ESTIMATING THE NODE-LEVEL BEHAVIORS IN COMPLEX NETWORKS FROM STRUCTURAL DATASETS

Zhenghui Sha
School of Mechanical Engineering
Purdue University
West Lafayette, Indiana 47906

Jitesh H Panchal∗
School of Mechanical Engineering
Purdue University
West Lafayette, Indiana 47906

ABSTRACT

There is an emerging class of networks that evolve endogenously based on the local characteristics and behaviors of nodes. Examples of such networks include social, economic, and peer-to-peer communication networks. The node-level behaviors determine the overall structure and performance of these networks. This is in contrast to exogenously designed networks whose structures are directly determined by network designers. To influence the performance of endogenous networks, it is crucial to understand a) what kinds of local behaviors result in the observed network structures and b) how these local behaviors influence the overall performance. The focus in this paper is on the first aspect, where information about the structure of networks is available at different points in time and the goal is to estimate the behavior of nodes that resulted in the observed structures. We use three different approaches to estimate the node-level behaviors. The first approach is based on the generalized preferential attachment model of network evolution. In the second approach, statistical regression-based models are used to estimate the node-level behaviors from consecutive snapshots of the network structure. In the third approach, the nodes are modeled as rational decision-making agents who make linking decisions based on the maximization of their payoffs. Within the decision-making framework, the multinomial logit choice model is adopted to estimate the preferences of decision-making nodes. The autonomous system (AS) level Internet is used as an illustrative example to illustrate and compare the three approaches.

Keywords: Complex networks, evolutionary systems, multinomial logit choice models, AS-level Internet.

1 INTRODUCTION - EVOLUTION OF COMPLEX NETWORKS

A number of complex social, economic, and peer-to-peer networks evolve endogenously based on the local characteristics and behaviors of nodes. The structures of these networks are not directly controlled by the designers, but evolve as a result of decisions and behaviors of individual self-directed entities. The key characteristic of an endogenously evolving network is that the local behaviors of nodes affect the network’s global structure, the structure affects the properties, and the properties affect the resulting performance. Consider the example of the Internet at an autonomous system (AS) level, where a node represents an AS and a link represents communication between two autonomous systems. The nodes make strategic decisions about linking with other autonomous systems in order to route data. These local decisions affect the global structure of the Internet. The global structure in turn affects its robustness and resilience to node failure (i.e., the performance). Thus, the node-level behavior is crucial to understanding the overall network performance. This is in contrast to exogenously designed networks whose structures are directly determined by network designers.

The underlying dynamics of complex endogenous networks can be understood by modeling the mappings across the following five levels, shown in Figure 1:

1. node-level preferences,
2. node-level behavior,
3. network structure,
4. network properties, and
5. system-level performance.

The network structure (level 3) emerges from the node-level
2 REVIEW OF EXISTING LITERATURE

Significant literature on complex networks has been focused on developing models to generate network structures and properties similar to the real-world networks. A class of models aims at achieving desired network structure or network performance by directly manipulating the network topology. In these models, the change of the network results from an external force, such as when a designer decides which node should be linked with whom or which edge should be added (or removed) between two nodes. For example, Beygelzimer et al. [2] proposed an approach to improve the network robustness by testing several different strategies that modify the network topology by rewiring a fraction of the edges or by adding new edges. Bornholdt [3] proposes a model constrained solely by the requirement of robustness from an evolutionary standpoint. The model evolves a new network from an old network by accepting rewiring mutation schemes. Shin and Namatame [4] present a model in which a network is optimized for low congestion and design cost. Network optimization is carried out using genetic algorithms. Such models do not directly determine the node-level behavior (level 2). Since the change in network topology is due to external forces, these models have limited applicability for complex networks that evolve in an endogenous way.

The literature on endogenously evolving networks aim at developing models of node-level behaviors. These models can be categorized into static models and dynamic models. Static models are based on a single snapshot of the network. Examples include Erdos and Renyi model [5], exchangeable graph models [6], and p* models [7]. Dynamic network models account for network evolution through the addition/removal of nodes and links. Examples of such models are the Barabasi-Albert (BA) model [8] designed for generating scale-free (SF) networks and its variants [9], continuous time and discrete time Markov chain models [10,11], and models for generating SF networks with tunable clustering coefficient proposed by Herrera [12], Holme [13], and Klemm and Eguuluz [14, 15]. In these models, a node-level behavior is considered to represent reality if the generated networks have structures similar to the real networks.

Kotegawa et al. [16] use logistic regression to deduce node-level connectivity behaviors within an airport connectivity network to create models for the dynamic evolution of airline transportation network to improve the airline forecasts. Within the domain of open-source product development, Le and Panchal [17] discuss how the behaviors of individuals within a community affect the growth of product and community networks.

Another group of models for endogenously evolving networks is derived based on decision-based frameworks. For example, Snijders [18] present an actor-oriented model to evaluate the dynamics in social networks. Actors try to achieve their individual goals, subject to constraints, which are extracted from available longitudinal network datasets. The model is executed as a continuous-time Markov chain in which the parameters are deduced from the observed data.

