Cluster Damage Robustness Analysis and Space Independent Community Detection in Complex Networks

A Thesis Submitted for the Degree of
Doctor of Philosophy

By

EMIL GEGOV

Mechanical Engineering Department
School of Engineering and Design
Brunel University
Uxbridge, UB8 3PH

September 2012

Abstract

This thesis investigates the evolution of two very different complex systems using network theory. This multi-disciplinary technique is widely used to model and analyse vastly diverse systems of multiple interacting components, and therefore, it is applied in this thesis to study the complexity of the systems. This complexity is rooted in the components' interactions such that the whole system is more than the sum of all the individual parts. The first novelty in this research is the proposal of a new type of structural perturbation, cluster damage, for measuring another dimension of network robustness. The second novelty is the first application of a community detection method, which uncovers space-independent communities in spatial networks, to airport and linguistic networks.

A critical property of complex systems – robustness – is explored within a partial model of the Internet, by demonstrating a novel perturbation strategy based on the iterative removal of clusters. The main contribution of this theoretical case study is the methodology for cluster damage, which has not been investigated by literature on the robustness of complex networks. The model, part of the Internet at the Autonomous System level, only serves as a domain where the novel methodology is demonstrated, and it is chosen because the Internet is known to be robust due to its distributed (non-centralised) nature, even though it is often subjected to large perturbations and failures.

The first applied case study is in the field of air transportation. Specifically, it explores the topology and passenger flows of the United States Airport Network (USAN) over two decades. The network model consists of a time-series of six network snapshots for the years 1990, 2000 and 2010, which capture bi-monthly passenger flows among US airports. Since the network is embedded in space, the volume of these flows is naturally affected by spatial proximity, and therefore, a model (recently proposed in the literature) accounting for this phenomenon is used to identify the communities of airports that have particularly high flows among them, given their spatial separation.

The second applied case study – in the field of language acquisition – investigates the word co-occurrence network of children, as they develop their linguistic abilities at an early age. Similarly to the previous case study, the network model consists of six children and three discrete developmental stages. These networks are not embedded in physical space, but they are mapped to an artificial semantic space that defines the semantic distance between pairs of words. This novel approach allows for an additional dimension of network information that results in a more complete dataset. Then, community detection identifies groups of words that have particularly high co-occurrence frequency, given their semantic distance.

This research highlights the fact that some general techniques from network theory, such as network modelling and analysis, can be successfully applied for the study of diverse systems, while others, such as community detection, need to be tailored for the specific system. However, methods originally developed for one domain may be applied somewhere completely new, as illustrated by the application of spatial community detection to a non-spatial network. This underlines the importance of inter-disciplinary research.

Declaration

The research presented in this thesis is the original work of the author except where otherwise specified, or where acknowledgements are made by references. This project was carried out at the School of Engineering and Design, Brunel University, under the supervision of Dr. M. Atherton and Prof. F. Gobet.

The work has not been submitted for another degree or award to any other institution.

Acknowledgements

I would like to thank my supervisors Dr. M. Atherton and Prof. F. Gobet for four years of continuous guidance and inspiration.

I would also like to thank my parents, Alex and Juliet, for encouraging me to do a Ph.D. and helping me get this far.

Finally, I am thankful to my partner Nadia for supporting me and believing in me all the time.

Table of Contents

Abstract ..... **i**
Declaration ..... **iii**
Acknowledgements ..... **iv**
Table of Contents ..... **v**
List of Figures ..... **viii**
List of Tables ..... **xiii**
Nomenclature ..... **xiv**
Glossary ..... **xvi**
Chapter 1 Introduction **1**
Chapter 2 Literature Review **8**
2.1 Network Theory 8
2.1.1 Definitions 9
2.1.2 Topology Classes 9
2.1.3 Community Structure 14
2.2 Robustness in Systems and Networks 16
2.2.1 Stability and Robustness 16
2.2.2 Robustness in Biological Systems 19
2.2.3 Robustness in Complex Networks 21
2.2.4 Mechanisms of Network Robustness 22
2.2.5 Cascades 25
2.3 Air Transportation 27
2.4 Language Acquisition 28
2.5 Research Problems 31
2.6 Research Solutions 34
2.7 Research Methodology 34
2.7.1 Cluster Damage 35
2.7.2 Detailed Network Modelling 35
2.7.3 Space-Independent Community Structure 36

2.7.4 Dynamics and Evolution 37
2.8 Summary 38
Chapter 3 Robustness to Cluster Damage 39
3.1 Methodology 39
3.2 The Internet 41
3.3 Results 42
3.4 Summary 47
Chapter 4 Air Transportation Networks 49
4.1 Domain Description 49
4.2 Data Set 50
4.3 Methodology 51
4.3.1 Network Modelling 52
4.3.2 Network Analysis 53
4.4 Results 56
4.4.1 Network Parameters 57
4.4.2 Community Structure 60
4.5 Summary 63
Chapter 5 Language Acquisition Networks 64
5.1 Domain Description 64
5.2 Data Sets 65
5.2.1 Mothers and Children 65
5.2.2 Model Of Syntax Acquisition In Children 66
5.2.3 Baseline 71
5.3 Methodology 74
5.3.1 Filtering and Reduction 74
5.3.2 Construction of Networks 75
5.3.3 Network Analysis 76
5.4 Results 82
5.4.1 Network Parameters 82
5.4.2 Community Structure 104
5.5 Summary 107
Chapter 6 Discussion 108

6.1 Air Transportation Networks 108
6.1.1 Network Parameters 108
6.1.2 Community Structure 110
6.2 Language Acquisition Networks 122
6.2.1 Network Parameters 122
6.2.2 Community Structure 125
6.3 Comparison and Generalities 134
6.4 Summary 137
Chapter 7 Conclusion 138
7.1 Research Contributions 138
7.2 Theoretical Implications 141
7.3 Future Work 143
Bibliography 145
Publications 154
Appendix A USAN Community Structure by Expert's Method 155
Appendix B Children's Community Structure by Expert's Method 158
Appendix C USAN Community Structure by Newman's Method 168
Appendix D Children's Community Structure by Newman's Method 171

List of Figures

Fig. 1.1. Client-server model (ibiblio, n.d.). 2
Fig. 1.2. Peer-to-peer model (ibiblio, n.d.)..... 2
Fig. 2.1. Random network (Albert, Jeong and Barabási, 2000). 10
Fig. 2.2. *P(k)* of random network (Jeong *et al.*, 2000). 10
Fig. 2.3. Evolution of a random graph consisting of 20 nodes: (a) initially the network is isolated; (b) with $p = 0.1$ there are three isolated nodes; (c) with $p = 0.2$ there is one isolated node (Australian National University, n.d.). 11
Fig. 2.4. Random rewiring procedure for differentiating between a regular ring lattice and a random network, without changing the number of nodes or links in the network (Watts and Strogatz, 1998). 12
Fig. 2.5. Scale-free network (Albert, Jeong and Barabási, 2000). 13
Fig. 2.6. *P(k)* of scale-free network (Jeong *et al.*, 2000)..... 13
Fig. 2.7. Global dynamics emerging from local interactions. 18
Fig. 2.8. Common methodology for air transportation and language acquisition based on network theory. 34
Fig. 3.1. Network of Autonomous Systems. 41
Fig. 3.2. Links *E* as a function of nodes *N* for 74 clusters. 43
Fig. 3.3. Average geodesic *L* as a function of nodes *N* for 74 clusters. 44
Fig. 3.4. Giant Connected Component *GCC* as a function of nodes *N* for 74 clusters..... 45
Fig. 3.5. Modularity as a function of the number of clusters. 45
Fig. 3.6. Robustness to cluster attacks as a function of the number of clusters. .. 46
Fig. 3.7. Robustness to cluster attacks as a function of modularity. 47
Fig. 4.1. US macro-regions and major airports in 2010 (Mackun *et al.*, 2011). .. 53
Fig. 4.2. In-degree probability distribution for Nov-Dec 2010 snapshot of the USAN. 54
Fig. 4.3. Ranked weight (normalised frequency of passengers) *W(r)* for Nov-Dec 2010 snapshot of the USAN. 55
Fig. 4.4. Number of airports as a function of time. 57
Fig. 4.5. Number of connections as a function of time. 57
Fig. 4.6. Number of connected airports as a function of time. 57
Fig. 4.7. Average connections per airport as a function of time. 57
Fig. 4.8. Average geodesic length as a function of time. 58
Fig. 4.9. Clustering coefficient as a function of time. 58
Fig. 4.10. Probability(0 connections in) as a function of time. 58
Fig. 4.11. Probability(0 connections out) as a function of time. 58
Fig. 4.12. Probability(1 connection in) as a function of time. 59
Fig. 4.13. Probability(1 connection out) as a function of time. 59
Fig. 4.14. Scaling factor a_{in} as a function of time. 59

