DECISION-CENTRIC FOUNDATIONS FOR COMPLEX SYSTEMS

ENGINEERING AND DESIGN

A Dissertation

Submitted to the Faculty

of

Purdue University

by

Zhenghui Sha

In Partial Fulfillment of the

Requirements for the Degree

of

Doctor of Philosophy

August 2015

Purdue University

West Lafayette, Indiana
This dissertation is lovingly dedicated to my wife, Shasha Liu, my son, Jinghang A. Sha, my mother, Jie Ma, and my father, Quanliang Sha. Their support, encouragement and unconditional love have sustained me throughout my life.
ACKNOWLEDGMENTS

It is my immense pleasure to thank my wife, my mother and my father for their continuous encouragement and support during my Ph.D. studies in the United States. With their continuous emotional and financial support, my Ph.D. studies could not have been smoother. Without their support, my resolve on the road of Ph.D. study would not be so firm. I would like to thank my son, for making my life happy, vibrant and meaningful. Such happiness motivates me to work hard for my career.

I would like to express my special appreciation for my advisor, Dr. Jitesh Panchal. Dr. Panchal played a vital role during my five-year Ph.D. life. I am grateful that Dr. Panchal introduced me to a research area I am passionate about, which led to Ph.D. study that was joyful work. I would like to thank him for his continuous support, guidance and advice on my research. Dr. Panchal has an open style of discussion and preferred to inspire me to learn rather than pushing me to learn. Also, I would like to thank him for the patient and professional training, which enabled me to hone my skills as well as create the mindset I need to be a successful scholar in the future. I would like to thank him for his continuous trust and frankness. Such qualities made Dr. Panchal not only a scholarly mentor but also a beneficial friend to me. His words and deeds have made me more mature, stronger and more confident. All in all, working with Dr. Panchal has influenced me significantly and there is too much to express here. Without his support, this dissertation and the academic achievement would not be the same.

I would like to thank my committee members: Dr. DeLaurentis, Dr. Zhao and Dr. Ramani. I really enjoyed working with them. All the committee members have given me valuable advice during my studies. They discussed with me broad research issues and inspired my research. I appreciate their patient guidance, advising styles and attitudes. Their perspectives and ways of solving research problems inspire me
deeply and help me form effective mindsets for the academic field. I would like to also thank Dr. Kanna; his guidance and advice notably boosted my research progress.

I would like to thank my lab mates: Qize Le, Bryant Hawthorne, Ahmad Taha, Murtuza Shergadwala, Joseph Thekinen, Siva Chaduvula, Pramod Kumar, Naman Mandhan and Piyush Pandita. Qize, Ahmad, Murtuza and Joseph provided valuable input on my research. Others provided lots of comments on my published papers and helped me polish my English. I enjoyed every minute I spent with them at DELP. They are not only lab mates, but also close friends and even like family. In addition, I would like to thank all of my friends and peers - both in China and the U.S. - who have helped me in the past five years in any way.

I would like to thank the School of Mechanical Engineering for providing the assistantship. I would like to thank everyone in ME who provided various academic consulting services on course registration, graduation requirements and procedures. My time in Mechanical Engineering at Purdue is a special period that I will never forget.

Finally, I gratefully acknowledge the fellowship provided by China Scholarship Council award No. 2010628122 and the financial support from the National Science Foundation grants 1360361, 1400050 and 0954447.
# TABLE OF CONTENTS

| LIST OF TABLES                              | ix     |
| LIST OF FIGURES                            | xi     |
| SYMBOLS                                    | xiv    |
| ABBREVIATIONS                              | xvi    |
| ABSTRACT                                   | xviii  |
| CHAPTER 1. OVERVIEW OF THE DISSERTATION     | 1      |
| 1.1 Research Background and Motivation      | 1      |
| 1.2 Research Overview                      | 4      |
| 1.3 Research Questions                      | 6      |
| 1.4 Validation Examples and Applications    | 10     |
| 1.5 Summary of Contributions                | 15     |
| 1.6 Overview of the Dissertation            | 17     |
| CHAPTER 2. BACKGROUND AND LITERATURE REVIEW | 19     |
| 2.1 Large-scale Complex Networked Systems   | 21     |
| 2.1.1 Existing Studies on Modeling Network Generation with Node-Level Behavioral Models | 22 |
| 2.1.2 Existing Studies on Models for Improving Network Performance | 24 |
| 2.1.3 Research Gaps                        | 25     |
| 2.2 Social Product Development - An Emerging Paradigm for Engineering System Design and Innovation | 27 |
| 2.2.1 Existing Studies on Crowdsourcing in Product Development | 28 |
| 2.2.2 Existing Studies on Game Theory in Engineering System Design | 30 |
| 2.2.3 Research Gaps                        | 31     |
| CHAPTER 3. A DECISION-CENTRIC APPROACH FOR MODELING AND ANALYZING COMPLEX NETWORKS | 34 |
| 3.1 Chapter Overview                        | 34     |
| 3.2 Theoretical Foundations                 | 36     |
| 3.3 The Decision-Centric Modeling Framework | 39     |
| 3.4 The Effect of Local Decision-making Preferences and Behaviors | 43 |
| 3.4.1 Decision-Centric Modeling of Network Evolution | 43 |
| 3.4.2 Theoretical Analysis of the Decision-Centric Model | 45 |
| 3.4.2.1 Numerical Solution of Degree Distribution | 46 |
### 3.4.2.2 Percolation Analysis of Network Robustness Against Random Failure and Targeted Attack
- Page 48

### 3.4.3 Verification
- Page 52
  - 3.4.3.1 Simulation Results of Network Structures
    - Page 53
  - 3.4.3.2 Simulation Results of Network Robustness
    - Page 56

### 3.5 Comparative Study between the Decision-Centric Model and Classical Network Models
- Page 59
  - 3.5.1 Analyzing the Generalized Preferential Attachment Model
    - Page 61
      - 3.5.1.1 Analysis of Degree Distribution
        - Page 61
      - 3.5.1.2 Analysis of Network Properties
        - Page 63
      - 3.5.1.3 Analysis of Network Robustness
        - Page 63
  - 3.5.2 Model Comparison and Discussion
    - Page 65

### 3.6 Closing Comments for Chapter 3
- Page 69

### CHAPTER 4. ESTIMATING LOCAL DECISION-MAKING PREFERENCES IN COMPLEX NETWORKED SYSTEMS
- Page 71
  - 4.1 Chapter Overview
    - Page 71
  - 4.2 Autonomous System (AS) Level Internet
    - Page 73
      - 4.2.1 Background of AS
        - Page 74
      - 4.2.2 Modeling AS-level Internet – State-of-the-Art
        - Page 75
  - 4.3 A Decision-Centric Approach for Estimating AS Peering Preferences and Behaviors
    - Page 78
  - 4.4 Data Collection
    - Page 80
      - 4.4.1 Factors Affecting AS Linking Decisions
        - Page 80
      - 4.4.2 Graph Data of AS-level Internet
        - Page 82
      - 4.4.3 Geographic Data of ASes
        - Page 83
      - 4.4.4 Traffic-related Data and Address Space of ASes
        - Page 83
      - 4.4.5 Business Type Data of ASes
        - Page 84
      - 4.4.6 Preparing Sample Observations
        - Page 85
  - 4.5 Data Analysis – Descriptive Statistics
    - Page 87
      - 4.5.1 Geographic Distance between ASes
        - Page 87
      - 4.5.2 Address Space of ASes
        - Page 87
      - 4.5.3 Business Types of ASes
        - Page 89
      - 4.5.4 Number of Customers and Providers of ASes
        - Page 90
  - 4.6 Model Implementation and Parameter Estimation
    - Page 91
      - 4.6.1 Correlation Analysis
        - Page 91
      - 4.6.2 Regression Analysis and Interpretation
        - Page 91
          - 4.6.2.1 Model 1 for Linking Activity 1 – C2P Connections between New ASes and Existing ASes
            - Page 93
          - 4.6.2.2 Model 2 for Linking Activity 2 – C2P Connections between Two Existing ASes
            - Page 96
          - 4.6.2.3 Model 3 for Linking Activity 3 - P2P Connections between Two Existing ASes
            - Page 97
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.6.3 Model Consistency</td>
<td>100</td>
</tr>
<tr>
<td>4.7 Closing Comments for Chapter 4</td>
<td>102</td>
</tr>
<tr>
<td>CHAPTER 5. THREE APPROACHES TO ESTIMATE LOCAL DECISION-</td>
<td></td>
</tr>
<tr>
<td>MAKING PREFERENCES - A COMPARATIVE STUDY</td>
<td>104</td>
</tr>
<tr>
<td>5.1 Chapter Overview</td>
<td>104</td>
</tr>
<tr>
<td>5.2 Approach 1: Generalized Preferential Attachment Model</td>
<td>106</td>
</tr>
<tr>
<td>5.2.1 Overview</td>
<td>106</td>
</tr>
<tr>
<td>5.2.2 Fitting Power-Law Degree Distribution</td>
<td>109</td>
</tr>
<tr>
<td>5.3 Approach 2: Statistical Regression Model</td>
<td>110</td>
</tr>
<tr>
<td>5.4 Approach 3: Discrete Choice Model</td>
<td>112</td>
</tr>
<tr>
<td>5.5 Estimating AS-Level Behaviors in the Internet Using Three Approaches</td>
<td>113</td>
</tr>
<tr>
<td>5.5.1 Results of Approach 1: Generalized Preferential Attachment</td>
<td>115</td>
</tr>
<tr>
<td>5.5.2 Results of Approach 2: Statistical Regression</td>
<td>118</td>
</tr>
<tr>
<td>5.5.3 Results of Approach 3: Multinomial Logit Choice Model</td>
<td>120</td>
</tr>
<tr>
<td>5.6 Comparison of Three Node-level Behavioral Models</td>
<td>122</td>
</tr>
<tr>
<td>5.6.1 Node-level Behaviors and Generated Networks</td>
<td>122</td>
</tr>
<tr>
<td>5.6.2 Generality and Extensibility</td>
<td>126</td>
</tr>
<tr>
<td>5.6.3 Computational Capability</td>
<td>127</td>
</tr>
<tr>
<td>5.7 Closing Comments for Chapter 5</td>
<td>128</td>
</tr>
<tr>
<td>CHAPTER 6. BEHAVIORAL EXPERIMENTATION AND GAME THEORY</td>
<td></td>
</tr>
<tr>
<td>IN SOCIAL PRODUCT DEVELOPMENT</td>
<td>129</td>
</tr>
<tr>
<td>6.1 Chapter Overview</td>
<td>129</td>
</tr>
<tr>
<td>6.2 Crowdsourcing in Engineering Systems Design</td>
<td>131</td>
</tr>
<tr>
<td>6.3 Behavioral Experimentation for Design Crowdsourcing</td>
<td>132</td>
</tr>
<tr>
<td>6.3.1 Modeling an Engineering Design Problem as a Search Problem</td>
<td>132</td>
</tr>
<tr>
<td>6.3.2 Details of Experimental Design</td>
<td>135</td>
</tr>
<tr>
<td>6.4 Game Theory in Crowdsourcing – A Theoretical Analysis</td>
<td>138</td>
</tr>
<tr>
<td>6.4.1 Modeling Tournaments as Non-Cooperative Games</td>
<td>138</td>
</tr>
<tr>
<td>6.4.2 Game-Theoretic Analysis</td>
<td>140</td>
</tr>
<tr>
<td>6.5 Assess Game-Theoretic Models through Experimental Data</td>
<td>144</td>
</tr>
<tr>
<td>6.5.1 Hypothesis 1: The Expected Number of Tries Decreases as the</td>
<td>144</td>
</tr>
<tr>
<td>Cost Per Trial Increases</td>
<td></td>
</tr>
<tr>
<td>6.5.2 Hypothesis 2: The Solution Quality Monotonically Increases</td>
<td>145</td>
</tr>
<tr>
<td>with the Number of Tries</td>
<td></td>
</tr>
<tr>
<td>6.5.3 Hypothesis 3: Increasing the Number of Tries Increases the</td>
<td>147</td>
</tr>
<tr>
<td>Probability of Winning</td>
<td></td>
</tr>
<tr>
<td>6.6 Closing Comments for Chapter 6</td>
<td>151</td>
</tr>
<tr>
<td>CHAPTER 7. CLOSING THOUGHTS AND FUTURE WORK</td>
<td>154</td>
</tr>
<tr>
<td>7.1 Conclusions and Contributions</td>
<td>154</td>
</tr>
<tr>
<td>7.1.1 Conclusions</td>
<td>154</td>
</tr>
<tr>
<td>7.1.2 Contributions to Design of Complex Networked Systems</td>
<td>156</td>
</tr>
<tr>
<td>Section</td>
<td>Page</td>
</tr>
<tr>
<td>----------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>7.1.3 Contributions to Design of Social Product Development</td>
<td>158</td>
</tr>
<tr>
<td>7.2 Research Challenges and Future Work</td>
<td>159</td>
</tr>
<tr>
<td>7.2.1 Research Challenges</td>
<td>159</td>
</tr>
<tr>
<td>7.2.2 Future Work</td>
<td>162</td>
</tr>
<tr>
<td>LIST OF REFERENCES</td>
<td>165</td>
</tr>
<tr>
<td>VITA</td>
<td>176</td>
</tr>
</tbody>
</table>
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Distinction between hierarchically-designed systems and bottom-up evolutionary systems.</td>
<td>3</td>
</tr>
<tr>
<td>1.2</td>
<td>Research hypothesis and research objective.</td>
<td>5</td>
</tr>
<tr>
<td>1.3</td>
<td>Overview of research questions and approaches.</td>
<td>8</td>
</tr>
<tr>
<td>1.4</td>
<td>Summary of research tasks in validation examples.</td>
<td>13</td>
</tr>
<tr>
<td>1.5</td>
<td>Summary of contributions.</td>
<td>16</td>
</tr>
<tr>
<td>1.6</td>
<td>Framework of the research tasks in the dissertation.</td>
<td>18</td>
</tr>
<tr>
<td>2.1</td>
<td>Mapping of literature review, research gaps and research tasks.</td>
<td>20</td>
</tr>
<tr>
<td>3.1</td>
<td>Critical fraction points for three critical values of $\beta_1$.</td>
<td>51</td>
</tr>
<tr>
<td>3.2</td>
<td>Fitting functions of the degree distribution of representative synthetic networks with three critical values of $\beta_1$.</td>
<td>55</td>
</tr>
<tr>
<td>3.3</td>
<td>Network properties and robustness against random failure and targeted attack for networks generated with critical $\beta_1$ values (values in bracket are the length of 95% confidence interval).</td>
<td>57</td>
</tr>
<tr>
<td>4.1</td>
<td>Summary of existing Internet topology generators.</td>
<td>78</td>
</tr>
<tr>
<td>4.2</td>
<td>Statistics for the new connection events in five consecutive evolution instances from Jan. 5th to Nov. 1st, 2004.</td>
<td>86</td>
</tr>
<tr>
<td>4.3</td>
<td>Descriptive statistics for the number of ASes in each business type on Jan. 5th 2004 and the first evolution from Jan. 5th to Mar. 1st.</td>
<td>89</td>
</tr>
<tr>
<td>4.4</td>
<td>Correlation analysis on the variables for regression analysis.</td>
<td>92</td>
</tr>
<tr>
<td>4.5</td>
<td>Mixed logit estimation results for Model 1 (Random parameters are normally distributed).</td>
<td>94</td>
</tr>
<tr>
<td>4.6</td>
<td>Mixed logit estimation results for Model 2 (Random parameters are normally distributed).</td>
<td>97</td>
</tr>
<tr>
<td>4.7</td>
<td>Mixed logit estimation results Model 3 (Random parameters are normally distributed).</td>
<td>98</td>
</tr>
<tr>
<td>Table</td>
<td>Page</td>
<td></td>
</tr>
<tr>
<td>-------</td>
<td>------</td>
<td></td>
</tr>
<tr>
<td>4.8 Mixed logit estimation results for five consecutive evolution instances from Jan. 5th to Nov. 1st, 2004.</td>
<td>101</td>
<td></td>
</tr>
<tr>
<td>5.1 An overview of three approaches presented in this chapter.</td>
<td>114</td>
<td></td>
</tr>
<tr>
<td>5.2 Comparison of linking probability.</td>
<td>123</td>
<td></td>
</tr>
<tr>
<td>5.3 Network metrics for evaluation.</td>
<td>126</td>
<td></td>
</tr>
<tr>
<td>5.4 Comparison of metrics in simulated networks.</td>
<td>126</td>
<td></td>
</tr>
<tr>
<td>5.5 Comparison of the three estimation approaches.</td>
<td>128</td>
<td></td>
</tr>
<tr>
<td>6.1 Number of participants and cost (c) settings in each session. Low cost = 10 tokens, High cost = 20 tokens.</td>
<td>136</td>
<td></td>
</tr>
<tr>
<td>6.2 The rational reaction sets and Nash equilibria for different choices of quality functions (QF) and contest success functions (CSF) for two-player games.</td>
<td>142</td>
<td></td>
</tr>
<tr>
<td>6.3 Mean ($\mu$) and standard deviation ($\sigma$) of number of tries ($e$) in the treatments with different cost settings.</td>
<td>145</td>
<td></td>
</tr>
<tr>
<td>6.4 Summary of four hypothesis tests for mean value of number of tries $\mu_e$.</td>
<td>145</td>
<td></td>
</tr>
<tr>
<td>6.5 Estimation for regression coefficients in Equation (6.18).</td>
<td>147</td>
<td></td>
</tr>
<tr>
<td>6.6 Average and standard deviation of the number of tries categorized by winning status with different cost settings.</td>
<td>148</td>
<td></td>
</tr>
<tr>
<td>6.7 Estimation for regression coefficients in Equation (6.21).</td>
<td>149</td>
<td></td>
</tr>
<tr>
<td>6.8 Welch two sample t-test for average number of tries in sessions with different initial costs.</td>
<td>150</td>
<td></td>
</tr>
</tbody>
</table>
## LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>2</td>
</tr>
<tr>
<td>1.2</td>
<td>5</td>
</tr>
<tr>
<td>1.3</td>
<td>6</td>
</tr>
<tr>
<td>1.4</td>
<td>11</td>
</tr>
<tr>
<td>1.5</td>
<td>18</td>
</tr>
<tr>
<td>3.1</td>
<td>35</td>
</tr>
<tr>
<td>3.2</td>
<td>40</td>
</tr>
<tr>
<td>3.3</td>
<td>48</td>
</tr>
<tr>
<td>3.4</td>
<td>49</td>
</tr>
<tr>
<td>3.5</td>
<td>53</td>
</tr>
<tr>
<td>3.6</td>
<td>54</td>
</tr>
<tr>
<td>3.7</td>
<td>55</td>
</tr>
<tr>
<td>3.8</td>
<td>57</td>
</tr>
<tr>
<td>3.9</td>
<td>59</td>
</tr>
<tr>
<td>3.10</td>
<td>62</td>
</tr>
<tr>
<td>3.11</td>
<td>63</td>
</tr>
<tr>
<td>3.12</td>
<td>64</td>
</tr>
<tr>
<td>3.13</td>
<td>65</td>
</tr>
<tr>
<td>4.1</td>
<td>72</td>
</tr>
<tr>
<td>Figure</td>
<td>Page</td>
</tr>
<tr>
<td>--------</td>
<td>------</td>
</tr>
<tr>
<td>4.2</td>
<td>A visualization of AS-level Internet graph from Cooperative Association for Internet Data Analysis (CAIDA) [96].</td>
</tr>
<tr>
<td>4.3</td>
<td>Overview of the approach.</td>
</tr>
<tr>
<td>4.4</td>
<td>Overview of the data collection process.</td>
</tr>
<tr>
<td>4.5</td>
<td>The geographic distribution of ASes on Jan. 5th, 2004 in continental US.</td>
</tr>
<tr>
<td>4.6</td>
<td>The distribution of geographic distances of AS connections on Jan. 5th, 2004.</td>
</tr>
<tr>
<td>4.7</td>
<td>The distribution of address space of ASes in the AS Internet graph on Jan. 5th, 2004.</td>
</tr>
<tr>
<td>4.8</td>
<td>The distribution of ASes connectivity in AS Internet graph on Jan. 5th, 2004.</td>
</tr>
<tr>
<td>5.1</td>
<td>Overview of Chapter 5: research questions, approaches and tasks.</td>
</tr>
<tr>
<td>5.2</td>
<td>Complementary cumulative degree distribution of networks generated by generalized BA model with different $\lambda$ values.</td>
</tr>
<tr>
<td>5.3</td>
<td>Data extraction for estimating the node-level behaviors.</td>
</tr>
<tr>
<td>5.4</td>
<td>Complementary cumulative degree distribution of AS-level Internet on Jan 5th, 2004.</td>
</tr>
<tr>
<td>5.5</td>
<td>Exponent ($\gamma$) in the degree distribution vs. network size.</td>
</tr>
<tr>
<td>5.6</td>
<td>Number of edges ($E$) vs network size ($J$).</td>
</tr>
<tr>
<td>5.7</td>
<td>The exponent ($\gamma$) of 122 AS-level Internet networks obtained with the Maximum Likelihood fitting with goodness-of-fit tests based on Kolmogorov-Smirnov (KS) statistic and likelihood ratio.</td>
</tr>
<tr>
<td>5.8</td>
<td>Node-level behavior for three network evolution steps.</td>
</tr>
<tr>
<td>5.9</td>
<td>Network size vs. parameters in the node’s decision model 2.</td>
</tr>
<tr>
<td>5.10</td>
<td>Network size vs. parameters in the node’s decision model 3.</td>
</tr>
<tr>
<td>5.11</td>
<td>Comparison of the node-level behavior of AS-level Internet on Jan. 5th, 2004 deduced by three approaches.</td>
</tr>
<tr>
<td>5.12</td>
<td>Comparison of complementary cumulative degree distribution between the real network and simulated network with three approaches.</td>
</tr>
<tr>
<td>5.13</td>
<td>KL divergence on the degree distribution of simulated networks.</td>
</tr>
<tr>
<td>6.1</td>
<td>Overview of Chapter 6: research questions, approaches and tasks.</td>
</tr>
<tr>
<td>Figure</td>
<td>Description</td>
</tr>
<tr>
<td>--------</td>
<td>-----------------------------------------------------------------------------</td>
</tr>
<tr>
<td>6.2</td>
<td>Relationship between parameters.</td>
</tr>
<tr>
<td>6.3</td>
<td>Screenshot of the interface used by the participants.</td>
</tr>
<tr>
<td>6.4</td>
<td>Relationship between number of tries and quality of solution.</td>
</tr>
<tr>
<td>6.5</td>
<td>Predicted winning probability as a function of number of tries ($e$) for four scenarios.</td>
</tr>
</tbody>
</table>
SYMBOLS

\( U \) Utility function
\( V \) Representative/Observed utility
\( \epsilon \) Unobserved utility
\( x \) A vector of alternative-specific variables
\( s \) A vector of decision-maker-specific variables
\( \tilde{x} \) A vector of alternative-specific, non-network-metric variables
\( \tilde{s} \) A vector of decision-maker-specific, non-network-metric variables
\( \hat{x} \) A vector of alternative-specific, network-metric variables
\( \hat{s} \) A vector of decision-maker-specific, network-metric variables
\( \beta \) Parameter vector that quantifies the individual preferences
\( P \) Probability
\( d \) Node’s degree
\( t \) Time
\( N \) Network size
\( \beta_1 \) Parameter that quantifies the node’s preference to degree
\( S \) Fraction of nodes that in the giant cluster
\( u \) The average probability that a node is not connected to a giant cluster
\( f_c \) Critical fraction point
\( \phi_c \) Percolation threshold
\( \phi \) The probability that a node is present in the network
\( \xi \) Threshold of LCC index for network fragmentation
\( A \) Node’s additional attractiveness
\( G \) Node’s fitness
\( S_d \) A group of nodes with the same degree \( d \)
\( n_d \)  The number of nodes in group \( S_d \)
\( L \)  The total number of links created in a specific period
\( l_d \)  The number of new links created with existing nodes in group \( S_d \)
\( \Pi \)  Fixed winning prize
\( c \)  Cost of performing a search in design crowdsourcing
\( e_i \)  The effort that subject \( i \) spends in design crowdsourcing
\( C_i \)  The total cost of subject \( i \)
\( q_i \)  The quality of solution of subject \( i \) in design crowdsourcing
\( K_i \)  The knowledge and expertise of subject \( i \) in design crowdsourcing
\( E \)  Expected payoff
\( \sigma \)  Standard deviation of a distribution
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>AS</td>
<td>Autonomous System</td>
</tr>
<tr>
<td>ABM</td>
<td>Agent-based Model</td>
</tr>
<tr>
<td>ACC</td>
<td>Average Cluster Coefficient</td>
</tr>
<tr>
<td>APL</td>
<td>Average Path Length</td>
</tr>
<tr>
<td>ARIN</td>
<td>American Registry for Internet Numbers</td>
</tr>
<tr>
<td>ASIM</td>
<td>AS Simulation Model</td>
</tr>
<tr>
<td>BA</td>
<td>Barabsi-Albert</td>
</tr>
<tr>
<td>BAU</td>
<td>Business Access Utility</td>
</tr>
<tr>
<td>BGP</td>
<td>Border Gateway Protocol</td>
</tr>
<tr>
<td>BRD</td>
<td>Business Requirements Document</td>
</tr>
<tr>
<td>BS</td>
<td>Business Service</td>
</tr>
<tr>
<td>C2P</td>
<td>Customer to Provider</td>
</tr>
<tr>
<td>CAIDA</td>
<td>Cooperative Association for Internet Data Analysis</td>
</tr>
<tr>
<td>CSF</td>
<td>Contest Success Function</td>
</tr>
<tr>
<td>DC</td>
<td>Decision-Centric</td>
</tr>
<tr>
<td>DCA</td>
<td>Discrete Choice Analysis</td>
</tr>
<tr>
<td>DCM</td>
<td>Discrete Choice Model</td>
</tr>
<tr>
<td>DBDC</td>
<td>Degree-based Decision-Centric</td>
</tr>
<tr>
<td>DFMS</td>
<td>Design for Market Systems</td>
</tr>
<tr>
<td>FANG</td>
<td>Fast Adaptable Next-Generation Ground Vehicle</td>
</tr>
<tr>
<td>GPA</td>
<td>Generalized Preferential Attachment</td>
</tr>
<tr>
<td>HOT</td>
<td>Highly Optimized Tolerance</td>
</tr>
<tr>
<td>IFV</td>
<td>Infantry Flight Vehicle</td>
</tr>
<tr>
<td>ISP</td>
<td>Internet Service Provider</td>
</tr>
<tr>
<td>LCC</td>
<td>Largest Connected Component</td>
</tr>
<tr>
<td>Acronym</td>
<td>Definition</td>
</tr>
<tr>
<td>---------</td>
<td>------------</td>
</tr>
<tr>
<td>MLE</td>
<td>Maximum Likelihood Estimation</td>
</tr>
<tr>
<td>NS</td>
<td>Network Service</td>
</tr>
<tr>
<td>OLS</td>
<td>Ordinary Least Square</td>
</tr>
<tr>
<td>P2P</td>
<td>Peer to Peer</td>
</tr>
<tr>
<td>PFP</td>
<td>Positive Feedback Preference</td>
</tr>
<tr>
<td>PLRG</td>
<td>Power Law Random Graph</td>
</tr>
<tr>
<td>PSO</td>
<td>Popularity × Similarity Optimization</td>
</tr>
<tr>
<td>QA</td>
<td>Quality Assurance</td>
</tr>
<tr>
<td>QF</td>
<td>Quality Function</td>
</tr>
<tr>
<td>RA</td>
<td>Residential Access</td>
</tr>
<tr>
<td>RQ</td>
<td>Research Question</td>
</tr>
<tr>
<td>RS</td>
<td>Retail Service</td>
</tr>
<tr>
<td>RRS</td>
<td>Rational Reaction Sets</td>
</tr>
<tr>
<td>RUT</td>
<td>Random Utility Theory</td>
</tr>
<tr>
<td>WH</td>
<td>Web Hosting</td>
</tr>
<tr>
<td>WS</td>
<td>Watts-Strogatz</td>
</tr>
</tbody>
</table>
ABSTRACT

Sha, Zhenghui Ph.D., Purdue University, August 2015. Decision-Centric Foundations for Complex Systems Engineering and Design. Major Professor: Jitesh H. Panchal, School of Mechanical Engineering.

During the past decade, there has been significant interest in bottom-up evolutionary complex systems, such as the Internet, smart transportation systems and social product development. Compared with hierarchically-designed systems such as automotive and aircraft, the architectures of such complex systems are not under the direct control of designers, but emerge in a bottom-up manner based on decisions made by individual entities. The design strategy for bottom-up evolutionary complex systems is to influence agents’ behavior at the micro-level in order to indirectly achieve the requirements and desired performance at the system-level. To this end, the research objective of this dissertation is to establish a framework to model, analyze and estimate the micro-level decision-making behaviors for facilitating complex systems engineering and design.

Existing studies have provided insights on modeling micro-level behaviors and understanding their effects on system-level performance. However, there is a lack of theoretical foundations for explaining why agents’ behaviors are modeled according to the rationality assumptions and whether or not such assumptions are appropriate. In other words, existing studies are primarily focused on the outcomes of decision-making (micro-level behavior) instead of the reasons for those decisions (decision-making preferences). There is a research gap in understanding the effects of agents’ decision-making preferences on system structure and dynamics. To address this research gap, a decision-centric framework is proposed in this dissertation. This framework provides theoretical foundations for explaining agents’ rational behavior, establishes a link
between decision-making preferences and the micro-level behaviors, and builds the relationship between agents’ preferences and system-level performance.

Towards establishing the decision-centric framework, three research questions regarding modeling, analyzing and estimating micro-level behavior are addressed. The approaches proposed for answering the questions are validated using two applications examples: the Internet and social product development. In the case of the Internet, an approach based on discrete choice random utility theory integrated with complex network analysis is proposed to obtain micro-level behavioral models. The primary outcome is a generic decision-centric approach for modeling the evolution of complex systems. In the case of social product development, by integrating game theory and behavioral experimentation, a rigorous framework is established for modeling the decision-making behaviors of individuals in crowdsourcing competitions, and understanding its effects on the design outcomes.

The overall contribution in this dissertation is a decision-centric framework to support bottom-up engineering and design of complex systems. The approaches established in this dissertation support the attainment of new knowledge on modeling, analysis and estimation of micro-level behaviors in complex systems. The results presented in this dissertation provide insights on directing the design of incentives and mechanisms to influence agents’ interactions at the micro-level to achieve desired system-level performance in complex networked systems, and to improve the crowdsourcing process for better design outcomes.
CHAPTER 1. OVERVIEW OF THE DISSERTATION

1.1 Research Background and Motivation

With our fast-evolving society, engineering is facing challenges, such as security in cyberspace, healthy environment and energy sustainability, which cannot be solved by a single engineering product or system. When seeking ways to address the challenges in order to meet fundamental human needs, the solutions often lead to large-scale complex systems that encompass various areas such as physical, psychological, economic and cultural domains. In complex systems, there are a large number of interactive entities (agents or components or parts) that work together to build a system of value greater than the sum of the individual parts.

Recently, there have been increasing interests in a special class of complex systems that evolve in a bottom-up manner. “Bottom-up” means that systems evolve as a result of interactions among individual entities. In such systems, the entities are heterogeneous in nature and each one has its own interests. Rational decisions are made to realize these private goals. Such decisions result in local interrelations and interactions. Apart from this, individual entities may be at different levels in a system hierarchy, such as policy makers vs. operators. Accordingly, interactions also exist between levels, resulting in recursive levels of integration. All these interactions and decision-making behavior at lower levels in turn affect the performance at the system level. A well-known example of such systems is the smart electric grid, which consists of a wide range of decision-makers at different levels including consumers, utilities, micro-grid operators, and other participants in the distribution infrastructure. The energy producers, distributors, and utilities independently make technical decisions within the rules and regulations to meet the system level objectives while maximizing

1In this dissertation, the individual entity, agent and component are used interchangeably.
their own profits. These objectives include reliability, security and load demand. The decisions made by the stakeholders affect the technical, social, economic, and environmental performance [1]. Similar relationships between local decisions and system-level performance also exist in other bottom-up evolutionary complex systems, such as the Internet, air-transportation systems, smart urban systems and social product development. Figure 1.1 displays some examples.

Design research on complex systems has been traditionally focused on engineered systems, such as automotive and aerospace systems. These systems are indeed complex in structure and large in scale. But they are hierarchically designed, which is fundamentally different than the bottom-up evolutionary complex systems. Table 1.1 shows a comparison between hierarchically designed systems and bottom-up evolutionary systems. The design of such systems starts from the system requirements analysis and is driven by top-down hierarchical decomposition, followed by the design of sub-systems and components. Thereafter, the component testing, subsystem testing, and finally systems testing are implemented. The component-level designs are integrated into a complete system and validated against system-level requirements. This general process is embodied in various existing systematic design methods (e.g.,
Pahl and Beitz [2]), systems engineering models (e.g., Systems Engineering Vee [3]), and systems engineering processes adopted by organizations such as NASA [4].

Table 1.1. Distinction between hierarchically-designed systems and bottom-up evolutionary systems.

<table>
<thead>
<tr>
<th></th>
<th>Hierarchically-Designed Systems</th>
<th>Bottom-up Evolutionary Systems</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Defining characteristics</strong></td>
<td>Set of interacting components forming an integrated whole by designers</td>
<td>System evolves as a result of decisions and behaviors of individual self-directed entities</td>
</tr>
<tr>
<td><strong>Local components/entities</strong></td>
<td>No decision-making ability</td>
<td>Self-directed, generally selfish</td>
</tr>
<tr>
<td><strong>Design variables</strong></td>
<td>Components' attributes such as dimensions, material, etc.</td>
<td>Incentives to local entities and interacting mechanisms</td>
</tr>
<tr>
<td><strong>Design strategy</strong></td>
<td>Top-down hierarchical design</td>
<td>Bottom-up design</td>
</tr>
<tr>
<td><strong>Examples of design problems</strong></td>
<td>Design of transportation network layout, traditional power grid design assignment, machine design etc.</td>
<td>Traffic mechanism design, protocol design for Internet, policy design for green energy, incentive design etc.</td>
</tr>
</tbody>
</table>

Due to these distinct characteristics, the traditional top-down design approaches are not suitable for bottom-up evolutionary systems. The focus of engineering design in the context of bottom-up evolutionary systems should be on the individual entities at the micro level. Particularly, the decision-making nature of entities calls for a decision-based design. However, existing studies in decision-based design (DBD), e.g., the one proposed by Hazerlrigg [5], heavily rely on the top-down design methodology particular to hierarchically designed systems. This manifests in two ways:

- Integrating decision-making into design methodology formulates a design problem as an optimization problem. For example, an optimization problem meets functional requirements subject to constraints with a specific objective, such as minimizing cost.
• The DBD framework is mainly focused on designing artifacts in which the components have no decision-making ability. Instead, designers or design teams have the overall control over the design process. Further, DBD is focused on integrating designers’ choices into the optimization framework.