In summary, the models based on direct manipulation of linking behavior (level 2), which is driven by the node-level preferences (level 1). The network structure in turn determines the network properties (level 4) and network performance (level 5). The process of traversing from the lower to the higher levels (i.e., 1 to 5) is analysis, in which the performance of the network is determined in terms of the node-level preferences and behaviors. On the other hand, achieving targeted performance by determining how to modify the node-level preferences can be viewed as a design problem [1]. To address the design problem in endogenous networks, it is crucial to understand the structure of node-level preferences that result in observed network structures, and how the node-level preferences and behaviors influence the overall structure. The focus in this paper is on the estimation of node-level behaviors and preferences, as shown in Figure 1. Understanding the node-level preferences and behaviors in real-world networks can help in:

i) accurately modeling the evolutionary dynamics of endogenous networks, and

ii) determining mechanisms for influencing the node-level behaviors and the provision of incentives to achieve targeted system performance.

In this paper, our objective is to illustrate and compare three approaches to estimate the node-level behaviors and preferences in complex endogenous networks from their structure at different time steps. The preferences of the nodes refer to the utility functions that the nodes maximize. In the following section, a review of the relevant network science literature is provided. In Section 3, the three approaches for estimating the node-level behavior and preferences from network structure data are presented. In Section 4, the AS-level Internet network is used as a case study to illustrate the approaches and to estimate the AS-level linking behavior that results in the observed evolution of the Internet. The results from the three approaches are compared. Closing thoughts are presented in Section 5.
the network topology are unsuitable for endogenously evolving networks. For endogenous networks, there are three categories of models for network evolution. The models in the first category generate networks based on hypothesized behavior of the nodes (e.g., preferential attachment). The models in the second category use the incremental evolution of the network between two instances. Finally, the models in the third category use decision-based frameworks to model network evolution (e.g., actor-oriented models). Our goal in this paper is to compare the models in three categories by applying them to the AS-level Internet network. In the following section, the three approaches for estimating the node-level behaviors and preferences are discussed.

3 ESTIMATING NODE-LEVEL BEHAVIORS FROM NETWORK STRUCTURE DATA

The first approach is based on the generalized preferential attachment (GPA) [8]. In this method, a relationship between the network structure and the degree-based node-level behavior (probability of linking) is derived using the continuum theory of network evolution presented by Albert et al. [9]. The node-level behavior is derived from the degree distribution of the network at a given time. In the second approach, the node-level behaviors are derived by analyzing how the networks change between two consecutive instances. The nodes and links created (or deleted) between the two instances are used to deduce the node-level behaviors using regression techniques.

While the first two approaches deduce the node-level behavior, the third approach can be used to determine both the behavior and the node-level preferences. In the third approach, network evolution is modeled as a decision-making process where the nodes are decision-making agents and their behaviors are based on utility maximization. The node-level preferences are determined using the multinomial logit choice model from discrete choice theory [19]. A comparison of the three methods is provided in the table. Further details are provided in the following sections.

3.1 Approach 1: Generalized Preferential Attachment

The preferential attachment model for complex networks was initially proposed by Barabasi and Albert [8]. In this model, a new node preferentially links to existing nodes based on certain characteristics of the target node. Network evolution is this approach is assumed to follow two mechanisms: growth and preferential attachment [20]. The growth mechanism prescribes that at each time step, one new node is added with \( m \) edges linking the new node to \( m \) existing nodes in the network. In the simple preferential attachment model initially proposed by Barabasi and Albert [8], the probability of link creation between a new node and an existing node is linearly proportional to the degree of an existing node.

The preferential attachment has been widely accepted in

The field of complex networks research and has been utilized for modeling real-world complex networks such as the Internet [21], the World Wide Web [22], and networks of metabolic reactions [23]. Existing literature [24, 25] has shown that the degree-based preferential attachment mechanism can be used for modeling real-world complex evolving networks with a power-law degree distribution. The degree-based linear preferential attachment model has been extended to generalized preferential attachment (GPA). An overview of GPA is presented next. 

3.1.1 Overview of generalized preferential attachment (GPA). In the GPA model, the affinity of a node to link with an existing target node \( i \) at time \( t \) is modeled as:

\[
V(i,t) = G(i,t)d^T(i,t) + A(i,t)
\]  

where, the \( V, G, A \) and \( d \) are functions of the node \( i \) and time \( t \). \( G(i,t) \) is the fitness value of node \( i \) at time \( t \), and \( A(i,t) \) is the additional attractiveness of node \( i \) at time \( t \). \( d(i,t) \) stands for the degree of node \( i \) at time \( t \), which is the number of neighbors of node \( i \). Using Equation (1), the probability of an arbitrary node \( j \) getting chosen to connect by another node \( n \) among \( J \) nodes is
equal to:

\[ P_n(j,t) = \frac{V(j,t)}{\sum_{i=1}^{J} V(i,t)} = \frac{G(j,t) d(j,t)}{\sum_{i=1}^{J} (G(i,t) + A(i))} + \frac{A(j,t)}{\sum_{i=1}^{J} (G(i,t) + A(i))} \]