Fig. 4.15. Scaling factor a_{out} as a function of time. 59
Fig. 4.16. Exponent n_{in} as a function of time. 60
Fig. 4.17. Exponent n_{out} as a function of time. 60
Fig. 4.18. Scaling factor *b* as a function of time. 60
Fig. 4.19. Intercept *c* as a function of time 60
Fig. 5.1. MOSAIC network after one appearance of *did he go* (Freudenthal, Pine and Gobet, 2006). 67
Fig. 5.2. MOSAIC network after three appearances of *did he go* (Freudenthal, Pine and Gobet, 2006). 68
Fig. 5.3. MOSAIC network after three appearances of *did he go* and two appearances of *he walks* (Freudenthal, Pine and Gobet, 2006) 69
Fig. 5.4. MOSAIC network with a generative link (arrow) (Freudenthal, Pine and Gobet, 2006). 71
Fig. 5.5. Steps of building the baseline co-occurrence networks. 72
Fig. 5.6. Simple word co-occurrence network. 75
Fig. 5.7. In-degree probability distribution for Ann stage 3 network. 79
Fig. 5.8. Ranked weight (normalised frequency) distribution *W(r)*. 80
Fig. 5.9. Analysis flowchart. 81
Fig. 5.10. Summary plot for *MLU*. 85
Fig. 5.11. Summary plot for *N*. 86
Fig. 5.12. Summary plot for *E*. 86
Fig 5.13. Summary plot for *GCC*. 87
Fig. 5.14. Summary plot for <*k*>. 88
Fig. 5.15. Summary plot for *L*. 89
Fig. 5.16. Summary plot for *C*. 90
Fig. 5.17. Summary plot for parameter *p* of the in-degree distribution. 94
Fig. 5.18. Summary plot for parameter *p* of the out-degree distribution. 95
Fig. 5.19. Summary plot for parameter *a* of the in-degree distribution. 96
Fig. 5.20. Summary plot for parameter *a* of the out-degree distribution. 97
Fig. 5.21. Summary plot for parameter *n* of the in-degree distribution. 97
Fig. 5.22. Summary plot for parameter *n* of the out-degree distribution. 98
Fig. 5.23. Summary plot for parameter *a*. 102
Fig. 5.24. Summary plot for parameter *n*. 103
Fig. 5.25. Community structure in aggregated children in stage 1. 106
Fig. 5.26. Community structure in aggregated children in stage 2. 107
Fig. 5.27. Community structure in aggregated children in stage 3. 107
Fig. 6.1. Normalised Mutual Information (NMI) of consecutive network snapshots. 112
Fig. 6.2. 1990 migration patterns among the four macro-regions: West, Midwest, Northeast and South. 118
Fig. 6.3. 2000 migration patterns among the four macro-regions: West, Midwest, Northeast and South. 119

Fig. 6.4. 2010 migration patterns among the four macro-regions: West, Midwest, Northeast and South. 119
Fig. 6.5. Community structure in USAN in NOV-DEC 2010 identified using Newman's method (same as Fig. C.54 in Appendix). 120
Fig. 6.6. Community structure in USAN in NOV-DEC 2010 identified using Expert's method (same as Fig. A.18 in Appendix). 120
Fig. 6.7. Normalised Variation of Information (NVI) among community structure identified using Expert's and Newman's null models. 122
Fig. 6.8. Normalised Mutual Information (NMI) of consecutive stage networks of individual children. 129
Fig. 6.9. Community structure in Gai 3 identified using Newman's method (same as Fig. D.72 in Appendix). 132
Fig. 6.10. Normalised Variation of Information (NVI) among community structure identified using Expert's and Newman's null models. 134
Fig. 6.11. Comparison and generalities of air transportation and language acquisition using network theory. 135
Fig. A.1. JAN-FEB 1990 community structure with Expert. 155
Fig. A.2. MAR-APR 1990 community structure with Expert. 155
Fig. A.3. MAY-JUN 1990 community structure with Expert. 155
Fig. A.4. JUL-AUG 1990 community structure with Expert. 155
Fig. A.5. SEP-OCT 1990 community structure with Expert. 155
Fig. A.6. NOV-DEC 1990 community structure with Expert. 155
Fig. A.7. JAN-FEB 2000 community structure with Expert. 156
Fig. A.8. MAR-APR 2000 community structure with Expert. 156
Fig. A.9. MAY-JUN 2000 community structure with Expert. 156
Fig. A.10. JUL-AUG 2000 community structure with Expert. 156
Fig. A.11. SEP-OCT 2000 community structure with Expert. 156
Fig. A.12. NOV-DEC 2000 community structure with Expert. 156
Fig. A.13. JAN-FEB 2010 community structure with Expert. 156
Fig. A.14. MAR-APR 2010 community structure with Expert. 156
Fig. A.15. MAY-JUN 2010 community structure with Expert. 157
Fig. A.16. JUL-AUG 2010 community structure with Expert. 157
Fig. A.17. SEP-OCT 2010 community structure with Expert. 157
Fig. A.18. NOV-DEC 2010 community structure with Expert. 157
Fig. B.19. Ann 1 community structure with Expert. 158
Fig. B.20. Ara 1 community structure with Expert. 159
Fig. B.21. Bec 1 community structure with Expert. 159
Fig. B.22. Car 1 community structure with Expert. 160
Fig. B.23. Dom 1 community structure with Expert. 160
Fig. B.24. Gai 1 community structure with Expert. 161
Fig. B.25. Ann 2 community structure with Expert. 161
Fig. B.26. Ara 2 community structure with Expert. 162

Fig. B.27. Bec 2 community structure with Expert. 162
Fig. B.28. Car 2 community structure with Expert. 163
Fig. B.29. Dom 2 community structure with Expert. 163
Fig. B.30. Gai 2 community structure with Expert. 164
Fig. B.31. Ann 3 community structure with Expert. 164
Fig. B.32. Ara 3 community structure with Expert. 165
Fig. B.33. Bec 3 community structure with Expert. 165
Fig. B.34. Car 3 community structure with Expert. 166
Fig. B.35. Dom 3 community structure with Expert. 166
Fig. B.36. Gai 3 community structure with Expert. 167
Fig. C.37. JAN-FEB 1990 community structure with Newman. 168
Fig. C.38. MAR-APR 1990 community structure with Newman. 168
Fig. C.39. MAY-JUN 1990 community structure with Newman. 168
Fig. C.40. JUL-AUG 1990 community structure with Newman. 168
Fig. C.41. SEP-OCT 1990 community structure with Newman. 169
Fig. C.42. NOV-DEC 1990 community structure with Newman. 169
Fig. C.43. JAN-FEB 2000 community structure with Newman. 169
Fig. C.44. MAR-APR 2000 community structure with Newman. 169
Fig. C.45. MAY-JUN 2000 community structure with Newman. 169
Fig. C.46. JUL-AUG 2000 community structure with Newman. 169
Fig. C.47. SEP-OCT 2000 community structure with Newman. 169
Fig. C.48. NOV-DEC 2000 community structure with Newman. 169
Fig. C.49. JAN-FEB 2010 community structure with Newman. 170
Fig. C.50. MAR-APR 2010 community structure with Newman. 170
Fig. C.51. MAY-JUN 2010 community structure with Newman. 170
Fig. C.52. JUL-AUG 2010 community structure with Newman. 170
Fig. C.53. SEP-OCT 2010 community structure with Newman. 170
Fig. C.54. NOV-DEC 2010 community structure with Newman. 170
Fig. D.55. Ann 1 community structure with Newman. 171
Fig. D.56. Ann 2 community structure with Newman. 172
Fig. D.57. Ann 3 community structure with Newman. 172
Fig. D.58. Ara 1 community structue with Newman. 173
Fig. D.59. Ara 2 community structure with Newman. 173
Fig. D.60. Ara 3 community structure with Newman. 174
Fig. D.61. Bec 1 community structure with Newman 174
Fig. D.62. Bec 2 community structure with Newman 175
Fig. D.63. Bec 3 community structure with Newman 175
Fig. D.64. Car 1 community structure with Newman. 176
Fig. D.65. Car 2 community structure with Newman. 176
Fig. D.66. Car 3 community structure with Newman. 177
Fig. D.67. Dom 1 community structure with Newman. 177
Fig. D.68. Dom 2 community structure with Newman. 178

Fig. D.69. Dom 3 community structure with Newman. 178
Fig. D.70. Gai 1 community structure with Newman. 179
Fig. D.71. Gai 2 community structure with Newman. 179
Fig. D.72. Gai 3 community structure with Newman. 180

List of Tables

Table 2.1. Parameters of interest in linguistic networks. 29
Table 3.1. Summary of different network partitions into clusters. 42
Table 5.1. Age of children at start of stages in years;months.days. 66
Table 5.2. Individual network parameters. 83
Table 5.3. Correlations of individual network parameters. 91
Table 5.4. Parameters of the best-fit of the in-degree and out-degree distributions. 92
Table 5.5. Correlations of best-fit parameters of the in-degree and out-degree distributions. 99
Table 5.6. Parameters of the best-fit of the ranked weight distribution. 100
Table 5.7. Correlations of best-fit parameters of the ranked weight distribution. 103
Table 6.1. Atlanta's connections. 116
Table 6.2. In-migration, representing the number of people migrating to specific US states in 2009-2010 (United States Census Bureau, n.d.). 117
Table 6.3. Main findings in MOSAIC and children. 124
Table 6.4. Number of words in children's networks and number of common words present in all three stages. 128
Table 6.5. Main findings in air transportation and children's language acquisition. 136