In the design of complex systems, a perspective of design that is both decision-centric and bottom-up is desired. Rather than directly controlling the system structure, the behaviors of the interacting entities must be influenced in a way that leads to desired system structure. Such influence requires improved coordination mechanisms and global policies, which can be realized through the provision of incentives, imposition of penalties or taxes and definition of rules or protocols.

1.2 Research Overview

In bottom-up design, the challenge is that the incentive structures should be carefully designed so that individual entities are pursuing their private goals and desired system structures are simultaneously obtained without compromising the system-level performance. The focus of this dissertation is the creation of knowledge that would help system designers and researchers in achieving successful bottom-up design of complex systems\(^2\). Table 1.2 shows the central hypothesis under which the research is conducted, and the overall research objective. The central hypothesis is that the observed evolutionary dynamics in complex systems result from the decision-making activities among agents. Under this hypothesis, the knowledge that facilitates the bottom-up design of complex systems can be acquired through the understanding of how individual entities make decisions. The overall research objective is to establish a framework to model, analyze and estimate the local decision-making preferences and behaviors for facilitating complex systems engineering and design.

Figure 1.2 provides an overview of the research. In order to achieve the research objective, a number of approaches are developed and a feasible integration of theories\(^2\)The complex systems in the rest of the dissertation refers to bottom-up evolutionary complex systems.
Table 1.2. Research hypothesis and research objective.

<table>
<thead>
<tr>
<th>Central hypothesis</th>
<th>The observed evolutionary dynamics in complex systems are a result of the decision-making activities among agents.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Research objective</td>
<td>To establish a framework to model, analyze and estimate the local decision-making preferences and behaviors for facilitating complex systems engineering and design.</td>
</tr>
</tbody>
</table>

from different domains is realized. These include analytical models and behavioral experiments developed by integrating knowledge areas in discrete choice analysis, complex networks, applied statistics, game theory, behavior economics and contest theory. The framework, approaches, models and tools developed in this dissertation are used to generate knowledge that facilitates the design of complex systems in different application areas, such as the Internet, the smart grid, air transportation systems and systems of social product development. In this dissertation, we use the Internet and social product development as application examples. In the following section, the specific research questions are presented and formulated in a decision-centric framework.

Figure 1.2. Research overview.
1.3 Research Questions

The underlying dynamics of complex systems can be understood by modeling the systems along five levels shown in Figure 1.3:

- Level 1: individual preferences,
- Level 2: micro-level behavior,
- Level 3: system structure,
- Level 4: system properties, and
- Level 5: system-level performance.

This framework places system analysis and design on a decision-centric foundation, and highlights the hierarchical interactions between individual components (local) and the system level (global). In system science, existing studies have provided theoretical insights on modeling micro-level behaviors (Level 2) and analyzing their effects on system structure (Level 3), properties (Level 4) and performance (Level 5). For example, in the field of complex networks, studies have focused on developing

Figure 1.3. Five-level decomposition and associated research questions.
models for nodes’ linking probability to describe the linking behavior. The models are capable of generating scale-free degree distribution that exhibits high cluster coefficient and low system robustness. Another set of examples are complex systems in which the human is a part of the system. Existing studies are focused on using models, such as game-theoretic models, to model the interactive behavior and analyze the resulting solution (e.g., Nash equilibrium), which is used to evaluate the system performance. In these studies, micro-level behavior models are put forward directly. Theoretical foundations that explain why agents’ behaviors are modeled according to the assumptions of rationality and whether such assumptions are appropriate or not are altogether lacking. In other words, previous decision-making studies place too much emphasis on the outcomes of decision-making (micro-level behavior) instead of the reasons for those decisions (decision-making preferences). The relationship between individual preference (Level 1) and the micro-level behavior (Level 2) is unexplored. This oversight makes existing models unsuitable for modeling complex systems engineering and design process in a bottom-up manner.

Therefore the research gap is that the effect of individual preferences on the system structure, properties and performance is not clearly understood. This research gap impedes our understanding of the system’s evolutionary dynamics and the ability to establish ways to direct their evolution. In order to address this fundamental research gap, it is crucial to explicitly model the decision-making preferences of individual entities and analyze the impact of preferences on system performance. Specifically, there are three research questions (RQs) to address:

- **RQ1**: How can individual decision-making preferences and behaviors be modeled?
- **RQ2**: How can the effects of local decision-making preferences and behaviors be understood?
- **RQ3**: How can unobserved decision-making preferences and behaviors of individual entities be estimated?
The central hypothesis is validated by addressing these three research questions. Table 1.3 summarizes the research approaches used to address these research questions.

Table 1.3. Overview of research questions and approaches.

<table>
<thead>
<tr>
<th>Research Questions</th>
<th>Research Approaches</th>
</tr>
</thead>
<tbody>
<tr>
<td>RQ1</td>
<td>1.1 integrating discrete choice random utility theory with complex networks analysis.</td>
</tr>
<tr>
<td>How can individual decision-making preferences and behaviors be modeled?</td>
<td>1.2 integrating game theory and behavioral experimentation.</td>
</tr>
<tr>
<td>RQ2</td>
<td>2.1 Performing quantitative simulation.</td>
</tr>
<tr>
<td>How can the effects of local decision-making preferences and behaviors be understood?</td>
<td>2.2 Designing and executing human-subject experiments.</td>
</tr>
<tr>
<td>RQ3</td>
<td>3.1 Using field data about the system structure to statistically estimate decision-makers’ preferences using discrete-choice models.</td>
</tr>
<tr>
<td>How can unobserved decision-making preferences of individual entities be estimated?</td>
<td>3.2 Using targeted data about human decisions in actual design to estimate the designers’ decision-making behaviors with game-theoretic models.</td>
</tr>
</tbody>
</table>

RQ1 focuses on modeling individual decision-making preferences and behaviors. The aim is to set up a relationship between individual preferences and micro-level behaviors (see Figure 1.3). In this dissertation, under the assumption that the individual entities behave rationally, random utility theory and contest theory are adopted as the mathematical frameworks to formulate decision-making preferences into probabilistic behavioral models. For example, random utility theory assumes that individuals seek to maximize their own utility, $U$, when making a choice from the choice set. $U$ is modeled as a linear combination of decision criteria, represented by $X = \{x_1, x_2, ..., x_n\}$, in which $x_i$ is the $i$’s decision criterion. Then, $U = \beta X$, where, $\beta = \{\beta_1, \beta_2, ..., \beta_n\}$ quantifies the preferences for the criteria. The utility-maximization assumption results in a closed form of choice probability (i.e., behavior) modeled as a function of
which incorporates preferences as parameters in the models (See Chapter 3 for more details).

RQ2 focuses on analyzing the effects of decision-making preferences. Through modeling the approaches proposed in RQ1, the effects of individual preferences (Level 1) on other levels in Figure 1.3 can be analyzed. Specifically, the research analyzes the effects on system topology, system robustness and system efficiency. Given different application contexts, technical systems or social systems, the analysis can be performed by quantitative simulation or human-subject experiments. For example, in complex networks, the functional relationship between individual decision-making preferences and system-level performance can be established. The effect on system performance can then be analyzed with quantitative simulation by changing the values of preferences. On the other hand, in systems involving humans, such as social product development with designed human-subject experiments, the effects of decision-making preferences of participants on the quality of design can be analyzed by changing the incentive structures, such as the cost or prize amount.

RQ3 focuses on estimating individual decision-making preferences. In systems design, the decision-makers’ preferences are often hidden from observers, and direct access to decision makers is very difficult. An alternate way is to infer preferences indirectly from the data regarding decisions that have been made in real choice situations. For example, in the air transportation system, the decision-making of airline companies on route planning is difficult to obtain because of commercial regulations and confidentiality. However, we can observe how routes are added and deleted every year. Changing routes reflect how airlines make decisions under specific social and economic circumstances. This research is focused on obtaining the unobserved preferences of individual entities through the data of decision criteria. Depending on the mathematical framework used in answering RQ1, the parameterized preferences can be estimated using data-driven approaches, such as maximum likelihood estimation and Bayesian estimation.
By answering the three RQs, insights and new knowledge for the bottom-up design of complex systems are obtained. For example, the results from RQ2 provide direct functional mapping from the micro-level preferences to system-level performance. Such mapping guides the design of incentives that influence individual entities to avoid critical preferences which result in inferior system performance. Similarly, the results obtained from RQ3 reveal the decision-makers’ unobserved preferences in systems design. These results help systems designers and policy makers better understand the local interactions and heterogeneity in agents, which are crucial to incentive/mechanism design. So the three research questions form a closed loop at the micro-level. The answers to the RQs work together as a holistic decision-centric framework to tackle research issues in complex systems engineering and design.

1.4 Validation Examples and Applications

In this dissertation, two applications are used to validate the approaches shown in Table 1.3. The first application area is complex networked systems in which the individual components are autonomous entities. The second application area is social product development in which the individual entities are humans. In complex networked systems, the autonomous system (AS) level Internet is chosen as an illustrative example. In social product development, design crowdsourcing is chosen for study. The purpose of choosing different applications is two-fold: 1) to validate the generality of the proposed models, approaches and computational techniques proposed in the research, and 2) to show how theories from different knowledge areas can be integrated to facilitate answering the research questions in different domains.

As shown in Figure 1.4, the mapping of each application onto the five-level decision-centric framework facilitates describing the system characteristics. Using the Internet as an example, the decision-centric framework captures the decision-making nature of ASes. Two objectives of each AS are to minimize the cost of building links with other ASes and to minimize the sum of the distances to all other ASes (to
minimize delay in sending and receiving packets) [6]. Since both objectives are in conflict, a trade-off is required. Each AS makes peering decisions to best route data based on commercial, contractual relationships. The node-level peering preferences
(Level 1) and behaviors (Level 2) have significant impact on the global structure and properties of the Internet (Level 3 and 4). This global structure in turn affects the overall system-level structure and performance (Level 5), such as transit speed and routing efficiency. Thus, node-level behavior is crucial to understand overall network performance. Similarly, the underlying dynamics of design crowdsourcing can also be modeled using the five-level framework shown in Figure 1.4(b).

The goal of each application is as follows:

- **AS-level Internet**: to estimate ASes linking preferences and behavior for modeling the evolution of Internet.

- **Design crowdsourcing**: to understand contestants’ decision-making behavior for facilitating the design of crowdsourcing competitions.

While both examples are related to the complex systems theme, the first example is focused on the study of complex systems design through the data on system structure (i.e., the design outcome), whereas the second example is focused on the study of complex systems design through the data on design process. With regard to the mapping of research questions, the study of AS-level Internet answer all three RQs, and the study of design crowdsourcing focuses more on RQ1 and RQ3. Specifically, in the context of AS-level internet, RQ1, RQ2 and RQ3 are answered using the approaches 1.1, 2.1 and 3.1 shown in the Table 1.3. In the context of design crowdsourcing, RQ1 and RQ3 are answered with approaches 1.2 and 3.2. RQ2 in this application is related to the incentives analysis and mechanism design, but outside the scope of this dissertation. More discussion regarding incentive design and analysis is presented in the future work in Chapter 7.

Table 1.4 illustrates the detailed research in each application study. In the example of AS-level Internet, each individual AS is a decision-maker that determines which target ASes to link with. By assuming that the motivation of ASes’ decision-making is to maximize utilities, discrete choice random utility theory can be used to model the ASes’ linking behavior. The random utility theory allows us to model the utilities in
Table 1.4. Summary of research tasks in validation examples.

<table>
<thead>
<tr>
<th>Theoretical Study on complex networks</th>
<th>1. Established a decision-centric framework for modeling node-level preferences and behavior in complex networked systems.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2. Modeled the network evolution within a decision-centric framework.</td>
</tr>
<tr>
<td></td>
<td>3. Analyzed the proposed decision-centric model theoretically.</td>
</tr>
<tr>
<td></td>
<td>4. Established the connection from node-level preferences to network structure, and to system-level performance.</td>
</tr>
<tr>
<td></td>
<td>5. Verified by analyzing the synthetic networks generated with proposed model.</td>
</tr>
<tr>
<td></td>
<td>6. Compared the proposed model with the classical network models.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Empirical Study on AS-level Internet</th>
<th>1. Established a stepwise framework to implement the decision-centric approach.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3. Applied the stepwise framework to estimate the AS linking preferences by addressing the economic, traffic and geographic constraints.</td>
</tr>
<tr>
<td></td>
<td>4. Explored other estimation approaches, and compared these approaches through modeling the system structures with estimated local behavioral models.</td>
</tr>
<tr>
<td></td>
<td>5. Established a decision-centric agent-based model for modeling AS-level Internet topology and evolutionary dynamics.</td>
</tr>
<tr>
<td></td>
<td>6. Validated the proposed decision-centric agent-based model by reconstructing the Internet topology and comparing it with other proposed different models.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Theoretical and Experimental Study on design crowdsourcing</th>
<th>1. Established game-theoretic models of contestants’ decision-making behavior in design crowdsourcing competitions.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2. Designed and developed an economic decision game (i.e., the behavioral experiment).</td>
</tr>
<tr>
<td></td>
<td>3. Executed the experiment; performed the data collection and data analysis.</td>
</tr>
<tr>
<td></td>
<td>4. Estimated contestants’ behaviors in engineering design with statistical models.</td>
</tr>
<tr>
<td></td>
<td>5. Compared the results obtained from experiments with the results predicted by the theoretical models to gain insights for refining the theoretical models.</td>
</tr>
</tbody>
</table>

terms of the observed characteristics of ASes and the uncertainty due to unobserved characteristics. Study of this application features two parts: theoretical study and empirical study. In the theoretical study, the focus is on studying the underlying dynamics in networks. The main task is to establish a decision-centric approach to model the evolution of complex networks. This task is accomplished in six steps.
The node-level linking preferences and behavior are first modeled as a probabilistic model using discrete choice models. Then, a modeling framework is established by integrating discrete choice models with network growth mechanisms. To analyze the effect of node-level preferences on the system-level performance, theoretical analysis and simulation are performed. The results are compared with a well-known network generation model - generalized preferential attachment model (See Chapter 3 for details).

In the empirical study, the focus is on the application of the decision-centric approach to model the evolution of AS-level Internet by using the estimated AS-level linking preferences. Similarly, this task is accomplished in six steps. Using the discrete choice models as the core, a stepwise framework for estimating AS-level linking preferences is presented. Through the investigation of peering types, the key decision-making criteria in economic, geographic and technological aspects are identified. The corresponding data are collected. The stepwise framework is applied using the collected data (see Chapter 4 for details). Finally, the proposed approach is evaluated through a comparative study. In the comparative study, the node-level behavioral models obtained from three different approaches are compared and evaluated. The validation is then performed by reconstructing the AS-level Internet topology with the estimated AS behavior and the proposed network evolution mechanism (see Chapter 5 for details). In summary, the approach developed in this study shows how discrete choice random utility theory can be integrated with complex network analysis to estimate the individual entities’ decision-making preferences and to model the evolutionary dynamics of the Internet.

In the example of design crowdsourcing, the aim is to establish a framework for understanding participants’ decision-making behavior in crowdsourcing within the context of engineering design. The outcomes help in answering the following questions: 1) how design in the competitive environment is performed; 2) what are the human motivations to participate; 3) what are the winning strategies in design competitions. The research is performed both theoretically and experimentally. For the
theoretical study, contestants’ strategic decision-making behavior in crowdsourcing is modeled as a non-cooperative game from contest theory. The theoretical models are validated by employing behavioral experimentation in which an economic decision game is designed and executed. Data on participants’ actual behavior are obtained from the experiment. Such data are then analyzed to estimate contestants’ decision-making preferences to validate the outcomes predicted from the purely theoretical models (See Chapter 6 for details). In summary, through the integration of contest theory and behavioral experimentation, the results obtained reveal deviations from rational human behaviors, e.g., biases that are not structurally included in the theoretical models. New insights on refining the theoretical models are obtained. Additionally, such insights facilitate the design of crowdsourcing mechanism for engineering design with the goal of realizing high quality of solution, small effort invested by the contestants, and high-efficiency/low-cost of running competitions.

1.5 Summary of Contributions

The overall contribution of this dissertation is a decision-centric framework to model, analyze and estimate individual decision-making preferences and micro-level behavior for facilitating complex systems engineering and design. The proposed decision-centric framework builds an appropriate level of abstraction of the complex systems, which ensures applicability across domains. With such a framework, fundamental research questions are identified and answered in two applications. Table 1.5 lists the specific contributions corresponding to the research questions.

By addressing RQ1, an agent-based decision-centric approach based on discrete choice models is developed for modeling the structure and evolution of complex networked systems. Also, a general approach based on contest theory and game theory is developed for modeling strategic decision-making of participants in design competitions. Both approaches illustrate how theories from different disciplines can be synergistically integrated to facilitate the bottom-up design of complex systems.
Table 1.5. Summary of contributions.

<table>
<thead>
<tr>
<th>Research Questions</th>
<th>Contributions</th>
</tr>
</thead>
</table>
| RQ1 | 1. An agent-based decision-centric approach based on discrete choice models for modeling complex networked systems and their evolution.  
2. A general approach based on contest theory and game theory for modeling strategic decision-making in design competitions. |
| RQ2 | 3. A theoretical framework based on continuum theory and percolation theory to analyze the effect of node-level preferences on network topology and robustness.  
4. A systematic experimentation for simulating design crowdsourcing and testing human behaviors and their effects in response to the experiment design options. |
| RQ3 | 5. A stepwise framework based on discrete choice models, complex network analysis and applied statistics for estimating node-level decision-making preferences in complex networked systems.  
6. A data-driven approach for eliciting participants’ decision-making preferences in crowdsourced design. |

By addressing RQ2, a theoretical framework based on continuum theory and percolation theory is established to analyze the effect of node-level preferences on topology and robustness of complex networked systems modeled as networks. In this approach, several methods/algorithms are proposed for overcoming the computational difficulties in large-scale systems. For example, a recursive ODE solving algorithm is developed for analyzing the evolution of complex networks with the proposed decision-centric degree based model (see Section 3.4.2.1 for details). Additionally, as part of answering RQ2 in the context of design crowdsourcing, systematic experimentation is performed for simulating the real situation in design crowdsourcing. The experiment can be used to test human behaviors and their effects as related to the experiment
design options. This approach helps researchers gain domain-specific insights on the actual behavior of humans in systems, which are crucial in the bottom-up design of complex systems.

By addressing RQ3, a stepwise framework based on discrete choice models, complex network analysis and applied statistics is established for estimating the node-level decision-making preferences in complex networked systems. In the example of design crowdsourcing, a data-driven approach for eliciting participants’ decision-making preferences is developed. This dissertation shows the appropriate combination of sets of methods in statistical regression, estimation and hypothesis testing to realize the estimation objective.

1.6 Overview of the Dissertation

In Chapter 2, the background on complex networked systems and social product development is presented. Existing studies for each application example are reviewed and the domain-specific research gaps are identified. Specifically, the literature review is divided into four parts: 1) network generation models using local behaviors; 2) models for improving the network performance; 3) crowdsourcing in product development; 4) game theory in engineering systems design.

From Chapter 3 to Chapter 6, the research activities presented in Table 1.6 are discussed in detail to answer the research questions. In Chapter 3, theoretical study on complex networked systems is presented. In this chapter, RQ1 and RQ2 are answered. In Chapter 4 and Chapter 5, empirical study in the context of AS-level Internet is presented. In Chapter 4, RQ3 is answered. In Chapter 5, RQ1 and RQ3 are answered by performing a comparative study in the context of AS-level Internet. In Chapter 6, RQ1, RQ2 and RQ3 in the context of design crowdsourcing are answered. Figure 1.5 displays the dissertation overview and roadmap.
Table 1.6. Framework of the research tasks in the dissertation.

<table>
<thead>
<tr>
<th>RQ Addressed</th>
<th>Research Tasks Fulfilled</th>
<th>Corresponding Chapters</th>
</tr>
</thead>
<tbody>
<tr>
<td>RQ1 and RQ2</td>
<td>Research tasks for theoretical study in the complex networked systems</td>
<td>Chapter 3</td>
</tr>
<tr>
<td>RQ1 and RQ3</td>
<td>Research tasks for empirical study in the Internet</td>
<td>Chapter 4 and 5</td>
</tr>
<tr>
<td>RQ1, RQ2 and RQ3</td>
<td>Research tasks for both theoretical study and experimental study in the design crowdsourcing</td>
<td>Chapter 6</td>
</tr>
</tbody>
</table>

Figure 1.5. Roadmap of the dissertation.
CHAPTER 2. BACKGROUND AND LITERATURE REVIEW

In this dissertation, two application examples are selected to apply the proposed framework and approaches for the purpose of validation. The first application is in complex networked systems and the second application is in social product development. Both examples are chosen to validate the approaches proposed for answering the three RQs, as shown in Table 1.3. The aim of this chapter is to introduce these two applications and identify research gaps in the existing literature. The structure of this chapter consists of two sections. Each section corresponds to one application. Table 2.1 shows the structure of literature review, research gaps and research tasks.

In Section 2.1, an introduction to complex networked system is first presented. To better understand such systems, it is crucial to model the systems’ structure and evolutionary dynamics. In Section 2.1.1, existing studies on complex network generation models are reviewed and summarized. In Section 2.1.2, the existing studies on models for improving network performance are discussed. The literature review in this section identifies the research gaps for addressing the RQ2. The research gaps are summarized in Section 2.1.3.

In Section 2.2, the introduction and literature review on social product development are presented. In particular, crowdsourcing - an important way of realizing social product development - is introduced, and the concept of crowdsourcing is presented. In Section 2.2.1, existing studies on crowdsourcing for product development are reviewed. In Section 2.2.2, a literature review of existing studies on game theory in engineering systems design is presented. This section is to identify the research gaps in game-theoretic models for modeling decision-making behaviors in design crowdsourcing. Finally, the research gaps are summarized in Section 2.2.3.
Table 2.1. Mapping of literature review, research gaps and research tasks.

<table>
<thead>
<tr>
<th>Literature review</th>
<th>Existing studies on complex network generation models (Section 2.1.1)</th>
<th>Existing studies on models for improving network performance (Section 2.1.2)</th>
<th>Existing studies on crowdsourcing in product development (Section 2.2.1)</th>
<th>Existing studies on game theory in engineering systems design Section (2.2.2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Research gaps</td>
<td>1) Theoretical foundations for nodes’ rational behavior are lacking;</td>
<td>1) The direct mapping between node-level preferences and the system performance is lacking;</td>
<td>1) Existing approaches do not account for nodes’ heterogeneity.</td>
<td>1) Purely analytical models do not capture the actual behaviors of humans;</td>
</tr>
<tr>
<td></td>
<td>2) Social and economic characteristics are not included in network formation.</td>
<td>2) Existing approaches do not account for nodes’ heterogeneity.</td>
<td></td>
<td>2) The data analysis lacks access to private information of the individuals.</td>
</tr>
<tr>
<td>Research tasks to address the gaps</td>
<td>1) Establishing decision-centric framework for complex networks; (Chapter 3)</td>
<td>1) Analyzing the effects of node-level preferences on system performance; (Chapter 3)</td>
<td>1) Establishing game-theoretic model of participants’ behavior by integrating contest theory; (Chapter 6)</td>
<td>1) Behavioral experimentation: designing and executing an economic decision games;</td>
</tr>
<tr>
<td></td>
<td>2) Modeling the nodes’ linking behavior by considering geographic, economic and social factors. (Chapter 4)</td>
<td>2) Estimating heterogeneous linking preferences in complex systems. (Chapter 4 and Chapter 5)</td>
<td>2) Perform game theoretical analysis and calculate system equilibrium. (Chapter 6)</td>
<td>2) Performing statistical analysis to investigate individuals’ decision-making behaviors. (Chapter 6)</td>
</tr>
</tbody>
</table>
2.1 Large-scale Complex Networked Systems

A specific class of bottom-up evolutionary complex systems is called complex networked systems. In such systems, the system components or individual entities can be represented by nodes, and the interaction between entities can be represented by links. Complex networked systems have played a significant role in both social and technological realms. The Internet, for example, has progressed from an emerging technology to one with a profound impact on commerce (e.g., distributing and shopping) and daily human life (e.g., communicating and entertaining) in just 20 years.

The complex networked systems can be studied effectively in terms of complex networks because networks represent two key characteristics: connectivity and interdependence. The study of networks has been at the core of systems science because the formation of network topology reveals underlying systems’ evolutionary dynamics. This dissertation takes a decision-centric perspective on complex networks. Specifically, the nodes are viewed as decision-makers, the link formation process is viewed as a decision-making process in which the nodes make individual decisions to link with target nodes.

In order to answer the research questions identified in Figure 1.3 within a decision-centric framework, a network generation model needs to be developed. Also, approaches for improving the network performance in a bottom-up manner are required. Therefore, a literature review on existing network generation models and the approaches for improving network performance is included. First, the research gaps are identified. In addition, theoretical foundations and modeling techniques that relate to the essentials of complex networks are reviewed. In this literature review, models for achieving desired network structure through direct manipulation of the network topology are not considered because these models are not suitable for bottom-up evolutionary complex networks as these models are only applicable within networks designed by external forces (e.g. the network designer).
2.1.1 Existing Studies on Modeling Network Generation with Node-Level Behavioral Models

Network generation models can be traced back to Erdos and Renyi [7] in the 1960s. The model, referred to as the ER model, assumes that the probability $p$ of a link between any pair of nodes is $p \in [0, 1]$. The presence of a link is independent of other links. An extended model of this ER model, called “exchangeable graph models” [8], introduces a weak form of dependence among the probability of sampling edges in the form of node-specific binary strings. Such models can be referred to as structure-based network models because they focus on generating the static structure of the networks instead of explicitly modeling how the network actually evolves. The exponential random graph model, also referred to as the $P^*$ model, is a widely used structure-based model [9–11]. In the $P^*$ model, it is assumed that the probability of realizing a specific network is given by $P(Y = y) = \frac{1}{k} \exp\{\sum_A[n_A g_A(y)]\}$ [10], where $P(Y = y)$ represents the probability that a network $y$ emerges, $k$ is a normalizing parameter that ensures the probability falls in a proper distribution, $A$ is the set of substructure configurations, $g_A(y)$ is the network statistic corresponding to the configurations $A$. Based on the observed network, the parameters $\eta_A$ are calculated using statistical estimation methods, such as pseudo-likelihood estimation [12] and maximum likelihood estimation with the Monte Carlo Markov Chain (MCMC) [13].

While static models focus on modeling the network structure at a single snapshot, dynamic network models explicitly model the evolutionary process. In dynamic network models, an initial network is chosen to represent the early stage of a real network. New nodes and links are gradually added (and removed) to simulate the evolution of the network [14]. The resulting network structure and properties are therefore the outcome of the evolutionary process. Such a process, i.e., the linking of new nodes, can be driven by different local properties of connecting nodes. The most widely used local property is a node’s degree, i.e. the number of neighbors a node connects with. A typical model representing this degree-based evolution is the Barabasi-Albert (BA)
model [15] which results in scale-free networks. Other examples of dynamic models are the Watts-Strogatz (WS) model [16] designed for producing local clusters triadic closure, and various variants of BA model for simulating unique structural characteristics of real networks, such as accelerating growth mechanism [17], rewiring of edges [18], copying mechanism [19], walking mechanism [20]. The models derived from the original BA model have been used to model various real-world networks including the World Wide Web, the Internet, and collaboration networks. For addressing the real-world networks with both scale-free and high cluster coefficient properties, models for generating scale-free networks with tunable cluster coefficient are proposed by Herrera [21], Holme [22], and Klemm and Eguluz [23, 24].

Dynamic network models have also been generated using the principles of Markov chains. Both continuous time and discrete time Markov chain models have been proposed [25–27]. These models are based on the assumption that networks evolve by modifying one edge at a time and the future state of the network is dependent on the current state only. The transition between states is dependent on node-level statistical parameters that can be estimated using longitudinal network data.

While the majority of the research on complex networks is focused on models for network formation where node behaviors are independent of each other, the theory of network formation games concerns an understanding of how networks arise from players’ (nodes’) strategic choices concerning link formation [28]. Network formation games provide a mathematical framework to model the node-level behaviors in order to minimize the cost they incur in building a network, but at the same time can benefit a maximal payoff from the system-level. Models, such as the local connection game [29], the global connection game [30], facility location game [31] and dynamic network formation game [32], provide insights on the basic principles of network formation. Network formation games have been utilized for modeling free trade networks, market sharing agreements, labor markets, and computer networks. A review of these methods is provided in [6] and [28].
Snijders [33] presents actor-oriented models to model the evolution of social networks as a result of independent actors making linking decisions to maximize their personal utilities. The models result in a probability of change in the network (addition/removal of links) based on the multinomial logit expression. The authors use various structural measures of the network to estimate the actors’ decision models. These measures include degree, number of reciprocated relations, sum of covariate, etc.

2.1.2 Existing Studies on Models for Improving Network Performance

While researchers have put forward various network generation models to capture real-world networks’ structures and properties, efforts have also developed models and approaches for achieving desired system-level performance, and analyzing how changes in the network topology affect the system-level performance. Beygelzimer et al. [34] propose 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. Schneider et al. [35] propose a simple modification scheme resulting in small changes to the network structure, but significantly increasing the robustness of diverse networks while having minimal impact on the functionality. Zhuo [36] proposes a strategy which removes a fraction of crashed hub nodes to improve network robustness against the coordinated attack.

While these models apply subjective modification on the network topology, other models adopt evolutionary ideas to guide the modification of edges and/or nodes. Bornholdt [37] proposes a model constrained solely by the requirement of robustness from an evolutionary standpoint. The network evolves a new single network from an old network by accepting rewiring mutation schemes. Some authors also consider multiple objectives. Shin and Namatame [23] present a model in which the network is optimized for two performance characteristics: low congestion and design cost. Network optimization is carried out using genetic algorithms. Shargel et al. [38]
propose a node-level mechanism for generating a so-called (1,0) network that features interconnectedness closer to that of a scale-free network, a robustness to attack closer to that of an exponential network, and a resistance to failure better than both of those networks.

2.1.3 Research Gaps

As discussed in Section 1.2, holistic frameworks for modeling networks must quantitatively capture the mappings across all the five levels shown in Figure 1.3. However, existing models and approaches have several limitations, as follows. These limitations pose a significant roadblock in the understanding of the network evolution and our ability to establish ways to direct the evolution of the networks for achieving the higher performance in a bottom-up manner:

1. Aforementioned structure-based and dynamic models, such as ER model, BA model and WS model, focus on matching the observed network structure and properties (Level 3 and 4) using hypothesized behavioral models (Level 2). Theoretical foundations for explaining why the individual node behaves the way they are modeled are generally lacking. Hence, the relationship between node-level preferences (Level 1) and node-level behaviors (Level 2) is absent.

2. Existing network generation models assume that the decision-making behaviors of all nodes are the same. However, in the case of complex systems such as the Internet, different types of nodes (e.g., customers and ISPs) have different objectives and linking behaviors. Similarly, existing approaches do not account for heterogeneity in the target nodes.

3. Even though the network formation game models provide an explanatory framework to describe the node-level decision-making behaviors, the main limitation of these models is that the effects of node-level preferences and behaviors on
network structure and performance (i.e., from Levels 1 and 2 to Levels 3, 4 and 5) cannot be quantitatively determined.

4. While the actor-oriented models [33] use a decision-making framework, they only model the linking behavior based on the network’s structural parameters (node degree, centrality, etc.). Other characteristics of the nodes (e.g., capacity and size in the case of airports) are not included in the network formation process.

5. Existing models and approaches which target the desired system-level performance (discussed in Section 2.1.2) are based on the direct modification of network topology (Level 3). The direct mapping between node-level preferences (Level 1) and the system performance (Level 5) is lacking.

Given the decision-centric framework, the approaches proposed in Chapters 3 and 4 address these research gaps. Specifically, the random utility discrete choice theory provides an explanatory framework for the node-level behavior by accounting for nodes’ decision-making preferences. The explanation of each node’s action is to maximize their own utilities. Accordingly, the research gaps 1 and 3 can be addressed. In the study, the mixed logit model is utilized. This model is generic and enables the heterogeneous preferences of individuals (see Chapter 3 for details). So, the assumption that all the nodes are the homogeneous (research gap 2) can be relaxed. This modeling characteristic makes the model itself more realistic to model real-world complex systems. With discrete choice models, a degree-based network generation model is established in Chapter 3. The effect of node-level preference on network performance is carried out; therefore, a direct mapping between node-level preferences and system-level performance is established. Thereby, research gap 5 is addressed. When applying the proposed approaches to the real-world cases, node-level preferences to non-network metrics, such as the social, technological factors are taken into consideration. For example, in the AS-level Internet application in Chapter 4, besides the network metrics, other factors including the traffic of ASes, the
geographical location, the price that an ISP offers and the role of ASes are considered. The addition of these non-network metrics overcome the limitation identified in research gap 4.

2.2 Social Product Development - An Emerging Paradigm for Engineering System Design and Innovation

Crowdsourcing is one emerging Web 2.0 based phenomenon. Traditionally, tasks in a company are assigned to, and performed by employees or designated agents or contractors. However, in the case of crowdsourcing, tasks are distributed to networked people [39]. The term *crowdsourcing* was first created by Howe in 2006 [40]. It relies on two simple but powerful concepts: 1) individuals online are no longer passive browsers, but active contributors; 2) virtually everyone has the potential to contribute valuable information [41].

A typical process of crowdsourcing is as follows: an individual/organization identifies the tasks and releases them online to a crowd of outsiders who are interested in performing these tasks on the requester’s behalf, for a stipulated fee or other incentive. A large number of individuals then offer to undertake the tasks individually or collectively. Upon completion, the participants involved submit their work to the crowdsourcing platform, and the requester assesses the quality of the work [42]. In this process, the crowdsourcing platforms, i.e., the websites, act as a third party for organizations to publish tasks and for participants to submit solutions. The websites distribute the award to the winners on behalf of the organization and may take a small portion of the award as a service fee. Therefore, crowdsourcing is evolving as a distributed problem-solving and business product realization model. A *system of crowdsourcing* is the integration of the platform, the requesters, the participants, the tasks, the incentives and the mechanism running on the platform. The system dynamics arise from interactions among these individual entities.
2.2.1 Existing Studies on Crowdsourcing in Product Development

So far, crowdsourcing has gained extensive attention and a lot of online platforms
have emerged, such as Amazon Mechanical Turk [43], Topcoder [44], Yahoo! An-
swers [45], taskcn [46], eLance [47], 99designs [48], etc. For example, taskcn.com, one
of the largest crowdsourcing websites in China, has 3,405,068 people registered and
53,832 tasks requested. An amount of 35,336,890 (Chinese Yuan, about 5.76 million
in US dollars) in monetary awards has been distributed to 263,725 winners (data
collected in September, 2013). Given its fast emergence, crowdsourcing studies are
still developing. Zhao and Zhu [42] perform a comprehensive survey of existing aca-
demic activities on crowdsourcing. The survey indicates that the theories applied to
study the crowdsourcing, such as auction theory [49], motivation crowding theory [50],
cognitive evaluation theory [50], game theory [51], strategic management theory [52]
etc., are more focused on providing explanations for existing crowdsourcing phenom-
ena rather than analysis of participants’ behaviors to provide design insights and
modeling behaviors to predict the evolutionary dynamics of crowdsourcing processes.
Additionally, the decision-making nature of the participants, which results in system
dynamics, is not addressed.