(2)

This probability function is assumed to result in the evolution of the network between two time steps. Furthermore, it is assumed that i) the network is undirected, ii) the fitness value for all the nodes is the same and is time independent, thus \( G(i,t) \) is constant, iii) the additional attractiveness for each node is time independent and a constant, thus \( A(i,t) \) is constant, and iv) the affinity \( V \) in Equation (1) is a linear function of the node degree, i.e., \( \tau = 1 \). Therefore, Equation (1) can be modeled as:

\[ V(j,t) = d(j,t) + A(j) \]

(3)

If \( G(j,t) \) is assumed to be the same for all nodes, its impact can be accounted for by scaling the additional attractiveness parameter as follows:

\[ P(j,t) = \frac{V(j,t)}{\sum_{i=1}^{J} V(i,t)} = \frac{Gd(j)}{\sum_{i=1}^{J} (Gd(i) + A(i))} + \frac{A(j)}{\sum_{i=1}^{J} (d(i) + A(i))} \]

(4)

The additional attractiveness \( A \) and the node’s degree determine the complete behavior model of a linking node. Based on the prior work by Sha and Panchal [1], it has been shown that additional attractiveness \( A \) has a significant impact on the network structure and network performance. The additional attractiveness is estimated through the degree distribution obtained by using the continuum theory of network evolution, discussed next.

For analyzing the evolutionary process in this model, the continuum theory approach proposed by Albert et al. [9] provides a bridge between the network structure and the node-level properties such as the degree. With the continuum theory, the effect of additional attractiveness \( A \) on the structure, specifically the degree distribution of the resulting network, can be analyzed. According to the growth mechanism described above and the model proposed in Equation (3), the rate of change of a node’s degree \( d_i \) is given by:

\[ \frac{\partial d_i}{\partial t} = m \frac{d_i + A}{\sum_{j=1}^{J} (d_j + A)} \]

(5)

where \( m \) is the number of edges linking to a new node in each timestep. Following the steps in [6], as the network becomes large,

\[ \lim_{t \to \infty} P[d_i(t) \geq d] = \left( \frac{d + A}{m + A} \right)^{-\gamma} \]

(6)

Thus, as \( t \to \infty \), the asymptotic cumulative degree distribution has the form:

\[ F(d) = P[d_i(t) \geq d] \propto d^{-\gamma} \]

(7)

where,

\[ \gamma = f(m, A) = \left( 2 + \frac{A}{m} \right) \]

(8)

Different degree distributions are associated with different \( A \) values, which can be used to differentiate the network structures. To illustrate, the degree distributions generated for representative values of \( A \) are shown in Figure 2. By fitting the power-law degree distribution, we can determine the exponent \( \gamma \) using regression techniques to deduce the values of additional attractiveness, \( A \), which defines the node-level behavior (Equation (4)). In Section 3.1.2, we introduce the techniques used for fitting the power law degree distribution.

**3.1.2 Fitting the Power-law Degree Distribution.**

A simple approach for fitting the power law degree distribution is the ordinary least square (OLS) regression [26]. The power law distribution in Equation (6) follows a straight line on a double
logarithmic plot. Therefore, a commonly adopted technique to estimate the power law behavior in empirical data is to measure the frequency of nodes with degree \( d \) in the network and plotting such frequency on the double logarithmic axis. Then, a linear model:

\[
y = \beta_0 + \beta_1 x + \varepsilon
\]  

(9)

can be used where \( \beta_0 \) is the intercept, \( \beta_1 \) is the slope and \( \varepsilon \) is the random error in the observation. The OLS regression can be utilized to fit the power law degree distribution with variable \( x \) equal to \( \ln(d) \) and observation value \( y \) equal to \( \ln(P) \). The estimation on the parameter \( \beta_1 \), which is the slope, \( \gamma \) is the exponent in the power law.

In practice, the power law often applies only for values greater than some minimum value \( x_{\text{min}} \). In such cases, the OLS regression method can produce inaccurate estimates of the parameters for power law distributions especially for the “tail” of the distribution where the values are under \( x_{\text{min}} \). To address this issue Clauset et al. [27] proposed an effective statistical framework for fitting the power law distribution to empirical data. The approach combines maximum likelihood fitting with goodness-of-fit tests based on the Kolmogorov-Smirnov (KS) statistic and likelihood ratio [28]. The key idea for estimating the exponent \( \gamma \) correctly is to first identify the lower bound \( x_{\text{min}} \) of power law behavior in the data. Hence, the parameter \( x_{\text{min}} \) is first chosen, and then \( \gamma \) of the power law is fit using maximum likelihood estimation. Then, with the estimated \( x_{\text{min}} \) and \( \gamma \) in the first step, the power law hypothesis is tested by calculating the p-value for goodness-of-fit test that quantifies the plausibility of the hypothesis. A power law hypothesis is considered plausible for the data if the resulting p-value is greater than 0.1. Finally, the power-law models derived using alternate values of \( x_{\text{min}} \) are compared via a likelihood ratio test [29]. If the calculated likelihood ratio is significantly different from zero, then its sign indicates whether an alternative is favored or not.