Nomenclature

Symbol	Description
A_{ij}	Adjacency matrix
b	Scaling factor in $W(r) = bLn(r) + c$
C	Average clustering coefficient in a network
c	Intercept in $W(r) = bLn(r) + c$
C_i	Clustering coefficient of node i
D	Network diameter is the length of the longest path from the set of all shortest paths between all pairs of nodes
DC	Degree Centralisation measures to what extent the links are centralised on a small number of high-degree nodes
d_{ij}	Distance between nodes i and j
D_{ij}	Distance matrix
E	Total number of network links
E_i	Number of edges among n_i nodes
$f(d_{ij})$	Deterrence function defining the expected level of interaction between nodes i and j that are separated by a distance d_{ij}
GCC	Giant Connected Component is the largest connected part of the network
$<k>$	Average degree (number of links connected to a node) in a network
k_i	Strength or degree of node i
L	Average geodesic length is the average length of all shortest paths between all pairs of nodes
$Length$	Total number of characters, words or utterances within a linguistic data set
m	Total links weight in the network
M	A constant (70,000)
MLU	Mean Length of Utterance is the average number of words in an utterance
N	Total number of network nodes

n	Exponent in $P(k) = ak^n$
NCC	Number of Connected Components is the number of network components that are disconnected from one another
n_i	Number of neighbours of node i
N_i	Importance (typically the strength) of node i
NN	Average nearest-neighbour degree of a node is the average degree of the nodes that are connected to the node
p	Probability of a node with 0 links, i.e. $P(0)$
$P(b)$	Probability distribution of a randomly chosen node with betweenness centrality b
$P(f)$	Probability distribution of a randomly chosen node with frequency f
$P(k)$	Probability distribution of a randomly chosen node with degree k
P_{ij}	Matrix of expected link weights between nodes i and j
Q	Modularity of network partition into communities
q	Probability of a node with 1 link, i.e. $P(1)$
Q_{ij}	Modularity matrix
R	Correlation coefficient measuring the quality of a best-fit
W	Length of a phrase in words
$W(r)$	Ranked weight distribution

Glossary

Term	Description
AS	Autonomous System
ATL	Atlanta
BTS	Bureau of Transportation Statistics
CHREST	Chunk Hierarchy and REtrieval STructure
DFS	Depth First Search
DISCO	extracting DIStributionally related words using CO-occurrences
EPAM	Elementary Perceiver And Memoriser
ER	Erdös and Rényi
Hsp90	Heat shock protein 90
IRS	Internal Revenue Service
LA	Los Angeles
MOSAIC	Model Of Syntax Acquisition In Children
NCP	Node Creation Probability
NG	Newman and Girvan's general community structure
NMI	Normalised Mutual Information
NVI	Normalised Variation of Information
SEM	Standard Error of the Mean
Spa	Expert's space-independent community structure
UG	Universal Grammar
USAN	United States Airport Network

Chapter 1
Introduction

This chapter introduces complex networks and their generality for modelling complex systems. Two applied case studies are briefly described, providing the motivations behind this thesis. The aims are outlined.

A common approach to dealing with complex systems is to make use of network theory to simulate symmetric or, more generally, asymmetric relations among discrete objects. Transportation, the Internet, mobile phone, power grid, social, and neural networks are just some examples of complex systems where network theory has been successfully applied. In fact, complex systems are often modelled as complex networks, i.e. graphs with non-trivial topological characteristics, because this provides a powerful abstraction that can eliminate the unnecessary complexity of the system while maintaining the key properties and interactions. Some networks have nodes and links arranged in physical space, in which case the topology alone does not contain all the necessary information to describe the network. Spatial constraints typically affect the structure and properties of spatial networks, and therefore, it is important to consider the physical distance between interacting nodes.

Complex networks typically consist of numerous nodes and links, where the whole is more than the sum of all the parts. This means that the interactions emerging from network structure are driving the network function. For example, in computer networks two popular configurations are the client-server (Fig. 1.1) and the peer-to-peer (Fig. 1.2) architectures. In the former, there are relatively few *server* nodes, which are very well connected and provide most of the services, which are required by the less well connected *client* nodes. In the latter, communication is established between two nodes by setting up a path between them, which consists of other nodes and links in the network. In this architecture, each node has roughly the same number of links. The client-server and peer-to-peer structures are very different in many ways. The former is a centralised

system where the network is completely dependent on the operation of one key node – the server, whereas in the latter all nodes are equal and if any one fails then the network re-configures itself. Therefore, it is more robust to node or link failure. In the context of complex systems, robustness (Atherton and Bates, 2005) is generally defined as the ability of a system to maintain its function in the presence of disturbances. Robustness is very important for both natural and engineered systems because all systems experience some sort of change or disturbance over their lifetime, and the better they are prepared to deal with this, the longer they will operate as required.

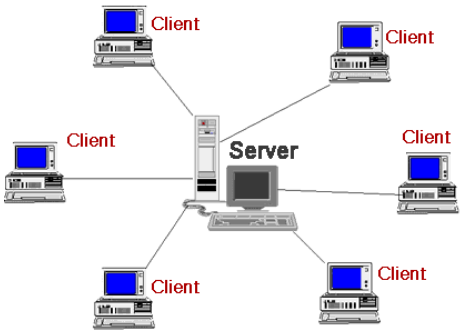

Fig. 1.1. Client-server model (ibiblio, n.d.).

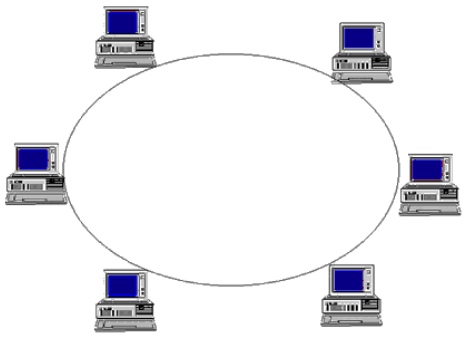

Fig. 1.2. Peer-to-peer model (ibiblio, n.d.).

Evolution-based modelling of a complex network can be defined as a process that takes as input some specific network data, and returns a complete network model of these data. In other words, all local interactions between pairs of nodes for some time period are mapped onto a global network model representing the structure and dynamics of the real complex network, for the period under study. In this way, it is possible to determine how the network is evolving over time, in terms of its topology and interactions. A network is essentially a set of nodes and links, so if data are in the form of node pairs (in most cases they are), it is easy to

build a network directly from the data: for each pair, insert a directed link from the source node to the target node, labelling the link with the given weight (representing strength of interaction), and hence, a snapshot of the evolving network is generated for each time slice of data. However, some complex systems may be more accurately described by networks containing nodes that interact not only in pairs but also in groups of multiple nodes. This network generalisation where a link connects more than two nodes is called a hypernetwork.

The air transportation network of a country or region is a critical component of its infrastructure, with huge impacts on its economy, the transportation of people, cargo, and mail, as well as the potential for propagating negative effects, such as globally spreading diseases (Guimerà *et al.*, 2005). Therefore, researchers from multiple disciplines have recently shown a lot of interest in this field, and with an abundance of available data, have made attempts to model and to analyse airport networks. This provides an understanding of how these networks operate; the critical airport nodes that connect otherwise distant locations; whether there are any naturally occurring community structures; and how the networks evolve over time. Depending on several key factors, such as geographical area, population, economic growth, tourism, and trade, the national airport network of a country may grow and change its topology considerably over time, driven mainly by the actions of the airlines that seek to increase their short-term profits. This means that an airport network is constantly developing, or more precisely, evolving in response to the growing demands of the people using the network either directly as passengers or indirectly as consumers of transported goods. Globalisation and the introduction of more long-distance direct connections between distant regions does however present a serious threat to public health, as a small outbreak of a disease in a remote region may quickly turn into a global epidemic. Hence, the aim of some transportation network models is to predict how a disease would propagate through the network and to identify critical regions that need to be isolated in order to prevent further spread.

The ability of humans to communicate effectively through the use of a common language is remarkable. Even more impressive is the fact that young children,

when learning their first language, are able to pick it up so well, even in the presence of noisy input. When a child produces an incorrect utterance, there is rarely any corrective feedback, so it is up to the child to filter those out over time. Two opposing approaches try to explain this remarkable capability of children, and specifically, their syntax acquisition. Universal Grammar (UG) (Chomsky, 1957) argues that certain rules and parameters for language are hard-wired in each and every one of us, and the process of acquisition is simply a tuning of those variables. A key argument for UG is *the poverty of the stimulus argument*: the input received by children contains numerous errors, false starts, unfinished sentences, and thus is not sufficient for inferring the rules of language. UG has generally been accepted as the *de facto* theory for syntax acquisition, but alternative theories (such as distributional analysis), loosely based on the *empiricist* tradition and developed in the 1990s, are gaining increasingly more support in recent years. There is increasing evidence that the environment provides much more information than had been assumed by Chomsky, and a number of simulation models have shown that much grammatical knowledge can be learnt from child-directed speech (Freudenthal *et al.*, 2007; Redington, Chater and Finch, 1998). If one does not subscribe to UG's assumptions of innate abstract knowledge, one has to identify the specific mechanisms employed when young children acquire their first language. This can be achieved by using models, such as *Model Of Syntax Acquisition In Children* (MOSAIC) (Freudenthal *et al.*, 2007), which are trained with maternal utterances, and then produce utterances that can be directly compared to children's utterances. By examining the quality of the obtained results, the aim is to identify which mechanisms account for the empirical data. However, data sets of utterances are typically very large, noisy, and difficult to compare directly. Therefore, a worthwhile approach is to extract key characteristics of corpora first. This is what network modelling offers. It holds the key to embed all the information contained in a data set in a network, which can be analysed and compared with similar networks. It is important to identify which network parameters represent a useful statistic of the raw data, so they can be extracted from the network for analysis and cross-comparison.

When children acquire their native language they have to deal with considerable complexity as the input is potentially noisy and inconsistent (Freudenthal, Pine and Gobet, 2006). Conventional linguistic theory has attempted to explain this phenomenon with the *nativist* theory, which proposes that children are born with a domain-specific knowledge of language (Pinker, 1984; Chomsky, 1981). However, recent work using computational modelling techniques has found that the level of detail that can be extracted by analysing the statistical properties of language is far greater than originally assumed by the nativist approach (Cartwright and Brent, 1997; Elman, 1993). In addition, research on infants' distributional learning abilities has shown that children pick-up the statistical properties of the language they are exposed to (Gomez and Gerken, 1999; Saffran, Aslin and Newport, 1996). In other words, it is possible that some of the phenomena considered as evidence for innate linguistic knowledge in children can also be explained by the children's distributional learning of statistical properties of the input language. The standard approach to explore this possibility is to test computational models that simulate linguistic development by *distributional analysis* of child-directed language. The main problem with this approach is that models need to simulate more directly the tasks carried out by children, using more realistic input data, thereby more precisely matching empirical observations (Christiansen and Chater, 2001).