Existing studies focus on investigating the participants’ behaviors in crowdsourc-
ing process for three purposes: exploration, explanation and prediction. For studies
focused on exploration, the main task is to explore participants’ behaviors in the
crowdsourcing process. By analyzing the data collected from the biggest crowdsourc-
ing website in China - Taskcn.com, Yang et al. [39] quantitatively showed that partic-
ipants tend to compete for the tasks with fewer opponents and with higher expected
rewards. In addition, most participants become inactive after a few submissions,
while a certain number of participants continue to attempt the tasks. Archak [53]
presented an empirical analysis on the participants’ behaviors in crowdsourcing com-
petitions using a dataset from the world’s largest competitive software development
community — TopCoder.com. The author observes that in order to soften the com-
petition, the contestants move first to the registration phase of the contest, signing up early for particular tasks they are interested in.

For the studies focusing on providing explanations, the main task is to investigate and identify factors that affect observed participants’ behaviors. These studies analyze the incentives implemented in existing crowdsourcing platform, the attributes of the tasks posted, and/or the participants’ intrinsic motivations. Intuitively, financial incentive is one of the most important factors for achieving effective crowdsourcing. Harris [54] found that financial incentives not only attract more participants, but also improve quality if the task is designed appropriately. Besides financial incentives, Archak [53] investigates reputation incentives for TopCoder.com. In this case, highly rated contestants face tougher competition in the contest. Zheng et al. [55] developed a model to explain the effect of task attributes on participants’ motivation based on the theory of extrinsic and intrinsic motivation as well as the theory of job design.

For the studies that focus on prediction, the main task is to develop models that mimic participants’ behaviors under a specific crowdsourcing mechanism in order to predict the future status of the crowdsourcing process. DiPalantino and Vojnovic [49] model crowdsourcing contests as all-pay auctions to demonstrate the relationship between incentives and participation in such a system. The model is validated by comparing empirical data from the Taskcn.com. Later, the all-pay auction mechanism is further validated by Liu et al [56] with a field experiment on Taskcn.com. Stewart et al. [57] propose a SCOUT (Super Contributor, Contributor, and Outlier) model for describing user participation inside the enterprise and show that it is possible to achieve a more equitable distribution of participants’ activities. Horton [51] et al. present a supply model to predict the smallest wage a worker is willing to accept for a task. The authors also show how the model is validated and applied to make predictions with two experiments. To the best of our knowledge, these three models are the only models developed for predicting participants behaviors in crowdsourcing process.
This literature review demonstrates that existing studies focus more on analysis and explanation, and less on modeling and predicting the evolution of crowdsourcing process. Existing studies such as [56] take steps to model crowdsourcing, but only evaluate models based on some statistical measures, e.g., the likelihood ratio and $\rho^2$; in other words, no systematical approaches are proposed to validate the model and evaluate the performance of the model.

2.2.2 Existing Studies on Game Theory in Engineering System Design

Game theory is becoming an increasingly important tool for engineering design research. One of the first models of the design process using game-theory was by Vincent [58]. In this model, a multi-objective design problem is assigned to several decentralized designers where each designer is responsible for a subset of objectives. The decentralized design problem is formulated as a non-cooperative game, with the Nash equilibrium as the solution concept, in which no player has incentive to unilaterally change their strategy. The model was extended by Lewis and Mistree [59] for modeling different types of interactions between decision makers—sequential, collaborative, and isolated. The resulting game-based design framework was later used for developing ranged sets of specifications using design capability indices [60].

Building on the non-cooperative game model of decentralized design, studies analyze the convergence characteristics of the design process and the quality of equilibrium solutions. Chanron and Lewis [61] assume that decision makers follow an iterative process of communicating best responses to the decisions made by other decision makers, and analyze the convergence of solutions to the Nash equilibria. Devendorf and Lewis [62] use the linear systems theory to analyze the dependence of system stability on solution process architecture.

Game theoretic models have inspired the development of efficient protocols for information exchange in design. As it is well known that the quality of Nash equilibria may be inferior to Pareto optimal solutions, various researchers have proposed
modifications to the standard information exchange (or best replies) and developed protocols for bringing the resulting design closer to Pareto optimal solutions. For example, Ciucci et al. [63] develop strategies for passing additional information to facilitate convergence to Pareto-optimal designs. Fernandez et al. [64] present a coordination mechanism for establishing shared design spaces and exploring regions of acceptable performance. Rao and Freiheit [65] model a multiobjective design problem as a game, and develop a modified game theoretic method to obtain Pareto optimal solutions. Takai [66] models a scenario where a design project has both team and individual components, and analyzes the conditions that lead to cooperation between the designers, thereby leading to good system-level outcomes.

Recently, game theoretic models have been used at the core of design for market systems (DFMS), where the focus is on the integration of demand models, market competition, and engineering performance of products in a comprehensive design optimization framework [67]. The models support the integration of decisions made by manufacturing firms competing in a profit maximizing game, manufacturer-retailer interactions, and consumer preferences in a unified framework. Design researchers are using game theoretic models for analyzing large-scale complex systems such as distributed energy generation systems [68,69].

2.2.3 Research Gaps

Human decision-making deviates from rational behavior due to cognitive limitations such as the inability to process large amounts of information and various constraints such as insufficient time to make decisions and lack of complete information regarding alternatives. Simon [70] describes this concept as bounded rationality. In addition to the assumption of rationality, analytical models of games are based on further assumptions about players' knowledge of the game and iterated reasoning. Human decisions in interactive decision-making scenarios deviate from these assumptions. For example, humans have a limited capability for iterative reasoning, which
may be assumed in arriving at Nash equilibria. Players' feelings about the fairness of payoff distribution affect their actions. Similarly, learning may play a significant role in their actions.

The need for considering these human attributes during the design process, particularly in design decision making, has been emphasized within the design research community [71]. Although there have been some efforts to consider the effects of bounded rationality on design decision making [72, 73], synergistic analysis of the analytical game theoretic models along with the deviations from the rationality assumptions is rare in design research.

The validity of game theoretic models in real decision making scenarios can be evaluated in two ways – empirical analysis of secondary data and via a controlled experiment. A secondary dataset is one that has been collected for purposes other than analysis (e.g., for operating a business). An example of the use of secondary data for crowdsourcing in order to analyze tournaments is presented by Boudreau et al. [74], who analyze secondary data from such tournaments on TopCoder. The primary limitation of secondary data analysis is that it lacks access to private agent information and, it may be difficult to explain how behaviors (e.g., an agent pretending to be another type) lead to observed outcomes.

In summary, the research gaps in the study of crowdsourcing and game theory in engineering design analysis are listed below.

1. Existing studies in crowdsourcing focus more on analysis and explanation, and less on modeling crowdsourcing process. In addition, validation methods have been primarily statistical in nature.

2. Humans do not always behave in a manner consistent with the rationality assumption of pure analytical game models because they are not entirely driven by extrinsic rewards and have cognitive limitations.

3. Empirical analysis of secondary data to validate game theoretic models is limited in its assessment of private information of individual agents.
The approaches presented in Chapter 6 are developed to address these research gaps. Specifically, game theory is employed to model and estimate the designer’s decision-making behavior with non-cooperative game models. In order to address the limitation of purely using game theoretical models, controlled laboratory experiments are adopted. Specifically, an economic decision game using behavioral game theory is designed and executed to investigate contestants’ decision-making in an actual crowdsourcing situation.

Controlled laboratory experiments can overcome many of the limitations of secondary data analysis. In the experimental economics methodology, experiments are conducted with human subjects under controlled settings [75], and subjects are rewarded based on their actions. Emphasis is placed on experimental designs to ensure that incentives in the real world are accurately represented in the experiment, and that participants maximize their payments. The emphasis on incentives distinguishes it from experimental psychology, where the focus is on contextual aspects [76] rather than incentives.
CHAPTER 3. A DECISION-CENTRIC APPROACH FOR MODELING AND ANALYZING COMPLEX NETWORKS

3.1 Chapter Overview

In this chapter, the objective is to answer RQ1 and RQ2 in the context of complex networked systems through the Approaches 1.1 and 2.1. The two RQs are:

- RQ1: how can individual decision-making preferences and behavior be modeled?
- RQ2: how can the effects of local decision-making preferences and behaviors be understood?

The framework and approaches established in this chapter address the main research gap (see Section 1.3) in existing studies of complex networked systems: there is a lack of theoretical foundations for explaining individual entities’ rational behavior. The established framework is based on the random-utility discrete choice theory and complex network theory. With the proposed framework, a degree-based linking model is first established. In the model, nodes’ linking behavior is characterized in terms of linking probability by structuring nodes’ decision-making preferences to degree, which is quantified by model parameter $\beta$. Second, the network growth mechanism is adopted from a well known network generation model – the Barabasi Albert (BA) model. Therefore, a degree-based decision-centric (DBDC) network generation model is established. The DBDC model is a specific example of the proposed decision-centric modeling framework to answer RQ1.

To answer RQ2, theoretical analysis of the DBDC model is performed. Specifically, physical theories, such as continuum theory and percolation theory, are used to analyze how nodes’ decision-making preferences impact the network structure and network robustness. Results obtained from the theoretical analysis are then verified...
by running an agent-based simulation to generate synthetic networks. The degree distribution analysis and fragmentation analysis are performed on the synthetic networks to investigate the effects of node-level decision-making preferences. The theoretical results can be verified if the same results are obtained from the analysis of synthetic networks obtained from agent-based simulation. Finally, the proposed DBDC model is evaluated by comparison with an existing network generation model: the generalized preferential attachment model. In summary, Figure 3.1 shows how the proposed framework and model answer both RQ1 and RQ2 in the application of complex networked systems.

Figure 3.1. Overview of Chapter 3: research questions, approaches and tasks.

The outline of this chapter is as follows. In Section 3.2, the random utility discrete choice theory is introduced. Given the discrete choice models, a decision-centric modeling framework is proposed in Section 3.3. In this section, the DBDC model is established as a specific implementation of the proposed framework. In order to execute the analysis of the node-level decision-making preferences, theoretical analysis and simulations are performed in Section 3.4. In Section 3.5, a comparative study is
performed to evaluate the proposed DBDC model. Insights are gained by comparing the DBDC model with the generalized preferential attachment model. Finally, closing comments are provided in the Section 3.6.

3.2 Theoretical Foundations

As discussed in Chapter 1, in complex networked systems, local decisions influence the overall system structure and performance. By assuming that a node maximizes utility, existing knowledge in discrete choice random utility theory from economics can be utilized to model preferences and resulting decisions. The reasons for choosing random utility discrete choice theory are that:

1. It quantitatively constructs the agents’ decision model with statistical techniques.
2. It enables the estimation of agent’s unobserved payoff in terms of utility function based on historical information about their choice.
3. It accounts for heterogeneity of decision makers' preferences,
4. It accounts for uncertainty in unobserved attributes, and
5. It captures the preferences of individual node rather than the group behavior.

Discrete Choice Analysis (DCA) [77] uses statistical techniques to relate a decision maker’s choices to the attributes of the available alternatives [78]. DCA has been widely used to model and forecast product demand by capturing individual customer’s choice behaviors, especially for demand estimation [79]. Earlier applications of DCA are in the field of transportation engineering, but extended to the field of product design to facilitate design with consumer preference and uncertainty [80]. DCA is conducted based on the Random Utility Theory (RUT) [81] which assumes that individuals seek to maximize their personal utility, $U$, when making a choice
from a choice set. Discrete Choice Model (DCM) can be derived under this utility-maximizing assumption. The derivation is briefly introduced as follows.

From the decision maker’s perspective – a decision maker, labeled \( n \), makes a choice among \( N \) alternatives. The utility that decision maker \( n \) obtains by choosing alternative \( i \) is \( U_{ni} \), \( i = 1, \ldots, N \). This utility is known to the decision maker but not to the researcher. With RUT, the behavior model is: choose alternative \( i \) rather than \( j \) if and only if \( U_{ni} > U_{nj}, \forall i \neq j \).

From the researchers’ perspective – the researchers do not know the decision maker’s utility completely, but is only able to observe the choices made by the decision maker. The researchers also observes some explanatory variables. Therefore, the utility function is random from the researchers’ perspective and is represented as a sum of two components [78]:

\[
U_{ni} = V_{ni} + \epsilon_{ni},
\]  
(3.1)

1. **Representative utility**, \( V_{ni} = V(x_{ni}, s_n | \beta_n) \) which is a function of explanatory variables that are either *alternative-specific* attributes, labeled as vector \( x_{ni} \), or *decision-maker-specific* attributes, labeled as vector \( s_n \), given parameter vector \( \beta_n = (\beta_{n1}, \ldots, \beta_{nk})' \), where, \( k \) is the total number of variables identified by the researchers. This parameter vector describes each decision-maker’s preferences to the explanatory variables. Representative utility is deterministic from the researcher’s point of view.

2. **Unobserved utility**, \( \epsilon_{ni} \) which can be represented as a random variable from the researcher’s point of view to quantify their unknown knowledge of the utility function.

The characteristics of \( \epsilon_{ni} \), such as its distribution, depend on the researchers’ specification, and are not defined by the choice situation per se. The joint density
function of the random vector \( \epsilon_n = (\epsilon_{n1}, \ldots, \epsilon_{ni})' \) is denoted as \( f(\epsilon_n) \). Following the RUT, the probability that decision maker \( n \) choose alternative \( i \) is

\[
P_{ni} = P(U_{ni} > U_{nj}), \forall j \neq i
\]

\[
= P(V_{ni} + \epsilon_{ni} > V_{nj} + \epsilon_{nj}), \forall j \neq i
\]

\[
= P(\epsilon_{nj} - \epsilon_{ni} < V_{ni} - V_{nj}), \forall j \neq i
\]

(3.2)

The result in Equation (3.2) indicates a cumulative distribution. Namely, the probability that each random term \( (\epsilon_{nj} - \epsilon_{ni}) \) is below the observed quantity \( (V_{ni} - V_{nj}) \).

The above equation turns to

\[
P_{ni} = P(\epsilon_{nj} - \epsilon_{ni} < V_{ni} - V_{nj}), \forall j \neq i
\]

\[
= \int I(\epsilon_{nj} - \epsilon_{ni} < V_{ni} - V_{nj}) f(\epsilon_n) d\epsilon_n, \forall j \neq i
\]

(3.3)

Where \( I(\cdot) \) is the indicator function,

\[
I(\cdot) = \begin{cases} 
1, & \text{if } \epsilon_{nj} - \epsilon_{ni} < V_{ni} - V_{nj} \\
0, & \text{otherwise}
\end{cases}
\]

(3.4)

With different assumptions about the distribution of the unobserved part of utility \( \epsilon_{ni} \), different types of DCMs are obtained. In this study, a multinomial logit model [78] is adopted. The logit model is derived under the assumption that \( \epsilon_{ni} \) is identical independent distributed (iid) type I extreme value, which is also called Gumbel distribution [77]. The cumulative distribution is \( F(\epsilon_{ni}) = e^{-e^{-\epsilon_{ni}}} \).

If the preferences to the explanatory variables are homogeneous for all the decision-makers, then \( \beta_n \) can be denoted as \( \beta \). With Equation (3.3), the logit form of the choice probability is obtained as follows (refer to [78] for details),

\[
P_{ni} = \frac{e^{V_{ni}}}{\sum_{j=1}^{N} e^{V_{nj}}} = \frac{e^{V(x_{ni}, s_n|\beta)}}{\sum_{j=1}^{N} e^{V(x_{nj}, s_n|\beta)}}
\]

(3.5)

If the preferences to the variables, i.e. \( x_{ni} \) and \( s_n \), are heterogeneous, which means the decision-makers have preference variation following a density function of \( f(\beta) \), the resulting choice probability takes the mixed-logit form as follows. The mixed logit
probability is a weighted average of the logit formula in Equation (3.5) evaluated at
different values of $\beta$, with the weights given by the density $f(\beta)$.

$$
    P_{ni} = \int \frac{e^{V(x_{ni}, s_n | \beta)}}{\sum_{j=1}^{N} e^{V(x_{nj}, s_n | \beta)}} f(\beta) \ d\beta
$$

(3.6)

The advantage of the models in Equations (3.5) and (3.6) is that the corresponding
log-likelihood function with the choice probability is globally concave so that param-
eters in this model can be estimated based on first order derivative method, such as
maximum likelihood estimation. If choice probability follows Equation (3.5), the vec-
tor $\beta$ is estimated. If heterogeneous preferences exist, the choice probability follows
Equation (3.6) and a set of parameters that determine the density function $f(\beta)$ is
estimated. For instance, if $f(\beta)$ is a normal distribution, the mean and variance are
estimated.

3.3 The Decision-Centric Modeling Framework

Using the discrete choice random utility theory, a decision-centric modeling frame-
work is proposed. The modeling framework is to answer RQ1. The framework consists
of three parts for the evolution of network at each time step. To facilitate the model
establishment process, a flow chart is shown in Figure 3.2 as below to help researchers
come up with their own model.

1. Evolution Mechanism

   (a) New node addition and edge addition between new node and existing node.

      At each time step, there will be $l$ new nodes entering into the network with
      $m$ edges associated with each new node. This means, for each newly added
      node, it will choose $m$ target nodes to link with. So, the decision-making
      nodes are newly added nodes. The target nodes whom the newly added
      nodes will link with will be determined with choice probability defined in
      part 3(a).
Figure 3.2. Modeling procedure with decision-centric framework.
(b) Edge addition and deletion between existing nodes.
At each time step, \( p \) new edges will be established between \( p \) pairs of existing nodes, and \( q \) existing edges will be deleted between \( q \) pairs of existing nodes. The nodes who are going to establish new edges and delete existing edges will be determined with the choice probability defined the part 2(a) and 2(b). The target nodes whom the decision-making nodes choose to establish new edges and the target nodes whom the decision-making nodes choose to delete existing edges will be determined with the choice probability defined in part 3(b) and 3(c).

(c) Nodes deletion and the associated edges’ deletion.
At each time step, \( s \) nodes will be deleted from the network, and their associated edges will be deleted as well. The nodes who will be deleted is determined with the choice probability defined in part 2(c).

2. Decision-making nodes determination - Choice probability construction to identify the nodes who are going to make a decision (The choice set is all the nodes in the current network, the size of choice set is equal to the network size, i.e. \( N \)).

(a) For part 1(b), construct the representative utility function \( V_1 \) by identifying the explanatory variables (here, the explanatory variables is only alternative-specific because the lack of information on how the decision-making nodes are selected) that may affect the utility. The probability that a node who will make decision on establishing new edge between existing node is selected follows Equation (3.5) (without knowing the decision-maker’s information, thus no preference variation).

(b) Similarly for part 1(b), construct the representative utility function \( V_2 \). The probability that a node who will make decision on deleting existing edges between existing pairs of nodes is selected follows Equation (3.5).
(c) For part 1(c), construct the representative utility function $V_3$. The probability that a node who will be deleted is selected follows Equation (3.5).

3. Target nodes determination - Choice probability construction for decision-making nodes to identify which nodes they are going to connect or disconnect (The choice set is all the nodes in the current network except the node who is going to make decision, the size of choice set is equal to the network size minus one, i.e. $N - 1$).

(a) For part 1(a), specify the explanatory variables (either alternative-specific or decision-maker-specific) that may affect the decision-making nodes’ utility, and construct the representative utility function $V_4$. The probability that a target node that the newly added nodes will select to link with follows Equation (3.5) if the decision-making nodes’ preferences are homogeneous to the specified variables, but follows Equation (3.6) if the researcher assumes the decision-making nodes’ preferences are heterogeneous.

(b) For part 1(b), similarly, construct representative utility function $V_5$ by identifying the explanatory variables. The probability that a target existing node that a existing node will select to link with follows the Equation (3.5) if the decision-making nodes’ preferences are homogeneous, but follows Equation (3.6) if heterogeneous.

(c) Similarly for part 1(b), construct representative utility function $V_6$. The probability that a target existing node that another existing node will select to disconnect follows the Equation (3.5) if the decision-making nodes’ preferences are homogeneous, but follows Equation (3.6) if heterogeneous.

The proposed modeling framework is quite flexible and general to establish models for complex networks in different contexts. It is not necessary that all the three evolution mechanisms be included in one model. The researchers can choose any combinations of the three mechanisms depending on difference cases. For each scheme,
the researchers just need to construct the corresponding choice probability function as defined in part 2 and part 3 according to the mechanism that is adopted. However, in most cases, the network size is increasing over time. So, the mechanism in part 1(a) is always adopted. If the mechanism part 1(c) is also included, to ensure the network size is increasing, it is required that \( l > s \).

In the following sections, the work presented is to answer the first research question, RQ2. With the proposed decision-centric modeling framework, a degree-based decision-centric (DBDC) network generation model is developed. How the node’s preferences to degree influence the network structure, properties and performance is analyzed. Finally, the proposed DBDC model is compared with the classical degree-based model — generalized preferential attachment (GPA) model. The results obtained from the comparative study provide unique perspective for explaining the evolution of complex networks and insights for optimal design of network with high robustness and resilience.

### 3.4 The Effect of Local Decision-making Preferences and Behaviors

#### 3.4.1 Decision-Centric Modeling of Network Evolution

Within the RUT discussed in Section 3.2, a node’s preferences are modeled as utility functions as defined by the researchers. A node \( i \) is a decision-maker and its decision on which target node \( j \) should be connected (edge addition) or disconnected (edge deletion) is motivated by maximizing the utility function \( U_i \) according to the network topology. With the DCM, given the network context, the explanatory variables are further specified as *alternative-specific network-metric attributes*, \( \hat{x}_{ni} \), *alternative-specific non-network-metric attributes*, \( \tilde{x}_{ni} \), *decision-maker-specific network-metric attributes* \( \hat{s}_n \) and *decision-maker-specific non-network-metric attributes* \( \tilde{s}_n \). For example, the network-metric attributes include degree, cluster coefficient, centrality, etc. The non-network-centric attributes are the social, economical, ecological and/or political factors, such as the income, price, gender, etc, depending on what the node
stand for in different applications. Hence, the resulting choice probability in Equation (3.5) and Equation (3.6) become

\[ P_{ni} = \frac{e^{V(\hat{x}_{ni}, \hat{x}_{n}, \hat{s}_{n})}}{\sum_{j=1}^{N} e^{V(\hat{x}_{nj}, \hat{x}_{n}, \hat{s}_{n})}} \quad (3.7) \]

\[ P_{ni} = \int \frac{e^{V(\hat{x}_{ni}, \hat{x}_{n}, \hat{s}_{n})}}{\sum_{j=1}^{N} e^{V(\hat{x}_{nj}, \hat{x}_{n}, \hat{s}_{n})}} f(\beta) \, d\beta \quad (3.8) \]

For illustrative purpose, when establishing the network generation model, only evolution mechanism part 1(a) is considered. This means we will not consider the evolutionary dynamics resulted from edges addition and deletion between existing nodes and the nodes’ deletion and associated edges deletion. The model is shown as follows.

1. **Growth mechanism**: at each time step, there is 1 new node entering into the network with \( m \) edges associated with it.

2. **Utility function formulation**: only the alternative-specific network-metric attribute - node’s degree \( d \) as the explanatory variable, i.e. \( \hat{x}_{ni} = d_i \), is considered. The utility function is a simple linear function of degree in Equation (3.9) (since one node’s degree is fixed from other nodes’ point of view, the subscript \( n \) is omitted).

\[ V_i = \beta_1 d_i + \beta_0 \quad (3.9) \]

3. **Choice probability**. It is assumed that the decision-making nodes have homogeneous preferences to the degree. The resulting choice probability that the node \( i \) is selected hence follows the form in Equation (3.7) by substituting the defined utility function,

\[ P_i = \frac{e^{\beta_1 d_i + \beta_0}}{\sum_{j=1}^{N} e^{\beta_1 d_j + \beta_0}} = \frac{e^{\beta_1 d_i}}{\sum_{j=1}^{N} e^{\beta_1 d_j}} \quad (3.10) \]
Other assumptions:

- The network is undirected, and only one edge exists between a pair of nodes.
- The network is initialized with an ER random model with 5 nodes linking with a probability of 0.5. \(^1\)
- The parameter \(\beta_1\) is time independent. This means the decision-maker’s preferences is unchanged with time.

The model developed is a simple degree-based decision-centric (DBDC) model because the only variable considered for the utility function is node’s degree. The reasons for developing the degree-based model are two-fold: 1) existing literature [82] has validated that the degree-based mechanism has better performance in modeling real-world complex networked systems than the structure-based models. 2) to keep consistency in comparison. A comparative study is performed in Section 3.5 by using a degree-based generalized preferential attachment (GPA) model. In this simple DBDC model, the parameter \(\beta_1\) captures the nodes’ preferences to the degree, and the resulting choice probability is not subjectively assigned but derived based on RUT.

Therefore, the RUT provides a good explanatory framework for establishing the relationship between node-level preferences, \(\beta_1\), and node-level behaviors, \(P_i\). This facilitates setting up the direct relationship between the node-level preferences and the system-level performance. Thus, the model proposed in this section completes the research Task 1.1 in RQ1, as shown in Table 1.3. In the following, the theoretical analysis and experimental verification for the proposed model is performed to analyze the effect of node-level preferences on network structure and performance.

### 3.4.2 Theoretical Analysis of the Decision-Centric Model

The work presented in this section corresponds to Task 1.2 in RQ1. Specifically, theoretical analysis is performed on the model proposed. The aim is to obtain the

\(^1\)With this assumption, it implicitly requires \(m \leq 5\)
theoretical solution to the question: how the node-level preferences, modeled as the parameter $\beta_1$, impact on the network structure in terms of degree distribution, and on the system performance in terms of the network robustness against random failure and targeted attack (see Section 3.4.2.2).

### 3.4.2.1 Numerical Solution of Degree Distribution

The network structure is manifested by degree distribution. The degree distribution $P(d)$ of a network is then defined to be the fraction of nodes in the network with degree $d$. The basis for the theoretical analysis of degree distribution is the continuum theory-based approach presented by Albert et al. [83]. According to the continuum theory, the degree of a given node $i$, $d_i$, is time dependent. This degree increases every time a new edge is established. Assuming $d_i$ is a continuous real variable, the rate at which $d_i$ changes is expected to be proportional to $P_i$ in Equation (3.10) with the growth mechanism in the DBDC model. Consequently, $d_i$ satisfies the following dynamical equation:

$$\frac{\partial d_i}{\partial t} = m P_i = m \frac{e^{\beta_1 d_i}}{\sum_{j=1}^{N} e^{\beta_1 d_j}}$$  \hspace{1cm} (3.11)

This differential equation is difficult to solve analytically because of the exponential item. But numerical solution can be obtained. An algorithm targeted at solving this problem numerically is proposed. At each time stage, a new node enters the network. At time $t = t_i$, it is a system of $i$ ODEs, where $i = 1 \ldots N$. The initial condition for $ith$ ODE at time $t_i$ is $d_i = m$. The initial conditions for the remaining $(i - 1)$ ODEs at time $t_i$ are determined by solving the ODEs at $t_{i-1}$. The iterative process is illustrated in the following pseudocode.

Using the proposed algorithm, $d_i(t)$ is obtained for each node $i$. At a certain time $t_i$, each node’s degree is determined and the resulting degree distribution of the network is evaluated by discretizing the continuous results. This approach is used to evaluate the degree distributions for different $\beta_1$ values. Figure 3.3 shows how $\beta_1$
influences the network degree distribution (complementary cumulative distribution (CCD)) for \( m = 3 \) and \( N = 2000 \). Based on the results, it is observed that the degree distributions are qualitatively different for ranges: (a) \( \beta_1 < -1 \), (b) \(-1 \leq \beta_1 \leq 0.04 \), and (c) \( \beta_1 > 0.04 \). Hence, degree distributions corresponding to these three ranges are plotted separately. The cutoff values for these ranges have been chosen through observation.

**Algorithm 1:** Pseudocode of numerically solving Equation (3.11).

<table>
<thead>
<tr>
<th>Result:</th>
<th>Each node’s degree ( d[i] ) as a function of ( t ), i.e. ( d_i(t) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initialization:</td>
<td>set the initial condition for node 1, ( d_{\text{init}}[1] = m );</td>
</tr>
<tr>
<td>for ( i = 1 ) to ( N ) do</td>
<td></td>
</tr>
<tr>
<td></td>
<td>construct ( i ) ODEs;</td>
</tr>
<tr>
<td></td>
<td>set ( Timespan = 1 ) for solving the ODE;</td>
</tr>
<tr>
<td></td>
<td>solve the system of ( i ) ODEs and get solution for each node ( i )'s degree ( d[i] );</td>
</tr>
<tr>
<td></td>
<td>for ( j = 1 ) to ( N - 1 ) do</td>
</tr>
<tr>
<td></td>
<td>Update the value of ( d_{\text{init}}[j] ) based on the solution at time ( i );</td>
</tr>
<tr>
<td></td>
<td>end</td>
</tr>
<tr>
<td></td>
<td>set the initial condition for the new coming node, ( d_{\text{init}}[i + 1] = m );</td>
</tr>
<tr>
<td>end</td>
<td></td>
</tr>
</tbody>
</table>

The analysis results on degree distribution reveal that as \( \beta_1 \) increases from negative to positive, the network structure transitions from “chain” \( \rightarrow \) small-world (truncated Gaussian) \( \rightarrow \) random (exponential) \( \rightarrow \) scale-free (power-law) \( \rightarrow \) giant-hub (for detailed discussion, please refer to [84]). To verify the theoretical solutions, synthetic networks at different values of \( \beta_1 \) are first generated in Section 3.4.3.1, and then the degree distributions are extracted to determine whether similar conclusions are obtained. In addition, the network properties of the synthetic networks are analyzed for different values of \( \beta_1 \) to see if the pattern observed in degree distribution analysis can be reflected as well.
3.4.2.2 Percolation Analysis of Network Robustness Against Random Failure and Targeted Attack

Recall the structure of endogenously evolving networks shown in Figure 1.3, the results obtained from the previous section established the connection between node-level preferences (Level 1) and network structure (Level 3). The theory and results presented in this section is for establishing direct connections between node-level preferences (Level 1) and system-level performance (Level 5).

The system-level performance considered in this study is the network’s robustness. Robustness means the persistence of a system’s characteristic behavior under perturbations [85]. In network science, the perturbations are mainly from the failure
of nodes. There are different ways that cause nodes failure. In this study, the network robustness against two processes are studied: random failure of nodes and targeted attack. In random failure, nodes are randomly selected and removed from the network. On the other hand, targeted attack involves elimination of nodes with certain properties. In this study, targeted attacks on nodes with the highest degree are considered. Such process can be used to represent a computer hacker trying to bring down the routers with the highest connectivity. In both scenarios, after a certain fraction of nodes (say $f_c$) is removed from the network (after $d$ steps of attacks), the network becomes fragmented. The robustness of the network is proportional to the fraction of nodes that need to be removed before fragmentation occurs. To quantify the robustness in terms of $f_c$, it is important to choose a measure that indicates when the network is fragmented. There are many such measures [85–88]. In this study, the size of largest connected component (LCC), i.e., the giant cluster, is adopted as the index. The rationale is that the presence of a giant cluster is an indicator of a network that is at least partly performing its intended function, while the size of the LCC tells us how much of the network is working.

The theoretical analysis on network robustness is based on percolation theory [89], which is briefly discussed as follows. In random failure scenario, let $\phi$ be the
probability that a node is present in the network, the nodes tend to be connected together in a giant cluster when $\phi$ is large, and they tend to be fragmented apart when $\phi$ is small. The formation or dissolution of a giant cluster in a network is called the percolation transition. The $\phi$ at which the percolation transition occurs is called the percolation threshold $\phi_c$. Percolation threshold is actually equal to one minus the critical fraction point $f_c$, i.e. $\phi_c = (1 - f_c)$. The lower the percolation threshold, the higher the network robustness is. Let $u$ be the average probability that a node that is not connected to the giant cluster via a particular neighbor, and consider a network with degree distribution $p(d)$. The percolation analysis indicates that the total fraction of nodes in the giant cluster, $S$, in random failure process is determined by:

$$S = \phi [1 - g_0(u)]$$

(3.12)

$$u = 1 - \phi + \phi g_1(u)$$

(3.13)

in which,

$$g_0(x) = \sum_{d=0}^{\infty} p(d)x^d$$

(3.14)

$$g_1(x) = \sum_{d=0}^{\infty} q(d)x^d$$

(3.15)

$g_0(x)$ is the generating function for the degree distribution [90]. $g_1(x)$ is the generating function for the excess degree distribution, in which $q(d) = \frac{(d+1)p(d+1)}{\langle d \rangle}$, and $\langle d \rangle$ is the network’s average degree. Given the degree distribution $p(d)$ and the value of $\phi$, Equation (3.13) can be numerically solved to get the solution, say $u^*$. Substituting $u^*$ into Equation (3.12), the complete solution is obtained for $S$. The percolation transition process, $S(\phi)$, is obtained.

In the targeted attack scenario, the nodes are preferentially removed according to their degree. Let $\phi_d$ be the probability that a node with degree $d$ is present in the network. Suppose at each step, the cut-off degree for nodes that are removed is $d_0$. Consequently, $\phi_d = 1$ for all nodes with degree $d < d_0$, and $\phi_d = 0$ for all nodes with
degree \( d \geq d_0 \). The total fraction of nodes in the giant cluster, \( S \), in random failure process can be numerically determined by:

\[
S = f_0(1) - f_0(u) \tag{3.16}
\]
\[
u = 1 - f_1(1) - f_1(u) \tag{3.17}
\]

where,

\[
f_0(x) = \sum_{d=0}^{d_0} p(d) \phi_d x^d \tag{3.18}
\]
\[
f_1(x) = \sum_{d=0}^{d_0} q(d) \phi_{d+1} x^d \tag{3.19}
\]

Given the cut-off degree \( d_0 \), \( f_0(x) = \sum_{d=0}^{d_0} p(d)x^d \), and \( f_1(x) = \sum_{d=0}^{d_0} q(d)x^d \). Equation (3.17) can be numerically solved, say the solution is \( u^* \). Substituting \( u^* \) into Equation (3.16) gives the complete solution for \( S \).