Once we have the estimation on the exponent \( \gamma \) in the power law, we can obtain the additional attractiveness \( (A) \) using Equation (8). The \( m \) values for different networks can be obtained by plotting the number of nodes \( (J) \) versus the number of edges \( (E) \) in the network over time. An OLS regression between the number of new nodes and the number of new links can be used to estimate \( m \). An illustrative example is presented in the case study in Section 4.3.1.

3.2 Approach 2: Statistical Regression-based Model

In the second approach, the linking behavior is determined by comparing two consecutive instances of the network structure. The node-level behavior is then obtained by fitting the node-level linking probability data using regression techniques. Consider a complex endogenous network that evolves from network \( N(t_0) \) to network \( N(t_1) \) during an interval \([t_0, t_1]\). In order to obtain the behavior of the added nodes, we calculate each target node’s probability of getting a connection from the newly added nodes. From the datasets \( N(t_0) \) and \( N(t_1) \), we obtain the number of new nodes entering the network during \([t_0, t_1]\). For each newly added node, the target nodes are identified. Based on the network structure from the dataset \( N(t_0) \), the degrees of these target nodes are extracted. In the following step, all nodes in \( N(t_0) \) are divided into different groups based on their degrees. All nodes with degree \( d \) are grouped into a group \( S_d \). The number of nodes within a group \( S_d \) is represented as \( n_d \). If the number of new links created with existing nodes in \( S_d \) is denoted by \( l_d \), and the total number of links created during the interval \([t_0, t_1]\) is \( L \), then the probability of a group \( S_d \) receiving a link is:

\[
P(S_d) = \frac{l_d}{L}
\]  

(10)

This is based on the assumption that each linking decision made by a node is a mutually exclusive event, the probability of all nodes getting a connection in the same group is the sum of the probability of each node in this group getting a connection. Considering all nodes to be identical, an individual node with degree \( d \) has the following probability of receiving a connection:

\[
P(d) = \frac{1}{n_d}P(S_d)
\]  

(11)

Once the probability of an individual node with degree \( d \) getting connections has been determined, the degree versus probability relationship is plotted. By using an appropriate fitting model using OLS regression, the node-level behavior can be obtained. The application of the proposed approach for the case study is presented in Section 4.3.2.

3.3 Approach 3: Multinomial Logit Choice Model

In the third approach, we model the network evolution using a decision-making framework. Here, each new node is considered as a decision maker that maximizes its own utility function \( u \). Say a node \( i \) is a decision maker at time \( t \), its decision on which target node \( j \) to link to is based on the maximization of \( u_l \) based on the network topology at time \( t \). At a given time, node \( i \) has \( J \) alternatives to choose from. The decision-maker node \( i \) selects a target node \( j \) for creation of an edge based on the utility function. The utility function, \( u_l \), can depend on the structural parameters of the nodes (e.g., degree), or non-structural factors (e.g., capacity, cost, etc.).

We use the random utility discrete choice models to estimate the utility functions that nodes maximize while selecting other nodes to link to. Specifically, we use the multinomial logit choice model to model the decisions of the nodes. A brief introduction to discrete choice models and multinomial logit is provided in Section 3.3.1. Modeling the network evolution using the multinomial logit choice model is discussed in Section 3.3.2.
3.3.1 Discrete Choice Model and Multinomial Logit  
Multinomial logit is a technique for discrete choice analysis (DCA) [19] in which the size of the choice set, \( J \), for the decision maker is greater than two. DCA has been widely used to model and forecast product demand by capturing individual customers’ choice behavior, especially for demand estimation [30]. Earlier applications were in the field of transportation engineering, but later, DCA was extended to the field of product design to model consumer preferences under uncertainty [31]. DCA is based on the assumption that individuals seek to maximize their personal utility, \( u \), when selecting an alternative from the choice set. The decision maker knows the utility function and uses it for making decisions. However, the observer is only able to observe the choices made by the decision maker. Therefore, from the observer’s perspective, the utility function is random. DCA assumes that the individual’s utility is a sum of two components [32]:

1. **Systematic component**, denoted by \( V_j \), which is a function of different observed attributes \( x_j \) of the alternatives which can be either alternative specific or decision-maker specific [33]. It is assumed that this component is a linear combination of the observed attributes: \( V_j = \beta_j^T \cdot x_j \), where \( \beta_j \) are the parameters corresponding to the observed attributes \( x_j \). The observed attributes can also be referred to as the explanatory variables that describe the decision maker’s utility function. It is important to note that this component is deterministic from the observer’s point of view.