This research has four main aims:

1. The first aim is to test a novel perturbation strategy, cluster damage, in a partial model of the Internet, in order to study a new dimension of network robustness. This is motivated by the lack of existing methods that investigate network robustness to the failure of various structural components.

2. The second aim is to develop a comprehensive model of the evolving US airport network. This is motivated by the fact that this network is critical for the mobility of millions of people, and hence, it has a huge economic and health impact.

3. The third aim is to investigate the extent to which MOSAIC is able to simulate language acquisition in children. It is hypothesised that there will be significant similarity between its output and children' utterances in terms of linguistic network models. This aim is motivated by the increasing evidence suggesting that the environment provides much more information than originally assumed by Chomsky, casting doubts on his well-established Universal Grammar theory.

4. The fourth aim is to identify a more cohesive community structure in air transportation and language acquisition networks. This is motivated by the inaccuracy of general community detection methods that are widely used in the literature.

The following is a brief outline of the remaining chapters.

Chapter 2 reviews the complex networks literature in terms of network theory; robustness in systems and networks; air transportation; and language acquisition. Research problems and solutions are presented. A research methodology section discusses the solutions in more detail.

Chapter 3 describes a novel approach to investigate network robustness by damaging entire clusters of nodes. A partial network model of the Internet at the Autonomous System level is presented. The robustness of the network to node and cluster damage is discussed.

Chapter 4 presents the first applied case study on air transportation networks. US air travel and migration are introduced. An evolution-based model of the network is proposed. The general properties and the community structure of the network are presented.

Chapter 5 presents the second applied case study on language acquisition networks. The main types of linguistic networks are introduced. A development-based model of the networks is proposed. The general properties and the community structure of the networks are presented.

Chapter 6 discusses the general properties and the community structure of air transportation networks and language acquisition networks. The two applied case studies are compared for possible generalities.

Chapter 7 concludes the thesis. The major contributions to research in the field of complex networks are summarised. Theoretical implications for each individual case study and for complex networks in general are drawn based on the analysis of the results. Recommendations for future work are provided.

Chapter 2
Literature Review

This chapter reviews the complex networks literature on network theory, robustness, air transportation and language acquisition. In addition, the research problems and solutions are presented. The chapter concludes with a research methodology that describes the solutions to the research problems in more detail.

In recent years, the availability of huge data sets has enabled researchers across many disciplines to model and understand exceedingly complex systems by using network modelling and analysis. For example, biological networks such as metabolic (Morine *et al.*, 2009) and gene co-expression (Carter *et al.*, 2004); technological networks such as the Internet (Alderson and Willinger, 2005), and the power grid (Carreras *et al.*, 2002); and social networks such as friendship (Girvan and Newman, 2002), and co-authorship (Barthélemy *et al.*, 2005), have been widely studied and interesting patterns have emerged. This research has shown that network modelling provides a powerful abstraction of networked complex systems in the real-world that is able to strip away the detail of individual systems, while retaining the core information, such as network structure (topology) and dynamics (link weights). Hence, it is possible to model the evolution of complex systems at a high level, and to identify common properties, as well as trends, over time. This leads to a better understanding of complex systems, with potential benefits to many areas, such as medicine, technology, and the social sciences, to name a few.

2.1 Network Theory

Network theory, also known as graph theory, is a powerful tool for analysing complex systems. One of the benefits of network modelling is that it facilitates inter-disciplinary research since systems from different domains are represented in the same way (as networks) and may be directly compared.

2.1.1 Definitions

This section defines the core concepts from network theory that are used throughout this thesis:

- A *network* is a set of nodes *N* and a set of links *E* that connect pairs of nodes of *N*.
- The *Giant Connected Component* (GCC) of a network is the largest connected subnetwork.
- A *cluster* of a network is a connected subnetwork.
- *Robustness* is the ability of a network to maintain its function in the presence of disturbances.
- A *community* of a network is a cluster with particularly strong internal and weaker external connections.

2.1.2 Topology Classes

In terms of the structure of connections, there are four classes of complex networks: regular, random, small-world and scale-free. Topology is defined by the connectivity distribution *P(k)*, giving the probability that a node in the network is connected to *k* other nodes.

- **Regular Networks**

In order to describe regular networks, an essential concept of complex networks needs to be introduced. The clustering coefficient C_i of a node i is the average fraction of pairs of neighbours n_i of i that are also neighbours of each other (Wang and Chen, 2003), e.g. in a friendship network, this would be the probability that two of your friends are also friends themselves:

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

Where E_i is the number of edges that actually exist among these n_i nodes from the total possible number $n_i(n_i - 1)/2$. Then, the clustering coefficient C of the whole

network is the average of all the clustering coefficients for all nodes. *C* is always less than or equal to 1, and equals 1 if and only if the network is globally coupled, i.e. every node is connected to every other node. A popular regular network model is the nearest-neighbour coupled network (also known as a lattice), which is a regular graph where every node is linked to a few of its nearest neighbours (leftmost network in Fig. 2.4). The clustering coefficient of this type of network is approximately ¾, which means that lattices are typically highly clustered (Wang and Chen, 2003). However, their diameter is large. The diameter *d* is defined as the largest number of links that need to be traversed to get from one node to another.

- **Random Networks**

The complete opposite of a regular network is a random graph (Fig. 2.1), studied first by Erdös and Rényi (ER) in 1959 (Erdös and Rényi, 1959). This network has a *P(k)*, which follows a Poisson distribution (Fig. 2.2), i.e. *P(k)* peaks at an average number of links <*k*> and decays exponentially for large or small *k* (<*k*> is standard notation for average node degree in network theory). This means that most nodes have <*k*> links and the rest have close to <*k*> links.

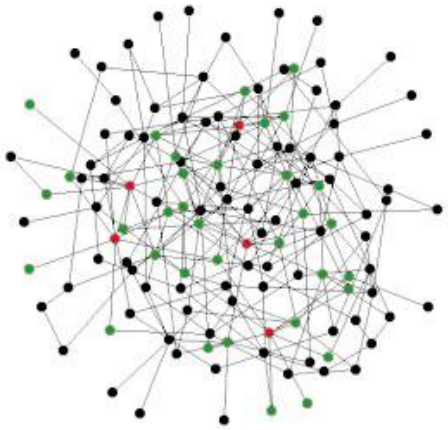

Fig. 2.1. Random network (Albert, Jeong and Barabási, 2000)**.**

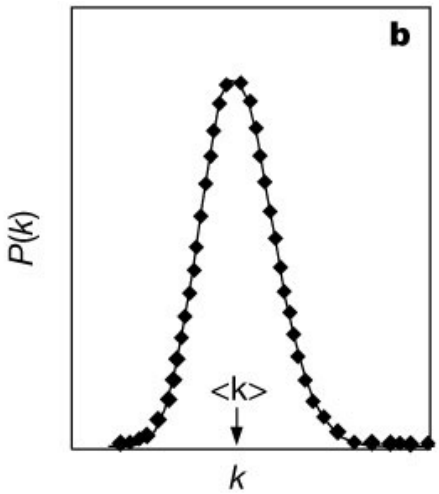

Fig. 2.2. *P(k)* of random network (Jeong *et al.*, 2000)**.**

A random graph is generated as follows (Fig. 2.3): by starting with a set of *N* isolated nodes (there are no links between nodes), one connects each pair of nodes with probability *p*, forming a network with connectivity proportional to *p*.

The idea is to determine at what connectivity p a particular property of a graph will arise (Wang and Chen, 2003). Erdös and Rényi discovered that if p is greater than a specific threshold $t \sim (\ln N)/N$, then almost every random graph is connected, i.e. there are no isolated nodes. In a random network, the probability that a node has two neighbour nodes, which are connected, is no greater than the probability that two randomly chosen nodes are connected. Therefore, the clustering coefficient of a random graph is $p = \langle k \rangle / N$. Hence, since $(\ln N)/N$ decreases as N increases, large-scale random networks may be connected, but they do not show high clustering like regular networks.

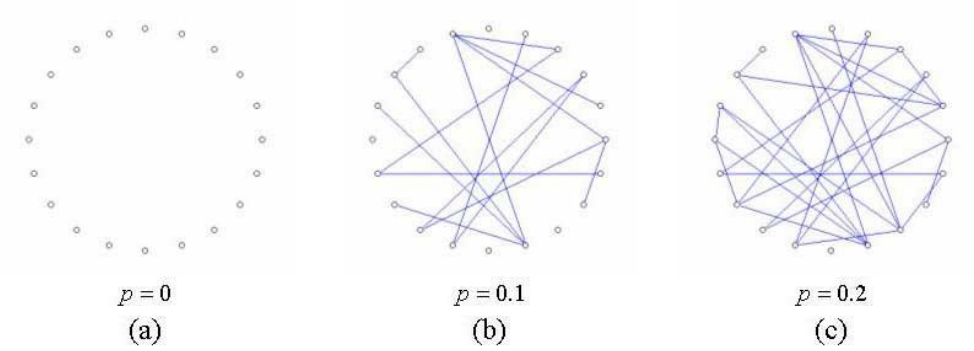

Fig. 2.3. Evolution of a random graph consisting of 20 nodes: (a) initially the network is isolated; (b) with p = 0.1 there are three isolated nodes; (c) with p = 0.2 there is one isolated node (Australian National University, n.d.)**.**

To investigate the structural robustness (to node removal) of random networks, Albert (Albert, Jeong and Barabási, 2000) observed the change in diameter when a small fraction f of nodes is removed. They found that the diameter increases monotonically with f. Furthermore, it was discovered that there is no significant difference between removing nodes at random (errors) and targeting the most connected nodes (attacks). Both of these results are due to the homogeneity of the network.