Since a closed form of the degree distribution for the network generated with the DBDC model is unavailable, our strategy is to investigate the robustness of representative networks with critical \( \beta_1 \) values. A prediction on how the network robustness changes can be obtained accordingly. Later, this prediction is verified with simulation results, and a complete picture of how the node-level preferences \( \beta_1 \) affect the network robustness is obtained in Section 3.4.3.2. The networks used for percolation analysis are listed in Table 3.1

<table>
<thead>
<tr>
<th>( \beta_1 )</th>
<th>Degree distribution</th>
<th>Network structure</th>
<th>( f_c ) random failure</th>
<th>( f_c ) targeted attack</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1</td>
<td>Truncated Gaussian: ( p(d) \sim N(5.5, 1.11) ) in [3, 8]</td>
<td>small world</td>
<td>0.78</td>
<td>0.969</td>
</tr>
<tr>
<td>0</td>
<td>Exponential: ( p(d) = 0.583e^{-0.307d} )</td>
<td>Random</td>
<td>0.82</td>
<td>0.398</td>
</tr>
<tr>
<td>0.04</td>
<td>Power-law: ( p(d) = 15.449d^{-2.986} )</td>
<td>Scale-free</td>
<td>0.74</td>
<td>0.0028</td>
</tr>
</tbody>
</table>

Table 3.1. Critical fraction points for three critical values of \( \beta_1 \).
Figure 3.4(a) shows numerical results of percolation transition of the three networks for random failure of nodes. The percolation threshold $\phi_c$ is the point where the fraction of giant cluster $S$ goes below $\xi = 1 \times 10^{-4}$. $\phi_c$ is used to calculate the network robustness, $f_c$ which is obtained for the three networks as: $f_c^{Gau} = 0.78$, $f_c^{Expo} = 0.82$, and $f_c^{Power} = 0.74$, see Table 3.1. Based on the results, we conclude that the overall network robustness stays high as $\beta_1$ changes. This claim is verified in Section 3.4.3 with experimental analysis.

In the targeted attack scenario, the nodes are preferentially removed according to their degree. Figure 3.4(b) shows percolation transition of targeted attack on three representative networks. In order to extract the percolation threshold $\phi_c$, $\xi = 1 \times 10^{-4}$ is chosen. Consequently, the fraction points for the three networks are $f_c^{Gau} = 0.969$, $f_c^{Expo} = 0.398$, and $f_c^{Power} = 0.0028$. The robustness against targeted attack of Gaussian network is higher than the exponential network, and that of the exponential network is higher than the power-law network. It is predicted that while $\beta_1$ increases from negative to positive, the network robustness decreases gradually following a sharp decrease due to the emergence of giant-hub network structure. This claim is verified in Section 3.4.3 with computational experiment.

### 3.4.3 Verification

In this section, the theoretical solution obtained is verified by analyzing the synthetic networks that are generated with the DBDC model, thus the Task 1.3 in RQ1 is addressed. In the experimental analysis, the synthetic networks are generated using the parameters $m = 3$, and $N = 2000$. Since the generation of synthetic network is a stochastic process, the model is executed 10 times for each $\beta_1$ value. After obtaining the network properties for each run, 95% confidence interval is calculated using the $t$-distribution.
3.4.3.1 Simulation Results of Network Structures

Figure 3.5 shows how $\beta_1$ affects the network structures. It visually verifies the conclusion drawn in theoretical analysis: as $\beta_1$ increases from negative to positive, the network structure transitions from “chain” → small-world → random → scale-free → giant-hub. In the following, the transitions of degree distribution and the network properties are discussed in detail.

The simulation results of the network degree distribution at different $\beta_1$ values are shown in Figure 3.6. The fitting function for the degree distribution at critical values of $\beta_1$ are presented in Table 3.2. As shown in Figures 3.6(b) and referring to Table 3.2, for $\beta_1 \in [0, 0.05]$, the degree distribution transitions from exponential to power-law. For $\beta_1 \in [0, -2]$, the degree distribution transitions from exponential to truncated Gaussian, and then transitions to a degree distribution in which all the nodes almost have the same degree around mean. As $\beta_1$ increases beyond 0.05, as shown in the 3.6(c), it is observed that a few nodes’ degrees are much larger than others, which indicates the emergence of giant-hub structure. This simulation

![Network visualization at different $\beta_1$ values (N = 200, $m = 3$ for above and $m = 1$ for below).](image-url)
result thus verifies the conclusions drawn from Figure 3.3 in the theoretical analysis (see Section 3.4.2.1). Besides, the simulation results provide two more insights: on one hand, if $\beta_1$ increases beyond 0.3 shown in Figure 3.6(d), the network degree distribution gradually transitions back to the experiential, i.e. random network. The fitting function for degree distribution at $\beta_1 = 15$ is $F(d) = 0.8544e^{-0.212d}$ ($R^2 = 0.85$). On the other hand, as shown in Figure 3.6(a), if $\beta_1$ decreases beyond -2, the resulting degree distribution gradually transitions back to the experiential network as well. The fitting function for degree distribution at $\beta_1 = -20$ is $F(d) = 4.5679e^{-0.397d}$ ($R^2 = 0.996$). In summary, by combining all the results obtained from degree distribution analysis, the effect of node’s preferences to degree, $\beta_1$, on network structure is shown

\begin{figure}[h]
\centering
\subfloat[$\beta_1 < -2$]{
\includegraphics[width=0.4\textwidth]{fig1a.png}} \hfill
\subfloat[$-2 \leq \beta_1 \leq 0.05$]{
\includegraphics[width=0.4\textwidth]{fig1b.png}}
\subfloat[$0.05 < \beta_1 \leq 0.32$]{
\includegraphics[width=0.4\textwidth]{fig1c.png}} \hfill
\subfloat[$\beta_1 > 0.32$]{
\includegraphics[width=0.4\textwidth]{fig1d.png}}
\caption{Simulation results of the degree complementary cumulative distributions (CCD) at different $\beta_1$ values.}
\end{figure}
Figure 3.7. Effect of node’s preferences to degree on network structure.

Table 3.2. Fitting functions of the degree distribution of representative synthetic networks with three critical values of $\beta_1$.

<table>
<thead>
<tr>
<th>Critical $\beta_1$</th>
<th>Fitting Function for CCD [ F(d) = p(d_i \geq d) ]</th>
<th>$R^2$</th>
<th>Network Structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_1 = 0$</td>
<td>$2.6233e^{-0.298d}$</td>
<td>0.996</td>
<td>Random</td>
</tr>
<tr>
<td>$\beta_1 = 0.05$</td>
<td>$46.444d^{-2.673}$</td>
<td>0.976</td>
<td>Scale-free</td>
</tr>
<tr>
<td>$\beta_1 = -2$</td>
<td>$\Phi(5.99, 0.817)$ truncated between [3, 8]</td>
<td>0.982</td>
<td>Small world</td>
</tr>
<tr>
<td>$\beta_1 = 15$</td>
<td>$0.8544e^{-0.212d}$</td>
<td>0.85</td>
<td>Random</td>
</tr>
<tr>
<td>$\beta_1 = -20$</td>
<td>$4.5679e^{-0.397d}$</td>
<td>0.996</td>
<td>Random</td>
</tr>
</tbody>
</table>

in Figure 3.7. The result sheds light on the overarching goal, i.e., through the design of node-level mechanisms, the node’s preferences and behaviors can be steered such that desired system structures can be achieved.

Besides the degree distribution, the effect of $\beta_1$ on network properties is analyzed as well. The average cluster coefficient (ACC) [91] and average path length (APL) [91] are chosen as two major indices for the network properties. We investigate how the ACC and APL change as $\beta_1$ changes. Figure 3.8(a) and 3.8(b) shows how the APL and ACC change as a function of $\beta_1$, respectively. Table 3.3 shows the APL and ACC for the key $\beta_1$ values identified in the previous section. Additional $\beta_1$ values have been included where network properties show transitional behaviors. The following insights are gained from the simulation results.

1. Within the range $0 \leq \beta_1 \leq 0.32$, the APL starts at 4.33 when $\beta_1 = 0$, and decreases monotonically until it reaches a minimum value of 1.998 at $\beta_1 = 0.32$. 
In contrast, the ACC increases from 0.0041 at $\beta_1 = 0$ to a maximum of 0.868 when $\beta_1 = 0.32$. As discussed in the previous section, the network corresponding to $\beta_1 = 0.32$ has $m$ giant hubs. The APL of such networks is low because any node can reach another node in at-most two steps through the giant hubs.

2. For $\beta_1 > 0.32$, i.e., the transition from giant-hub to exponential random structure, the APL increases and the ACC declines as $\beta_1$ increases. For large $\beta_1$, APL and ACC converge to the values corresponding to an exponential random network (e.g., APL = 4.34 and ACC = 0.0043 for $\beta_1 = 15$).

3. Within the range $-5 \leq \beta_1 \leq 0$, i.e., the transition from exponential random to small world to “chain” structure, both APL and ACC increase. At $\beta_1 = -5$, the APL reaches a maximum value of 5.976, whereas the ACC reaches a local maximum of 0.2. The simultaneous increase of APL and ACC in this range indicate the emergence of small-world network similar to the WS model [16].

4. For $\beta_1 < -5$, i.e., the transition from “chain” to exponential random network, decreasing $\beta_1$ results in a gradual decrease in the values of APL and ACC. For example, the APL and ACC are 4.49 and 0.0034 respectively at $\beta_1 = -20$.

Therefore, the changing of network properties well reflect the phenomenon as observed in the degree distribution.

3.4.3.2 Simulation Results of Network Robustness

The synthetic networks analyzed in this section follow the same parameters settings, i.e., $m = 3$, $N = 2000$. To simulate network fragmentation process, 1% nodes are randomly selected in random failure. While in the targeted attack, 1% nodes with highest degree are selected. While nodes are being removed, their associated edges are removed as well. The fraction point at which the LCC reaches 2% is the critical faction point, $f_c$, that quantifies the network robustness. To address the stochastic process of the simulation, each network goes through the random failure process and
Figure 3.8. Simulation results of network properties as a function of $\beta_1$ values.

Table 3.3. Network properties and robustness against random failure and targeted attack for networks generated with critical $\beta_1$ values (values in bracket are the length of 95% confidence interval).

<table>
<thead>
<tr>
<th>$\beta_1$</th>
<th>Average Path Length (APL)</th>
<th>Average Clustering coefficient (ACC)</th>
<th>Robustness against Random Failure</th>
<th>Robustness against Targeted Attack</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_1 = -20$</td>
<td>4.49 (0.0085)</td>
<td>0.0034 (0.00094)</td>
<td>0.844 (0.027)</td>
<td>0.457 (0.0097)</td>
</tr>
<tr>
<td>$\beta_1 = -5$</td>
<td>5.98 (0.068)</td>
<td>0.19 (0.0033)</td>
<td>0.772 (0.031)</td>
<td>0.529 (0.0045)</td>
</tr>
<tr>
<td>$\beta_1 = -2$</td>
<td>4.86 (0.013)</td>
<td>0.015 (0.0020)</td>
<td>0.806 (0.025)</td>
<td>0.511 (0.0081)</td>
</tr>
<tr>
<td>$\beta_1 = 0$</td>
<td>4.33 (0.0094)</td>
<td>0.0041 (0.0011)</td>
<td>0.854 (0.025)</td>
<td>0.382 (0.011)</td>
</tr>
<tr>
<td>$\beta_1 = 0.32$</td>
<td>1.998 (0.0017)</td>
<td>0.868 (0.126)</td>
<td>0.681 (0.317)</td>
<td>0.02 (0.02)</td>
</tr>
<tr>
<td>$\beta_1 = 2$</td>
<td>3.952 (0.013)</td>
<td>0.026 (0.0035)</td>
<td>0.813 (0.058)</td>
<td>0.373 (0.014)</td>
</tr>
<tr>
<td>$\beta_1 = 15$</td>
<td>4.34 (0.012)</td>
<td>0.0043 (0.00070)</td>
<td>0.843 (0.032)</td>
<td>0.387 (0.014)</td>
</tr>
</tbody>
</table>

targeted attack 10 times for each $\beta_1$ value. The 95% confidence interval is calculated using the $t$-distribution.

Simulation results of network robustness against random failure and targeted attack are shown in Figure 3.9. It is observed that robustness against random failure is higher than the robustness against targeted attack, and overall maintains at a high level. This verifies the conclusion from theoretical analysis (see Section 3.4.2.2).
However, it is observed that the network robustness in the range of $\beta_1 \in [0.05, 2]$ is fluctuating. The values are within the range from 0.658 to 0.859. This is because, as discussed in Section 3.4.3.1, $\beta_1$ in this range corresponds to the network structure with several hubs. Since these hubs occupy edges in the network, the robustness is highly dependent on when these hubs are removed. Since nodes are randomly selected, the variation in network robustness in this range is high.

For targeted attack, the network robustness increases logarithmically $^2$ as $\beta_1$ decreases from 0 to -5, which corresponds to the transition of network structure from random to small-world network. The robustness reaches maximum at 0.533 at $\beta_1 = -5$. The robustness at $\beta_1 = 0$ is 0.382 which corresponds to the robustness of a random network. Then, the network robustness decreases for $\beta_1 \in [0, 0.32]$, and reaches its minimum value of 0.02 at $\beta_1 = 0.32$ where the largest giant-hub network structure with minimum APL and maximum ACC is formed. However, the robustness shows a sharp increase from 0.02 to 0.373 for $0.32 \leq \beta_1 \leq 2$. In summary, the network robustness continually decreases as network degree distribution transitions from truncated Gaussian to exponential and to the giant-hub. This result verifies the theoretical prediction (see Section 3.4.2.2).

Additionally, for $\beta_1 \geq 2$, the robustness is almost constant, with an average value of 0.383. For $\beta_1 < -5$, the network robustness gradually decreases to 0.457 at $\beta_1 = -20$. It is observed that both values are close to the robustness of a random network, indicating that the network robustness returns to the robustness of the random network as $\beta_1 \to +\infty$ and $\beta_1 \to -\infty$. Hence, a complete picture of how the node-level preferences $\beta_1$ affect the network robustness in Figure 3.9 is obtained.

In summary, the theoretical and experimental analysis presented in Section 3.4.2 and 3.4.3 illustrate how the connection between node-level preferences and network structure, and to the network performance is set up within the decision-centric model. The phenomenon observed and the conclusion obtained from theoretical part and

$^2$The fitting function is $y = 0.0311\ln(x) + 0.4908$ ($R^2 = 0.9803$), and is obtained if plotting the robustness over the absolute value of $\beta_1$ and exclude the point at $\beta_1 = 0$
verified from experimental part give us the complete picture of how the node-level preferences influence the network structure, properties and performance. In order to show the unique aspects of the DBDC model and highlight the difference between the DBDC model with existing degree-based network generation models, a comparative study is performed in the following Section 3.5 to accomplish Task 1.4 in RQ1.

3.5 Comparative Study between the Decision-Centric Model and Classical Network Models

In order to quantify the differences between the DBDC model and existing degree-based network generation models, a comparative study is presented in this section. The model adopted for the comparison is a degree-based model extended from the BA model, called generalized preferential attachment (GPA) model [17]. The model defines that each node $i$ has an attribute called total attractiveness $U_i$, which follows Equation (3.20)

$$U_i = G_i d_i + A_i$$  \hspace{1cm} (3.20)
where the $G$ and $A$ are node-specific parameters, called node fitness and additional attractiveness. $d_i$ is node $i$’s degree. $\tau$ is an exponent to describe whether the total attractiveness is linear or nonlinear to degree. The model prescribes that in a network with $N$ nodes, the probability for a new link to be attached to an existing node $i$ is proportional to

$$P_i = \frac{U_i}{\sum_{j=1}^{N} U_j} = \frac{G_i d_i^\tau + A_i}{\sum_{j=1}^{N} (G_j d_j^\tau + A_j)} \quad (3.21)$$

Other assumptions on the GPA model are:

1. The network is undirected, initialized as an initial ER random network contains 5 nodes linking with probability equals to 0.5.

2. The total attractiveness $U$ is a linear function of node degree $d$, i.e., $\tau = 1$.

3. The fitness value is the same for all the nodes and is time independent, i.e., $G_i$ is constant.

4. The additional attractiveness for all the nodes is the same and is time independent, i.e., $A_i$ is constant.

With the assumption added, if $G_i = constant$, its impact is accounted for in $A$ by scaling $G$ in both numerator and denominator in Equation (3.21) as follows:

$$P_i = \frac{U_i}{\sum_{j} U_j} = \frac{G d_i + A}{\sum_{j} (G d_j + A)} = \frac{d_i + \frac{A}{G}}{\sum_{j} (d_j + \frac{A}{G})} \quad (3.22)$$

So, the total attractiveness is just described as:

$$U_i = d_i + A \quad (3.23)$$

In all, the model for comparison is shown as follows:

1. **Growth mechanism:** At each time step, one new node is added with $m$ edges that link to $m$ different nodes already present in the network.
2. Linking probability: \[ P_i = \frac{d_i + A}{\sum_{j=1}^{N} (d_j + A)} \]

In this model, the additional attractiveness \( A \) describes the average impact of factors other than degree on a node’s total attractiveness. It affects the preference of a node being connected. If the additional attractiveness of all nodes increases simultaneously, the effect of a node’s degree on the linking probability is reduced. In contrast, reducing the additional attractiveness of all nodes increases the preference to a node’s degree in determining the linking probability. Therefore, \( A \) can be viewed as a parameter influencing the network structure, properties and performance. The same analysis method utilized in the DBDC model is followed. Detail analysis of the GPA model can also be referred in [92]. The main purpose of this comparative study is to find the fundamental differences between the GPA model and DBDC model which gives totally different perspective and explanation on the evolution of complex networks.

### 3.5.1 Analyzing the Generalized Preferential Attachment Model

#### 3.5.1.1 Analysis of Degree Distribution

With the continuum-based approach [83], the rate of change of a node’s degree \( d_i \) is given by

\[ \frac{\partial d_i}{\partial t} = \frac{m d_i + A}{\sum_{j=1}^{N} (d_j + A)} \quad (3.24) \]

Unlike Equation (3.11), this system of differential equations can be analytically solved to find a closed form of the network degree distribution. The asymptotic degree distribution as \( N \to \infty \) is

\[ p(d) = \alpha (m + A)^a (d + A)^{-(\alpha+1)} \quad (3.25) \]
where, $\alpha = (2 + \frac{A}{m})$. For different $A$ values, different degree distributions are achieved. For $A = 0$, the model becomes a BA model which yields the scale-free network structure with power-law degree distribution. As $A \to \infty$, intuitively from Equation (3.22), the probability of getting connection is equal to $\frac{1}{N}$ for all nodes. This results in a random network model with an exponential degree distribution. The degree distribution in Equation (3.25) transitions from power-law to exponential as $A \to \infty$,

$$\lim_{A \to \infty} p(d) = \lim_{A \to \infty} \alpha (m + A)^{\alpha} (d + A)^{-(\alpha + 1)} = \frac{e}{m} \exp(-\frac{d}{m})$$

(3.26)

This conclusion is verified using simulation analysis, as shown in Figure 3.10. It shows the degree distributions of networks generated with the GPA model using $m = 3$ and $N = 2000$ for a range of $A$ values. It is observed that the networks’ degree distributions transition from power-law to exponential network as $A$ increases from 0, which verifies the conclusion drawn from continuum theory.
3.5.1.2 Analysis of Network Properties

Figures 3.11(a) and 3.11(b) show the simulation results of the effect of $A$ on network APL and ACC. As $A$ increases, the network’s ACC decreases, whereas the APL increases monotonically. This is because increase in $A$ reduces the impact of degree on the linking probability. Accordingly, nodes with lower degrees have greater probability of linking. As $A \to \infty$, the network structure converges to an exponential network. Both ACC and APL converge to the values corresponding to the exponential random network. The effect of $A$ on network properties verifies the conclusion from the degree distribution analysis.

![Figure 3.11](image)

(a) Average path length (APL)  
(b) Average clustering coefficient (ACC)

Figure 3.11. Effect of additional attractiveness ($A$) on average path length and clustering coefficient.

3.5.1.3 Analysis of Network Robustness

Using the degree distribution derived in Equation (3.25), the percolation transition of the GPA model against random failure is obtained from Equations (3.12) and (3.13). The percolation threshold $\phi_c$ is evaluated, and the network robustness quantified by the critical fraction point $f_c = (1 - \phi_c)$ is determined. Figure 3.12(a) shows the network robustness for different values of $A$. The critical fraction points
for three $m$ values are plotted. It is observed that as $A$ increases, $f_c$ is maintained in the range $[0.75, 0.9]$, which shows that the robustness of the network against random failure is high for all $m$ values.

In the targeted attack case, the percolation transition in the GPA model is obtained using Equations (3.16) and (3.17). The percolation transition for different $A$ values are determined, thus the critical fraction points ($f_c$) are obtained. The critical fraction points are plotted in Figure 3.12(b). It is observed that the network robustness increases as $A$ increases.

The theoretical results are verified by performing fragmentation analysis on synthetic networks with different values of $A$ and calculating the LCC index. Figure 3.13 shows the results for both random failure and targeted attack scenarios. No significant changes in the network robustness are observed for random failure. However, as $A$ increases, the robustness against targeted attacks increases, as shown in Figure 3.13, which is also revealed in the theoretical analysis. The increase is more prominent for small $A$. For example, at $A = 0$ the fraction $f_c$ is 0.263, and increases to 0.346 at $A = 10$. The network robustness does not increase significantly after $A = 10$.

In this model, the additional attractiveness $A$ describes the average impacts of the factors that have not been identified besides the degree on a node’s total attrac-
Figure 3.13. Effect of additional attractiveness ($A$) on network robustness against random failure and targeted attack.

tiveness. The primary effect of the additional attractiveness is that it affects the preference of a node being connected. If the additional attractiveness of all nodes increases simultaneously, the effect of a node’s degree on the linking probability is reduced. In contrast, reducing the additional attractiveness of all nodes makes the attachment more preferential to the node’s degree. Thereby, the $A$ value is the tunable parameter that can influence the network structure, properties and performance (detailed discussion of this model is in [92]). The main purpose of this comparative study is to find the fundamental differences between the GPA model and DBDC model which gives totally different perspective and explanation on the evolution of complex networks.

3.5.2 Model Comparison and Discussion

In this section, the DBDC model is compared with the GPA model. The fundamental differences are listed below.

1. Difference vs. ratio

The first main difference between the DBDC and GPA model is in the linking
probabilities. By transforming the choice probability (i.e., Equation (3.10)) of
DBDC model and the linking probability (i.e., Equation (3.22)) of GPA model
to the following:

\[ P_i = \frac{e^{\beta_1 d_i}}{\sum_{j=1}^{N} e^{\beta_1 d_j}} = \frac{1}{1 + \sum_{j=1,j\neq i}^{N} e^{\beta_1 (d_j - d_i)}} \] (3.27)

\[ P_i = \frac{d_i + A}{\sum_{j=1}^{N} (d_j + A)} = \frac{1}{1 + \sum_{j=1,j\neq i}^{N} \frac{1+A}{\frac{d_i}{d_j} + \frac{d_j}{d_i}}} \] (3.28)

we observe in DBDC model, the difference between two nodes’ degrees affect
the choices of nodes, see Equation (3.27). However, in the GPA model, the
ratio of two nodes’ degrees affects the linking probability, see Equation (3.28).
The distinction between the two models results in different explanations on
how nodes make decisions. Using different models for modeling the real-world
networked system lead to totally different explanation on the evolution of the
systems.

2. Fitness vs. additional attractiveness

With the same interpretation of parameters in both models, parameters \( \beta_1 \)
and \( \beta_0 \) in utility correspond to the fitness \( G \) and additional attractiveness \( A \)
respectively in the GPA model. However, \( \beta_0 \) is canceled out in DBDC model
when deriving the choice probability (i.e., Equation (3.10)), thus only \( \beta_1 \) (i.e.,
fitness) determines the choice probability. In contrast, in the GPA model, only
additional attractiveness \( A \) matters since the \( G \)’s impact on choice probability
is accounted for in \( A \) (i.e., Equation (3.22)). The difference observed at this
point gives us insights on incentive design to affect node-level behaviors and the
evolution of complex networked systems: one way is to provide incentives to
change individuals’ preferences, such as the preference to degree, \( \beta_1 \). Whereas
another way is to give incentives to change individuals’ characteristics, such as
the node’s additional attractiveness, \( A \).
3. A richer model

As the results indicates, the DBDC model is simple, yet generic. $\beta_1$ and $A$ are the two tunable parameters in each model that influence the network structure, properties and performance. But, $\beta_1$ has broader impact than $A$ does. By changing $\beta_1$, a wide range of network structure, such as line graph, “star” graph, small-world network and scale-free network, may be obtained. The richer model give us the capability of capturing the structures of more types of complex networked system. On the other hand, the GPA model does not have this capability.

4. Generality and Extensibility

The generality is embodied by providing decision-making interpretation to existing model. For example, the choice probability of the DBDC model in Equation (3.10) transforms to the GPA model if the utility function defined in Equation (3.9) is changed to $V_j = \ln(\beta_1d_j + \beta_0)$. Furthermore, the major advantage of the DBDC model is its extensibility. The model can be extended by including other network-metrics and/or non-network-metrics, linear or nonlinear utility functions, and by considering the preferences variations (i.e., multinational logit model or mixed logit model). It provides a modeling framework for different network generation models.

5. Insights on design of network with high robustness and resilience

The scale-free networks are “paradoxically both robust and fragile” [93]. This characteristic is one of the factors that result in the malfunction of the real-world complex systems, such as the paralysis of Internet and the large-scale blackouts in power grid. Thereby the question is raised: how to design a system whose structure is robust against both random failure and targeted attack? The analysis results obtained from the two models shed light on this question assuming the models could represent the mechanism underlying the real systems.
As a guiding principle, $\beta_1$ values resulting in giant-hub networks should be avoided for the low robustness against targeted attack. There are three critical values of $\beta_1$ for acquiring a network with high robustness: $\beta_1 = \{0, 2, -2\}$, as shown in Figure 3.9. For $\beta_1 = 0$, a random network is obtained with high robustness against targeted attack. The second critical value is at $\beta_1 = 2$ where the resulting network has the robustness value of random network and at the same time, the network robustness does not increase but is maintained after this value. The third critical value is in the transitions of network structure from random to small-world as $\beta_1$ decreases from 0 to -5. The increase of robustness follows logarithmic in this range, as obtained in Section 3.4.3.2. Below $\beta_1 = -2$, the robustness does not increase significantly. To interpret these results into the real-world scenario, incentives should be designed to influence the individuals’ preferences to degree so that a network structure that is invulnerable to intentional attack can be achieved.

In the GPA model, as shown in Figure 3.13 the maximum possible value of $A$ should be chosen to get maximum robustness against targeted attack [92]. However, there may be tradeoffs associated with choosing the maximum value of the additional attractiveness, e.g., when additional attractiveness is attained by providing monetary incentives. Based on the simulation results, the increase in the robustness diminishes after the critical value of additional attractiveness at about $A = 10$. It is found that the critical value of $A$ is highly dependent on the average degree which follows Equation (3.29) based on the simulation.

$$A = \mu \cdot \langle d \rangle$$

(3.29)

where $\mu \in [1, 2]$ is the design coefficient obtained from the simulation results, and $\langle d \rangle$ is the average degree of the network. This equation can be used to help identify the critical value of $A$ to guide the design of the network with high robustness.
So far, the results presented in Section 3.4 are for answering the RQ1. Referring back to the Figure 1.3. The outcomes of this section accomplish three tasks. The DBDC model proposed in Section 3.4.1 provides a specific example to show how the random-utility discrete choice theory can be integrated with the network to model the node-level preferences and to establish the relationship between the node-level preferences and behaviors. The analysis results obtained theoretically in Section 3.4.2 and verified experimentally in Section 3.4.3 provide a direct mapping between the node-level preferences and system-level performance in terms of the network robustness. Finally, the comparative study performed in Section 3.5 reveals the fundamental differences between the decision-centric model and the classical degree-based model. Design insights are gained for achieving networks with high robustness and resilience by obtaining the critical values of parameters representing the node-level preferences.

3.6 Closing Comments for Chapter 3

In this chapter, a degree-based decision-centric (DBDC) model is proposed for modeling the evolution of complex networked systems. The dynamic characteristics of the model are analyzed theoretically and verified with simulation. In addition, the robustness of networks generated with the proposed DBDC model against random failure and targeted attack is analyzed. The results discussed in this chapter provide insights on how node-level preferences influence the network structure, properties and network robustness. Utilizing the random utility theory, discrete choice analysis, together with continuum theory and percolation theory establishes a bridge for connecting the node-level preferences to the system-level performance. Therefore, a holistic framework that quantitatively captures the mappings across all the five levels (shown in Figure 1.3) is provided. The DBDC model is compared with a degree-based model - the generalized preferential attachment model. The results obtained from the comparative study provide insights and guidelines for the ways in which incentives
can be designed to influence the node-level preferences and behaviors, to steer the whole system and to achieve desired properties and performance.

One major limitation of the DBDC model is that it only models a situation where all decision-makers have the same decision-making preferences. In addition, the model only accommodates a scenario where decision-makers make independent decisions. The model assumes that the decision-makers’ behaviors will not affect with each other. Therefore, the DBDC model is not suitable to model individuals’ strategic decision-making behaviors. Furthermore, the model does not take into account domain-specific knowledge and instead only considers degree in the network generation process. This limits the application of the model in real-world systems. In the next chapter, the discrete choice modeling approach is applied to a real complex networked systems, the autonomous system (AS) level Internet. The goal is to estimate ASes’ peering preferences in the Internet. In order to take into account heterogeneous preferences among ASes, the mixed logit model is employed by assuming that decision-makers’ preferences follow a certain type of distribution. Given the Internet context, domain-specific decision criteria, such as geographic location of ASes, the traffic of each AS transits and ASes’ business type, are considered in the estimation.
CHAPTER 4. ESTIMATING LOCAL DECISION-MAKING PREFERENCES IN COMPLEX NETWORKED SYSTEMS

4.1 Chapter Overview

In this chapter, the objective is to answer RQ3 in the context of complex networked systems through the Approach 3.1. RQ3 is: How can the nodes’ unobserved preferences and behaviors be estimated?

As shown in Figure 1.3, in order to influence the structure and performance of complex networked systems, it is crucial to understand local decisions. The associated research gap is that there is a lack of systematic approach for estimating node-level decision-making preferences in complex networked systems. The proposed approach is based on the discrete choice models introduced in Section 3.2. With the proposed approach, the decision-making preferences are estimated with statistical techniques including Maximum Likelihood Estimation and Bayesian Estimation.

This approach is utilized for an illustrative example of complex networked systems – the Autonomous System (AS) level Internet. First, different types of AS linking activities are categorized. For each type of activity, there are multiple criteria for making linking decisions. So, the decision criteria are investigated. Finally, the geographic distance, number of customers and providers, business types, etc. are identified as the key factors affecting ASes’ linking decisions. To perform the preferences estimation, the data related to these factors as well as data related to the network metrics are collected and analyzed. The aim is to estimate ASes’ heterogeneous peering preferences for these factors in different linking activities. The outcomes of this study are three peering behavior models that quantify the linking probability for three different linking activities. Specific insights regarding ASes’ peering strategies are gained. In addition, the proposed approach uncovers hidden behavioral patterns that existing
approaches cannot reveal. Below, Figure 4.1 displays how the proposed approaches answer RQ3 in the context of complex networked systems.

Figure 4.1. Overview of Chapter 4: research questions, approaches and tasks.

The outline of this chapter is as follows. In Section 4.2, a background of AS-level Internet is presented. In particular, the state-of-art modeling of Internet topology and evolution is reviewed. In Section 4.3, the decision-centric approach for estimating AS peering preferences and behavior is proposed. In Section 4.4, the factors affecting AS linking decisions are explored and identified. Then, the corresponding data is collected. With the obtained datasets, preliminary analysis, such as descriptive statistics and distribution analysis, is performed in Section 4.5. In Section 4.6, the model parameters are estimated using Bayesian estimation techniques with the assumptions that different ASes have different peering preferences. The estimation results are discussed and interpreted in detail for each model. Finally, the closing comments are presented in Section 4.7.
4.2 Autonomous System (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. Figure 4.2 shows a snapshot of the AS-level Internet network visualized as complex network. The Internet is a typical example of complex networked system in both physical and virtual aspects. Modeling the Internet topology has significance in many aspects. For example, it is important for developing routing protocols (e.g. the Border Gateway Protocol (BGP)), determining the factors that affect its robustness, and understanding the interplay between technology, topology, and economic forces [94]. The topology of the Internet can be studied at three different levels [95]:

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. This physical infrastructure is the one over which information is routed.

3. **AS level**, which describes how ASes are interconnected. The Internet is composed of thousands of domains interconnected with each other.

![Image of AS-level Internet graph](image-url)

Figure 4.2. A visualization of AS-level Internet graph from Cooperative Association for Internet Data Analysis (CAIDA) [96].
4.2.1 Background of AS

An AS is defined as “a connected group of one or more IP prefixes run by one or more network operators which has a single and clearly defined routing policy” [97]. Examples include Internet Service Provider (ISP), corporate networks, universities, etc. An ISP can have one or more ASes. An AS has its own set of routers and routing policies [98], and is connected with other ASes via dedicated links or public network access points. A link between two ASes represents a contract to forward data to each other. Two main objectives of each AS are to minimize the cost of building links with other ASes and to minimize the delay in sending and receiving packets [6]. Since both of these objectives are conflicting in nature, a tradeoff is required. Each AS can choose its policy to select the best route for data based on commercial contractual relationships. These contracts and policies play a significant role in determining the structure of the Internet and its overall performance [98]. The AS-level topology also influences the definition of routing protocols such as the BGP [99]. Hence, it is an important and appropriate level of abstraction to model the decisions that influence the structure of the Internet.

The formation of AS-level network is a dynamic process [100]. A new AS is added into the network when a new ISP or a large institution enters the network. New links are added when customer ASes purchase Internet access from ISP ASes, i.e. costumer to provider (C2P), or when two ASes agree to share information among each other, i.e., peering to peering (P2P). Links and nodes are deleted from the network when the corresponding administrative domains cease to exist or merge with other ASes.

The dynamic process reflects the decision-making nature of an AS. These evolutionary dynamics, i.e., the determination of which ISP should be selected, the agreement on which AS should be peered with, and whether an AS should cease to exist or not, are all decision-making behaviors. However, the existing Internet topology generators are either random graph generators, structural generators, or power law generators [99] that do not account for this decision-making characteristic.
To capture the evolutionary dynamics, in this dissertation, the evolution of Internet structure is modeled as a function of the decisions made by the individual AS. With a decision-centric topology generator, this case study provides evidence that the observed AS-level Internet structure has resulted from the local decision-making process of individual ASes.

4.2.2 Modeling AS-level Internet – State-of-the-Art

The research on AS-level Internet network reconstruction has been intensive in the last decade, and a number of network generators have been proposed. Waxman [101] in 1988 first proposed a random graph generator to simulate the protocol in the Internet by extending the classical ER model. It incorporates the information about distance into the linking probability function. Later, researchers find that the real network topology does not have a random structure, but exhibit hierarchy. Models that address this hierarchical structure in the Internet are referred to as structure-based model in which the Transit-Stub [102] and Tiers [103] are the two symbol generators.