2. **Unobserved component**, \( \varepsilon_j \), which can be represented as a random variable from the observer’s point of view. This error term includes the impact of all unobserved variables that affect the utility of a specific alternative. Thus,

\[
\begin{align*}
  u_j &= \beta_j^T \cdot x_j + \varepsilon_j
\end{align*}
\]  

(12)

Based on utility maximization, the probability that alternative 1 is chosen by the decision-maker \( n \) from a choice set containing two alternatives, is equal to the probability that the utility of alternative 1 is greater than the utility of alternative 2. This can also be represented as:

\[
\begin{align*}
  P_n(1|1,2) &= P_n(u_{1n} > u_{2n}) \\
  &= P_n(V_{1n} + \varepsilon_{1n} > V_{2n} + \varepsilon_{2n}) \\
  &= P_n(\varepsilon_{1n} - \varepsilon_{2n} > V_{2n} - V_{1n})
\end{align*}
\]  

(13)

In order to predict the choice probability, methods such as binary logit, probit, multinomial logit and mixed logit [32] can be used. The primary difference between these models is the assumption about the probability distribution of the unobserved component. In this paper, we use the multinomial logit model where the error terms \( \varepsilon_j \) are assumed to be independent and identically distributed across choice alternatives and observations (decision-maker), and follow a Gumbel distribution [19]. Then, the probability that decision-maker \( n \) chooses alternative 1 over alternative 2 is:

\[
P_n(1|1,2) = P_n(u_{1n} > u_{2n}) = \frac{e^{V_{1n}}}{e^{V_{1n}} + e^{V_{2n}}}
\]  

(14)

The binary alternatives scenario has been extended to the Multinomial Logit (MNL) model, see Equation (15), that describes the probability of alternative \( j \) being chosen by decision-maker \( n \) from among a choice set containing \( J \) alternatives.

\[
P_j(C_j) = \frac{e^{V_{jn}}}{\sum_{i=1}^{J} e^{V_{in}}}
\]  

(15)

Estimation techniques such as the maximum likelihood and Bayesian estimation can be used to determine the coefficients \( \beta \) in Equation (12) such that the model’s prediction of choices matches the observed choices as closely as possible. In practice, existing statistical analysis software can be used for estimation of parameters in a multinomial logit model. In this paper, the mlogit package [34] for R [35] is used.

3.3.2 Describing the Network Evolution using the Multinomial Logit Model  
If the complex network evolution is based on node-level decision-making process, the principles from discrete choice theory can be utilized to estimate the utility function, and the resulting choice probability (i.e., the probability of a newly added node in \( N(t_1) \) choosing an existing node from network \( N(t_0) \)). The choice probabilities of individual nodes can then be aggregated to estimate the aggregate node-level behaviors.

In the multinomial logit choice model, the observations from a researcher’s point of view are the newly added nodes that choose a target node to link to. The alternatives are the existing nodes during the previous time step. For the selection of each node’s utility, if the observed variable \( x \) in the systematic component is only alternative specific, e.g., the node’s degree, then the systematic component is:

\[
V_j = \beta_{0j} + \beta_{1j}d_j
\]  

(16)

This corresponds to equation (1) in which \( \beta_{0j} \) stands for the additional attractiveness and \( \beta_{1j} \) is the node fitness. The resulting probability of the node alternative \( j \) that gets connection from the decision-making node \( n \) is given by Equation (15). Note that this is fundamentally different from Equation (4). Finally, the parameters \( \beta_{0j} \) and \( \beta_{1j} \) can be estimated using maximum likelihood estimation techniques.

For large sized networks, the choice set may be large (e.g., over 10,000 nodes). To reduce the complexity of parameter estimation, the size of the choice set can be reduced by grouping...
the nodes with same degree together as one alternative. The resulting probability is the one that a group is chosen over other groups by a newly added node. The probability of connecting to a node within a group can then be obtained by randomly choosing a node from the group to which that individual node belongs.

The utility function can be further refined by considering other structural and non-structural parameters of the network. By considering more attributes of the alternatives, such as, the clustering coefficient [36], and betweenness centrality [37] different hypotheses about the utility functions can be generated and tested. Through this approach, accurate models of choices that match the observed choices can be obtained, and the factors (besides node’s degree) that constitute the additional attractiveness of a node can be investigated. Hence, this approach is richer than the two approaches described in Sections 3.1 and 3.2.

4 CASE STUDY: AUTONOMOUS SYSTEM LEVEL INTERNET NETWORK

In this section, the approaches discussed in Section 3 are applied to the autonomous system (AS) level Internet network. The goal here is to illustrate how these approaches can be implemented in practice to deduce the AS-level behaviors and decisions that result in the observed evolution of the Internet. We provide a brief introduction to the AS-level Internet in Section 4.1. The dataset of the AS-level Internet network is described in Section 4.2. The results from the different approaches are presented in Section 4.3. Finally, a discussion of the results is presented in Section 4.4.

4.1 Introduction to AS-level Internet

The Internet is a network of interconnected computers consisting of private, public, academic, business, and government networks linked by various networking technologies. The Internet network can be treated as an endogenously evolving complex network because of the decentralized governance in usage and access policies. The topology of the Internet can be studied at three different levels [38]:

1. **IP level**, which is composed of the interfaces of routers that exchange information because each interface owns an IP on the Internet.
2. **Router level**, which is the interconnection of routers on the Internet. It represents cables, satellite or radio links, etc. This physical infrastructure is the one over which information is routed.
3. **AS level**, which models the way autonomous systems are interconnected. The Internet can be divided into thousands of domains connected with each other. Each domain is a collection of hosts connected via routers and switching facilities.