- **Small-World Networks**

The small-world phenomenon (also known as six degrees of separation) is the idea that any two people are connected by an average of six friendship links, i.e. the global social network (with over six billion individuals) has a geodesic length of around six (Milgram, 1967). By analogy with this phenomenon, a small-world

network is a network in which any two nodes are connected by a relatively short path (Jeong *et al.*, 2000), and hence, its diameter is small. As a consequence, infectious diseases are predicted to spread much more easily in a small world. A small-world network can be generated by the random rewiring of a regular network (Watts and Strogatz, 1998), where each node is rewired at random with probability *r* (Fig. 2.4). The key point is that for intermediate values of *r* this yields a small-world network with high clustering like a regular graph, yet with small diameter like a random graph. Examples of small-world networks include the neural network of the worm *Caenorhabditis elegans*, the power grid of the western United States, and the collaboration graph of film actors.

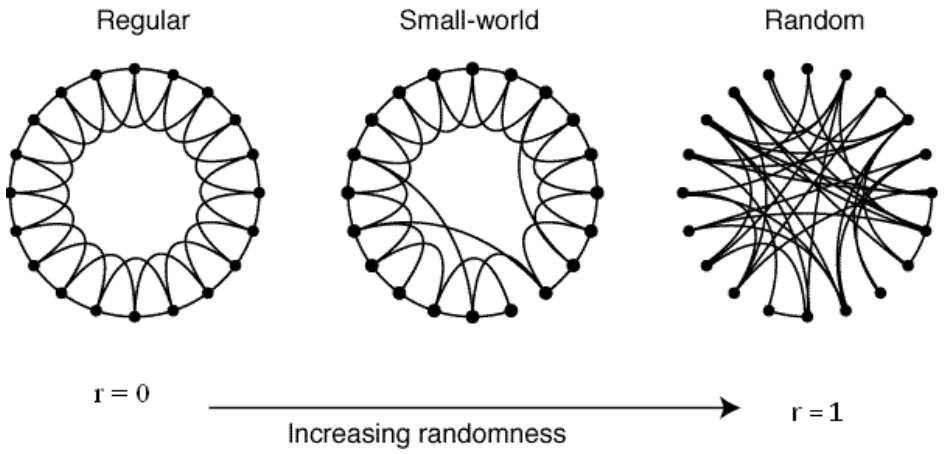

Fig. 2.4. Random rewiring procedure for differentiating between a regular ring lattice and a random network, without changing the number of nodes or links in the network (Watts and Strogatz, 1998)**.**

- **Scale-Free Networks**

The stability of complex systems is often attributed to redundant wiring of the network, i.e. by introducing more additional links the network connectivity is increased and hence, if a node is removed (along with its links), it is highly unlikely that the network will be affected since there are plenty of links left. However, error tolerance is not displayed by all redundant systems (Albert, Jeong and Barabási, 2000); it is shown mainly by a class of heterogeneously wired networks called scale-free networks (Fig. 2.5). These have a *P(k)*, which decays as a power-law (Fig. 2.6), so most nodes have few links but some nodes have many links.

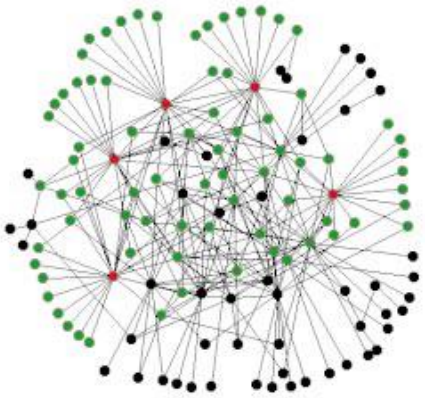

Fig. 2.5. Scale-free network (Albert, Jeong and Barabási, 2000)**.**

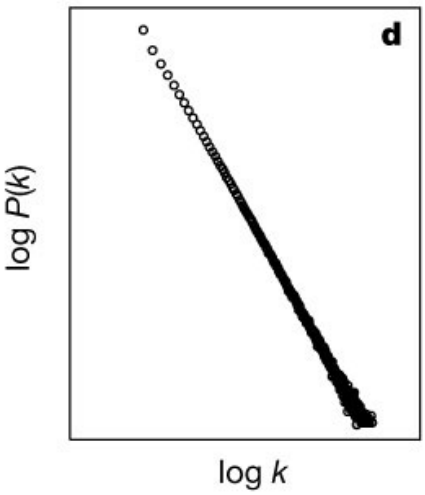

Fig. 2.6. *P(k)* of scale-free network (Jeong *et al.*, 2000)**.**

Scale-free networks demonstrate a surprising degree of robustness to node failure (Albert, Jeong and Barabási, 2000), i.e. even when nodes are damaged, these local failures rarely lead to large-scale problems within the network. The Internet (Faloutsos, Faloutsos and Faioutsos, 1999), the World-Wide Web (Albert, Jeong and Barabási, 1999; Huberman and Adamic, 1999; Kumar *et al.*, 1999), social networks (Wasserman and Faust, 1994) and metabolic networks (Jeong *et al.*, 2000) are particularly good examples of this phenomenon. In fact, all of these networks share similar scale-free topological properties with remarkable similarities in their organisation. However, as non-biological networks grow their diameter increases logarithmically with the addition of new nodes (Barabási and Albert, 1999; Barthélémy and Nunes Amaral, 1999; Watts and Strogatz, 1998), but the diameter of the metabolic networks of 43 organisms is found to be constant (Jeong *et al.*, 2000), irrespective of the number of substrates (nodes) found in the given species. In other words, *d* is fixed for both small and large metabolic networks. This is only possible if with increasing organism (network) complexity, individual substrates (nodes) become increasingly connected to maintain a fixed *d*. This unique feature of metabolic networks is the key to designing non-biological networks with an optimal structural organisation with respect to network efficiency. For example, if a given system is expected to grow significantly in the near future, it is important to calculate the necessary increase

in connectivity that would be necessary in order to maintain its current performance.

Although scale-free networks are robust against errors (removing random nodes) due to their heterogeneous connectivity distribution, it is this same property that makes them extremely vulnerable to attacks (targeting important nodes) (Albert, Jeong and Barabási, 2000). This is because when removing at random, there is a high probability of choosing a node with one link, so removing it will have no effect on the rest of the network. On the other hand, when targeting the most connected node, its removal will also knock out many links, resulting in decreased connectivity or even network fragmentation into disconnected clusters.

2.1.3 Community Structure

Concerning complex networks, it is of interest to look at their *community structure*, which is a prominent feature in many biological (Meunier, Lambiotte and Bullmore, 2010), social (Blondel *et al.*, 2008) and technological (Blondel *et al.*, 2008) complex systems. Community structure is defined as the presence of highly *intra*-connected *modules* of nodes that are loosely *inter*-connected to the rest of the network. In other words, nodes are organised in clusters and most links are inside those clusters. The reason for this phenomenon is that nodes that share functional similarity and/or dependency tend to interact more and therefore they should be more connected. There are two main advantages of this community architecture: the first is efficiency, as most interactions are within modules which are internally well-connected, thereby reducing the path length (the number of links that separate a pair of nodes); and the second is robustness, as entire modules may fail autonomously, without severely affecting the operation of other modules, and hence, the function of the entire network. Therefore, the emergence of community structures in complex networks has implications for their efficiency and robustness, as well as their particular characteristics.

In recent years, research on complex networks has proposed many community detection methods (Lancichinetti and Fortunato, 2009) that aim to discover the most sensible partition of a network into communities. Most of them work on the principle of modularity (Newman and Girvan, 2004) optimisation, aiming to

maximise the modularity benefit function describing the quality of a network partition. The more links that fall within a community compared to an ensemble of benchmark random networks with the same community structure, then the more bias there is for links to connect to nodes belonging to the same community, and therefore the higher the modularity Q (Eq. 2-2). In essence, modularity measures how sharply the modules are defined.

$$Q = (fraction\ of\ links\ within\ communities) - (expected\ fraction\ of\ such\ links) \tag{2-2}$$

The expected fraction of links within communities is calculated from an ensemble of random networks that resemble the network under scrutiny in terms of its strength (total weight on all adjacent links) distribution. In addition, it is necessary to quantify the average level of interaction between a pair of nodes, and this is achieved by defining a null model matrix P_{ij} that describes the expected weight of a link between nodes *i* and *j*, over the ensemble. The standard choice for P_{ij}, defined by Newman and Girvan (2004), preserves the strength of nodes in the random networks:

$$P_{ij}^{NG} = \frac{k_i k_j}{2m} \tag{2-3}$$

where k_i is the strength of node i and m is the total weight in the network. A limitation of this null model, and of community detection methods that use it, is that only network topology and traffic are considered, but this is insufficient for networks embedded in space, such as the USAN. The reason for this is that most spatial networks (excluding the Internet for example) are very biased towards short-range connections due to the cost involved in long-range interactions in physical space. In terms of topology, an airport network is not a typical spatial network, as long-range connections are common. However, in terms of traffic, the higher financial and temporal costs involved in long-range travel play an important role for passengers, thereby affecting the flow on the network. Hence, standard community detection methods (typically based on the NG null model) discover communities of nodes that are spatially close, as opposed to communities that have particularly strong internal interactions (Ball, Karrer and

Newman, 2011; Calabrese *et al.*, 2011; Lancichinetti *et al.*, 2011; Almendral *et al.*, 2010; Estrada and Hatano, 2009). To address this, Expert (2011) proposed an alternative null model for P_{ij} that takes into account the effect of space by favouring communities of nodes i and j that are more connected than expected, given the physical distance d_{ij} between them:

$$P_{ij}^{Spa} = N_i N_j f(d_{ij}) \tag{2-4}$$

where N_i is the importance (typically the strength) of node i and $f(d_{ij})$ is the function that incorporates the effect of space. This so-called deterrence function describes the expected level of interaction between nodes i and j that are separated by some distance d_{ij}. In other words, the function defines how interaction decays, analogous to gravity, as distance between objects increases.