These structural generators had been popular until the renowned power-law degree distribution phenomenon was first discovered in AS-level Internet network by Faloutsos et al. [104]. Subsequently, structural generators itself proved to be unsuitable for modeling the Internet since the inability of producing the power-law distribution. In contrast, a large number of topology generators, which primarily focus on matching the power-law degree distribution of the Internet, were proposed. These Internet topology generators are referred to as degree-based models. For example, the classical degree-based models include the power-law random graph (PLRG) [105], BRITE generator [106], Barabsi-Albert (BA) Model [15], various BA variants [18, 19, 107], generalized preferential attachment (GPA) model [99], Internet topology generator (Inet) [108], positive feedback preference (PFP) model [109].
Preferential attachment has been regarded as the foundation for the degree-based models. However, it has been recently found that not only the popularity but also the similarity works as a strong force shaping the power-law network structure by Papadopoulos et al. [110]. They propose a popularity $\times$ similarity optimization (PSO) model where new connections are formed by optimizing certain trade-offs between popularity and similarity. Also with the optimization-based perspective, Alderson et al. [111] propose a highly optimized tolerance (HOT) model and argues the Internet topology can be understood in terms of trade-offs between network performance and technological and economic factors. Later, Fabrikant et al. [112] show a similar phenomenon suggesting that power law tends to arise as a result of complex, multi-objective optimization. They propose a toy model in which the Internet growth is simulated as a result of optimization along two objectives: “last mile” connection costs and transmission delays measured in hops.

Agent-based models (ABM) define a set of agents and interaction rules that mimic the dynamic of real systems. Lodhi et al. [113] propose an agent-based network formation model called GENESIS. GENESIS is developed based on realistic strategies of transiting and peering at AS-level. In GENESIS, ASes behave selfishly in a decentralized manner to optimize a cost-related fitness function. In order to understand the intervention strategies for controlling malware over the Internet, Holme et al. [114] propose an ABM called ASIM to reproduce key features of AS-level Internet by modeling each AS as an economic agent. Each agent manages traffic over a geographically sub-network and profits from the traffic that flows through its network.

Another group of models explain the evolution of Internet as games played by pair of ASes. There have been several local connection games to model peering relations between ASes. Fabrikant et al. [29] propose a game that models creation of Internet topology by modeling selfish nodes who pay for the links that they establish, and benefit from short paths to all destinations. Johari et al. [115] develop a game model to investigate the traffic over Internet by considering cost incurred for routing.
A summary of existing Internet generation models is provided in Table 4.1. Although a number of Internet generation models have been proposed in the past two decades, there are mainly two limitations of the existing models. First, existing models of the decision-making behaviors of ASes are not quantitatively generated from historical data. They are either assumed or established based on experience and/or empirical analysis. The second limitation is that existing models consider all ASes as homogeneous entities, but do not account for the heterogeneity of different decision-makers. For example, Sha and Panchal [116] developed a decision-centric model with historical data, their model fails to address heterogeneity of the utility function that different ASes want to maximize.

The approach proposed in this chapter utilizes the observable Internet topology as the input and discrete choice analysis as the means for estimating the decision-making preferences and modeling the behaviors of ASes by considering not only the network metrics but also the geographical and economic aspects of each AS. The model provides an interpretation of AS decisions through the lens of utility-maximization principles. The statistical model, i.e., mixed logit model, adopted does not only support estimating the AS with same decision-making preferences, but also enables the estimation of heterogeneous preferences for different ASes. To the best of our knowledge, this is the first study on analyzing the decision-making strategy at the AS-level of Internet with the discrete choice models, specifically, with mixed logit model. The overarching goal of this study is to provide a modeling approach that helps derive the model for network generation. The model is derived statistically from a historical dataset, and is validated through statistical measurements on the overall fit of the model. In the next section, the approach is presented. The performance of the proposed approach and other commonly used network generating approaches are compared in detail in Chapter 5.
Table 4.1. Summary of existing Internet topology generators.

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Model Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random graph mode</td>
<td>Waxman generator [101]</td>
</tr>
<tr>
<td>Structure-based model</td>
<td>Transit-Stub [102]</td>
</tr>
<tr>
<td></td>
<td>Tiers [103]</td>
</tr>
<tr>
<td>Degree-based mode</td>
<td>PLRG [105]</td>
</tr>
<tr>
<td></td>
<td>BRITE [106]</td>
</tr>
<tr>
<td></td>
<td>BA &amp; BA-variant [15, 18, 19, 107]</td>
</tr>
<tr>
<td></td>
<td>GPA [99]</td>
</tr>
<tr>
<td></td>
<td>Inet [108]</td>
</tr>
<tr>
<td></td>
<td>Xenofontas's framework [117]</td>
</tr>
<tr>
<td></td>
<td>PFP [109]</td>
</tr>
<tr>
<td>Optimization-based model</td>
<td>PSO [110]</td>
</tr>
<tr>
<td></td>
<td>HOT [111]</td>
</tr>
<tr>
<td></td>
<td>Toy Model [112]</td>
</tr>
<tr>
<td></td>
<td>Univariate HOT model [118]</td>
</tr>
<tr>
<td></td>
<td>Economically-principled generative model [119]</td>
</tr>
<tr>
<td>Agent-based model</td>
<td>GENESIS [113]</td>
</tr>
<tr>
<td></td>
<td>ASIM [114]</td>
</tr>
<tr>
<td></td>
<td>Decision-based HOT [120]</td>
</tr>
<tr>
<td></td>
<td>ITER [121]</td>
</tr>
<tr>
<td></td>
<td>GDNG [122]</td>
</tr>
<tr>
<td>Network formation games</td>
<td>Local connection game model [123]</td>
</tr>
</tbody>
</table>

4.3 A Decision-Centric Approach for Estimating AS Peering Preferences and Behaviors

An overview of the approach is presented in Figure 4.3. The approach consists of three steps: 1) decision-making criteria exploration, 2) data collection and analysis, and 3) model implementation. The core of the approach is a statistical model derived from Discrete Choice Analysis (DCA) presented in Section 3.2. The assumption is that each AS tries to maximize the utility function in the form of Equation (3.1). The
outcome of implementing this framework is the model parameters $\beta$ in Equation (3.5) or Equation (3.6) estimated by regression analysis. In Section 4.4.1, the aim is to identify the key decision criteria used by ASes. The identified criteria guide the collection of corresponding data. Specifically, three different types of linking activities in AS-level Internet are identified. For each type of peering activity, we investigate the preferences of ASes in making peering decisions. In Section 4.4, the ways of collecting data are introduced. The data are processed, and the evolutionary dynamics between two consecutive network instances are analyzed for acquiring sample observations that can be used for estimation. Specifically in this step, the trade-off in determining appropriate time interval for two consecutive network instances is proposed. The analysis of the descriptive statistics for the data is performed in Section 4.5. The results from descriptive statistics provide the distribution of data and help identify variables that are potentially important for model estimation. In Section 4.6, correlation analysis is performed to reduce the number of explanatory variables. Preliminary regression is then developed and appropriate models are selected. The models are refined and finally determined by evaluating the performance of estimation and by performing hypothesis testing.

![Figure 4.3. Overview of the approach.](image)

The approach is extensible to scenarios where more decision criteria and the associated data can be obtained. While this approach is proposed for estimating AS peering preferences and behaviors, the structure is general and can be applied to other
networked systems to elicit the local linking preferences. The overall approach can be used for other types of networks.

4.4 Data Collection

4.4.1 Factors Affecting AS Linking Decisions

The ASes play different roles depending on the business relationships with other ASes, which results in a multi-tier hierarchy of transit providers. Backbone providers, also known as Tier-1 ISPs, are at the top of the hierarchy. The regional ISPs are often regarded as Tier-2 ISPs, which are the customers of Tier-1 ISPs. The residential and small business access providers are Tier-3 ISPs. They are typically the customers of Tier-2 ISPs. The hierarchical structure places the content providers (CPs) at the lower layer as the customers of Tier-1, Tier-2 and/or Tier-3 ISPs. The rest of ASes placed at the bottom of this hierarchy are various customers [121].

The business relationships and the traffic among ASes are based upon commercial agreements which can be broadly classified into Internet transit and Peering. In Internet transit, one AS (the provider) sells Internet connectivity to the other (the customer). In peering, two ASes bilaterally agree to exchange their local and customer routes for free, called the settlement-free agreement. The two categories of commercial agreements lead to two major types of business relationships between pairs of ASes: a) Customer to Provider (C2P) relationship, and b) Peer-to-Peer (P2P) relationship.¹

The criteria used by an AS to make linking decisions are inherently dependent on the business-type. In addition, each AS has its own unique set of decision criteria. For example, the factors involved in a link establishment between new ASes and existing ISPs are different from the establishment of a new link between two existing ASes.

¹Sibling-to-sibling (S2S) relationship is not considered here because: 1) both ASes in S2S are administered by the same decision making organization, and 2) the S2S linking percentage in real Internet is low, e.g., there were only 0.9% S2S links (298 out of 32955 AS connections) in the AS Internet graph in Jan. 5th 2004.
Similarly, the criteria are different for ASes forming C2P and P2P connections. In the following, we summarize the key criteria for three different types of linking activities.

1. **Activity 1: C2P connections between new ASes and existing ASes:** When a new AS decides to engage in a C2P relationship, the decision about which ISP to buy the Internet service from depends on the quality of access services offered by the ISPs. Specific factors include the transit price, cost, reliability, performance of the provider, number of value-added services, geographic reach, projected network built out, prior acquaintance between the parties involved, customer service, and technical support [120].

2. **Activity 2: C2P connections between two existing ASes:** With the increase in the traffic volume or the number of customers, existing ASes may seek to buy additional Internet service to enlarge their bandwidth. Their overall goal is to minimize their operational expenses, maximize their transit revenue, and/or improve performance and reliability. In addition to the criteria mentioned for new ASes, real-time factors such as traffic to the potential ISP candidates, and the number of their current customers are also important in the decision.

3. **Activity 3: P2P connections between existing ASes:** An existing C2P relationship between two ASes may evolve into a P2P relationship due to the customer’s growing market share [120]. Alternatively, an existing P2P relationship can be terminated by either of the two parties. The top 4 reasons to peer, as listed in [124], include lower transit costs, lower latency, usage-based traffic billing, and marketing benefits. In the current ISP market, the unit bandwidth cost of a P2P connection is much lower than the corresponding upstream bandwidth cost. By shifting traffic from expensive upstream links to relatively cheaper peer links, ASes can reduce their network operating costs [120].

In this study, the proposed approach is applied to model the decision-making behavior for the three linking activities. Existing literature has shown that geographic, economical and traffic factors have the most influence on the evolution of
Therefore, we collect data related to geographic reach, traffic, transit cost, business role of AS, and data related to AS business relationship such as number of customers, and the number of providers.

There are two rounds of data collection: raw data collection and secondary data derivation. The raw data are either acquired from existing sources or by querying online databases. In this study, raw data include the AS Internet graph, AS geographic data, AS residential access utility and AS web utility. The secondary data are derived from raw data either by functional mapping or by transformation. The secondary data include AS-link geographic distance, AS business access utility, AS business-type, the number of providers and number of customers of each AS. Figure 4.4 provides an overview of the data collection process.

![Data Collection Process Diagram](image)

Figure 4.4. Overview of the data collection process.

### 4.4.2 Graph Data of AS-level Internet

Publicly available data sources are available for Internet network data. Skitter, Archipelago (Ark) from Cooperative Association for Internet Data Analysis (CAIDA) [96] and the RoutView [125] 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 [96].

The dataset adopted is from CAIDA AS Relationship Dataset from January 2004 to November 2007. There are 122 files in total, each file containing a full AS graph annotated with AS relationships derived from a set of BGP table snapshots.

### 4.4.3 Geographic Data of ASes

Obtaining the geographic coordinates of ASes involves two steps. First, the AS geographic location such as the street, city, state, country, etc., is extracted by querying WHOIS [126]. Second, the geographic location is used to query the online map database, OpenStreetMap API [127]. The response is analyzed and the longitudes and latitudes are extracted. The coordinates are within the city level, i.e., the distance between two ASes in the same city is considered zero. For this study, only the geographical information of ASes belonging to American Registry for Internet Numbers (ARIN), i.e., ASes’ in North America, is obtained. Figure 4.5 shows the geographic distribution of ASes in the US mainland in 2004. The distribution shows that most ASes are located around the major cities of each state. To acquire the geographic distance between two ASes, the haversine equation [128] is adopted to calculate the shortest distance over the earth’s surface.

### 4.4.4 Traffic-related Data and Address Space of ASes

Chang et al. [129] proposed an asymmetric gravity model to calculate the traffic demand between each pair of ASes. To implement the model, a normalized rank vector for each AS is essential. The rank vector is a $3 \times 1$ vector consisting of the rank of each AS’ web hosting (WH) utility, residential access (RA) utility and business access (BA) utility in decreasing order. The WH and RA utility values \(^2\) are directly accessible

\(^2\)WH utility and RA utility are collected in Sep. 2004. For better estimation, in the analysis in Section 4.6, only the Internet graphs data in 2004 is adopted. Meanwhile, by using these datasets
through [130]. The BA utility depends on the AS graph investigated. We adopt the algorithm in [129] for calculating the BA utility for each AS. After the data for three utilities are obtained, we develop three ranked lists of ASes, each corresponding to a utility. If two ASes have the same BA utility, Chang et al. [129] suggest the use of AS address space as the second sorting reference.\textsuperscript{3} The AS address space defines how many IP addresses may be administrated by an AS. We again query WHOIS and get the information regarding the announced prefixes of each AS.

4.4.5 Business Type Data of ASes

The business role that each AS plays in the Internet inherently determines the AS activities. The adopted taxonomy for categorizing the AS business type is based on the business model proposed by Chang et al. [129]. The model utilizes the normalized BA, WH and RA rank to categorize the ASes in seven different business types – Tier-1, retail service (RS), business service (BS), network service (NS), web hosting (WH), residential access (RS) and business access (BA).

\textsuperscript{3}Since the focus of this study is on decisions made by new ASes, there is no traffic between the new ASes and the target ASes. So, traffic demand calculation is not necessary.
4.4.6 Preparing Sample Observations

To acquire the observation set which contains information regarding target ASes that are selected by the decision-making ASes, we analyze the evolutionary edge dynamics between two consecutive graphs to extract the new C2P/P2P links. Monthly snapshots of the Internet network topology data are collected in 2004. There are several options to determine the two consecutive graphs with different time intervals. Appropriate month interval should be determined for extracting the observations dataset. The trade-off is that the narrower the time interval, the better the data captures the network evolution, therefore, more the accurate estimation results. On the other hand, there may be only a few new C2P/P2P links established during the narrow time interval. This results in fewer observations for performing the statistical analysis, and impedes obtaining significant statistical results. For example, there are only 41 valid P2P links newly established from January to February. For estimating with the logit models in Equation (3.6), only trivial results are obtained. With trial and error, we determine that 2-months is an appropriate time interval for obtaining the evolution instances, and therefore the observations for the model.

Table 4.2 provides the statistics of edge dynamics related to five evolution instances from Jan. 5th to Nov. 1st, 2004. The newly established edges are categorized according to the three different linking activities. With linking activity 1 as an example, it is observed that there are 986 new edges established between new ASes and existing ASes from Jan. 5th 2004 to Mar. 1st 2004. Among them, there are 959 C2P events, 17 P2C events, 6 P2P events, and 4 S2S events. Out of the 959 new C2P events, there are 378 edges with both source AS and target AS have available data. Therefore, we have 378 available observations for model estimation. From the five evolution instances, it is shown that most newly established connections are C2P connections, but only a few P2Cs. The new P2P and S2S connections are rare. This is because most P2P connections are established between two existing ASes, and the few S2S connections is due to the small population of ASes working in S2S relationship.
Table 4.2. Statistics for the new connection events in five consecutive evolution instances from Jan. 5th to Nov. 1st, 2004.

<table>
<thead>
<tr>
<th>Variable Description</th>
<th>Jan. 5th – Mar. 1st</th>
<th>Mar. 1st – May. 3rd</th>
<th>May. 3rd – Jul. 5th</th>
<th>Jul. 5th – Sep. 9th</th>
<th>Sep. 9th – Nov. 1st</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of new links between new ASes and existing ASes</td>
<td>986</td>
<td>1351</td>
<td>1094</td>
<td>1240</td>
<td>1204</td>
</tr>
<tr>
<td>Number of C2P from new ASes to existing ASes</td>
<td>959</td>
<td>1255</td>
<td>1052</td>
<td>1184</td>
<td>1157</td>
</tr>
<tr>
<td>Number of P2C from new ASes to existing ASes</td>
<td>17</td>
<td>88</td>
<td>35</td>
<td>39</td>
<td>37</td>
</tr>
<tr>
<td>Number of P2P from new ASes to existing ASes</td>
<td>6</td>
<td>7</td>
<td>6</td>
<td>11</td>
<td>7</td>
</tr>
<tr>
<td>Number of S2S from new ASes to existing ASes</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>Number of observations (ASes with available data)</td>
<td>378</td>
<td>494</td>
<td>439</td>
<td>519</td>
<td>475</td>
</tr>
<tr>
<td>Number of AS alternatives</td>
<td>141</td>
<td>155</td>
<td>126</td>
<td>153</td>
<td>133</td>
</tr>
<tr>
<td>Number of decision-making ASes</td>
<td>280</td>
<td>345</td>
<td>330</td>
<td>368</td>
<td>332</td>
</tr>
<tr>
<td>Number of new links between two existing ASes</td>
<td>2665</td>
<td>3779</td>
<td>3148</td>
<td>3528</td>
<td>2692</td>
</tr>
<tr>
<td>Number of C2P between two existing ASes</td>
<td>2247</td>
<td>2777</td>
<td>2487</td>
<td>2639</td>
<td>2313</td>
</tr>
<tr>
<td>Number of P2P between two existing ASes</td>
<td>409</td>
<td>992</td>
<td>656</td>
<td>886</td>
<td>374</td>
</tr>
<tr>
<td>Number of S2S between two existing ASes</td>
<td>9</td>
<td>10</td>
<td>5</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>Number of C2P observations (ASes with available data)</td>
<td>914</td>
<td>1026</td>
<td>793</td>
<td>1057</td>
<td>902</td>
</tr>
<tr>
<td>Number of C2P AS alternatives</td>
<td>223</td>
<td>220</td>
<td>210</td>
<td>228</td>
<td>221</td>
</tr>
<tr>
<td>Number of C2P decision-making ASes</td>
<td>867</td>
<td>966</td>
<td>727</td>
<td>975</td>
<td>845</td>
</tr>
<tr>
<td>Number of P2P observations (ASes with available data)</td>
<td>73</td>
<td>151</td>
<td>83</td>
<td>214</td>
<td>61</td>
</tr>
<tr>
<td>Number of P2P AS alternatives</td>
<td>43</td>
<td>79</td>
<td>49</td>
<td>94</td>
<td>25</td>
</tr>
<tr>
<td>Number of P2P decision-making ASes</td>
<td>44</td>
<td>63</td>
<td>46</td>
<td>96</td>
<td>46</td>
</tr>
</tbody>
</table>
4.5 Data Analysis – Descriptive Statistics

In this section, the AS graph on Jan. 5th, 2004, and the evolution from Jan. 5th to Mar. 1st, 2004 are adopted as illustrative examples to present the descriptive statistics of collected data. The descriptive statistics provide initial guidance for model selection and identification.

4.5.1 Geographic Distance between ASes

The graph for AS Internet on Jan. 5th, 2004 has 16301 nodes and 32955 edges. Among these, there are 14675 edges from North America which occupy 44.5% of all edges in the graph. With the approach discussed in Section 4.4.3, a total of 7927 geographic coordinates of ASes in ARIN are collected. Figure 4.6 shows the distribution of geographic distance of these 14675 edges. The minimum distance is 0, indicating two ASes in the same city, while the maximum distance is 7954 km. It is the distance from AS7018 (AT&T Service Inc. in Middletown, New Jersey) to AS11560 (Sandwich Isles Communications Inc. in Honolulu, Hawaii). The average distance is 927.42 km. 63.4% of the connections in the US are between ASes less than 1000 km away, and 87% of the connections in the US are less than 2000 km away. This indicates ASes prefer to link with nearby ASes.

4.5.2 Address Space of ASes

In total, 12461 ASes have the data for address space from the WHOIS entry. Figure 4.7 shows the distribution of the address space. The figure indicates that 89% of the ASes have small address spaces ranging from 0 to 131072, while only a few ASes have large address space, indicating large ASes such as Tier-1 ISPs and regional ISPs. Hence, most ASes are small and medium sized ASes. It is also observed that the address space distribution is segmented by regional peaks at 256, 66536, 131072, 196608, and 262144. This is probably because there are different categories of ASes.
with different functionality. ASes with address space between 0 to 66536 can be described as small business company or university who do not need to administer many IP address. In each category, the distribution roughly follows a power-law, indicating that in each category, most ASes administer small number of IP address, whereas only a few ASes are large with large address spaces.
4.5.3 Business Types of ASes

Based on the AS business type taxonomy, ASes are categorized into 7 different types. As summarized in Table 4.3, about 65.8% ASes have the role of providing business access, while only 0.5% and 3.6% ASes belong to the category of residential service and residential access, respectively. The number of ASes in the Tier-1, business service and web hosting are roughly equal. From Jan. 5th, 2004 to Mar. 1st, 2004, there are 378 new C2P linking events between new ASes and existing ASes. Within these observations, there are 280 decision-making ASes, and 141 target ASes. Most decision-making ASes (about 80.7%) belong to the group of business access. However, in the set of chosen ASes in this period, about 45.4% are Tier-1 ISPs. This phenomenon reflects the preferences of ASes in choosing target ISPs. Similar trends are observed in the new C2P links between two existing ASes as shown in the Table 4.3. However, it is observed that in the new P2P links between two existing ASes, both decision-making ASes (about 45.5%) and chosen ASes (about 46.5%) are from Tier-1 ISPs, indicating that the private P2P connections are favored by Tier-1 ISPs.

Table 4.3. Descriptive statistics for the number of ASes in each business type on Jan. 5th 2004 and the first evolution from Jan. 5th to Mar. 1st.

<table>
<thead>
<tr>
<th></th>
<th>Tier-1</th>
<th>RS</th>
<th>BS</th>
<th>NA</th>
<th>WH</th>
<th>RA</th>
<th>BA</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASes on Jan. 5th</td>
<td>1094</td>
<td>82</td>
<td>1422</td>
<td>1328</td>
<td>1062</td>
<td>593</td>
<td>10720</td>
</tr>
<tr>
<td>Decision-making ASes</td>
<td>4</td>
<td>3</td>
<td>13</td>
<td>9</td>
<td>20</td>
<td>5</td>
<td>226</td>
</tr>
<tr>
<td>Chosen ASes</td>
<td>64</td>
<td>0</td>
<td>27</td>
<td>17</td>
<td>1</td>
<td>1</td>
<td>31</td>
</tr>
<tr>
<td>Decision-making ASes</td>
<td>84</td>
<td>4</td>
<td>112</td>
<td>48</td>
<td>59</td>
<td>22</td>
<td>538</td>
</tr>
<tr>
<td>Chosen ASes</td>
<td>97</td>
<td>0</td>
<td>35</td>
<td>41</td>
<td>1</td>
<td>0</td>
<td>50</td>
</tr>
<tr>
<td>Decision-making ASes</td>
<td>20</td>
<td>0</td>
<td>7</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>12</td>
</tr>
<tr>
<td>Chosen ASes</td>
<td>20</td>
<td>0</td>
<td>6</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>13</td>
</tr>
</tbody>
</table>
4.5.4 Number of Customers and Providers of ASes

Figure 4.8 shows the distribution of ASes’ number of customers, number of providers and the total neighbors. It is observed that all the three distributions follow the power-law, which indicates the effect of degree on the linking preference, and also the possible imbalance in distributing the traffic and resources. Existing degree-based models have been regarded as useful tools for developing Internet topology generators. In this study, we focus specifically on the ASes’ preferences on each composition of the degree, number of customers (i.e., the in-degree) or number of providers (i.e., the out-degree), to determine which type of degree serves as the key factor that affects the ASes’ decision-making. Furthermore, instead of purely including or excluding the degree in the model, the proposed approach provides utility-maximization interpretation to individual ASes’ linking behavior.

Figure 4.8. The distribution of ASes connectivity in AS Internet graph on Jan. 5th, 2004.
4.6 Model Implementation and Parameter Estimation

4.6.1 Correlation Analysis

Before estimating the parameters in the model, we first perform correlation analysis on the candidate variables to avoid multicolinearity in the model. The analysis helps reduce the number of explanatory variables, and also helps in managing the complexity of the model selection process. The correlation coefficient adopted is based on the Pearson correlation coefficient [131]. The results are a symmetric matrix as shown in the Table 4.4. The correlation coefficients between the address space and the number of customers, and the BA utility, are 0.67 and 0.6 respectively, indicating a high correlation among these three variables (a criterion for high correlation coefficient can be referred in [132]). High correlation is also observed between AS’ number of customers and BA utility, and WH utility. The correlation coefficients are 0.95 and 0.5, respectively. We avoid using highly correlated variables together in the model. In addition to the correlation analysis, the choice of the variable is determined in the following section using the $t$-statistics of the estimated parameters.

4.6.2 Regression Analysis and Interpretation

In this section, we present the results of parameter estimation in the DCA models. Bayesian estimation technique is used to obtain the coefficient $\beta$ in Equation (3.6) such that the model’s predictions match the observed choices as closely as possible. In the following discussion, unless otherwise listed, the $t$-statistics discussed in this section are one-sided (i.e., only the absolute value of $t$-statistic is utilized), and the level of significance is set as 0.1, which corresponds the $t$-statistic of 1.64. It is assumed that the random effects are normally distributed, i.e., $\beta \sim N(\mu, \sigma^2)$. If the standard deviation of estimated parameter is statistically significant, then random effects are considered to exist.
Table 4.4. Correlation analysis on the variables for regression analysis.

<table>
<thead>
<tr>
<th></th>
<th>Distance</th>
<th>Address spaces</th>
<th># of customers</th>
<th># of providers</th>
<th>BA rank</th>
<th>RA rank</th>
<th>WH rank</th>
<th>BA utility</th>
<th>RA utility</th>
<th>WH utility</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distance</td>
<td>1</td>
<td>-0.17</td>
<td>-0.19</td>
<td>0.22</td>
<td>0.12</td>
<td>0.11</td>
<td>0.12</td>
<td>-0.16</td>
<td>-0.04</td>
<td>-0.1</td>
</tr>
<tr>
<td>Address spaces</td>
<td>-0.17</td>
<td>1</td>
<td>0.67</td>
<td>-0.29</td>
<td>-0.18</td>
<td>-0.26</td>
<td>-0.1</td>
<td>0.6</td>
<td>0.26</td>
<td>0.27</td>
</tr>
<tr>
<td># of customers</td>
<td>-0.19</td>
<td>0.67</td>
<td>1</td>
<td>-0.37</td>
<td>-0.22</td>
<td>-0.32</td>
<td>-0.17</td>
<td>0.95</td>
<td>0.20</td>
<td>0.50</td>
</tr>
<tr>
<td># of providers</td>
<td>0.22</td>
<td>-0.29</td>
<td>-0.37</td>
<td>1</td>
<td>0.12</td>
<td>0.15</td>
<td>0.09</td>
<td>-0.36</td>
<td>-0.1</td>
<td>-0.29</td>
</tr>
<tr>
<td>BA rank</td>
<td>0.12</td>
<td>-0.18</td>
<td>-0.22</td>
<td>0.12</td>
<td>1</td>
<td>0.26</td>
<td>0.14</td>
<td>-0.21</td>
<td>-0.15</td>
<td>-0.19</td>
</tr>
<tr>
<td>RA rank</td>
<td>0.11</td>
<td>-0.26</td>
<td>-0.32</td>
<td>0.15</td>
<td>0.26</td>
<td>1</td>
<td>0.31</td>
<td>-0.31</td>
<td>-0.30</td>
<td>-0.29</td>
</tr>
<tr>
<td>WH rank</td>
<td>0.12</td>
<td>-0.15</td>
<td>-0.17</td>
<td>0.09</td>
<td>0.14</td>
<td>1</td>
<td>-0.12</td>
<td>-0.16</td>
<td>-0.16</td>
<td>-0.32</td>
</tr>
<tr>
<td>BA utility</td>
<td>-0.16</td>
<td>0.6</td>
<td>0.95</td>
<td>-0.36</td>
<td>-0.21</td>
<td>-0.31</td>
<td>-0.12</td>
<td>1</td>
<td>0.14</td>
<td>0.52</td>
</tr>
<tr>
<td>RA utility</td>
<td>-0.04</td>
<td>0.26</td>
<td>0.20</td>
<td>-0.1</td>
<td>-0.15</td>
<td>-0.30</td>
<td>0.16</td>
<td>0.14</td>
<td>1</td>
<td>0.12</td>
</tr>
<tr>
<td>WH utility</td>
<td>-0.1</td>
<td>0.27</td>
<td>0.50</td>
<td>-0.29</td>
<td>-0.19</td>
<td>-0.29</td>
<td>-0.32</td>
<td>0.52</td>
<td>0.12</td>
<td>1</td>
</tr>
</tbody>
</table>
4.6.2.1 Model 1 for Linking Activity 1 – C2P Connections between New ASes and Existing ASes

The final model achieved for the AS linking decision behaviors between new ASes and existing ASes consists of four variables:

1. $\tilde{x}_1$ - geographic distance,
2. $\tilde{x}_2$ - normalized business access rank,
3. $\tilde{x}_3$ - normalized residential access rank, and
4. $\tilde{x}_4$ - number of customers, i.e., the in-degree.

$\tilde{x}_{ni}$ refers to alternative-specific non-network-metric attributes and $\tilde{x}_{ni}$ refers to alternative-specific network-metric attributes. See Section 3.4 for details. Considering a linear form of utility function, the utility of AS $n$ choosing alternative AS $i$ is:

$$U_{ni} = V_{ni} + \epsilon_{ni} = \beta_1 \tilde{x}_1 + \beta_2 \tilde{x}_2 + \beta_3 \tilde{x}_3 + \beta_4 \tilde{x}_4 + \epsilon_{ni}$$

(4.1)

where $\beta_1$, $\beta_2$, $\beta_3$ and $\beta_4$ are the parameters estimated from the datasets. If these parameters exhibit randomness, the standard deviation is also estimated. Using the parameter values, the decisions of new ASes can be determined using the probability function in Equation (3.6).

Table 4.5 shows the estimation results for the mixed logit model. First, the likelihood ratio test is performed to check whether the inclusion of random parameters significantly improves the model. As shown in Table 4.5, the log-likelihood without random effects is $-1588.3$, and the log-likelihood for the mixed logit model is $-1580.5$. The test statistic is $D = -2 \times (-1588.3 + 1580.5) = 15.6$. The log-likelihood of the random parameter model is smaller than that of the fixed parameter model indicating a better fit. The degree of freedom is the number of random parameters, so $dof = 1$. With the Chi-square distribution, the p-value is less than 0.001. The test indicates that the random effect is superior to the fixed parameter model. Furthermore, the
high $t$-statistic on the standard deviations of the parameter of the normalized business access rank, as shown in Table 4.5, is an indication that random effects for the parameter are warranted. Among the four variables listed above, it is observed that the geographic distance and the number of customers have the higher $t$-statistics, therefore, a stronger relationship with the dependent variable, i.e., the utility. In order to quantify the overall goodness of fit, the Estrella and adjusted Estrella [133] measures are adopted. The measure must take values in $[0,1]$, where 0 represents no fit and 1 corresponds to perfect fit. Table 4.5 shows the mixed logit model has an adjusted Estrella of 0.806. In the following, the specific insights obtained from each estimated parameter in the mixed logit model is discussed:

1. **Parameter $\beta_1$**: The $t$-statistic of $\beta_1$ is $-8.284$. The high value of $t$-statistic indicates that the geographic distance has a significant impact on how new ASes make decisions to choose ISPs. The negative sign of $\beta_1$ indicates that

Table 4.5. Mixed logit estimation results for Model 1 (Random parameters are normally distributed).

<table>
<thead>
<tr>
<th>Variable Description</th>
<th>Estimated Parameter</th>
<th>$t$ statistic</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Geographic distance</td>
<td>$-4.086 \times 10^{-4}$</td>
<td>$-8.284$</td>
<td>$&lt;0.001$</td>
</tr>
<tr>
<td>Normalized business access rank</td>
<td>$-8.366$</td>
<td>$-3.889$</td>
<td>$&lt;0.001$</td>
</tr>
<tr>
<td>(standard deviation of parameter distribution)</td>
<td>(6.627)</td>
<td>(4.545)</td>
<td>($&lt;0.001$)</td>
</tr>
<tr>
<td>Normalized residential access rank</td>
<td>$-0.714$</td>
<td>$-3.654$</td>
<td>$&lt;0.001$</td>
</tr>
<tr>
<td>Number of customers</td>
<td>$1.262 \times 10^{-3}$</td>
<td>$17.293$</td>
<td>$&lt;0.001$</td>
</tr>
<tr>
<td>Number of observations</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Log likelihood at zero</td>
<td>$-1871$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Log likelihood at convergence (without random effect)</td>
<td>$-1588.3$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Log likelihood at convergence (with random effect)</td>
<td>$-1580.5$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Likelihood ratio test statistics</td>
<td>$15.6$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Estrella</td>
<td>$0.812$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adjusted Estrella</td>
<td>$0.806$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
ASes prefer to link with ASes that are nearer. This may result from the fact
that acquiring long local loops for far away customers may incur a high cost
to the provider. Specifically, with a 1 km increase in the distance between the
decision-making AS and the target AS, the utility obtained by choosing that
AS would decrease by $-4.068 \times 10^{-4}$.

2. **Parameter $\beta_2$:** The estimation results show that $\beta_2$ exhibits random effects. The mean is $-8.366$, and the standard deviation is 6.627. The $t$-statistic of the standard deviation is 4.545 indicating that $\beta_2$ varies across the decision-makers. The $t$-statistic of the mean is $-3.889$, indicating high significance. Assuming normal distribution of the parameter, it is found that 89.66% of the estimated $\beta_2$ are less than zero and 10.34% are greater than zero, indicating heterogeneity among of new market entrants. For example, it is likely that only a fraction of new market entrants are focused on accessing Internet service instead of buying and reselling. In such cases, the ISPs with high business access utility may be less likely to be chosen. The BA rank is sorted according to the address space of AS if two ASes’ BA utilities are the same. Since the address space of AS captures the scale of an AS, the results indicate that on average more ASes incline to connect to large-scale ASes which can provide good quality of service and a reliable connection.

3. **Parameter $\beta_3$:** The $t$-statistic for $\beta_3$, corresponding to normalized residential access rank, is $-3.654$. This indicates its statistical significance at a level of 0.001, and the negative value highlights that ASes with lower normalized residential access rank are more likely to receive new links. This indicates that the higher the utility in providing residential Internet access, the greater the probability of receiving new links. This can be explained as follows: retail Internet businesses that directly deal with residential customers have existed since the inception of the commercial Internet. With advances in Internet access technologies, along with access speed, the number of residential users equipped with Internet access
has increased steadily. The increasing needs for accessing Internet from residential users call for the increased linking with ASes that could provide residential access service.