An autonomous system has its own set of routers and routing policies. Examples of autonomous systems include ISPs, corporate networks, and universities. An ISP can have one or more autonomous systems. Autonomous systems are connected via dedicated links or public network access points. A link between two autonomous systems represents a contract to forward data to each other over the link. Each AS can choose its policy to select the best route for data based on commercial contractual relationships. These contracts and AS-level policies play a significant role in determining the structure of the Internet and its overall performance [39]. The AS-level topology also influences the definition of routing protocols such as the Border Gateway Protocol (BGP). Hence, it is an important and appropriate level of abstraction to model the decisions that result in the structure of the Internet.

4.2 Data source

Publicly available data sources are available for Internet network data. Skitter, Archipelago (Ark) from Cooperative Association for Internet Data Analysis (CAIDA) [40] and the RoutView [41] from the University of Oregon are the three main projects for collecting the Internet topology data at the AS level. Specifically, the Ark project is an upgraded version of the previous Skitter project operated by CAIDA after Skitter served nearly a decade and was retired on Feb. 8th, 2008 [40].

The dataset adopted in this paper is from CAIDA AS Relationships Dataset from January 2004 to November 2007. There are 122 files in total, each file containing a full AS graph derived from a set of BGP table snapshots used to exchange routing information between autonomous systems.

4.3 Estimating the AS-Level Behavior in Internet Network Using the three Methods

In this section, we present the results from the three approaches. We first apply the method discussed in Section 3.1 to the AS-level Internet network to deduce the AS-level behavior.

4.3.1 Results from Approach 1: Generalized Preferential Attachment. The first step in this method is to develop a fit for the degree distribution of the network. Figure 3 shows an example degree distribution for the AS-level Internet on Jan. 5th, 2004, along with the OLS regression model. The figure shows that a power law [42] is a good fit for the degree distribution of the network. Since the degree distribution is plotted on double logarithmic axes, the slope of the fitting line is the exponent $\gamma$ in the power law relation (see Equation (9)).

To determine how the power law distribution changes with time, we extract the exponents of the degree distribution for all 122 snapshots of the network from 2004 to 2007. Since the network size increases monotonically over time, the exponent is plotted against the network size that corresponds to each network at each time in the $x$-axis. The exponents are plotted in Figure 4. It is observed that this exponent $\gamma$ increases with the network size.

Based on Equation (8), the additional attractiveness ($A$) in the node-level behavior model can be evaluated using the expo-
nents ($\gamma$) and the number of new links added in each time step ($m$). The $m$-value can be identified by plotting the number of nodes ($J$) vs. the number of edges ($E$) as the network grows. This plot is shown in Figure 5. It is observed from the figure that the number of edges increases linearly with the number of nodes. The slope of the line shows that for each new node, about 2 new edges are added. This indicates $m \approx 2$.

The additional attractiveness, $A$, is calculated using $m$ and $\gamma$. Figure 6 shows the additional attractiveness, $A$, for the 122 networks. It is observed that as the network evolves, the $A$-values increases from $-1.78$ to $-1.73$. This indicates that as the Internet grows, the additional attractiveness in the network increases, which impacts the node’s linking preference. The impact of additional attractiveness on the linking behavior is discussed in detail by Sha and Panchal [1]. As $A$ increases, more nodes have the opportunity to be connected.

We also used the approach suggested by Clauset et al. [29] (discussed in Section 3.1.2) to fit the degree distribution using the maximum likelihood estimator. The resulting exponents for the 122 networks are shown in Figure 7. By performing the t-test on the slope in the figure, the p-value corresponding to $\{H_0 : S_\gamma = 0$ vs. $H_1 : S_\gamma \neq 0\}$, where $S_\gamma$ is the slope of the parameter $\gamma$, is 0.14. Hence, there is no statistically significant change in $\gamma$. Note that the exponents in the power-law shown in Figure 7 are also greater than those in Figure 4. This can be explained as follows. The main difference in this method is that a minimum bound value $x_{min}$ is estimated beyond which the fit is close to power law, and the “tail” of the distribution with values of degree lower than $x_{min}$ are not considered in the fitting. Therefore, the resulting power law curve is only for the part of the data that is regarded as a true power law. Since the change in degree for the nodes that have low degree in the network is not substantial, the fitting for that part of data does not change significantly. Because of the positive linear relationship between the exponent $\gamma$ and the $A$ value, the $A$ value in turn remains unchanged as the network grows.
In this section, we utilize the approach presented in Section 3.3 to the AS-level Internet to estimate the AS-level behavior. Figure 8 shows an example of the node-level linking behavior (i.e., probability of new node linking to an existing node) for the AS-level Internet network as it evolves from Jan. 5th, 2004 to Feb. 2nd, 2004. The plot is shown on a log-log scale. We use the degrees of existing nodes based on the previous snapshot. They are not updated after each link is added. We fit the data with a power function \( y = \alpha x^\beta \) where \( y \) is the probability of linking to a node, and \( x \) is the degree of the target node. The parameters \( \alpha \) and \( \beta \) are estimated using OLS regression of \( \ln(x) \) vs \( \ln(y) \).