2.2 Robustness in Systems and Networks

2.2.1 Stability and Robustness

The idea of stability originates from celestial mechanics and the study of the stability of the solar system in particular (Jen, 2003). A given state of a dynamic system is defined to be *stable* if small perturbations to the state result in a new state that is close to the original. In this context, perturbations refer to emerging changes in the actual state of the system. Hence, stability ultimately depends on the magnitude of these changes. In addition, a dynamic system is *structurally stable* if small perturbations to the system result in a new system with the same qualitative dynamics. Here, perturbations can take the form of changes in the external parameters of the system. Structural stability requires that certain dynamical features of the system, such as orbit structure, are preserved and no qualitatively new features emerge. A classic example of structural stability is the flow on the surface of a river. Assuming that flow depends on a single external parameter, such as wind speed, the flow is structurally stable if small changes in wind speed do not change the qualitative dynamics of the flow, i.e. do not produce an eddy for example.

It is widely accepted that both stability and robustness are only defined for *specified features* of a given system, with *specified perturbations* being applied to the system. Both concepts are meaningless without prior definition of the features and perturbations of interest. Also, both concepts are concerned with the persistence, or lack of, those features under the specified perturbations. Hence, the level of persistence is a direct measure of the level of stability or robustness. However, robustness is broader than stability for two reasons. Firstly, robustness may apply to a more varied class of systems, features and perturbations. Secondly, robustness naturally leads to concepts, which are beyond the scope of stability, such as:

- The organisational architecture of a system
- The relationship between organisation and dynamics
- The link between evolvability and robustness
- The ability of a system to switch among multiple functions
- The anticipation of multiple perturbations in multiple dimensions

Robustness is the ability of a system to maintain its *function* in the presence of *structural disturbances*, e.g. mutational robustness in biology. There are numerous classes of systems, which cannot be studied effectively using stability theory, and therefore require the concept of robustness. Firstly, systems that cannot be quantified (describing the dependence on numerical variables), cannot be associated with a numerical metric, such as a mathematical function, and hence, the level of persistence cannot be specified. Secondly, systems where the specified perturbations are not changes in internal or external parameters but changes in system composition, system structure, or in the assumptions regarding the environment in which the system operates, cannot be analysed in terms of stability. In addition, robustness is particularly good at capturing the behaviour of systems, which are dependent on the relationship between organisational architecture and dynamics. Furthermore, Krakauer and Plotkin (2005) suggest that in stability theory it is typical to concentrate only on a single perturbation, as opposed to robustness, which inevitably has to take into account multiple

perturbations in possibly multiple dimensions. For example, a biological signalling system may be robust to a whole set of disturbances, including fluctuations in molecular concentrations and the removal of entire groups of genes, which at first sight appear to be essential for the functioning of the system.

There are also numerous classes of networks, which are difficult to represent using a stability framework. For example, heterarchies are interconnected, overlapping, usually hierarchical networks, with individual entities simultaneously belonging to multiple networks. The dynamics (behaviour) of the entire network of heterarchies, both emerges from, and controls the complex interactions between the individual networks. This idea can be easily illustrated by a closed feedback loop, as shown in Fig. 2.7. Standard examples of heterarchies are social networks where individuals are simultaneously members of many networks, such as familial, friendship, political, economic and professional, and the entire network of heterarchies represents human society as a whole. In this case, the economic growth of society depends on and also influences the intricate relationships between political and economic networks. A good example of this is the global financial crisis of 2008.

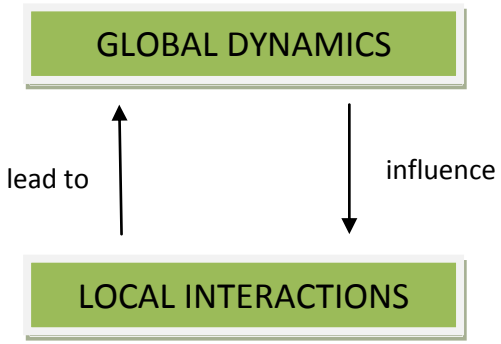

Fig. 2.7. Global dynamics emerging from local interactions.

Robustness in hierarchical and heterarchical systems makes no sense without a *specified level* of interest. For example, robust species may form an ecosystem, which is not robust itself due to competing interactions between the individual species. Conversely, fragile species may form a robust ecosystem, such as a school of fish or a flock of birds, through co-operative interactions. This means that robustness at any one level does not necessarily imply robustness at any other level. Typically, robust systems composed of fragile components need

some sort of self-organisation (Prehofer and Bettstetter, 2005) and are often called *complex adaptive systems* or *self-adapting systems*. However, the application of robustness to systems with strategic options (for responding to perturbations), requires a more context-specific definition of robustness. In this case, robustness can in fact be interpreted in two somewhat different ways. Firstly, it can be defined as the fitness of the possible strategic options, which have either emerged bottom-up or have been designed top-down for the system. Secondly, it can be defined as the ability of a system to switch among multiple strategic options, i.e. to perform multiple functions without any change in structure (this is also known as phenotypic plasticity).

2.2.2 Robustness in Biological Systems

Robustness does not mean staying unchanged in the presence of perturbations, such that the structure, components and operation of the system are unaffected. More precisely, robustness is the maintenance of specific functions of the system against perturbations, which often requires the system to change its mode of operation in a flexible way, i.e. robustness allows changes in the structure and components of the system as long as the functions are maintained (Kitano, 2004).

- **General Robustness**

According to Kitano (2004; 2002) there are five key mechanisms that enhance the robustness of a biological system: system control, redundancy, diversity, modularity and decoupling. System control is based on negative and positive feedback, which ensures that the system is constantly monitored and any changes in the output are instantly considered. Negative feedback is the main method for robust adaptation to perturbations, whereas positive feedback amplifies the stimuli, often resulting in bi-stability. For example, negative feedback is essential in bacterial chemotaxis (a phenomenon whereby organisms direct their movements according to certain chemicals in their environment), and positive feedback is used in signal transduction to form switch-like behaviour, such that there is a transition to a new state, which is more robust to noise and perturbations.

Redundancy (discussed earlier) is a simple concept, which generally refers to a situation where several identical or very similar components can replace one another if one of them fails. This method, however, is not practical in most engineered systems and tends to be costly and inappropriate. Diversity is the other extreme, which refers to a population of heterogeneous components. In this scenario, a given function can be achieved by different means available by utilising a range of components. This mechanism is essentially the same as distributed functionality. Modularity involves the partition of a system into somewhat separate modules in order to keep noise and damage localised, minimising the effect on the system. Modules are often observed in a range of organisms and systems, functioning as virtual design principles within biology or essential elements in engineering. An example of a module is a cell, which interacts with the environment and other cells within a multi-cellular system. Modules are often hierarchically organised, e.g. an organ is made up of tissues, tissues consist of cells, and cells are composed of organelles. Finally, decoupling isolates low-level variation from high-level functionalities in order to maintain the robustness of a system. For example, Hsp90 (heat shock protein 90) is a molecular chaperone (a protein that assists the non-covalent folding/unfolding in molecular biology), which decouples the genotype from the phenotype in order to cope with mutation while maintaining a degree of genetic diversity.

- **Mutational Robustness**

A biological system is robust to mutations if it continues to function normally after genetic changes in its parts, i.e. after permanent alteration in the wiring of the system (Wagner, 2005a). These changes can occur in two distinct ways. Firstly, one can perturb a part of an organism through intentional mutations. Secondly, there exist naturally occurring perturbations, i.e. mutations that occur by evolution over time. There are two principal mechanisms of mutational robustness in biological systems: distributed functionality and redundancy (a gene may be dispensable if the genome contains back-up copies). Furthermore, biological systems have evolved robustness for two reasons. Firstly, robust systems are easier to find in the search of evolution because the neutral space

associated with them is larger (a neutral space is a collection of equivalent solutions to the same biological problem). Secondly, natural selection further increases robustness by incremental evolution of a system within a neutral space. For example, genetic algorithms use evolutionary principles from biology to find optimum solutions to engineering problems. Wagner (2005a) investigates an application to integrated circuits to demonstrate mechanisms for evolved robustness.

2.2.3 Robustness in Complex Networks

There is a large body of literature investigating the robustness of numerous real-world and artificial networks to various types and strategies of damage. For example, type typically refers to node or link damage, and strategy refers to the way in which components (nodes or links) are selected to be damaged (such as random failures or targeted attacks). As chapter 3 discusses a novel damage type called cluster damage, some related work dealing with various types of damage is briefly reviewed here first. In fact, the idea for cluster damage and the strategies employed in this thesis were motivated by this work.