4. Parameter $\beta_4$: The $t$-statistic for $\beta_4$ is 17.293, which is the highest among the four parameters. It shows a strong regression relationship with the AS utility, and therefore the linking probability. The positive sign indicates that the greater the number of customers of an AS, the higher the probability of connection. The ASes with a large number of customers are often large ISPs. The observed result is intuitive because the new ASes get access to high-quality networks by connecting with large ISPs. This explains why degree-based models are effective in generating the AS Internet topology.

4.6.2.2 Model 2 for Linking Activity 2 – C2P Connections between Two Existing ASes

The model for AS decision-making behaviors of the C2P connections between two existing ASes is a mixed logit model that consists of five variables, as shown in Table 4.6. The high adjusted Estrella of 0.935 indicates a high overall goodness of fit. Different from the C2P model for the linking decision behavior between the newly entered ASes and existing ASes, it is observed that besides the four independent variables identified in Model 1, the number of providers (i.e., out-degree) significantly affects the decision behavior of existing ASes. In addition, it is found that the estimated parameter shows random effect. This indicates that while existing ASes choose ISPs, the preferences towards the number of providers of target ASes’ is heterogeneous. The mean is $-0.213$, and the standard deviation is $-0.517$. With the normal distribution, it indicates that 65.98% of the estimated $\beta_1$ are less than zero and 34.02% are greater than zero. On average, the lower the number of providers, the higher the probability of the ASes to be chosen as a provider. This is because in most cases, while existing ASes plan to setup a new C2P connection, their current
Table 4.6. Mixed logit estimation results for Model 2 (Random parameters are normally distributed).

<table>
<thead>
<tr>
<th>Variable Description</th>
<th>Estimated Parameter</th>
<th>t statistic</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Geographic distance</td>
<td>$-3.996 \times 10^{-4}$</td>
<td>$-12.482$</td>
<td>$&lt;0.001$</td>
</tr>
<tr>
<td>Normalized business access rank (standard deviation of parameter distribution)</td>
<td>$-11.467$</td>
<td>$-6.902$</td>
<td>$&lt;0.001$</td>
</tr>
<tr>
<td>Normalized residential access rank</td>
<td>$-0.853$</td>
<td>$-5.240$</td>
<td>$&lt;0.001$</td>
</tr>
<tr>
<td>Number of customers</td>
<td>$8.899 \times 10^{-4}$</td>
<td>$11.783$</td>
<td>$&lt;0.001$</td>
</tr>
<tr>
<td>Number of providers (standard deviation of parameter distribution)</td>
<td>$-0.213$</td>
<td>$-5.368$</td>
<td>$&lt;0.001$</td>
</tr>
<tr>
<td>Number of observations</td>
<td>914</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Log likelihood at zero</td>
<td>$-4942$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Log likelihood at convergence (without random effect)</td>
<td>$-4005$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Log likelihood at convergence (with random effect)</td>
<td>$-3829.8$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Likelihood ratio test statistics</td>
<td>350.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Estrella</td>
<td>0.937</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adjusted Estrella</td>
<td>0.935</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

providers no longer meet their requirements, and they seek to buy additional Internet service to enlarge their bandwidth and satisfy their increasing traffic volume. The ASes with less providers are often Tier-1 backbone carriers on the top layers of Internet hierarchy which only provide Internet service such that they are more likely to be chosen.

4.6.2.3 Model 3 for Linking Activity 3 - P2P Connections between Two Existing ASes

The final model achieved for the AS P2P linking decision behaviors is a mixed logit model that consists of two variables.

1. $\tilde{x}_1$ - geographic distance,
2. $\bar{x}_2$ - normalized business access rank.

With the linear form of utility function, the utility of AS $n$ choosing alternative AS $i$ is:

$$U_{ni} = V_{ni} + \epsilon_{ni} = \beta_1 \bar{x}_1 + \beta_2 \bar{x}_2 + \epsilon_{ni}$$  \hspace{1cm} (4.2)

where $\beta_1$ and $\beta_2$ are the parameters estimated from the datasets which quantify the AS preference to the target ASes’ geographic distance and normalized business access rank. Similarly, the likelihood ratio test is performed, and the test statistic of 14.04 indicates that the random effect is significantly superior to the fixed parameter model. The estimation results obtained are shown in Table 4.7. The results provide specific insights regarding the decision-making behaviors differences between C2P linking and P2P linking. In the following, these specific insights and the rationale for the parameters identified in the mixed logit model are discussed.

<table>
<thead>
<tr>
<th>Variable Description</th>
<th>Estimated Parameter</th>
<th>t statistic</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Geographic distance (standard deviation of parameter distribution)</td>
<td>$-2.096 \times 10^{-4}$</td>
<td>$-1.394$</td>
<td>$0.16$</td>
</tr>
<tr>
<td>Normalized business access rank (standard deviation of parameter distribution)</td>
<td>$-4.692$</td>
<td>$2.179$</td>
<td>$0.029$</td>
</tr>
<tr>
<td>Number of observations</td>
<td>73</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Log likelihood at zero</td>
<td>$-274.57$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Log likelihood at convergence (without random effect)</td>
<td>$-271.33$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Log likelihood at convergence (with random effect)</td>
<td>$-267.55$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Likelihood ratio test statistics</td>
<td>14.04</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Estrella</td>
<td>0.18</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adjusted Estrella</td>
<td>0.13</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
1. **Parameter $\beta_1$:** The mean is $-2.096 \times 10^{-4}$. However it is observed that the t-statistic of the mean is 0.16, indicating that on average, this variable is not significant. This is because the parameter of the geographic distance exhibits random effects. The random effect is validated by the high statistics for the standard deviation, which corresponds to the p-value at 0.002. Specifically, the standard deviation is $8.067 \times 10^{-4}$. The normal distribution shows that 59.87% of the estimated $\beta_1$ are less than zero, whereas 40.13% of the estimated $\beta_1$ are greater than zero. The positive effect offsets most of the negative effect, but on average the estimated parameter indicates that shorter distance between a pair of ASes increases the likelihood of the ASes to be selected. The random effect of the ASes’ preferences to the geographic distance reflects the heterogeneity of AS linking behaviors in selecting the target AS for P2P connection, and is the major difference between the P2P model from the C2P model. The heterogeneous preferences in geographic distance reflect two unique AS peering decision modes: *private peering* and *public peering* [134]. In private peering, two associated ASes establish a direct point-to-point circuit. In public peering, on the other hand, an AS co-locates itself with other ASes and connects to them at a public exchange point (IXP) via a shared medium or Ethernet. Private peering is favored by large ISPs or backbone operators, typically Tier-1 ISPs. Through the P2P connections, the main purpose of Tier-1 ISPs is to form a clique and all enjoy settlement-free access to all destinations to maintain the global connectivity and reachability. So, in private peering, even if two ISPs are far from each other, they may still prefer to set up a direct connection. On the other hand, in public peering, all co-located ASes connect to the same IXP. An IXP is treated similarly to a regular provider AS in that it provides connectivity and bandwidth to its co-located customer ASes. Unlike a provider AS, however, it does not provide transit service, but rather provides P2P based connectivity among its co-located members. Therefore, in this type of connections, ASes which need to setup P2P connections are all connected to the nearby IXPs.
Therefore, the two P2P models result in the random effect observed from the data, and is well captured by the proposed model.

2. Parameter $\beta_2$: The $t$-statistic for $\beta_2$ is $-2.179$. This indicates its statistical significance at a level of 0.03, and the negative value highlights that on average, the higher the utility of AS in providing business Internet access, the greater the probability of receiving new links. This can be explained as follows: no matter which types of peering, private peering or public peering, the chosen ASes are either large ISPs or IXPs. One common characteristic for ASes in these two categories is that they all have a large number of customers. The large number of customers results in a large business access utility of AS as indicated by the algorithm in [129]. Therefore, the higher the ranking in business access utility, the larger the probability of ASes being chosen for connecting. The major difference in P2P model is that the number of customers (in-degree) and the number of providers (out-degree) are not significant. This finding gives us new insights on refining the degree-based model for Internet topology generator by modeling the C2P and P2P connections separately. Specifically, it is better to not introduce degree in the decision model of P2P connection directly, but in terms of business access utility.

4.6.3 Model Consistency

The proposed approach is applied to five consecutive AS Internet evolution instances from Jan. 5th to Nov. 1st, 2004 with a two-month interval. The aim is to check whether the obtained models consistently capture the AS decision-making behaviors over time. Table 4.8 lists the three groups of estimated results of each model for the five evolution instances.
Table 4.8. Mixed logit estimation results for five consecutive evolution instances from Jan. 5th to Nov. 1st, 2004.

<table>
<thead>
<tr>
<th>Var. Desc.</th>
<th>Model 1</th>
<th></th>
<th>Model 2</th>
<th></th>
<th>Model 3</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Evolution 1</td>
<td>Evolution 2</td>
<td>Evolution 3</td>
<td>Evolution 4</td>
<td>Evolution 5</td>
<td>Evolution 4</td>
</tr>
<tr>
<td>Geo. distance</td>
<td>$-4.1 \times 10^{-4}$ $-8.3$</td>
<td>$-3.4 \times 10^{-4}$ $-8.5$</td>
<td>$-3.5 \times 10^{-4}$ $-7.9$</td>
<td>$-4.1 \times 10^{-4}$ $-9.7$</td>
<td>$-2.7 \times 10^{-4}$ $-6.1$</td>
<td></td>
</tr>
<tr>
<td>Norma. BA rank</td>
<td>$-8.4$ $-3.9$</td>
<td>$-2.7$ $-1.8$</td>
<td>$-3.69$ $-1.7$</td>
<td>$-2.4$ $-1.2$</td>
<td>$-10.1$ $-3.9$</td>
<td></td>
</tr>
<tr>
<td>(stdev of para.)</td>
<td>(6.6) (4.5)</td>
<td>(2.2) (1.3)</td>
<td>(3.1) (1.3)</td>
<td>(2.9) (1.4)</td>
<td>(6.8) (3.8)</td>
<td></td>
</tr>
<tr>
<td>Norm. RA rank</td>
<td>$-0.7$ $-3.7$</td>
<td>$-1.1$ $-5.9$</td>
<td>$-0.5$ $-2.3$</td>
<td>$-0.9$ $-4.9$</td>
<td>$-0.9$ $-4.1$</td>
<td></td>
</tr>
<tr>
<td>No. of customers</td>
<td>$1.3 \times 10^{-3}$ 17.3</td>
<td>$1.3 \times 10^{-3}$ 19.7</td>
<td>$1.3 \times 10^{-3}$ 19.1</td>
<td>$1.5 \times 10^{-3}$ 23.3</td>
<td>$1.3 \times 10^{-3}$ 19.3</td>
<td></td>
</tr>
<tr>
<td>Geo. distance</td>
<td>$-4.0 \times 10^{-4}$ $-12.5$</td>
<td>$-4.2 \times 10^{-4}$ $-14.6$</td>
<td>$-2.9 \times 10^{-4}$ $-8.9$</td>
<td>$-2.7 \times 10^{-4}$ $-9.0$</td>
<td>$-3.6 \times 10^{-4}$ $-10.7$</td>
<td></td>
</tr>
<tr>
<td>Norma. BA rank</td>
<td>$-11.5$ $-6.9$</td>
<td>$-29.0$ $-10.3$</td>
<td>$-12.5$ $-7.6$</td>
<td>$-28.8$ $-9.6$</td>
<td>$-29.8$ $-10.1$</td>
<td></td>
</tr>
<tr>
<td>(stdev of para.)</td>
<td>(8.9) (7.6)</td>
<td>(18.6) (8.5)</td>
<td>(8.5) (7.2)</td>
<td>(18.2) (8.0)</td>
<td>(19.0) (7.6)</td>
<td></td>
</tr>
<tr>
<td>Norm. RA rank</td>
<td>$-0.9$ $-5.2$</td>
<td>$-0.8$ $-5.6$</td>
<td>$-0.3$ $-1.8$</td>
<td>$-1.2$ $-7.9$</td>
<td>$-0.8$ $-5.5$</td>
<td></td>
</tr>
<tr>
<td>No. of customers</td>
<td>$8.9 \times 10^{-4}$ 11.8</td>
<td>$12.0 \times 10^{-4}$ 16.4</td>
<td>$8.6 \times 10^{-4}$ 12.0</td>
<td>$7.9 \times 10^{-4}$ 11.6</td>
<td>$4.7 \times 10^{-4}$ 7.2</td>
<td></td>
</tr>
<tr>
<td>No. of providers</td>
<td>$-0.2$ $-5.4$</td>
<td>$0.1$ $6.7$</td>
<td>$-0.4$ $-8.9$</td>
<td>$-0.2$ $-5.0$</td>
<td>$-0.3$ $-10.0$</td>
<td></td>
</tr>
<tr>
<td>(stdev of para.)</td>
<td>(0.5) (12.2)</td>
<td>(0.04) (0.3)</td>
<td>(0.4) (8.7)</td>
<td>(0.1) (3.3)</td>
<td>(0.3) (8.0)</td>
<td></td>
</tr>
<tr>
<td>Geo. distance</td>
<td>$-2.1 \times 10^{-4}$ $-1.4$</td>
<td>$-2.0 \times 10^{-4}$ $-1.6$</td>
<td>$-3.4 \times 10^{-4}$ $-2.0$</td>
<td>$-2.7 \times 10^{-4}$ $-3.2$</td>
<td>$-2.4 \times 10^{-4}$ $-1.1$</td>
<td></td>
</tr>
<tr>
<td>(stdev of para.)</td>
<td>$(8.1 \times 10^{-4})$ (3.1)</td>
<td>$(9.4 \times 10^{-4})$ (3.4)</td>
<td>$(9.8 \times 10^{-4})$ (3.0)</td>
<td>$(5.0 \times 10^{-4})$ (2.3)</td>
<td>$(10.9 \times 10^{-4})$ (2.0)</td>
<td></td>
</tr>
<tr>
<td>Norm. BA rank</td>
<td>$-4.7$ $-2.2$</td>
<td>$-2.3$ $-2.2$</td>
<td>$-2.8$ $-1.8$</td>
<td>$-7.6$ $-4.7$</td>
<td>$-46.9$ $-3.4$</td>
<td></td>
</tr>
<tr>
<td>(stdev of para.)</td>
<td>(4.7) (2.2)</td>
<td>(2.2) (1.4)</td>
<td>(3.1) (1.6)</td>
<td>(6.8) (4.5)</td>
<td>(35.2) (2.9)</td>
<td></td>
</tr>
</tbody>
</table>
In Model 1 (i.e., the decision model for C2P connections between new ASes and existing ASes), the four estimated parameters remain significant for all the five evolution steps. The number of customers (in-degree) consistently has the highest $t$-statistic indicating its high significance in formulating the ASes’ utility function. However, the estimation results for evolution 2, 3 and 4 shows that $t$-statistic of the standard deviation of the parameter for normalized business access rank is less than 1.64, indicating the random effect of this variable is not significant in these evolution instances. In Evolution 4, the results show that $\beta_2$ is neither significant nor has a random effect. Therefore, the normalized business access rank is not consistently significant as compared with other three variables in Model 1. Similarly, the number of providers in Model 2 (i.e., the decision model for C2P connections between two existing ASes) shows the variability of significance in Evolution 2 as well. Therefore, a conservative way to use the models is to only include the variables that are stable. On the other hand, to use the models for prediction or for developing Internet topology generators, it may be better to run more evolution instances and to use the average value of all estimated parameters to offset the variability.

4.7 Closing Comments for Chapter 4

In summary, the work presented in this chapter answers the RQ3. The main task accomplished in this chapter is to solve the estimation problem for complex networked systems. The approaches developed in this chapter could benefit complex systems analysis and design in three ways. First, the approaches are general enough to be used in other complex networked systems where network structure datasets are available at each time step in the evolution. Second, an understanding of the underlying decisions in complex networked systems helps to create better models to abstract real-world systems. Third, understanding node-level behaviors in complex networked systems helps guide the evolution of systems. The models help predict how systems evolve,
and how future structure and performance of systems can be directed by influencing node-level decision-making preferences.

Compared to existing models for Internet topology, the model presented in this chapter is different in three aspects. First, the proposed model enables the estimation of unobserved payoffs in terms of utility functions using historical network structural data. The resulting model is thus data-driven rather than hypothesized. This process allows us to identify the intrinsic factors that affect ASes’ decisions. Second, the model relies on random utility theory, resulting in a richer interpretation. Third, the model establishes a connection between AS decisions and preferences parameterized by the model coefficients. Moreover, the model accounts for uncertainty in decision-making process and the heterogeneity of ASes’ preferences.

The resulting statistics, e.g., the adjusted Estrella, shows the overall fit of the model. However, the adjusted Estrella for the model of P2P connections between two existing ASes is only 0.13 as too few significant variables can be identified from the available dataset. This calls for the future work of collecting data related to other aspects, such as transit price, measurement of service quality and number of value-added services. However, as the approach presented, the work is straightforward once those data are available. The inclusion of other significant variables would also help in improving the accuracy of estimation. Furthermore, the possibility of specification errors may exist in the models. Specification errors may arise from the endogeneity. For example, the normalized residential rank and the normalized business rank may be endogenous in nature. Some ASes who provide large business access may also provide residential access service.
CHAPTER 5. THREE APPROACHES TO ESTIMATE LOCAL DECISION-MAKING PREFERENCES - A COMPARATIVE STUDY

5.1 Chapter Overview

In this chapter, the objective is still to answer RQ3: how can nodes’ unobserved preferences and behaviors be estimated? However, the work presented here has a focus different from the preceding chapters. The research presented in this chapter is focused on exploring other estimation approaches for the node-level preferences in complex networked systems, as then compared to the proposed decision-centric estimation approach in Chapter 4.

In this comparative study, two additional estimation approaches are proposed. The first approach is based on hypothesized node-level behaviors. Specifically, we adopt the generalized preferential attachment (GPA) [15] as the hypothesized node-level behaviors model. Instead of using a statistical method, the parameters of this model can be estimated based on the functional relationship between the network structure and the degree-based linking probability, derived from the continuum theory presented by Albert et al. [83]. In the second approach, node-level behaviors are derived by analyzing how networks change between two consecutive instances. The nodes and links created (or deleted) between the two instances are extracted, and the node-level behavior is deduced by statistical regression. In this study, a new approach based on a linear regression model is developed. These two approaches are compared with the proposed decision-centric estimation approach in which nodes are modeled as decision-making agents with behaviors motivated by maximizing utility.

The rationale for choosing GPA is that it is one of the widely used approaches for modeling complex networks. GPA has also been used for modeling Internet topology and its properties. Statistical regression techniques are chosen for comparison pur-
poses because they are widely used within design literature for data-driven modeling in complex systems. The comparative study is performed in the context of the AS-level Internet. To ensure that the comparison among the models is consistent, we use the same input data in all approaches. In addition, rather than considering all the possible decision-making factors, we use degree as the only variable for the three models. In this study, the main purpose is not acquiring the model with best predicting performance, but instead evaluating the decision-centric estimation approach and gaining insight on decision-centric modeling. Figure 5.1 shows how the proposed approaches which answer RQ3 in the context of complex networked systems.

Figure 5.1. Overview of Chapter 5: research questions, approaches and tasks.

The outline of this chapter is as follows. An overview of the three approaches is presented in Sections 5.2, 5.3 and 5.4. In Section 5.5, the three approaches are applied to the AS-level Internet to estimate the ASes’ peering preferences. As a result, three different node-level degree-based behavioral models are obtained. These models are then used to construct three network generation models to regenerate synthetic Internet topology in Section 5.6. The three models are compared using the networks’ structural metrics, including average path length and cluster coefficient.
This comparison evaluates the pros and cons of each model. Finally, closing comments are made in Section 5.7.

5.2 Approach 1: Generalized Preferential Attachment Model

5.2.1 Overview

The preferential attachment model for complex networks was initially proposed by Barabasi and Albert [15]. In this model, a new node preferentially links to existing nodes based on certain characteristics of the target node. Network evolution in this approach is assumed to follow two mechanisms: growth and preferential attachment [17]. 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 [15], the probability of link creation between a new node and an existing node is linearly proportional to the degree of an existing node.

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 [104], the World Wide Web [135], and networks of metabolic reactions [136]. Existing literature [82] has shown that the degree-based preferential attachment mechanism has better performance in 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.

In the GPA model, the affinity of a node to link with an existing target node \( j \) at time \( t \) is modeled as:

\[
V_j(t) = G_j(t)d_j^\tau(t) + A_j(t) \tag{5.1}
\]

where, the \( V, G, A \) and \( d \) are functions of the node \( j \) and time \( t \). \( G_j(t) \) is the fitness value of node \( j \) at time \( t \), and \( A_j(t) \) is the additional attractiveness of node \( j \) at time \( t \). \( d_j(t) \) stands for the degree of node \( j \) at time \( t \), which is the number of neighbors of
node $j$. Using Equation (5.1), the probability of an arbitrary node $j$ getting chosen for connection among $J$ nodes is equal to:

$$P_j = \frac{V_j(t)}{\sum_{i=1}^{J} V_i(t)} = \frac{G_j(t)}{\sum_{i=1}^{J} V_i(t)} d_j(t) + \frac{A_j(t)}{\sum_{i=1}^{J} V_i(t)}$$  (5.2)

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 (5.1) is a linear function of the node degree, i.e., $\tau = 1$. Therefore, Equation (5.1) can be modeled as:

$$V_j(t) = d_j(t) + A_j$$  (5.3)

These assumptions are made to enable direct comparison with the other two approaches. By relaxing these assumptions, detailed models with more parameters can be generated. For example, if a directed network is used, then two separate probabilities are needed to model the creation of incoming and outgoing links. If different fitness values are used for different nodes, an additional parameter is added for each node, which increases the data requirements for parameter estimation. Similarly, considering time varying additional attractiveness also adds additional parameters in the model. Hence, for the purpose of this comparative study, we decided to limit the number of parameters. In the future studies, we will investigate the effects of relaxing these assumptions.

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:

$$P_j = \frac{V_j(t)}{\sum_{i=1}^{J} V_i(t)} = \frac{Gd_j + A_j}{\sum_{i=1}^{J} (Gd_i + A_i)} = \frac{d_j + \frac{A_j}{G}}{\sum_{i=1}^{J} (d_i + \frac{A_i}{G})}$$  (5.4)
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 [92], 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 function 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. [83] 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 (5.3), the changing rate of a node \(j\)'s degree \(d_j\) is given by:

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

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,

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

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

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

where,

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

Equation (5.8) indicates that different degree distributions are associated with different \(A\) values. Hence, the \(A\) value can be used to differentiate the network structures. The degree distributions generated for representative values of \(A\) are illustrated in Figure 5.2. 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 (5.4)). In Section 5.2.2, we introduce the techniques used for fitting the power law degree distribution.

5.2.2 Fitting Power-Law Degree Distribution

A simple approach for fitting the power law degree distribution is the ordinary least square (OLS) regression [137]. The power law distribution in Equation (5.7) 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 to plot such frequency on the double logarithmic axis. Then, a linear model:

$$y = \beta_0 + \beta_1 x + \epsilon$$  \hspace{1cm} (5.9)

can be used where $\beta_0$ is the intercept, $\beta_1$ is the slope and $\epsilon$ is the random error in the observation. The OLS regression can be utilized to fit the power law degree distribution with variable $x$ equals to $\ln(d)$ and observation value $y$ equals to $\ln(P)$. The estimation on $\beta_1$ corresponds to the $-\gamma$ in the power distribution in Equation (5.7).

![Figure 5.2. Complementary cumulative degree distribution of networks generated by generalized BA model with different $A$ values.](image-url)
In practice, the power law often applies only for values greater than some minimum value $x_{min}$. 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 $x_{min}$. To address this issue Clauset et al. [138] 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 [139]. The key idea for estimating the exponent, $\gamma$, correctly is to first identify the lower bound $x_{min}$ of power law behavior in the data. Hence, the parameter $x_{min}$ is first chosen, and then $\gamma$ of the power law is fit using maximum likelihood estimation. Then, with the estimated $x_{min}$ 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 $x_{min}$ are compared via a likelihood ratio test [140]. 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 (5.8). 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 Section 5.5.1.

### 5.3 Approach 2: Statistical Regression 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:

$$P(S_d) = \frac{l_d}{L} \quad (5.10)$$

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:

$$P(d) = \frac{1}{n_d} P(S_d) \quad (5.11)$$

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 5.5.2.
5.4 Approach 3: Discrete Choice Model

If the complex network evolution is based on node-level decision-making process, the principles from discrete choice theory introduced in the Section 3.2 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 who 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 observed utility is:

$$V_j = \beta_{0j} + \beta_{1j}d_j$$  \hfill (5.12)

The resulting probability of the node alternative $j$ that gets connection from the decision-making node $n$ is given by Equation (3.5). Note that this corresponds to Equation (5.1) in which $\beta_{0j}$ stands for the additional attractiveness and $\beta_{1j}$ is the node fitness. But the resulting probability is fundamentally different from Equation (5.4).

It has been shown that the corresponding log-likelihood function of the choice probability model in Equation (3.5) is globally concave [78], the parameter vector $\beta$ is estimated using maximum likelihood estimation in this study. To implement estimation, observation data about the decision-makers nodes and the alternatives needs to be extracted. From a researchers’ point of view, the decision-makers are the newly added nodes who choose a target node to link to. The alternatives are the existing nodes during the previous time step. Thus, two consecutive network datasets are used to extract the evolutionary data for estimation, see Figure 5.3.

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 [16], and betweenness centrality [141] 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 5.2 and 5.3.

As a summary, the proposed three approaches are presented in the Table 5.1

5.5 Estimating AS-Level Behaviors in the Internet Using Three Approaches

In this section, the approaches discussed are applied to the autonomous system (AS) level Internet network. While the approaches are applicable to a variety of
Table 5.1. An overview of three approaches presented in this chapter.

<table>
<thead>
<tr>
<th>Inputs</th>
<th>Generalized preferential attachment (Section 5.2)</th>
<th>Statistical Regression-based approach (Section 5.3)</th>
<th>Multinomial-logit choice model (Section 5.4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Outputs</td>
<td>Single Instance of Network</td>
<td>Consecutive instances of the network</td>
<td>Consecutive instances of the model</td>
</tr>
<tr>
<td>Outputs</td>
<td>Linking probability</td>
<td>Linking probability</td>
<td>Linking probability and node’s utility</td>
</tr>
<tr>
<td>Approach</td>
<td>Continuum theory</td>
<td>Regression</td>
<td>Discrete choice</td>
</tr>
<tr>
<td>Behavior model</td>
<td>$P_j(t) = \frac{G_j(t)}{\sum_{i=1}^{n} V_i(t)} d_j(t) + \frac{A_j(t)}{\sum_{i=1}^{n} V_i(t)}$</td>
<td>Best fitting model.</td>
<td>$P(j</td>
</tr>
<tr>
<td>Parameters in the model</td>
<td>$G$: Node’s fitness; $A$: node’s additional attractiveness</td>
<td>$\alpha$: coefficient; $\beta$: exponent</td>
<td>Utility function: $V_j = \beta_0 + \beta_1 d_j$</td>
</tr>
<tr>
<td>Estimation technique</td>
<td>Function mapping</td>
<td>Ordinary least square</td>
<td>Maximum likelihood</td>
</tr>
</tbody>
</table>

complex networked systems, the Internet is chosen as an example because of the availability of data. 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. The Internet is an ideal example of a complex evolving system that has emerged based on the decisions made by independent decision making entities. Estimation of local decision-making behaviors of the entities is important for understanding how the structure of the network evolves. The knowledge of the local behaviors can help in providing incentives to the autonomous systems to direct their linking behavior towards structures with desired performance characteristics such as robustness and resilience.
The dataset 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 ASes. In the following sections, we present the results from the three approaches, starting with the Generalized Preferential Attachment (GPA) approach.

5.5.1 Results of Approach 1: Generalized Preferential Attachment

The first step in this approach is to develop a fit for the degree distribution of the network. Figure 5.4 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 [93] 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 (5.9)).

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 5.5. It is observed that this exponent $\gamma$ increases with the network size.

Based on Equation (5.8), the additional attractiveness ($A$) in the node-level behavior model can be evaluated using the exponents ($\gamma$) 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 5.6. 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, $A$, can be calculated using $m$ and $\gamma$ based on Equation (5.8). We obtain that the $A$-value increases from -1.78 to -1.73. One-tail test
Figure 5.4. Complementary cumulative degree distribution of AS-level Internet on Jan 5th, 2004.

Figure 5.5. Exponent ($\gamma$) in the degree distribution vs. network size.

on the slope of the fitting function for the parameter $\gamma$ (i.e., $S_\gamma$) is performed. The t-statistic corresponding to $\{H_0 : S_\gamma = 0$ vs. $H_1 : S_\gamma > 0\}$ is 21.34, resulting in the p-value < 0.001. Hence, we claim that the slope of $\gamma_1$ is statistically significant. 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 at-
tractiveness on the linking behavior is discussed in detail by Sha and Panchal [92]. As $A$ increases, more nodes have the opportunity to be connected.

We also used the approach suggested by Clauset et al. [140] (discussed in Section 5.2.2) to fit the degree distribution using the maximum likelihood estimator. The resulting exponents for the 122 networks are shown in Figure 5.7. By performing
the t-test on the slope in the figure, the p-value corresponding to \( \{ H_0 : S_\gamma = 0 \text{ vs. } H_1 : S_\gamma \neq 0 \} \), 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 5.7 are also greater than those in Figure 5.5. This can be explained as follows. The main difference in this method is that a minimum bound value \( x_{\text{min}} \) is estimated beyond which the fit is close to power law, and the “tail” of the distribution with values of degree lower than \( x_{\text{min}} \) 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 fit 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.

5.5.2 Results of Approach 2: Statistical Regression

In this section, we utilize the approach presented in Section 5.4 to the AS-level. Figure 5.8 shows the node-level linking behavior (i.e., probability of new node linking to an existing node) in three pairs of consecutive network snapshots:

- a) Jan. 5\(^{th}\), 2004 (N1) - Feb. 2\(^{nd}\), 2004 (N2),
- b) Aug. 28\(^{th}\), 2006 (N59) - Sep. 4\(^{th}\), 2006 (N60), and
- c) Nov. 5\(^{th}\), 2007 (N120) - Nov. 12\(^{th}\), 2007 (N121).

The plot is shown on a log-log scale. We use the degrees of existing nodes based on the previous snapshot of the network structure. 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. Thus, the linking probability of the node \( j \) is:

\[
P_j = \alpha d_j^\beta
\] (5.13)

The parameters \( \alpha \) and \( \beta \) are estimated using OLS regression on \( ln(d) \) vs \( ln(P) \). Furthermore, as shown in the figure, the parameters of the three fitting functions are
close to each other, which indicates 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 Figure 5.5). However, the result is in agreement with the fit using the maximum likelihood estimation (see Figure 5.7).

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 $\beta$). Figures 5.9(a) and 5.9(b) show the two parameters

![Figure 5.8](image1.png)

**Figure 5.8.** Node-level behavior for three network evolution steps.

![Figure 5.9](image2.png)

(a) Network size vs. $\alpha$  
(b) Network size vs. $\beta$

**Figure 5.9.** Network size vs. parameters in the node’s decision model 2.
for the 121 timesteps. We performed two separate hypothesis tests to detect whether there have been any statistical significant changes in $\alpha$ and $\beta$ over the 121 evolutions. The p-value corresponding to $\{H^1_0 : S_\alpha = 0 \text{ vs. } H^1_1 : S_\alpha \neq 0\}$ and $\{H^2_0 : S_\beta = 0 \text{ vs. } H^2_1 : S_\beta \neq 0\}$ are 0.03 and 0.04 respectively. Here, $S_\alpha$ and $S_\beta$ are the slopes corresponding to parameters $\alpha$ and $\beta$ in Figures 5.9(a) and 5.9(b) respectively. Hence, we claim that there has been no statistically 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 is consistent during 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 $(d)$ using Equation (5.13).

5.5.3 Results of 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 (3.5). 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 the 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,

\[ V_j = \beta_1 \ln(d_j) + \beta_2 \ln(n_j) \]  \hspace{1cm} (5.14)

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 \( \beta_2 \) denote the preferences of the decision-making nodes on degree and group size. Thus the utility function based on Equation (3.1) is:

\[ u_j = \beta_1 \ln(d_j) + \beta_2 \ln(n_j) + \varepsilon_j \]  \hspace{1cm} (5.15)

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

\[ P_n(j|C_J) = \frac{d_j^{\beta_1} n_j^{\beta_2}}{\sum_{i=1}^{J} d_i^{\beta_1} n_i^{\beta_2}} \]  \hspace{1cm} (5.16)

Figure 5.10. Network size vs. parameters in the node’s decision model 3.

(a) Network size vs. parameter \( \beta_1 \)  \hspace{1cm} (b) Network size vs. parameter \( \beta_2 \)

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 (5.15) 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 5.10(a) and 5.10(b). It is observed in Figure 5.10(a) that the parameter ($\beta_1$) has an average value of 0.672 for network size smaller than 21000 nodes (corresponding to the Internet network on 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. The p-values corresponding to $\{H_0^1 : S_{\beta_1} = 0$ vs. $H_1^1 : S_{\beta_1} \neq 0\}$ for network size, $J \leq 21000$ and $J > 21000$ are 0.03 and 0.11 respectively. We cannot reject the null hypothesis at a 2% level of significance. Hence, we claim that there has been no statistically significant change in slopes of $\beta_1$ within ranges $J \leq 21000$ and $J > 21000$. The parameter ($\beta_2$) follows a similar trend. The average value of $\beta_2$ for $J \leq 21000$ is 0.661 (p-value = 0.12) and for $J > 21000$, the average value of $\beta_2$ is 0.525 (p-value = 0.06).

5.6 Comparison of Three Node-level Behavioral Models

In this section, we show a comparison on the node-level behaviors obtained by the three approaches and perform an evaluation on the generated network structures for a given time. We also discuss the generality, extensibility and computational capability of each approach.

5.6.1 Node-level Behaviors and Generated Networks

Figure 5.11 shows an example of the node-level linking behavior of the AS-level Internet network on Jan. 5th, 2004, estimated using three different approaches. In Approach 1, the probability that a node is chosen is determined by Equation (5.4) with estimated additional attractiveness ($A$). In Approach 2, the node-level behavior is described by Equation (5.13), 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 (5.16) and parameters $\beta_1$ and $\beta_2$ are estimated using the multinomial logit choice models. A comparison of the estimated linking behaviors is shown in Table 5.2.
Table 5.2. Comparison of linking probability.