Figure 9 shows the node-level linking behavior in three different pairs of consecutive network snapshots:

a) Jan. 5th, 2004 (N1) - Feb. 2nd, 2004 (N2),

b) Aug. 28th, 2006 (N59) - Sep. 4th, 2006 (N60),

c) Nov. 5th, 2007 (N120) - Nov. 12th, 2007 (N121).

As shown in the figure, the parameters of the three fitting functions are close to each other, which indicate that the node-level behavior is consistent over time. This conclusion about the node-level behavior is different from the one obtained using the first approach (see Figures 4 and 7). However, the result is in agreement with the fit using the maximum likelihood estimation (see Figure 7).

To further validate this conclusion, we extract the node-level behaviors from all the 121 changes in the network from Jan. 2004 and 2007, and then determine the parameters of the fit (\( \alpha \) and \( \beta \)). Figure 10 and Figure 11 show the two parameters for the 121 timesteps. We performed two separate hypothesis tests to detect whether there have been any significant changes in \( \alpha \) and \( \beta \) over the 121 evolutions. The p-value corresponding to \( H_0^\alpha : S_\alpha = 0 \) vs. \( H_1^\alpha : S_\alpha \neq 0 \) and \( H_0^\beta : S_\beta = 0 \) vs. \( H_1^\beta : S_\beta \neq 0 \) are 0.04 and 0.03 respectively. Here, \( S_\alpha \) and \( S_\beta \) are the slopes corresponding to parameters \( \alpha \) and \( \beta \) in Figures 10 and 11 respectively. Hence, we claim that there has been no significant change in slopes of these two parameters at a 2% level of significance. Hence, we conclude that the node-level behavior of the ASes was consistent during the periods 2004 and 2007. The average values of the parameters \( \alpha \) and \( \beta \) are 1.97 \times 10^{-5} \) and 0.959 respectively. Using these two parameters, we can determine the linking behavior in terms of the probability of linking to a node with degree \( d \):

\[
P = \alpha d^\beta
\]  

(17)

4.3.3 Results from Approach 3: Multinomial Logit Choice Model. In this section, we apply the multinomial logit model to deduce the node-level utility given the assumption that the node-level decision follows the form of Equation (15). In the multinomial logit choice model, each existing node is an
alternative. The size of the network is large (e.g., the number of nodes in the network of Jan. 5th, 2004 is 16301). Hence, to reduce the computational burden, the size of choice set is reduced by grouping the nodes with same degree together as one alternative. The resulting probability is that of a newly added node selecting a given group representing a particular degree. An individual node’s probability of getting a connection can then be obtained by assuming that all nodes within a group have the same probability.

In order to use the multinomial logit model, the first step is to identify the attributes to be considered in the systematic component of the utility function. We consider two aspects in the utility function: the node’s degree \( d_j \) and the number of nodes with degree \( n_j \). Instead of using these parameters directly in the utility function, we use the natural logarithms of these parameters as the attributes of the nodes. Hence,

\[
V_j = \beta_1 \ln(d_j) + \beta_2 \ln(n_j) \tag{18}
\]

where \( \beta_1 \) is the parameter corresponding to degree \( d_j \), \( \beta_2 \) is the parameter corresponding to the size of group size \( n_j \). This choice of the functional form is used because it results in a node-level behavior that can be directly estimated using existing multinomial logit algorithms. The parameters \( \beta_1 \) and \( \beta_1 \) denote the preferences of the decision-making nodes on degree and group size. Thus the utility function based on Equation (12) is:

\[
u_j = \beta_1 \ln(d_j) + \beta_2 \ln(n_j) + \epsilon_j \tag{19}\]

The resulting probability that the group \( j \) is chosen by node \( n \) from \( J \) existing nodes in a network is:

\[
P_n(j|C_J) = \frac{d_j^{\beta_1} n_j^{\beta_2}}{\sum_{i=1}^{J} d_i^{\beta_1} n_i^{\beta_2}} \tag{20}\]

The parameters \( \beta_1 \) and \( \beta_2 \) can be estimated by using the information from the observed network structure datasets. Thus the utility function in Equation (19) can be determined. We estimate the parameters \( \beta_1 \) and \( \beta_2 \) for all the 122 network datasets. The parameters are plotted against the network size in Figures 12 and 13. It is observed in Figure 12 that the parameter \( \beta_1 \) has an average value of 0.672 for network size smaller than 21000 nodes (corresponding to the Internet network at Jan. 2, 2006), and an average value of 0.428 afterwards. This is verified through hypothesis tests on the slope, as discussed in the previous section, resulting in \( p \)-values of 0.03 and 0.11. The parameter of group size \( \beta_2 \) follows a similar trend. The average value for network size smaller than 21000 is 0.661 \( (p \text{-value} = 0.12) \) and the average value is 0.525 \( (p \text{-value} = 0.06) \) afterwards.