Newman (2003) systematically reviewed developments in the field of complex systems, including concepts such as the small-world effect, degree distributions, clustering, network correlations, random graph models, models of network growth and preferential attachment, and dynamical processes taking place on networks. Specifically, it is mentioned that "a particularly thorough study of the resilience of both real-world and model networks has been conducted by Holme (2002), who looked not only at vertex removal but also at removal of edges and considered some additional strategies for selecting vertices based on so-called betweenness (p. 190)."

Holme studied the response of complex networks, which are subjected to targeted attacks on nodes (vertices) and links (edges). Four existing complex network models and two real-world networks are numerically investigated, and network performance is measured by the average inverse geodesic (shortest path) length (L) and the size of the largest connected component (GCC). Furthermore, four different attacking strategies are used: removing either by descending order of

degree or betweenness centrality, calculated either for the initial network or the current network. The correlation between the betweenness centrality and the degree in those networks is also studied.

Criado (2005) considered the security and stability of complex networks, and reducing the risk and consequences of attacks or dysfunctions. They suggest that the concept of vulnerability helps to measure the response of complex networks subjected to attacks, and allows the identification of critical components of a network in order to improve its security. Hence, they introduce a definition of network vulnerability, which is directly connected with its topology and they analyse its basic properties.

Whereas many studies have investigated specific aspects of robustness, such as molecular mechanisms of repair, Kaiser and Hilgetag (2004) focus more generally on how local structural features in networks may give rise to their global stability. In many networks the failure of single connections may be more likely than the extinction of nodes, and yet no analysis of edge importance has been provided so far for biological networks. They tested several measures for identifying vulnerable edges and compared their prediction performance in biological and artificial networks. Specifically, they say that "from theoretical studies it has been proposed for scale-free networks that edges between hubs are most vulnerable (Holme *et al.*, 2002) (p. 316)." However, in the networks they analysed, both in the scale-free yeast protein interaction network and in cortical networks, edges that connected nodes possessing many connections (large product of degrees) were not particularly vulnerable.

2.2.4 Mechanisms of Network Robustness

There are four general mechanisms of network robustness: structure, redundancy, distributed functionality, and self-organisation.

- **Structure**

The structure, or topology, of a network defines the way in which nodes are interconnected by links. In other words, it describes the explicit relationships, which exist between the individual entities. For example, a random graph is a

very homogeneous network with a Poisson degree distribution (Fig. 2.2). Consequently, there is no *centralisation* and no successful way to damage specific network components with the intent to cause maximum network disturbance. In contrast, the scale-free network is the exact opposite in terms of this centralisation: it is built with preferential attachment (new nodes attach preferentially to high degree nodes), which results in a heterogeneous network with a power-law degree distribution, free of a characteristic scale. Hence, the hub nodes are responsible for keeping the network interconnected and if they are removed the entire structure would collapse.

- **Redundancy**

Redundancy ensures that if a component fails there is another identical or similar component, which can carry out the function of the former. In other words, redundancy is the duplication of critical network components with the aim of increasing network reliability. This mechanism is sometimes called fail-safe, but this is not entirely true. For example, if the network needs a critical node in order to function and nodes have a probability of failing *p*, then each redundant node merely decreases the probability of network failure. Specifically, if there are *n* redundant nodes the probability of network failure would be equal to the product of *n* individual probabilities *p*, assuming that components are independent of each other.

- **Distributed Functionality**

Distributed functionality involves multiple heterogeneous components with overlapping (distributed) functions. In other words, many nodes contribute to network function but all of them have different roles, i.e. no two nodes are identical copies of each other. When some node fails other nodes can compensate by modifying their behaviour in an appropriate way. Examples of systems with distributed functionality include biological systems, such as a gene regulatory network (a collection of DNA segments in a cell) and a metabolic network (the set of processes that determine the properties of a cell); and technological systems, such as the global telephone network and the Internet. Distributed

functionality and redundancy are the two core mechanisms of mutational robustness in genetic networks (Wagner, 2005b) (discussed later in detail). A genetic network is robust to mutations if it continues to function normally after genetic changes in its components, i.e. after permanent alteration in the wiring of the network. In addition, metabolic networks do not contain any redundant reactions, and yet, over 50% of these reactions can be removed without affecting the metabolic output (Wagner, 2005a). The reason for this is that the network is able to re-route metabolic flux through unaffected parts. Hence, robustness is due to the co-operation of enzymes with different, possibly overlapping activities; not due to redundancy.

An important concept in control theory is the transfer function, which specifies how a system's output behaves as a function of its input. General theorems in the subject imply that any redundant system can be re-designed into a non-redundant system with an identical transfer function (Leigh, 1992). This means that in theory any system with redundant parts can be replaced by a cheaper, more optimised system, which relies on distributed functionality as opposed to redundancy.

- **Self-Organisation**

A system is *organised* if it has a certain structure and functionality (Staab *et al.*, 2003). Structure means that the entities are arranged in a particular manner and interact in some way. Functionality means that the overall system fulfils a certain purpose. A system is *self-organised* if it is organised without any external or central dedicated control entity, i.e. the individual entities interact directly in a distributed, peer-to-peer fashion (usually localised). In other words, connections between pairs of entities give rise to a robust and self-adapting network. For example, ant pheromone trails are a classic model of self-organised animal behavior (Sumpter, 2006). However, self-organisation is more than just distributed and localised control: it is about the relationship between the behaviour of the individual entities and the resulting structure and functionality of the entire system. In self-organised systems the simple behaviour at the *microscopic* level leads to sophisticated organisation at the *system* level. This

phenomenon is known as *emergent behaviour*. Self-organisation can be defined as the emergence of system-wide adaptive structure and functionality, from simple local interactions between individual entities (Prehofer and Bettstetter, 2005).

2.2.5 Cascades

Cascades in networks are an example of the *robust yet fragile* concept associated with many complex systems. A network may seem stable (robust) for long periods of time and then suddenly and unpredictably exhibit a large cascade (fragile), which may damage the network considerably. This concept is rooted in the infrastructure of the network. Even when the properties of individual components are well understood, cascades are very difficult to predict. For example, a single component failure will generally affect the network in some way but the precise possibility of subsequent failures (possibly leading to a global cascade) cannot be specified because of the *dynamics of redistribution of flows* on the network (Crucitti, Latora and Marchiori, 2004). However, two generic features of cascades can be explained with respect to the connectivity of the network: they occur rarely, but by definition are large when they do (Watts, 2002). Cascading failures are common in the Internet, electrical power grids, social systems, and most communication and transportation networks.

- **Cascading Failures in Electrical Power Grids**

When a power line is damaged its load is automatically shifted to nearby lines if they are able to handle the extra load. However, if those lines are already operating at maximum capacity they must redistribute the extra load to their neighbours, and so on, until the power is distributed properly, or else a cascade occurs. In the latter case, a large number of power lines are overloaded and this will probably result in blackouts. This is a good example of the *robust yet fragile* concept because it illustrates the advantages of distributed functionality and the associated disadvantages of possible cascades. Crucitti, Latora and Marchiori (2004) propose a model based on dynamic redistribution of flow, which is triggered by the initial breakdown of a single component of the network. A key

concept of the model is the introduction of a tolerance parameter, which specifies the amount of extra load that each node can handle. In most cases, the removal of a node changes the shortest paths between nodes and consequently the distribution of the loads, which may create overloads on some nodes. In some cases, this overloading can trigger an avalanche, covering the whole network. Crucitti suggest that small values of the tolerance parameter causes a decrease in network efficiency, but below a critical tolerance the network collapses. This is intuitive since the tolerance parameter is directly related to the capacity of the network to handle excess load. They also show that random networks are more resistant to cascading failures than scale-free networks, which is a very important result because it suggests that the robustness of networks may be enhanced by introducing random links. Furthermore, in scale-free networks random removals are far less likely to trigger cascades than load-based removals. This is due to the heterogeneous wiring of the network, and is also related to the error and attack tolerance of scale-free networks. Finally, they show that the removal of a single key node is sufficient to cause the entire grid to blackout.

- **Cascade Control and Defence**

In order to prevent a cascade from propagating through a network, Motter (2004) introduced a costless strategy of defence based on the selective further removal of nodes and links, after an initial attack or failure of a small fraction of nodes. According to Motter, a cascade consists of two parts: (1) the initial attack, where a fraction of nodes is removed; and (2) the propagation of the cascade, where another fraction of nodes is removed due to overloading. The intentional removal of components is performed between (1) and (2), resulting in a drastically reduced cascade with size proportional to the size of the largest connected component. This is achieved by removing nodes with small load and links with large excess of load, where load is defined as the total number of shortest paths that pass through a node or a link.

2.3 Air Transportation

Transportation networks are a good example of spatial networks. Their network topology is not only characterised by spatial aspects such as the location of nodes and the length of links but also by the association of a *transport cost* to the link length; implying that longer links are typically balanced by some benefit, such as connecting to a high-degree node, or a node in an attractive location. Transportation networks typify the specific nature of spatial networks particularly with regard to issues such as congestion, fast-growing urban sprawl and disease propagation. Network structure and dynamics play a key role in most, if not all, of these challenges. Transportation networks can be planar, as in road and rail networks, or non-planar, as in airport networks. In addition, transportation networks are usually weighted, where the link weight describes the intensity of some form of interaction, e.g. the amount of traffic. Air transportation networks are an important example of spatial networks. Nodes identify airports and links represent the existence of a direct air service among them. Weights on links may represent the number of passengers flying on that connection, and the distribution of weights is an initial indication of the existence of possible strong heterogeneities.