<table>
<thead>
<tr>
<th>Item</th>
<th>Approaches</th>
<th>Node-level Behavior Models</th>
</tr>
</thead>
<tbody>
<tr>
<td>Approach 1</td>
<td>DC Model</td>
<td>$P_n(i</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\beta_1 = 0.672$ and $\beta_2 = 0.661$ before Jan. 2$^{nd}$, 2006. $\beta_1 = 0.428$ and $\beta_2 = 0.525$ afterwards.</td>
</tr>
<tr>
<td>Approach 2</td>
<td>GPA Model: A increasing</td>
<td>$P_i = \frac{d_i + A}{\sum_{j=1}^{J} (d_j + A)}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>where $A$ changes over time according to $5 \times 10^{-6} J - 1.86$, and $J$ is the number of nodes in the network at time $t$</td>
</tr>
<tr>
<td>Approach 2</td>
<td>GPA Model: A unchanged</td>
<td>$P_i = \frac{d_i - 0.96}{\sum_{j=1}^{J} (d_j - 0.96)}$</td>
</tr>
<tr>
<td>Approach 3</td>
<td>Regression-based Model</td>
<td>$P_i = 1.97 \times 10^{-5} d_i^{0.959}$</td>
</tr>
</tbody>
</table>

Based on the node-level behavior models derived by using the three approaches, the network topology of Internet on Nov. 12$^{th}$, 2007 is simulated using the real network topology on Jan. 5$^{th}$, 2004 as the initial network. Since the $A$ value (additional attractiveness) in Approach 1 has different trends using OLS estimation and maximum likelihood estimation, we use the $A$-value from both methods and compare the resulting networks also. Figure 5.12 shows the degree distribution of the simulated
network structure based on different node-level behaviors. It is observed that all the four degree distribution functions are close to each other. To quantify the differences between the four simulated degree distributions from the original network, the Kullback-Leibler (KL) divergence [142] measure is adopted. The KL divergence is non-negative and zero if the distributions match exactly. It is the expectation of the logarithmic difference between two probability distributions. The larger the value, the less likely it is that the two distributions are the same. The KL divergence is calculated based on the probability mass instead of the cumulative distribution. Figure 5.13 shows the logarithmic difference between the simulated distribution and the true distribution at each degree point. The KL divergence values are as follows:

1. GPA with OLS approach: 0.098
2. GPA with maximum-likelihood approach: 0.192
3. Statistical regression-based approach: 0.141
4. DCM based approach: 0.271
Figure 5.13. KL divergence on the degree distribution of simulated networks.

The results show that the degree distribution of the network, which is generated with the node behavior model estimated by Approach 1 with GPA model and OLS estimation, is more likely to match the degree distribution of the true Internet AS network compared with other three approaches. However, to compare and evaluate the three approaches, we also compare other commonly used network measures, shown in the Table 5.3, to evaluate the differences among the generated networks. As shown in Table 5.4, the number of nodes and edges added during each step are the same for all the simulated networks because the network formation process is the same and the difference is in the linking probability only. The average path lengths (APL) of the generated networks are close to the true value. Specifically, the APL in Approach 1 with OLS estimation is slightly lower than the true value, but the APL values with other approaches are slightly higher than the true network. We observe that the clustering coefficients of all the simulated networks are less than that of the true network. This can be explained as follows. High clustering coefficient in the AS-level Internet network results from a large number of peer-to-peer relationships between ASes. However, in the approaches used in this study, the peer-to-peer mechanism is not explicitly included. We also observe a significant difference in the diameters of the true network and simulated networks. The diameters of the simulated networks
Table 5.3. Network metrics for evaluation.

<table>
<thead>
<tr>
<th>Metrics</th>
<th>Relationship to Internet</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average Path Length (APL)</td>
<td>Related to routing efficiency</td>
</tr>
<tr>
<td>Cluster Coefficient</td>
<td>Related to peering structure and route resilience</td>
</tr>
<tr>
<td>Diameter</td>
<td>Related to the span of Internet</td>
</tr>
</tbody>
</table>

Table 5.4. Comparison of metrics in simulated networks.

<table>
<thead>
<tr>
<th>Networks</th>
<th>Number of Nodes</th>
<th>Number of Edges</th>
<th>Cluster Coefficient</th>
<th>APL</th>
<th>Diameter</th>
</tr>
</thead>
<tbody>
<tr>
<td>True Network</td>
<td>26475</td>
<td>53381</td>
<td>0.208</td>
<td>3.876</td>
<td>17</td>
</tr>
<tr>
<td>Approach 1</td>
<td>26475</td>
<td>53303</td>
<td>0.104</td>
<td>4.043</td>
<td>10</td>
</tr>
<tr>
<td>Approach 2 with OLS</td>
<td>26475</td>
<td>53303</td>
<td>0.199</td>
<td>3.649</td>
<td>10</td>
</tr>
<tr>
<td>Approach 2 with Maximum likelihood</td>
<td>26475</td>
<td>53303</td>
<td>0.108</td>
<td>3.973</td>
<td>9</td>
</tr>
<tr>
<td>Approach 3</td>
<td>26475</td>
<td>53303</td>
<td>0.118</td>
<td>3.903</td>
<td>10</td>
</tr>
</tbody>
</table>

are around 10, whereas the diameter of the true network is 17. The potential reason for this difference in the diameter is that the models in the three approaches do not account for the geographic aspects. In the real world, when a new customer AS joins the network, it prefers to purchase service from a nearby provider in order to minimize the linking and routing costs. As a summary, some differences are observed between the real Internet network and the generated networks. These differences are due to a) the estimation process itself, and b) assumptions made about the network formation process for the specific case study.

5.6.2 Generality and Extensibility

Out of the three approaches, the first approach has the advantage of being simple and easy to evaluate because of the direct function mapping between degree distribution and node-level attributes, $A$, i.e., the node’s additional attractiveness. However,
it is based on an assumed behavior model of preferential attachment, and can be applied only if the network degree distribution follows the power-law form. This is the major limitation for the first approach. 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 has the advantage of providing an explanatory framework for the relationship between the node-level preferences, node-level behaviors and the network structure. This is a fundamental aspect that the other two approaches fail to address. Additionally, it provides a framework to integrate existing decision-centric models, such as ABM and network formation game models, with the available network structure datasets. As we discussed before, the DCM proposed in this study does not account for other attributes of ASes, such as economic, traffic, geographic attributes. Existing ABMs for the Internet have included these attributes. Thus, with the DCM, a more realistic model for reconstructing the Internet topology can be established if information about these attributes is available.

5.6.3 Computational Capability

A barrier in implementing the decision-centric 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. Therefore, if the network size is small (e.g. hundreds or thousands of nodes) and explanations are needed for interpreting the nodes’ behaviors, the third approach is a better choice, otherwise Approaches 1 and 2 are better.
Table 5.5. Comparison of the three estimation approaches.

<table>
<thead>
<tr>
<th>Item</th>
<th>Approach based on DCM</th>
<th>Approach based on GPA model</th>
<th>Approach based on Regression model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Advantages</td>
<td>An explanatory framework is provided</td>
<td>Simple and easy to evaluate</td>
<td>No presumed model on the node-level behavior</td>
</tr>
<tr>
<td>Limitations</td>
<td>Computation expensive</td>
<td>Hypothesized behaviors - preferential attachment; Network structure needs following power-law</td>
<td>No rationale for the node-level behavior model</td>
</tr>
</tbody>
</table>

5.7 Closing Comments for Chapter 5

To conclude, the research presented in this chapter is an extension of Chapter 4’s answer to RQ3. Beyond the decision-centric approaches introduced in the previous chapter, two additional estimation approaches based on a regression model and GPA model are developed. The developed approaches are then applied to estimate the ASes’ behaviors. The topology of Internet is then regenerated with these estimated ASes' behaviors. Future work should focused on relaxing some of the assumptions made in the models. For example, we only consider the addition of nodes in the network evolution. However, in reality, removal of nodes and link re-direction also occur during network evolution. Thus, consideration of decisions to remove nodes and redirect links is essential for further improvement of the models. Like any other data-driven modeling activity, the proposed approaches cannot be utilized if the system is new or impractical in the collection of network structure data. In such cases, an alternative approach is to interview decision makers to generate decision models. In scenarios where partial information about network structure is available, data can be used in conjunction with surveys and interviews to build a more reliable model. Finally, by utilizing theories from other disciplines, possible incentives can be designed and introduced into the network to achieve the desired network properties or performance.
CHAPTER 6. BEHAVIORAL EXPERIMENTATION AND GAME THEORY IN SOCIAL PRODUCT DEVELOPMENT

6.1 Chapter Overview

This chapter presents the second validation example of the dissertation. This example validates the performance of complex systems design as a result of local interactions and the decision-making behavior of design participants. In this chapter, RQ1, RQ2 and RQ3 are answered in the context of social product development through Approach 1.2, 2.2, and 3.2 (refer to Figure 1.3).

Specifically, the study of social product development is placed in the context of design crowdsourcing. In design crowdsourcing, the final design objective is realized through an open tournament in which participants submit their own design solutions in a competitive environment. In this open innovation process, participants are incentivized by awards, though their efforts incur a cost. Understanding an individual participant’s behavior in crowdsourcing (i.e., RQ1) is crucial for solution quality and efficiency in satisfying the design objectives. To answer RQ1, game theoretical analysis is performed. The individuals’ decision-making behaviors are modeled as a non-cooperative game. To understand the effect of participants’ decision-making behaviors in the crowdsourcing process (i.e., RQ2), an economic decision game is designed and executed. By changing the design options in the game, the effects of the participants’ decisions are investigated. To answer RQ3, statistical analysis is performed by using targeted data on individuals’ decisions in actual design process. The study reveals deviations from rational human behaviors, e.g., biases not structurally included in the game-theoretic models (See [143] for details). Figure 6.1 shows how the proposed approaches answer RQ1, RQ2 and RQ3 in the context of design crowdsourcing.
The outline of this chapter is as follows. In Section 6.2, crowdsourcing in engineering systems design is introduced. To simulate the crowdsourcing process, a computer-based experimental setup is designed for human subjects and presented in Section 6.3. The experiment is an economic decision game in which participants are grouped to compete in order to minimize a given convex function. In Section 6.4, such a game is modeled by employing contest theory and game-theoretic models. The Nash equilibrium is calculated in this section as well. In Section 6.5, the established game-theoretic models are assessed by performing an analysis of the data on participants’ actual behavior. The results obtained from data analysis validate the hypotheses drawn from the theoretical models, and provide more insights. Finally, the closing comments are presented in Section 6.6.
6.2 Crowdsourcing in Engineering Systems Design

Crowdsourcing is defined as the practice of outsourcing tasks, traditionally performed by employees or suppliers, to a large group of people in the form of open tournaments [144]. There is an increasing interest in utilizing crowdsourcing tournaments for engineering design and innovation. Examples involving successful use of crowdsourcing for product development include LocalMotors challenges [145], Innocentive for scientific innovation [146], Quirky [147] for product innovation, and TopCoder for software development [44].

Crowdsourcing is being viewed as an effective way of performing tasks that are otherwise difficult to automate on computers. Examples of such tasks include image annotation and locating a target object in an image. Within engineering design, such tasks include idea generation, problem solving, classification and evaluation of designs. While the benefits of crowdsourcing come from diversity of inputs from a large number of people, quality control of such information coming from the “crowd workers” [148] can be challenging. First, prior knowledge of all the possible outcomes may not be available, and some of these tasks demand a free response outcome space. In such scenarios we require techniques that could posit task-specific questions when deemed necessary to infer correct solutions and achieve quality [149]. Second, the inputs from the crowd may be dominated by low-quality or irrelevant solutions. There is also a possibility of spamming, such as when an individual provides solutions or assigns ratings to concepts randomly without carefully considering the specificity of the problems [150]. Therefore, it is important to incorporate the ability of each crowd worker and the bias of each individual in the crowd [151].

Recently, there have been a few studies in the design community to address these challenges. Green and coauthors [152] assess the use of crowdsourcing for evaluating engineering creativity by asking undergraduate students to rate the originality of solutions to design problems. The authors show that expert-level ratings can be extracted from non-expert students. Gerth and coauthors [153] hypothesize that
there is a relationship between task characteristics, individual expertise, and quality of the design outcome. The authors use a simple agent-based model to validate the hypothesis. Kudrowitz and Wallace [154] crowdsource the task of rapidly evaluating large sets of product ideas on Amazon Mechanical Turk. The ideas are presented in the form of sketches, and individuals in the crowd provide ratings of creativity, novelty and usefulness of the ideas.

These studies focus on utilizing the inputs from the participants. Since everyone is rewarded equally for their input (irrespective of the quality of their input), the questions such as a) whether someone would participate in the study, and b) how much effort they would put into the problem, etc. do not arise. However, if individuals are participating in a crowdsourcing contest, in which they may not win the prize, the issue of incentives becomes important. An example of such a scenario is DARPA’s Fast Adaptable Next-Generation Ground Vehicle (FANG) challenge [155]. DARPA explored the use of crowdsourcing for designing and developing a heavy, amphibious infantry fighting vehicle (IFV) through prize-based design competitions. Such challenges require significant effort (time and money) from the participants. In general, participation in such contests is high if the probability of winning is high. However, if more people participate in the crowdsourcing contest, the probability of winning of an individual decreases. Game theoretic models can be used to understand the effects of incentives on individual decisions and their interactions with other decision makers. In the following section, we discuss a foundational model of tournaments from non-cooperative game theory, which can be used to model a crowdsourcing tournament.

6.3 Behavioral Experimentation for Design Crowdsourcing

6.3.1 Modeling an Engineering Design Problem as a Search Problem

While there are many different views of design, one commonly accepted perspective is that designing is a process of searching through the space of design parameters, which characterize the structure of an artifact, to find a point that satisfies the need
in the best possible way [156]. Papalambros and Wilde [157] argue that design problems can be viewed as optimization problems, where the designers strive to satisfy the needs within the constraints of available resources. They highlight that the entire process of design can be viewed from the perspective of optimization. In their words, “… optimization suggests a philosophical and tactical approach during the design process. It is not a phase in the process, but rather a pervasive viewpoint. Philosophically, optimization formalizes what humans (and designers) have always done. Operationally, it can be used in design, in any situation where analysis is used…” [157, p. 12]. We consider this view of design in developing the experiment, and account for the following characteristics of real design problems:

1. A designer’s goal is to find the best design (defined by certain criteria).
2. Designers need to evaluate the performance of candidate designs, either through simulation models or physical experiments.
3. Experiments incur costs. Here, the term “cost” is used more generally. Costs can either be monetary cost or effort (computational, personnel, etc.).
4. Greater number of experiments result in a better understanding of the design space, and therefore, better quality of the design.

If a design problem is being solved by individuals in a crowdsourcing tournament, the expected payoff is determined by the quality of the designs by other designers also, see Equation (6.3). For a contestant, greater experimentation implies better quality, which in-turn results in a higher probability of winning, but also greater cost. If there are two participants in a tournament, the dependence on the parameters is as shown in Figure 6.2.

To simulate this scenario, we designed a behavioral economics experiment, where the participants are asked to optimize a design characterized by a single parameter \((x \in [-100, 100])\), whose performance is quantified by an unknown function, \(F(x)\). Each participant can request the value of the function for a specified \(x\) at a cost of \(c\).
tokens. Each player plays the game against one other player, randomly selected during each period. At the end of each period, the participant whose design parameter is the closest to the best possible design wins the fixed prize (Π). In this specific case, we asked the participants to minimize $F(x)$. Further details of the experimental setup are provided in Section 6.3.2.

This simple optimization problem is an abstraction of many complex design problems. It models the four characteristics of real design problems, listed above. Although it does not account for all the nuances of complex systems design problems, it is complete in the sense that it allows us to test behavioral assumptions and predictions from the game-theoretic model presented in this paper, and facilitates in gaining insights about participants’ decision making behavior. There are many advantages of starting with such a simple model. First, it is general enough to represent a wide variety of design problems. Second, the domain independent nature of the problem significantly reduces the variations among subjects due to diversity in knowledge and expertise, thereby reducing the resulting noise in the results. Such a simple design is preferred so that when the effects are observed, researchers can attribute them to the differences in the treatments. In a complex experiment, additional variables introduce additional sources of noise, which make it difficult to directly attribute the effects to the treatments. Therefore, so long as the missed variables do not interact with the treatments in influencing outcomes in reality, it is preferred to employ such simplistic designs from the behavioral economics standpoint.
The optimization problem in the experiment could have been posed as a specific design optimization problem, e.g., optimizing the dimensions of a beam under certain dimensional and strength constraints. However, we avoided such a problem because of two reasons. First, it would have restricted the participation of student subjects to specific groups who are familiar the problem domain. The second but more important reason is that it would have introduced additional noise in the data because of the varying knowledge related to the specific problem among the subjects. The goal in behavioral economics experiments is to control for noise from the parameters that have not been considered in the theoretical model to the extent possible.

6.3.2 Details of Experimental Design

As briefly mentioned earlier, the human subjects participated in a function minimization game. The subjects competed in groups of two to find the minimal point of a randomly generated convex function, $F(x) = (x - a)^2 + b$, where $a$ and $b$ are randomly chosen for each period from a uniform random distribution, $unif[-70, 70]$. Even though $F(x)$ had only one minimum because of the quadratic nature of the expression, that information was never revealed.

In this game, the subjects were allowed to sample multiple $x$ values. After every sampling attempt, the subjects were informed of the corresponding $F(x)$ values. The number of times the subject $i$ chooses to sample corresponds to the effort $e_i$. After both subjects finished sampling, the winner was picked as the subject choosing the $x$ that generates the minimal $F(x)$. Ties are broken randomly. If $i$ is the winner, the payment he/she receives is ($\Pi = 200$) tokens minus the cost ($c \times e_i$) incurred for each sampling of $x$. If however $i$ loses, he/she gains zero tokens but still incurs the cost ($c \times e_i$) for the samplings.

Notice that this game portrays the design innovation context well because often designers put in the effort to predict the outcome. Also the winner-takes-all and the loser-takes-nothing reward structure is reflective of the crowdsourcing context. These
characterizations are also consistent with prior related game theoretic models [158]. In particular, effort is proportional to the number of tries. Further, the quality of the solution is dependent on the number of tries, and the strategies for the tries.

We conducted the experiment in four sessions. Multiple subjects participated in each session. However, a subject was allowed to participate in only one session. Since our interest was in characterizing how the participants’ behaviors change with varying costs, we consider two different cost levels for sampling. Therefore, every session involved two treatments: a) a low-cost treatment, where the per unit sampling cost was \( c = 10 \) tokens; and b) a high-cost treatment, where the per unit sampling cost was \( c = 20 \) tokens. The subjects played 15 iterations of the function minimization game under each treatment. Each iteration is hereafter referred to as a period. Table 6.1 describes in detail the number of subjects in each session and the sequence within each session. Note that treatments were staggered differently across the sessions. This is done so that we can minimize any potential order effects.

<table>
<thead>
<tr>
<th>Session No.</th>
<th>Cost in the First Treatment</th>
<th>Cost in the Second Treatment</th>
<th>Number of Participants</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Low</td>
<td>High</td>
<td>10</td>
</tr>
<tr>
<td>2</td>
<td>High</td>
<td>Low</td>
<td>6</td>
</tr>
<tr>
<td>3</td>
<td>High</td>
<td>Low</td>
<td>14</td>
</tr>
<tr>
<td>4</td>
<td>Low</td>
<td>High</td>
<td>14</td>
</tr>
</tbody>
</table>

Observe that within each session, there could be multiple groups. Usually, when experiments involve groups, experimenters have to choose between the strangers or partners matching protocols. Our experiment employed a strangers-matching protocol. So, before each period began, subjects were randomly re-matched with a different group of subjects. Moreover, subjects were never informed of their groups’ composition or about individual effort levels of their group members. We chose the strangers
matching for two reasons. First, it allows us to experimentally capture the equilibrium for the one-shot game [159]. Second, it minimizes the reputation effects and also the potential for collusions.

Toward the end of each session, we paid the subjects as follows. The subjects were paid their earnings from 10 randomly chosen periods. One of the main reasons to do that is to ensure that the subjects play truthfully in each period and don’t subject themselves to end of experiment effects in later periods. The tokens were converted to actual dollars at a rate specified before the experiment began.

The experiment was conducted at Purdue University in the Fall of 2014. Subjects were recruited by email, and using the laboratory’s on-line recruitment system. The

Figure 6.3. Screenshot of the interface used by the participants.
computerized experimental environment was implemented using z-Tree [160]. Subjects were randomly assigned to individual computers, and communication was not allowed during the sessions. Copies of the experiment instructions were provided to each subject and were read aloud by the experiment administrator. A screenshot of the interface used by the participants is shown in Figure 6.3. The average earning per subject was $10.

6.4 Game Theory in Crowdsourcing – A Theoretical Analysis

Existing work related to tournaments has been carried out within the framework of contest theory. The theory of contests is a part of economics dealing with situations such as sports contests, rent seeking contests, conflicts and litigations where individual decision makers expend resources so as to increase their probability of winning [161]. The general approach within contest theory is to model a tournament as a game and to study the equilibrium and its stability. Various game-theoretic models of tournaments have been developed to provide insights about specific tournament design options. Terwiesch and Xu [158] present a classification of crowdsourcing projects: expertise based, ideation based, and trial-and-error projects. Different types of projects result in different analytical models of the game. The model presented in this section relates to the trial-and-error projects.

6.4.1 Modeling Tournaments as Non-Cooperative Games

A tournament is modeled as a non-cooperative game between contestants, assuming that contestants are self-interested agents whose goal is to maximize their own payoff. A participant’s payoffs ($\pi_i$) is dependent on the prize amount for the tournament and the probability of winning the prize. The probability of winning ($P_i$) is dependent on the quality of the submissions from all the participants, which in turn are dependent on the participants’ characteristics (e.g., expertise) and inputs (e.g., effort, time investment). Within the non-cooperative game, the players’ strategies
are the inputs. Hence, a general model of a tournament has three main parts: quality functions, contest success functions, and payoff functions, as discussed below.

**Quality functions** quantify the quality of the solution \( q_i \) proposed by the actors, as a function of the participant characteristics, such as knowledge and expertise \( K_i \) and the inputs, such as effort \( e_i \). It is assumed that the quality of each solution is independent of the characteristics and the inputs of other participants:

\[
q_i = q_i(e_i, K_i) \tag{6.1}
\]

**Contest success functions** provide each contestant’s probability of winning as a function of the quality of all solutions \[162\],

\[
P_i = P_i(q_1, q_2, \ldots, q_N) \tag{6.2}
\]

Conceptually, contest success functions are similar to probabilistic choice functions used within the design literature \[163\]. Generally, the functional form of the contest success functions is assumed in the formulation. The probability is assumed to have the following additive form:

\[
P_i = \begin{cases} 
\frac{f(q_i)}{\sum_{j=1}^{N} f(q_j)} & \text{if } \sum_{j=1}^{N} f(q_j) > 0 \\
\frac{1}{2} & \text{otherwise}
\end{cases} \tag{6.3}
\]

where \( f(q_i) \) is a non-negative increasing function. Two functional forms of \( f(q_i) \) are commonly used. The first form, referred to as the “power form” is: \( f(q_i) = q_i^m \) with \( m > 0 \), resulting in \( P_i = \frac{q_i^m}{\sum_{j=1}^{N} q_j^m} \). For a two-player scenario,

\[
P_1 = \frac{q_1^m}{q_1^m + q_2^m} = \frac{1}{1 + \left(\frac{q_2}{q_1}\right)^m}. \tag{6.4}
\]

Hence, the probability of winning is dependent on the ratio of the quality of the submitted solutions. The second form is \( f(q_i) = e^{kq_i} \) with \( k > 0 \), which results in
a multinomial “logit form” of the contest success function \( P_i = \frac{e^{kq_i}}{\sum_{j=1}^{N} e^{kq_j}} \). For two players, this reduces to

\[
P_1 = \frac{1}{1 + e^{k(q_2 - q_1)}}.
\]  

(6.5)

Hence, the probability of winning depends on the difference in the quality. The latter formulation can be derived both axiomatically and stochastically [162, 164].

**Payoff functions** relate the tournament design variables to the individual payoffs. For example, in a winner-takes-all contest, the payoff of an individual can be defined as the expected value of the prize,

\[
E(\pi_i) = \Pi P_i - C_i
\]

(6.6)

where \( \Pi \) is the amount of the prize and \( C_i \) is the cost incurred in developing the solution.

Using the quality function, contest success function, and the payoff function, the non-cooperative game is formulated. The Nash equilibrium of the game is generally used as the solution of the game. At the Nash equilibrium, player \( i \) chooses the input \( (e_i) \) that is a best response to other players’ best responses. The formulation can be used to quantify the effects of different tournament design concepts on the equilibrium effort invested by the players as a function of the exogenous parameters such as the prize, endogenous parameters such as the expertise and effort, and the structure of the game, such as winner-takes-all or auction style.

In the following section, we discuss a simple design crowdsourcing scenario, analyze it using this game theoretic model of contests, and generate different hypotheses. The hypotheses are tested using a behavioral economics experiment.

### 6.4.2 Game-Theoretic Analysis

In the context of the model discussed in Section 6.4.1, the participants are engaged in two-player contests during each period. The effort corresponds to the number of
different points for which a participant requests the value of the function (i.e., the number of tries). For participant \( i \), we represent the number of tries as \( e_i \). Therefore, the total cost incurred by the \( i \)th participant is

\[
C_i = c \times e_i \tag{6.7}
\]

Since the problem does not require any domain-specific knowledge, we assume that the participants are homogeneous in terms of expertise and knowledge \((K_i)\). The quality of a design \((q_i)\) generated by designer \( i \) is a monotonically decreasing function of the absolute difference between the optimum point and the best value of \( F(x) \) generated by the participant. We assume that the quality is only dependent on the number of tries, i.e.,

\[
q_i = q_i(e_i) \tag{6.8}
\]

The probability of winning the contest for each player is given by Equations (6.4) or (6.5), depending on the form of the contest success function. Since this is a winner-takes-all contest, the expected value of the prize for each player is given by Equation (6.6).

The commonly used functional forms for contest success functions (CSF) are the power form, and the logit form. If the power form is used, the expected payoffs of the two participants are:

\[
E(\pi_1) = \Pi \left( \frac{1}{1 + \left( \frac{q_2(e_2)}{q_1(e_1)} \right)^m} \right) - ce_1 \tag{6.9}
\]

\[
E(\pi_2) = \Pi \left( \frac{1}{1 + \left( \frac{q_1(e_1)}{q_2(e_2)} \right)^m} \right) - ce_2 \tag{6.10}
\]

where \( m > 0 \). On the other hand, if the logit form is used, the expected payoffs are:

\[
E(\pi_1) = \Pi \left( \frac{1}{1 + e^{k(q_2(e_2) - q_1(e_1))}} \right) - ce_1 \tag{6.11}
\]

\[
E(\pi_2) = \Pi \left( \frac{1}{1 + e^{k(q_1(e_1) - q_2(e_2))}} \right) - ce_2 \tag{6.12}
\]
where \( k > 0 \). Participant \( i \) chooses the value of \( e_i \) to maximize \( E(\pi_i) \) in response to \( e_{-i} \) where \(-i\) represents the other participant. If the functional form of \( q_i = q(e_i) \) is known, the Nash equilibrium for the game can be calculated. Note that the Nash equilibrium would depend on the functional form of the quality functions, \( q_i \), the contest success function chosen, and the parameters \( m \) and \( k \). These in turn depend on the specific problem and the characteristics of individuals participating in the contest.

Considering that the quality of the solution is monotonically increasing with the effort, two possible functional forms for the quality function (QF) are the linear form \( q_i = \alpha e_i \) and the exponential form \( q_i = \alpha \exp(\beta e_i) \), where \( \alpha, \beta, \) and \( \gamma \) are constants. The combinations of these CSFs and QFs result in four possible game formulations, as shown in Table 6.2. In a two-player scenario, each player maximizes the corresponding expected payoff function.

### Table 6.2. The rational reaction sets and Nash equilibria for different choices of quality functions (QF) and contest success functions (CSF) for two-player games.

<table>
<thead>
<tr>
<th>CSF/QF</th>
<th>Linear</th>
<th>Exponential</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Power Form</strong></td>
<td>( q_i = \alpha(e_i) )</td>
<td>( q_i = \alpha \exp(\beta e_i) )</td>
</tr>
<tr>
<td>( P_i = \frac{q_i^m}{\sum_j q_j^m} )</td>
<td>( \frac{e_i^m - e_{-i}^m}{(e_i^m + e_{-i}^m)^2} = \frac{c}{\Pi m} )</td>
<td>( \frac{\exp(m\beta(e_i + e_{-i}))}{\exp(m\beta e_{-i}) + \exp(m\beta e_i)} = \frac{c}{\Pi m \beta} )</td>
</tr>
<tr>
<td>Unique Nash eq.: ( e_i = e_{-i} = \frac{1}{4c} )</td>
<td>Multiple Nash equilibria.</td>
<td>All points that satisfy the equation.</td>
</tr>
</tbody>
</table>

| **Logit Form** | \( P_i = \frac{\exp(k \alpha e_i)}{\sum_j \exp(k \alpha e_j)} \) | \( \frac{\exp(\beta e_i) \exp(k \alpha (\sum_j \exp(\beta e_j)))}{\left( \sum_j \exp(\beta e_j) \right)^2} = \frac{c}{\Pi k \alpha \beta} \) |
| \( P_i = \frac{\exp(k \alpha e_i)}{\sum_j \exp(k \alpha e_j)} \) | \( \left( \sum_j \exp(\beta e_j) \right)^2 = \frac{c}{\Pi k \alpha \beta} \) | Unique Nash eq.: \( e_i = e_{-i} = \frac{1}{\beta} \ln \left( \frac{4c}{\Pi k \alpha \beta} \right) \) |

Consider the first example where the quality function is linearly dependent on the effort, i.e., \( q_i(e_i) = \alpha e_i \) where the constant \( \alpha \) is the same for all participants.
Assuming the power form of the contest success function, Equations (6.9) and (6.10) can be written as

\[ E(\pi_1) = \Pi \left( \frac{e_1^m}{e_1^m + e_2^m} \right) - ce_1 \]  
(6.13)

\[ E(\pi_2) = \Pi \left( \frac{e_2^m}{e_1^m + e_2^m} \right) - ce_2 \]  
(6.14)

The payoffs of the participants are used to derive the rational reaction sets (RRS). A participant’s RRS consists of the optimal strategies in response to the strategies of other participants. The rational reaction sets of the two players, derived using the first order optimality conditions, are

\[ \text{RRS}_1 : \frac{e_1^{m-1} e_2^m}{(e_1^m + e_2^m)^2} \cdot \frac{c}{\Pi m} = 0 \]  
(6.15)

\[ \text{RRS}_2 : \frac{e_2^{m-1} e_1^m}{(e_1^m + e_2^m)^2} \cdot \frac{c}{\Pi m} = 0 \]  
(6.16)

and the Nash equilibrium for this case, which is obtained by simultaneously solving RRS$_1$ and RRS$_2$, is

\[ e_1 = e_2 = \frac{\Pi m}{4c} \]  
(6.17)

In equilibrium, the individual effort \( e_i \) is inversely proportional to the cost \( c \) per trial. Through the experiments, we would like to validate whether individual actions are consistent with this conclusion from the model. Therefore our first hypothesis is:

**Hypothesis 1**: As the cost per trial, \( c \), increases, the expected number of tries, \( E(e) \), decreases.

The second hypothesis is related to the assumption about the quality function, and its dependence on the number of tries \( e_i \). The third hypothesis is related to the contest success function, which quantifies the probability of winning \( P_i \) in terms of the quality of the solution (and therefore, the effort). Specifically,

**Hypothesis 2**: The solution quality \( q_i \) monotonically increases with the number of tries \( e_i \).

**Hypothesis 3**: Increasing the number of tries, \( e_i \), increases the probability of winning, \( P_i \).
An experiment was set up to validate these hypotheses. We discuss the details of the experimental design in the following section.

6.5 Assess Game-Theoretic Models through Experimental Data

To minimize the effect of learning on our analysis, we discard the initial 5 periods under each treatment. Hence, the analyses henceforth will only consider the agents’ behaviors in the remaining 10 periods. Specifically, we use the results to test the hypotheses outlined in Section 6.4.2.

6.5.1 Hypothesis 1: The Expected Number of Tries Decreases as the Cost Per Trial Increases

The mean and standard deviation of the number of tries \( e \) are shown in Table 6.3. The second and third columns show the statistics for the low-cost treatment and high-cost treatment, respectively. The sessions where the low cost treatment was performed before the high-cost treatment (i.e., Sessions 1 and 4) resulted in average values of \( \mu_{e_{LL}} = 8.48 \) for low cost and \( \mu_{e_{LH}} = 6.25 \) for high cost. There were 24 participants in Sessions 1 and 4 combined. Each participant played for 15 periods for each treatment. Since the first five periods in each treatment are ignored, the sample size for each treatment is 240 (= 24 \( \times \) 10). The average values of \( e \) in sessions where high-cost treatment preceded the low-cost treatment are \( \mu_{e_{HL}} = 6.99 \) and \( \mu_{e_{HH}} = 4.18 \) for low and high costs, respectively.

Note that each individual participated in both the low and high-cost treatments. It is observed that the average number of tries \( e \) for the low cost treatment is greater than average in the high cost treatment in both experimental settings. To validate this, we conducted a paired two-sample t-test. The results of the test are shown in Table 6.4. Since the p-value is less than the level of significance (\( \alpha = 0.01 \)), we conclude that the mean value of the number of tries in the low-cost treatment is higher.
Table 6.3. Mean ($\mu$) and standard deviation ($\sigma$) of number of tries ($e$) in the treatments with different cost settings.

<table>
<thead>
<tr>
<th>Experiment settings</th>
<th>Low cost</th>
<th>High cost</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$c = 10$ tokens</td>
<td>$c = 20$ tokens</td>
</tr>
<tr>
<td>Low cost first</td>
<td>$\mu_e^{LL} = 8.48$</td>
<td>$\mu_e^{LH} = 6.25$</td>
</tr>
<tr>
<td>Sample size = 240</td>
<td>$\sigma_e = 2.92$</td>
<td>$\sigma_e = 2.48$</td>
</tr>
<tr>
<td>High cost first</td>
<td>$\mu_e^{HL} = 6.99$</td>
<td>$\mu_e^{HH} = 4.18$</td>
</tr>
<tr>
<td>Sample size = 200</td>
<td>$\sigma_e = 2.81$</td>
<td>$\sigma_e = 1.62$</td>
</tr>
</tbody>
</table>

Table 6.4. Summary of four hypothesis tests for mean value of number of tries $\mu_e$.

<table>
<thead>
<tr>
<th>Alternative hypothesis</th>
<th>t stat.</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paired two sample t-test</td>
<td>$\mu_e^{LL} &gt; \mu_e^{LH}$</td>
<td>11.25</td>
</tr>
<tr>
<td></td>
<td>$\mu_e^{HL} &gt; \mu_e^{HH}$</td>
<td>16.60</td>
</tr>
<tr>
<td>Welch two sample t-test</td>
<td>$\mu_e^{LL} &gt; \mu_e^{HH}$</td>
<td>19.51</td>
</tr>
<tr>
<td></td>
<td>$\mu_e^{HL} &gt; \mu_e^{LH}$</td>
<td>2.94</td>
</tr>
</tbody>
</table>

Further, to test whether the conclusion is independent of the initial conditions (i.e., whether they start with the low cost treatment or the high cost treatment), we conduct a Welch two sample t-test. Based on the results shown in Table 6.4, we conclude that the null hypothesis is rejected with the level of significance ($\alpha = 0.01$). Hence, the experimental results are consistent with the prediction that a higher cost per trial ($c$) results in a lower expected number of tries ($e$).