In order to evaluate the effectiveness of the three approaches in generating the network structure, we compare the simulated...
node-level behavior with the result obtained from each approach in the following section.

4.4 Comparison of the Node-level Behaviors and Resulting Networks Generated by the Node-Level Behaviors

In this section, we show a comparison among the node-level behaviors obtained by the three approaches and among the resulting network structures for a fixed time. Figure 14 shows an example of the node-level linking behavior of the AS-level Internet network on Jan. 5th, 2004, estimated using the three approaches. In the first approach (continuum theory-based approach), the probability that a node is chosen is determined by Equation (4) with estimated additional attractiveness ($A_i$). In the second approach, the node-level behavior is described by Equation (17), where the parameters $\alpha$ and $\beta$ of the power function are estimated using the ordinary least square regression (OLS). In the third approach, the node-level behavior is obtained by Equation (20) and parameters $\beta_1$ and $\beta_2$ are estimated using the multinomial logit choice models.

Based on these node-level behavior models derived by different approaches, the network topology of Internet on Nov. 12th, 2007 can be simulated with a real network topology at Jan. 5th, 2004 as initial network. Figure 15 shows the degree distribution of the simulated network structure based on three approaches. It is observed that all the three degree distribution functions are close to each other. To quantify the deviation of the three simulated degree distributions from the original network, the Kullback-Leibler (KL) test [43] is performed. Figure 16 shows KL divergence at each degree point of three degree distributions. The KL test values for i) the GPA approach with OLS regression, ii) GPA with maximum-likelihood approach, iii) statistical regression-based approach, and iv) the multinomial logit-based approach are $1.091 \times 10^{-2}$, $3.526 \times 10^{-2}$ and $2.414 \times 10^{-2}$, and $4.162 \times 10^{-2}$ respectively. Since all the values are small, we conclude that all the methods perform well in generating the networks with degree distribution closer to the true network.

Out of the three approaches, the first approach has the advantage of being simple and easy to evaluate. However, it is based on an assumed behavior model of preferential attachment. Both the second and third approaches have potential generality to be used in other similar problems in complex networks, without placing any assumption on the node-level behavior in advance. The strategy of regarding the network evolution as a decision-making process is a promising approach to model the evolution of endogenous networks. It provides information about the preferences of the nodes, in addition to their linking behavior. Such information cannot be obtained in the other approaches. However, the barrier in implementing the third approach is the computational burden. Since the number of alternatives is large, the estimation problem becomes computationally expensive if each
node is treated as an alternative. A potential approach to manage this complexity is to reduce the size of the choice set by grouping the nodes with similar characteristics (e.g., degree) into one group, and treating each group as an alternative as shown in the case study. This is only valid if each node within a group is equivalent. The appropriate approach to be chosen depends on various factors such as the dataset of the network structure, the size of the network, the attributes in the network, and the type of the network. The possibility of combining these approaches requires further investigation.

5 CLOSING COMMENTS

Complex endogenous networks are important because of their potentially wide applications in real-world complex systems. An understanding of the node-level behaviors and preferences is important to influence the design of complex endogenous networks. The contribution of this paper does not only help researchers gain a better understanding about how a real-world complex network evolves, but also provides an approach for obtaining realistic network topologies. The approaches discussed in this paper have potential applications in the area of complex network analysis and design in three aspects. First, the proposed approaches are general enough to be used for other complex networks where longitudinal network structure datasets are available. Second, gaining an understanding of the underlying behaviors can help in creating better network models to describe the real-world network. Third, obtaining a network’s node-level behavior can help in guiding the evolution of endogenous networks.

In summary, we propose three approaches to deduce the node-level behavior and preferences from network structure datasets. An illustrative example, AS level Internet network, is used to show how the proposed approaches can be applied to deducing the AS-level behavior. Future work would focus on relaxing some of the assumptions made in the models. For example, in this paper, we only consider addition of nodes in the network evolution. However, in reality, removal of nodes and link re-direction also occurs during network evolution. Thus, consideration of decisions to remove nodes and to redirect links would be an essential aspect for further improvement of the model. Additionally, other attributes of the nodes that may influence the node’s utility will be further investigated. For example, in the Internet case study, economic factors such as the link creation costs and profits, geographic location, type of AS, etc. play an important role in the creation of new links [44, 45]. Therefore these variables will be investigated in the future to refine the discrete choice based model. Finally, by utilizing theories from other disciplines such as psychology and economics, possible incentives can then be designed and introduced into the network to achieve desired network properties or performance.

6 ACKNOWLEDGEMENTS

The authors gratefully acknowledge the financial support from the National Science Foundation grants 1265622 and 1201114, and China Scholarship Council award No. 2010628122. The authors also thank Jyotishka Datta for helping with the statistical analysis.

REFERENCES


Kullback, S., and Leibler, R., 1951. “On information and
