fields, especially in the area of social network research. The Barabási-Albert and the Competitive Model are basically top-down approaches to network growth, citing a generalized behavior pattern which leads to scale-free networks over time. Pujol [1] modeled the various factors that would lead to the emergence of complex social networks from a local, bottom-up perspective. A theoretical extension of the bottom-up approach from the sociological perspective may also be applicable for ISN. Similarities between both fields include the rise of communities of interests (COI) as well as geographical or contextual dispersion factors.

Bibliography

1. Pujol, J.M., et al., *How can Social Networks Ever Become Complex? Modelling the Emergence of Complex Networks from Local Social Exchanges.* Journal of Artificial Societies and Social Simulation, 2005. **8**(4): p. 1-18.
2. Solé, R.V., et al., *Selection, Tinkering, and Emergence in Complex Networks.* Wiley Periodicals, 2003. **8**(1): p. 20-33.
3. Barabási, A.-L., R. Albert, and H. Jeong, *Mean-field theory for scale-free random networks.* Elsevier, 1999. **272**(1-2): p. 173-187.
4. Barabási, A.-L., et al. *Scale-free and hierarchical structures in complex networks.* in *AIP Conference Proceedings*. 2003.
5. Bianconi, G. and A.-L. Barabási, *Competition and multiscaling in evolving networks.* Europhysics Letters, 2001. **54**(4): p. 436-442.
6. Farshad, H. and G. Andreas, *Resolving semantic heterogeneity in schema integration*, in *Proceedings of the international conference on Formal Ontology in Information Systems - Volume 2001*. 2001, ACM Press: Ogunquit, Maine, USA.
7. Batini, C., M. Lenzerini, and S. Navathe, *A Comparative Analysis of Methodologies of Database Schema Integration.* ACM Computing Surveys, 1986. **18**(4): p. 323-364.
8. Reddy, M.P., et al., *A Methodology for Integration of Heterogeneous Databases.* IEEE Transactions on Knowledge and Data Engineering, 1994. **6**(6): p. 920-933.
9. Dayal, U. and H. Hwang, *View definition and genralization for database integration in MULTIBASE: A system for heterogeneous distributed dabases.* IEEE Trans. Software Eng., 1984. **SE-10**(6).
10. ElMasri, R., J. Larson, and S.B. Navathe, *Integration algorithms for federated databases and logical database design.* Tech. Rep., Honeywell Corporate Res. Cent., . 1987.
11. Mannino, M.V. and W. Effelsberg, *A methodology for global schema design*, in *Tech. Rep. TR-84-1, Comput. Inform. Sci. Dept.,* . 1984, Univ. of Florida.
12. Wino, M.V. and W. Effelsberg, *A methodology for global schema design* Tech. Rep. , 1984. **84**(1).
13. Reddy, M.P., B.E. Prasad, and P.G. Reddy. *A model for resolving semantic incompatibilities and data inconsistencies in integrating heterogeneous databases* in *Proc. Int. Conf. Management Data.* 1989.
14. Goh, C.H., *Representing and Reasoning about the Semantic Conflicts in Heterogeneous Information Systems*, in *Sloan School of Management*. 1997, Massachusetts Institute of Technology: Cambridge. p. 113.
15. Goh, C.H., et al., *Context interchange: new features and formalisms for the intelligent integration of information.* ACM Trans. on Information Systems, 1999. **13**(3): p. 270-293.
16. Sabbouh, M., et al., *Using Semantic Web Technologies to Enable Interoperability of Disparate Information Systems*. 2005, The MITRE Corporation.

17. Johansson, J.M., S.T. March, and J.D. Naumann, *Modeling Network Latency and Parallel Processing in Distributed Database Design.* Decision Sciences Journal, 2003.
18. Salton, G. and C. Buckley, *Parallel text search methods.* Commun. ACM, 1988. **31**(2): p. 202-215.
19. Yook, S.H., H. Jeong, and A.-L. Barabási, *Weighted Evolving Networks.* Physics Review E, 2001. **86**(25): p. 5835-5838.
20. Granovetter, M., Am. J. Soc, 1973. **78**(1360).
21. Koide, S. and M. Kawamura, *SWCLOS: A Semantic Web Processor on Common Lisp Object System.* 2004.
22. Watts, D.J. and S.H. Strogatz, Nature (London), 1998. **393**: p. 440.
23. Batini, C., S. Ceri, and S.B. Navathe, *Conceptual Database Design: An Entity Relationship Approach*, ed. B. Cummings. 1992, Redwood City, California.
24. Elmasri, R. and S.B. Navathe, *Fundamentals of database systems (2nd ed.).* 1994: Benjamin-Cummings Publishing Co., Inc. 873.
25. Jones, T.H. and I.-Y. Song, *Binary equivalents of ternary relationships in entity-relationship modeling: A logical decomposition approach.* Journal of Database Management, Apr-Jun 2000. **11**(2): p. 1063-8016.
26. Ling., T.W. *A normal form for entity-relationship diagrams.* in *4th Int. Conference on Entity-Relationship Approach.* 1985.
27. McAllister, A.J. and D. Sharpe, *An approach for decomposing N-ary data relationships.* Softw. Pract. Exper., 1998. **28**(2): p. 125-154.
28. Song, I.-Y., T.H. Jones, and E.K. Park, *Binary relationship imposition rules on ternary relationships in ER modeling*, in *Proceedings of the second international conference on Information and knowledge management.* 1993, ACM Press: Washington, D.C., United States.
29. Teorey, T.J., *Database modeling and design.* 1994, San Mateo, California: Morgan Kaufmann Publishers.
30. Thalheim, B., *Entity-Relationship Modeling: Foundations of Database Technology.* 2000: Springer-Verlag New York, Inc. 635.
31. Dahchour, M. and A. Pirotte. *The semantics of reifying n-ary relationships as classes.* in *4th Int. Conf. on Enterprise Information Systems, ICEIS'02.* April 2002. Ciudad Real, Spain.
32. *W3C Working Paper: Defining N-ary Relations on the Semantic Web: Use With Individuals W3C Working Draft 21 July 2004.*
33. Barabási, A.-L. and R. Albert, *Statistical Mechanics of Complex Networks.* Reviews of Modern Physics, 2002. **74**(1): p. 47-97.
34. Watts, D.J. and S.H. Strogatz, *Collective dynamics of `small-world' networks.* Nature, 1998. **393**(6684): p. 440-442.
35. Pastor-Satorras, R. and A. A. Vespignani, *Epidemic Spreading in Scale-Free Networks.* Phys. Rev. Letters, 2001. **86**(3200).
36. Dorogovtsev, S.N., J.F.F. Mendes, and A.N. Samukhin, *Structure of Growing Networks with Preferential Linking.* Phys. Rev. Letters, 2000. **85**(21): p. 4633–4636.
37. Krapivsky, P.L., S. Redner, and F. Leyvraz, *Connectivity of Growing Random Networks.* Phys. Rev. Letters, 2000. **85**(21): p. 4629-4632.

38. Albert, R., H. Jeong, and A.-L. Barabási, *The Diameter of the World Wide Web* Nature, 1999. **401**: p. 130-131.
39. Albert, R. and A.-L. Barabási, *Statistical mechanics of complex networks.* Reviews of Modern Physics, 2002. **74**(1): p. 47-95.