6.5.2 Hypothesis 2: The Solution Quality Monotonically Increases with the Number of Tries

The best value of the function, $f(\tilde{x}_i)$, submitted by participant $i$ at the end of each period is used to evaluate whether he/she wins during that period. The quality of the solution is a monotonically decreasing function of the distance between this
best value and the optimum value $f(x^*)$. Hence, we use $\Delta = f(\tilde{x}_i) - f(x^*)$ to test this hypothesis.

Out of the 880 total observations, 12 observations had $e_i = 0$, i.e., the participants did not try at all. We ignored these observations. The values of $\Delta$ for each $e$ in the low-cost first sessions (Sessions 1 and 4) and the high-cost first sessions (Sessions 2 and 3) are plotted in Figures 6.4(a) and 6.4(b) respectively on a semi-log scale. The regression indicates a decreasing exponential relationship between expected value of $\Delta$ and $e$.

![Figure 6.4](image)

(a) Sessions 1 and 4 (Low cost first)  
(b) Sessions 2 and 3 (High cost first)

Figure 6.4. Relationship between number of tries and quality of solution.

We perform a single predictor linear regression with log transformation. The model is

$$\ln \Delta = \beta_0 + \beta_1 e$$

(6.18)

where $\beta_0$ and $\beta_1$ are regression coefficients. The estimation results are shown in Table 6.5. Note that the $\beta_1$ values are negative. The p-values in both settings for the coefficient show that $\beta_1$ values are statistically significant, thereby validating the hypothesis.
Table 6.5. Estimation for regression coefficients in Equation (6.18).

<table>
<thead>
<tr>
<th>Experimental settings</th>
<th>Estimated parameter</th>
<th>t stat.</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low cost first</td>
<td>( \beta_0 = 3.14 )</td>
<td>8.06</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td></td>
<td>( \beta_1 = -0.75 )</td>
<td>-15.46</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>High cost first</td>
<td>( \beta_0 = 5.24 )</td>
<td>14.99</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td></td>
<td>( \beta_1 = -0.996 )</td>
<td>-17.68</td>
<td>&lt; 0.001</td>
</tr>
</tbody>
</table>

The regression model in Equation (6.18) can be compared to the exponential quality function in Table 6.2, i.e., \( q_i = \alpha \exp(\beta e_i) \). If we assume an inverse relationship between \( \Delta \) and the quality \( q \) (i.e., \( \Delta = \frac{1}{q} \)), then

\[
\beta_1 = -\beta \tag{6.19}
\]

\[
\beta_0 = -\ln \alpha \tag{6.20}
\]

Hence, the exponential quality function is a good choice for the analytical model, and the parameters of the quality function are directly available from the experimental results.

6.5.3 Hypothesis 3: Increasing the Number of Tries Increases the Probability of Winning

To test this hypothesis, we separated the datasets according to whether or not a participant won during a given period. The average values and the standard deviations of the number of tries, categorized by the winning status are presented in Table 6.6. From the table, we observe that the participants who won, tried more times on average than the ones who lost. For example, in the low-cost first setting, the average number of tries for winners (\( \mu_{LW} = 8.21 \)) is greater than the average number of tries for participants who did not win (\( \mu_{LN} = 6.52 \)).

The quantitative impact of the number of tries on the winning probability is assessed by performing a logistic regression between \( e \) and a boolean variable \( b \) which
Table 6.6. Average and standard deviation of the number of tries categorized by winning status with different cost settings.

<table>
<thead>
<tr>
<th>Experiment settings</th>
<th>Winning Status</th>
<th>Average number of tries</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low cost first</td>
<td>Win</td>
<td>(\mu^{LW}_e = 8.21)</td>
<td>(\sigma^{LW}_e = 2.69)</td>
</tr>
<tr>
<td>Sample size = 240</td>
<td>Lose</td>
<td>(\mu^{LN}_e = 6.52)</td>
<td>(\sigma^{LN}_e = 2.93)</td>
</tr>
<tr>
<td>High cost first</td>
<td>Win</td>
<td>(\mu^{HW}_e = 6.29)</td>
<td>(\sigma^{HW}_e = 2.83)</td>
</tr>
<tr>
<td>Sample size = 200</td>
<td>Lose</td>
<td>(\mu^{HN}_e = 4.89)</td>
<td>(\sigma^{HN}_e = 2.34)</td>
</tr>
</tbody>
</table>

is 1 if the participant wins, and 0 otherwise. The probability of winning during a period is modeled as

\[
P(b = 1) = \frac{\exp(\beta'_0 + \beta'_1 \tilde{e})}{1 + \exp(\beta'_0 + \beta'_1 \tilde{e})}
\]

(6.21)

where \(\beta'_0\) and \(\beta'_1\) are regression parameters. Here, \(\tilde{e} = \frac{e}{10}\) for the high cost treatment, and \(\tilde{e} = \frac{e}{20}\) in the low cost treatment. The scaling is carried out because any number of tries beyond the scaling factors would result in a negative expected payoff, see Equation (6.6).

The estimated parameters for the model in Equation (6.21) are shown in Table 6.7. The statistical significance of the \(\beta'_0\) and \(\beta'_1\) is tested using the Wald chi-square statistic. As shown in the table, the number of tries \(e\) is a significant predictor of participants’ winning probability (p-value < 0.01) for both experimental settings. To interpret the results, we use the odds ratio, which is the ratio of the probability of stopping to the probability of continuing. Using the low cost first setting as an example, for a unit\(^1\) increase in the number of tries, the odds of stopping (versus not stopping) increase by a factor of 11.98.

Using the estimated parameter values \((\beta'_0\) and \(\beta'_1\)) in Table 6.7, we calculate the following winning probabilities as a function of the number of tries \(e\).

1. \(P^{HH}\): High cost treatment in high-cost first setting

\(^1\)Since the number of tries has been scaled, one unit means 10 tries in low cost setting and 20 tries in high cost setting.
Table 6.7. Estimation for regression coefficients in Equation (6.21).

<table>
<thead>
<tr>
<th>Experiment setting</th>
<th>Estimated parameter</th>
<th>z stat.</th>
<th>p-value</th>
<th>Odds ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low cost first</td>
<td>( \beta_0 = -1.30 )</td>
<td>-5.19</td>
<td>&lt; 0.001</td>
<td>0.27</td>
</tr>
<tr>
<td></td>
<td>( \beta_1 = 2.48 )</td>
<td>5.60</td>
<td>&lt; 0.001</td>
<td>11.98</td>
</tr>
<tr>
<td>High cost first</td>
<td>( \beta_0 = -1.56 )</td>
<td>-5.27</td>
<td>&lt; 0.001</td>
<td>0.21</td>
</tr>
<tr>
<td></td>
<td>( \beta_1 = 4.07 )</td>
<td>5.61</td>
<td>&lt; 0.001</td>
<td>58.52</td>
</tr>
</tbody>
</table>

2. \( P^{HL} \): Low cost treatment in high-cost first setting

3. \( P^{LH} \): High cost treatment in low-cost first setting

4. \( P^{LL} \): Low cost treatment in low-cost first setting

As expected, the plots show that for a given number of tries, the predicted probability of winning is higher in the high-cost treatment compared to the low-cost treatment, for both settings. In addition, we observe that the probability of winning is also dependent on the initial cost setting. Specifically, for \( e > 3 \), the predicted
winning probability for a given number of tries is higher in the high-cost first setting, as compared to the low-cost first setting.

\[ P^{HH}(e) \geq P^{LH}(e) \quad (6.22) \]
\[ P^{HL}(e) \geq P^{LL}(e) \quad (6.23) \]

One possible reason for this trend is that individuals may be trying less number of times if they participate in the high cost first sessions. To evaluate this assumption, we performed a Welch two sample t-test with a null hypothesis that the average number of tries remains the same, regardless of the initial cost setting; results are shown in Table 6.8. It is observed that \( \mu^{LH}_e > \mu^{HH}_e \) and \( \mu^{LL}_e > \mu^{HL}_e \), clearly indicating that when participants start with a high cost setting, they try less number of times as compared to participants who start with a low cost setting. One possible explanation for the behavior is the anchoring bias [165]. Because subjects faced intense competition in the initial treatments when the costs were low, they possibly carried the same tendencies to sample lesser number of times even after transitioning to the high cost setting. The result in Equation (6.23) is striking, and has implications on multi-stage crowdsourcing competitions. Participant behavior in the initial stages, where the competition may be intense, may have implications on the later stages also.

Table 6.8. Welch two sample t-test for average number of tries in sessions with different initial costs.

<table>
<thead>
<tr>
<th>Alternative hypothesis</th>
<th>t stat.</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Welch two sample t-test</td>
<td>( \mu^{LH}_e &gt; \mu^{HH}_e )</td>
<td>10.49</td>
</tr>
<tr>
<td></td>
<td>( \mu^{LL}_e &gt; \mu^{HL}_e )</td>
<td>5.44</td>
</tr>
</tbody>
</table>

Finally, the parameters estimated in Table 6.7 can be related to the parameters in the analytical model presented in Table 6.2. Equation (6.21) can be written as

\[ P(b = 1) = \frac{\exp(\beta_1 \tilde{e})}{\exp(-\beta_0) + \exp(\beta_1 \tilde{e})} \quad (6.24) \]
Comparing this to the probability of winning (for player $i = 1$) from the analytical model with power form of contest success function and exponential quality function

$$P_1 = \frac{q_1^m}{q_1^m + q_2^m} = \frac{\exp(\beta me_1)}{\exp(\beta me_1) + \exp(\beta me_2)},$$

(6.25)

we observe that $\beta me_1 = \beta'_1 \tilde{e}_1$. Using Equation (6.19), we get $m = -\frac{\beta'_1 \tilde{e}_1}{\beta_1 e_1}$. For the high cost treatment, this reduces to $m = -\frac{\beta'_1 \tilde{e}_1}{10\beta_1}$ and for the low cost treatment, $m = -\frac{\beta'_1}{20\beta_1}$. The parameters $m$ and $\beta$ completely describe the analytical model listed in Table 6.2.

The hypotheses have important implications for structuring crowdsourcing contests for engineering design. Hypothesis 2 illustrates that the greater the effort from the participants, the higher the solution quality. Hence, it is in the interest of the crowdsourcing contest designer to incentivize participants to invest greater amounts of effort. At the same time, Hypothesis 1 indicates the effect of cost of gathering information on the effort invested by the participants. Lower cost results in greater effort. Hence, if the contest designer wants the participants to generate and evaluate a large number of alternate designs, the cost of developing and evaluating each design must be low. Finally, the results from Hypothesis 3 provide insights on how participants’ behaviors (e.g., the effort invested) may depend on the task sequence.

### 6.6 Closing Comments for Chapter 6

In summary, the work presented in this chapter answers the three RQs in the context of social product development by using design crowdsourcing as an example. Contest theory is employed to model contestants’ winning probability as a function of design effort. The contestants’ strategic decision-making is then modeled as a non-cooperative game by integrating winning probability models. Additionally, individuals’ preferences for design cost are also structured into the game theoretic model. In order to validate this theoretic model and to better understand participants’ behavior in a real design process, a human-subject experiment is designed and conducted.
Through the experiment, the data regarding contestants’ actual behavior is collected. This data is then used to analyze how design variables in crowdsourcing - such as cost - affect the design effort and to analyze how design effort affects winning probability. The results provide a better understanding of contestants’ preferences for spending effort and their preferences for design cost.

In this study, the crucial aspect for examination is the degree to which the environment is controlled. An important concern is the internal validity of an experiment. The experiments should be robust, replicable, and should enable us to confidently draw causal conclusions from the research [166]. When designing experiments, simplicity is generally preferred [167]. This is an essential aspect because it enables the researcher to explain differences in observations of the manipulations (or treatments) of the experimental environment. Simple games provide bold predictions about the outcome, and it is easier to control the environment to satisfy the game. Hence, a complex game theoretic model is simplified into a simple game (e.g., a 2-player game with a small strategy set). The simplified game is mapped onto an experimental design, which involves [168]:

1. identification of control variables, variables to be measured, and fixed variables,
2. providing formalized instructions,
3. deciding on an incentive structure, and
4. deciding on the information provided during each round.

Details on how the experiment will be run, including the number of players, the participant pool, number of runs, etc. are decided. The games are generally repeated to study convergence to equilibria and the learning of players over time.

Behavioral experiments within engineering design often involve students from academic institutions. Students are ideal subjects for many reasons. One, their expertise in specific technical areas can be easily assessed based on the courses they have taken. Two, it is easier to reach them for the purposes of the experiment. The natural
question that arises is: are the results obtained be representative of the actual population? It is true that the actual measures of outcomes will be different depending on whether students are in undergraduate or graduate programs, depending on their cultural backgrounds, etc. However, when the models are used primarily for the comparisons of policies, these actual differences have a minimal impact on the outcome. Additionally, the experimental economics literature has expanded its investigation on this issue.

The approach developed for modeling strategic decisions in design crowdsourcing has a limitation: the results obtained from game theoretical model, i.e., the Nash equilibrium, only describe the aggregate behavior of contestants. For example, the Nash equilibrium of the proposed model indicates that the higher the design cost, the less effort the contestant spends. In reality, however, individuals may not care about the cost, and simply be eager to win. The game theory fails to model these kinds of behavioral variations among individuals. In the future, statistical models that account for random effects can be developed to solve this problem.
CHAPTER 7. CLOSING THOUGHTS AND FUTURE WORK

In this chapter, it summarizes the accomplished research work and contribution of this dissertation. The research challenges of the study of complex systems engineering and design are identified. These challenges lead to thoughts for future research.

7.1 Conclusions and Contributions

7.1.1 Conclusions

Research in the design of bottom-up evolutionary complex systems is still in its early stages. Successful design of such systems requires a deeper understanding of the dynamics of local interactions. This calls for a framework which helps to synthesize knowledge from different domains to develop appropriate approaches and tools. In this dissertation a modest attempt is made in this direction to develop such a framework. The developed decision-centric framework facilitates a synergistic integration of theories from various fields, such as the utility theory, game theory and complex network theory. It is also shown that this framework provides a platform to better understand local interactions that result from decision-making processes in complex systems engineering and design. Specifically, this dissertation is focused on answering three research questions.

- RQ1: How can individual decision-making preferences and behavior be modeled?

- RQ2: How can the effect of local decision-making preferences and behavior be understood?

- RQ3: How can unobserved decision-making preferences and behaviors of individual entities be estimated?
These RQs are approached through a set of models and computational techniques. These approaches are validated in two application areas – the Internet and Crowdsourcing Design – to show the generality. Within the two application contexts, three RQs are answered. From the results, we conclude that:

1. The proposed decision-centric framework is capable of modeling, analyzing and estimating the individual preferences in bottom-up evolutionary complex systems. Within such a framework, the micro level interactions and system-level dynamics can be better understood.

2. The decision-making behavior of individual entities in complex systems have a significant impact on system structure and performance. By influencing individual preferences, desired system structures can be achieved. For example, by changing the preferences to nodes’ degree, a variety of network topologies ranging from scale-free to giant-hubs can be obtained.

3. Human behaviors have deviations (biases) distinct from the prediction of the theoretical models, and the proposed behavioral experimentation and analysis approach are capable of discovering those biases. For example, in design crowdsourcing, without considering the anchoring bias (see Chapter 6 for details) of human behavior observed from experiments, the results obtained directly from the game theoretical model may lead to an incomplete explanation. This brings attention to engineering problems involving strategic decisions, such as design of market systems and supply chain systems, where only analytical game-theoretic models have been used.

4. It is observed that heterogeneity of preferences exists among decision makers, and the proposed approach is based on mixed effect models capable of estimating such heterogeneity. It models the heterogeneous preferences as a certain type of distribution, e.g., the normal distribution. It is essential to model heterogeneous preferences within complex systems in order to better model the inherent system dynamics.
7.1.2 Contributions to Design of Complex Networked Systems

In the study of a complex networked system, a decision-centric modeling framework based on discrete choice random utility theory is proposed. The framework is domain-independent and enables modeling, analyzing and estimating node-level decision-making preferences and behaviors in complex networked systems. In terms of methodology, key contributions include:

1. A decision-centric approach for modeling node-level decision-making preferences and behaviors in complex networked systems. The approach helps obtain a variety of networks as surrogate models for real-world complex networked systems.

2. An approach that enables the integration of economic behavioral models with the theory of complex networks to estimate decision-making preferences and behaviors. The approach enables the estimation of node-level preference in terms of coefficients in discrete choice models. To facilitate the implementation of the approach, a stepwise framework is established.

3. The remaining two approaches estimate node-level linking behaviors. One is based on a generalized preferential attachment model and the other is developed based on the best fitting regression. To the best of our knowledge, the generalized preferential attachment model is the first to be used for estimation purposes.

4. A method for extracting decision-making observations and identifying choice set. The method is used to perform the analysis of edge dynamics with two consecutive networks annotated with types of nodes.

5. A method for handling large choice sets in discrete choice analysis. The dissertation proposes a nested method to decrease the size of the choice set by merging agents with same type of attributes.
6. A set of computational techniques for estimating and analyzing model parameters that represent individuals’ decision-making preferences. These techniques include correlation analysis, statistical estimation and hypothesis testing.

Besides the methodological aspects, the studies performed also make contributions in the field of complex networks. First, the dissertation establishes a general framework using discrete choice models as a core for building network topology generators. Second, while analyzing the network dynamics, a numerical method for solving the degree of nodes at a given time is established. This method is applicable when the results of continuum theoretical models of node’s degree dynamic cannot be analytically solved. Third, an approach for integrating continuum theory with percolation theory is developed to establish a direct mapping of node-level preference to system robustness in degree-based network evolution. The results provide insights on designing incentives and avoiding critical decision-making preferences for generating complex networks with improved robustness and resilience.

As for the applications of this study, the discrete choice analysis on Internet example makes contributions in modeling the topology and evolution of the Internet. First, existing studies generally claim degree to be an important factor affecting the Internet topology. This dissertation provides detailed insights on the influence of degree. For example, in the C2P connections between new ASes and existing ASes, only the in-degree affects the network evolution, but in the C2P connections between two existing ASes, both in-degree and out-degree affect the ASes’ peering preferences. Additionally, in the P2P connections between two existing ASes, the degree does not directly influence ASes’ decision-making. Instead, a variation of degree, i.e., the business access rank, plays a significant role in determining the linking probability. Secondly, it is found that AS decision-making preferences to geographical distance in C2P linking are homogeneous, but heterogeneous in P2P linking. Third, the normalized business access rank is consistently significant and exhibits random effects in all three types of linking activities, which indicates that AS linking activities in 2004
are mostly focused on accessing Internet service for the purpose of business use. All these insights facilitate more realistic surrogate models of the Internet.

7.1.3 Contributions to Design of Social Product Development

To the best of our knowledge, the decision-making study in design crowdsourcing in this dissertation is the first attempt to synergistically use analytical game theory (which describes what players with different cognitive capabilities would do) with behavioral game theory (which describes what people actually do [168]) in an engineering design context. In this dissertation, the study of this application area makes contributions in two aspects. First, the dissertation shows how analytical models of game theory can be used for designing experiments and generating testable hypotheses, and how experiments can be used to inform theory (e.g., specific functional forms of the quality functions and contest success functions). Second, the dissertation features the integration of behavioral experiments into the game theoretic modeling of design decision-making.

Within behavioral experiments, the goal of experimentation is not to test a full theory, but to test certain hypotheses used in (or generated from) theory. Some of the commonly used hypotheses in game theoretic contest models include rational decisions, and decision makers playing at the Nash equilibrium. Similarly, theoretical models can generate hypotheses such as “with the increase in the number of participants, the effort invested by individuals decreases”, and “increase in award amount increases the effort”. Additionally, the goal is not to disprove game theory but to improve it by establishing psychological regularities [168]. For example, it helps in identifying other aspects of human behaviors such as biases that are not structurally included in the theoretical models. In summary, the contributions of the dissertation in this study are:

1. A unique approach towards integrating game theory and behavioral experimentation for better understanding of the design situation in crowdsourcing.
2. Systematic experimentation using economic decision game to simulate design crowdsourcing.

3. A general approach for data-driven elicitation of contestants’ behavior in design crowdsourcing.

4. Domain-specific insights in the design of crowdsourcing tournament for better design solutions.

Besides contributions at the methodological level, specific insights regarding designers’ decision-making in design crowdsourcing are obtained. For example, the results show that (i) contestants’ preferences are heterogeneous in nature, and (ii) their behavior can be attributed to three factors: the cost of solving a problem, the quality of that solution, and the future uncertainty of the problems. These insights are critical for refining the analytical models and informing improvement for the design of crowdsourcing and better attracting, organizing and managing the participants.

7.2 Research Challenges and Future Work

7.2.1 Research Challenges

From the research, numerous additional challenges of decision-making in the design of complex systems have been identified.

*Data availability and applicability.* The estimation of decision-making preferences in complex systems relies upon statistical approaches which are data-driven in nature. The data is thereby key in order to perform the estimation. However, data collection is challenging due to the lack of organizational data and temporal information about the system’s evolution. This is because some organizations, such as defense departments, will not share data due to security concerns. Other organizations, such as commercial organizations are unwilling to share such data due to competition. Although sometimes we are able to collect the data, the data is incomplete, and results
in a missing value problem. Traditional approaches like bootstrapping and imputation are usually practiced to solve this problem. However, many assumptions are required for implementing these techniques. These assumptions will affect the estimation results. The challenge exists in making the assumptions contextually reasonable such that the data best represents reality.

*Computational ability.* The nature of complex systems introduces a large computational burden into analysis. For example, in the Internet study, there are more than 16,000 nodes and more than 30,000 edges in the network. When performing the discrete choice analysis, the large size of the choice set impedes the statistical regression analysis. In the dissertation, the approach used for solving this issue is to cluster all nodes with the same degree, and treat one cluster as a choice. Then, it is assumed that the probability of choosing a node in that cluster is uniform. However, this assumption itself may not reflect reality. So, the challenge exists in proposing effective computational techniques to overcome these kinds of issues.

*Simplicity vs. complexity.* While establishing models, a common issue is to make trade-off between simplicity and complexity. Compared with simple models, complex models capture more realistic components. When building the models, we need to make a trade-off based on the specific circumstances. In this dissertation, for example, the network generation model, i.e., the DBDC model, is very simple. Such simple models are capable of generating a variety of networks representing most types of typologies exhibited by existing systems. However, the dynamics of node deletion and edge deletion are not considered in the model. Such a model fails to precisely capture how the real systems’ dynamics evolve over time. So, the challenge exists in identifying the appropriate level of complexity in creating the models of decision-making behavior for complex systems.

*Multi-disciplinary nature.* The study of complex systems requires existing knowledge from different disciplines. For example, in this dissertation, in order to perform the analysis of complex networked systems, we integrate the discrete choice random utility theory and network theory. Similarly, when performing the analysis of de-
signers’ behaviors in crowdsourcing, we integrate the game theory and behavioral experimentation. The integration of theories is challenging because sometimes there is conflict in fundamental assumptions between different knowledge areas. For example, when doing the random effects analysis for the preferences of decision-makers, one assumption is that the heterogeneity of preferences is modeled as a normal distribution, which may not be the case in its real application.

Cognitive perceptive. Like most existing studies on human behavior in the engineering context, this dissertation models human behavior in competitive design by using normative models, e.g., analytical models from contest theory. However, design thinking itself is a cognitive process. When doing behavioral analysis, it is important to have models to capture the cognitive process of individuals, such as recognizing patterns, directing attention, forming concepts, visualizing space, etc. The lack of cognitive models and theories for the study of design thinking impedes the development of standards for collecting and analyzing data of decision-making behavior in design. The challenge exists in creating a framework that could couple the cognitive modeling into normative models while using behavioral experimentation as a complement for validation.

Validation framework. In this dissertation, two validating approaches are adopted. The first approach is to use statistics of model fitting to validate the behavioral models obtained from data-driven methods in complex networked systems. The way of validating uses the obtained model to regenerate the data and compare the accuracy of the synthetic data with real data. The second approach is to use experiments, which involve human-subject experiments to see if the same results can be observed from the actual behavior of humans. While validation is performed in the dissertation, both approaches are not perfect. For example, in the statistical validation method, the results from past data do not guarantee future behavior. Challenges exist in creating a standard framework that contains several validation techniques to ensure the results are validated within one domain as well as across applications.
7.2.2 Future Work

These challenges help in identifying research gaps and guiding future research in this field. This dissertation provides a platform to resolve the aforementioned challenges. Some of these are short-term and others are long-term in nature. Both are discussed in detail in this section.

First, future research can focus on developing computer-based data collection and interface for automated analyses with either local computing or cloud computing. Progress in this area requires collaboration with experts in the field of computer science.

This dissertation establishes a modeling framework for generating network topologies. In the future, the framework can be used to construct complex network topology generators by considering more detailed network growth mechanisms, such as the evolution mechanisms (b) and (c) discussed in Section 3.3. Furthermore, in the network generation model proposed, the nodes make independent decisions. In future work, this assumption can be relaxed by integrating discrete choice models with game theory to build network formation games.

In the Internet study, future work could be focused on forecasting Internet evolution based on the model developed in this dissertation. Since the calculation of business access utility is computationally expensive for large networks, this dissertation does not model the evolution of Internet using the obtained model. With the increasing computational capability, the study in forecasting the evolution of large-scale complex networked systems can be realized. Upon the completion of system forecasting, the validation of the model can be performed.

To model more realistic decision-making scenarios in systems, decision-making of stakeholders at different levels should be considered in the future research. The design of successful systems rests with the decisions of stakeholders at different levels, such as users, manufacturers, distributors, market, and environment. Therefore, it is crucial to understand how stakeholders compromise, coordinate and cooperate with
other stakeholders. Research opportunities exist in how decisions made by different stakeholders can be synthesized to facilitate the study of the entire system. Future research can be focused on creating such decision-centric design framework by coupling decisions of stakeholders at different design phases.

For the study of decision-making behaviors in the systems with humans, the subjects used in the experiment could be diverse. In this dissertation, only students are recruited in the design of experiment for the study of social product development. But in many other contexts, the focus may be on the nature of behavior rather than the actual outcome. In those cases, the behaviors may be inconsistent between students and non-students. For example, Lichtenstein and Slovic [169] studied how choices made in a gambling context were different for students and non-students. Non-students chose to evaluate the casino-goers in Las Vegas but observed the same set of results on choices that they had identified earlier [170] using student subjects. In the future, studies can be performed with experts from industry and results can be compared to assess the impact of the subject pool.

Furthermore, in the study of social product development, the experiment can be extended to more realistic design problems. First, we have only considered a simple abstraction of design problems as optimization problems. Real engineering systems design problems may involve complexities such as i) uncertainties in cost estimates, and ii) tradeoff between developing new high-performance designs and choosing less risky but low-performance options. Second, understanding crowdsourcing is essential to designing effective crowdsourcing competitions. There are many parameters that can be varied in a design competition, e.g., instead of winner takes all strategy, there could be multiple prizes with different values. The competitions can be either single stage or multistage, where the initial stages are used to shortlist participants who go to the later stages. A single competition can be carried out for a system-level design problem, or a problem can be decomposed into sub-problems with multiple contests. Crowdsourcing competitions can also allow team formation. The dynamics of team formation could affect the outcome of the contest. Future experiments will
involve analyzing the effects of different designs of crowdsourcing contests on the effectiveness of competitions for engineering systems design. Finally, we only focus on the aggregate/expected behavior of the participants. In the future work, we will investigate the individual decision making strategies and the effects of diversity in the group.

Finally, future research can focus on mechanism design for social product development. During the last decade, there have been many successful cases of social product development, and intensive research in analyzing the participants’ behavior and community structure in this new paradigm. However, there is little understanding of how incentive structures should be designed to improve design efficiency of the process and the quality of resulting designs. Research opportunities exist in creating knowledge of the design of mechanisms and/or incentive structures in social product development. The research objective will be to (a) use analytical game theory to analyze the impact of mechanism design options and incentive structures on quality of design outcomes, (b) verify the improved incentives structures through behavioral experiments, and (c) evaluate the effectiveness of the resulting mechanisms. The study would benefit other engineering design problems involving strategic decisions where only analytical game-theoretic models have been used.
LIST OF REFERENCES


VITA
VITA

ZHENGHUI SHA

School of Mechanical Engineering

585 Purdue Mall, West Lafayette, IN 47907

https://sites.google.com/site/zhhsa

RESEARCH INTERESTS

My research is focused on theoretical and experimental studies of decision-centric modeling of evolutionary dynamics of complex systems and the decision-making in engineering design. Specifically, my dissertation is focused on: (a) estimating micro-level decision-making behavior for modeling complex networked systems, and (b) game-theoretic analysis, stochastic modeling and behavioral experimentation in open engineering design (e.g., the crowdsourced design and open-source product development).

EDUCATION

Ph. D. in Mechanical Engineering

Graduate Certificate in Applied Statistics

Purdue University

M. S. in Mechanical Engineering

Xi’an Jiao Tong University

B. Eng. in Mechanical Engineering

Minor in Computer Science

Xi’an University of Technology

Xi’an, China

Xi’an, China

West Lafayette, IN, USA

Expected August 2015

August 2014

April 2010

July 2006

March 2006
PUBLICATIONS

Ph.D. Dissertation

Z. Sha, “Decision-Centric Foundations for Complex Systems Engineering and Design” presented to the faculty of the School of Mechanical Engineering, May 2015, Purdue University, West Lafayette, Indiana, USA.

Advisor: Dr. Jitesh H. Panchal.

M.S. Thesis

Z. Sha, “Theoretical and Experimental Studies on Design of Numerical Controlled Die Spotting Press” presented to the faculty of the School of Mechanical Engineering, April 2010, Xi’an JiaoTong University, Xi’an, China.

Advisor: Dr. ShengDun Zhao.

Journal Articles


**Conference Papers**


Paper number: DETC2014-34492. (Awarded with Purdue COE Travel Funds)


C11. J. Xie, S. Zhao, **Z. Sha**, and J. Liang, “An Online Adaptive Fast Control Approach Based on Local Linearization and Its Application to a Large Inertia


RESEARCH EXPERIENCE

Design Engineering Laboratory, Purdue University, West Lafayette, IN.

Research Assistant, Jul. 2012 - Present.

  - Contributed to proposal drafting, including reporting preliminary results and scoping research objectives.
  - Established a data-driven framework to infer the local decision-making preferences in large-scale networked systems.
  - Applied the proposed framework to two fields. First, the framework was applied to estimate airlines’ decision-making preferences on routes panning in air transportation network. Second, the framework is applied to estimate autonomous systems’ linking preferences in the Internet.
• Contributions to Research Grant – Crowdsourcing for Engineering Systems Design: Theoretical and Experimental Studies (NSF CMMI #1400050). PIs: Dr. Panchal and Dr. Kannan.
  – Contributed to proposal drafting and preliminary results reporting.
  – Developed rigorous framework to analyze design options for design crowdsourcing contest.
  – Modeled individuals’ decision-making behaviors in design crowdsourcing as a non-cooperative game. Conducted human-subject experiments to collect target data. Performed statistical analysis to estimate decision-making preferences and validate theoretical models.

• Contributed to Enterprise Project – Integrated Part Classification for Product Cost and Complexity Reduction. Collaborators: Tata Consultancy Service, Columbia OH.
  – Cooperated with two consultants in TCS to conduct comprehensive analysis on classifying part-level information and dynamic.
  – Performed requirement analysis on cross-functional model and common database for part information management.
  – With one keyword searching, the probability of successful searching with new part classification method can be improved by 7.68 times as compared to the old classification.

Collective System Laboratory, Washington State University, Pullman, WA.


• Contributions to Research Grant – Collective Innovation: Transforming the Realization of Complex Engineering Systems (NSF CMMI #0954447). PI: Dr. Panchal.
– Modeled the structural evolution of Open source 3D Printer, RepRap. Proposed genetic algorithm-based approach to design networks that match the community network of real-world open source products such as Drupal and Apache. Developed a Java applet to visualize the evolution of community networks.

– Developed an approach for building conceptual representations of agent-based models using SysML. Translated the conceptual model in SysML to an executable program in Java.

Intelligent Control Laboratory, Xi’an Jiao Tong University, Xi’an, China.

**Graduate Research Assistant**, Sep. 2007 - Mar. 2010

- Theoretical and Experimental Studies on Design of Numerical Controlled Die Spotting Press. Sponsored by: Chery Automobile Company
  - Developed and simulated dual-cylinder, synchronous electro-hydraulic servo control system and position servo control system for the press.

TEACHING EXPERIENCE

**Lab Instructor & Teaching Assistant**

**ME475: Automatic Control Systems**, Fall 2014, Purdue University.

- Lab Instructor for a section of 12 on-campus students. Was responsible for lab lectures throughout the semester. Oversaw the use of lab instruments and LabView-FPGA programming, graded lab reports, designed lab quizzes and guided final projects.
Guest Lecturer

**ME597: Decision Making in Engineering Systems Design**, Fall 2014, Purdue University.

- Presented two guest lectures for a graduate course consisting of 13 on-campus students and 1 distance-learning student from the University of Oklahoma. Co-authored course content and designed homework.

INDUSTRIAL EXPERIENCE

Heavy Industry & New-Tech Limited Company of Lanshi Group, Lanzhou, China


- Collaborated with other engineers to design a monitoring system of process status and operations for a 45MN high-speed forging hydraulic-press.

AWARDS


A2. Purdue COE Conference Travel Funds for excellent Ph.D. candidate, awarded by College of Engineering (COE), Purdue, 2014.

A3. ASME CIE Travel Stipend Award, awarded by ASME Computers and Information in Engineering division, 2013.

A4. PGSG Travel Grant, awarded by Purdue Graduate Student Government (PGSG), 2013.

A5. NSF CMMI Conference Fellowship, awarded by NSF division of Civil, Mechanical and Manufacturing Innovation (CMMI), 2012.

A7. **GPSA Travel Grant**, awarded by Graduate and Professional Student Association (GPSA) at Washington State University, 2011.

A8. **Excellent Performance Award** for volunteers in the International Orientation Week, awarded the Office of International Programs at Washington State University, 2011.

A9. **Si-Yuan Innovation Scholarship**, awarded by Xi’an Jiao Tong University, 2007 - 2010.

A10. **Outstanding Graduate Student Leader Award**, awared by School of Mechanical Engineering at Xi’an JiaoTong University, 2008

**ADDITIONAL CONFERENCE PRESENTATION AND POSTERS**


SCIENTIFIC COMMUNITY SERVICE


PROFESSIONAL MEMBERSHIPS

- Student Member of IEEE Computer Society, since 2014.
- Student Member of IEEE System, Man and Cybernetic Society, since 2014.
- Member, Golden Key International Honour Society, since 2013.
- Student Member of Society for Industry and Applied Mathematics (SIAM), since 2012.
- Student Member of American Society of Mechanical Engineers (ASME), since 2010.