The existence of links among airports depends on factors related to both airline strategies and passenger demand. Airlines decide to operate at a given airport on the basis of a significant demand, allowing them to reach satisfactory load factors. Location and socio-economic characteristics of the airport catchment area are the key factors generating air traffic demand. The airport choice made by both airlines and travellers depends on factors that can be ultimately reduced to time and monetary costs. For example, reduced airport charges may help airlines to offer lower air fares to potential travellers, and hence to induce more flights. The airport network is an example of a heterogeneous network where the hubs have high connectivity, high weight (in terms of traffic) and long-distance links (Barrat *et al.*, 2004).

In recent years, the analysis of complex transport networks has received considerable attention, mainly in terms of commuting networks (De Montis *et al.*,

2007; Patuelli *et al.*, 2007; Rouwendal, 2004). Airport networks have also been studied to characterise their level of degree correlations and clustering, their evolution in time, and their potential scale-free properties (Wuellner, Roy and D'Souza, 2010; Guimerà *et al.*, 2005; Amaral *et al.*, 2000). In terms of network robustness, network failure due to external factors such as bad weather conditions, volcanic eruptions, and political or security issues, may have significant impact on the air traffic depending on the criticality of the involved nodes and the extent of their influence. In terms of socio-economic characteristics, the emergence of community structure depends on the location and distribution of relevant activities. Concentration of activities in a given area generally means concentration of short trips in that area, and this is a typical commuting pattern. For medium-long distance trips, the main contributing factor is mass migration rather than commuting, and air transportation plays an important role in facilitating easier migration of workers. Within larger countries, such as the United States, a new kind of commuting by air can be identified, as people working in different parts of the country during the week return home at weekends. The changes in the availability, frequency and cost of air travel facilitate trips for migrants located far from traditional gateways (large airports with hub functions (hub-and-spoke) and inter-continental links) (Button, 2010).

2.4 Language Acquisition

Research in children's language acquisition has recently benefited from the application of network theory to large sets of empirical data, which has illuminated interesting patterns and trends. Network theory is an extremely powerful modelling and analysis tool, and its full potential in terms of extracting useful information from raw data has yet to be exploited. Researchers modelling language using network theory have experimented with a wide range of parameters, but there seems to be no consensus on which parameters are essential for understanding language acquisition, and which are merely providing some additional network information. Table 2.1 summarises the parameters used by recent research involving the network modelling of language, to give an idea of the most relevant network properties. The publications in the table were collected

by identifying seven key papers (1. (Ke and Yao, 2008); 2. (Cancho and Solé, 2001); 3. (Motter *et al.*, 2002); 4. (Solé *et al.*, 2010); 5. (Corominas-Murtra, Valverde and Solé, 2010); 6. (Adamo and Boylan, 2008); and 11. (Liang *et al.*, 2009) in Table 2.1) on the modelling of language using network theory, and papers (7. (Haitao and Fengguo, 2008); 8. (Li and Zhou, 2007); 9. (Zhou *et al.*, 2008); and 10. (Shi *et al.*, 2008) in Table 2.1) on the syntax of language that cite any one of the initial seven. The shaded entries correspond to the papers that investigate children's language acquisition, as opposed to language in general. The research summarised in Table 2.1 typically focuses on three types of networks: co-occurrence, syntactic and semantic; and two languages: English and Chinese.

Table 2.1. Parameters of interest in linguistic networks.

Ref.	Network Type	Language	Length	MLU	N	E	GCC	NCC	<*k*>	AC	DC	NN	*D*	*L*	*C*	*P(k)*	*P(f)*	*P(b)*
1	co-occurrence	English		x	x	x			x									
2	co-occurrence	English			x	x			x					x	x	x		
3	semantic	English			x	x			x					x	x	x		
4	co-occurrence syntactic semantic	English			x	x			x					x	x	x		
5	syntactic	English		x	x	x			x					x	x	x	x	
6	co-occurrence	English		x	x	x			x		x							
7	dependency syntactic	English			x	x			x				x	x	x	x		
8	char. structure	Chinese			x	x			x	x				x	x	x		
9	co-occurrence	Chinese			x	x			x	x		x		x	x	x	x	x
10	co-occurrence	Chinese	x		x	x	x	x	x				x	x	x	x		
11	co-occurrence	Chinese English	x		x	x	x	x	x				x	x	x	x		

A total of sixteen unique statistical parameters were investigated in these publications and they are briefly described below:

1. *Length* of a linguistic data set may refer to the total number of characters, words, or utterances within the sample.

2. Mean Length of Utterance (*MLU*) is the average number of words in an utterance. To be precise, *Length* and *MLU* are in fact data set parameters, but they are nevertheless treated like network parameters.

3. Total number of network nodes *N* is the first most basic network measure.

4. Total number of network links *E* is the second most basic network measure.

5. Giant Connected Component (*GCC*) represents the largest connected part of the entire network, i.e., the cluster with the highest number of nodes. In a connected network (where there are no isolated nodes or clusters of nodes), the GCC is identical to the original network.

6. Number of Connected Components (*NCC*) is simply the number of network components that are disconnected from one another.

7. Average degree $<k>$ is the average number of links adjacent to a network node. This key parameter reflects the overall connectivity of the network.

8. Assortativity Coefficient (*AC*) measures how assortative the network is, i.e., to what extent high-degree nodes are connected to other high-degree nodes.

9. Degree Centralisation (*DC*) measures to what extent the links are centralised on a small number of high-degree nodes.

10. Average nearest-neighbour degree *NN* of a node is the average degree of the nodes that are connected to the given node.

11. Network diameter *D* is the length of the longest path between a pair of nodes, when the shortest possible paths are considered, i.e., containing the fewest links.

12. Average geodesic length *L* is the average length of the shortest paths between all pairs of nodes.

13. Clustering coefficient *C*, averaged among all nodes, measures the likelihood of the neighbours of a node being connected themselves.

14. Node degree distribution *P(k)* is the probability distribution of a randomly chosen node with degree *k*.

15. Similarly, node frequency distribution *P(f)* is the probability distribution of a randomly chosen node with frequency *f*.

16. Finally, node *betweenness* distribution *P(b)* is the probability distribution of a randomly chosen node with betweenness centrality *b*.

The paper by Ke and Yao (2008) is particularly relevant for this research because it provides a comprehensive study of children's word co-occurrence networks. Two kinds of networks are considered: accumulative networks that accumulate the data over time; and stage networks that model five independent (incremental) time-slices of the data. Accumulative networks are built for twelve children's data and stage networks are constructed for four of them (two of whom are also modelled in this thesis: Carl and Anne), and their respective mothers. In addition, the authors propose the concept of hubs and authorities as measures for nodes' importance, where a hub node has many outgoing links and an authority node has many incoming links.

2.5 Research Problems

Based on the findings of the literature review, this section describes the research problems to be solved.

Possibly the most basic limitation of any network model is its inability to encode the properties of the real system with sufficient accuracy and detail. Models are typically developed to be as simple as possible due to lack of knowledge, uncertainty or complexity. However, although the latter two are more difficult to address, recent advances in data collection, storage and availability has significantly increased our knowledge, thereby allowing more detailed models to be developed. Data sets of greater breadth and depth can be used to build models incorporating more knowledge about the system under scrutiny.

- **Robustness**

Robustness, generally defined as the ability of a system to maintain its function in the presence of disturbance, is critical for the reliable operation of real-world networks. Since, the operating environment is typically uncertain and may also change considerably over time, it is important to consider many types of disturbances, in order to measure many dimensions of network robustness. The key problem here is that current research only focuses on node and link damage

as possible disturbances, but there may be more complex structural disturbances in a given context, such as cluster damage, that may affect the system very differently.

- **Air Transportation**

Air transportation has received considerable attention by the research community in recent years but most studies usually focus on a specific feature, when there are in fact multiple dimensions that should be modelled and studied in parallel, as this would reveal a more comprehensive picture of the huge complexity in the system. In other words, the emphasis of current research is more on depth instead of breadth, but it is important to have the breadth before going into more depth. For example, Bounova (2009) has used simple (unweighted and undirected) networks to model US airlines but these networks are insufficiently detailed for some analysis techniques, such as community structure detection. Moreover, since the USAN is embedded in space, topology alone cannot be used for the reliable detection of community structure, since the effect of space is non-trivial. However, no work addresses this issue so far apart from Expert (2011), who proposes a spatial null model for accurately detecting community structure in spatial networks. Another major issue in the study of networks developing over time is evolution. Clearly, systems that are constantly changing need to be studied in terms of these changes, in order to understand the dynamics of the system over time, at different time scales. However, most research tends to underestimate the significance of this important dimension, resulting in poor models that only capture a fixed snapshot of the continuous nature of dynamic systems. For example, Guimerà (2005) models the traffic in the world-wide airport network for a period of one year using a single network.

Researchers working on airport networks have typically focused on the modelling of a national airport network, such as the Airport Network of China (Li and Cai, 2004), and the Airport Network of India (Bagler, 2008); or the World Airport Network (Guimerà *et al.*, 2005), which is the global network of all airports. However, most studies so far have either investigated the evolution of the network over a not significantly long time period (Xu and Harriss, 2008;

Barrat *et al.*, 2004; Amaral *et al.*, 2000), or have not modelled in detail by ignoring link directionality and link weights (Bounova, 2009). Therefore, there is clearly a need for a more detailed model of the evolution of a complex airport network over a significant time period. The aim is to explore the development of the network, in order to expose growth patterns, and changes in structure as well as passenger demand.

- **Language Acquisition**

Research on children's language acquisition that is based on network modelling and analysis generally fails to exploit the full potential of network theory, by using simple network models and analysis techniques. For example, network models often neglect links' weight or directionality, and the analysis of the networks only considers some of the parameters highlighted in Table 2.1, without employing more advanced techniques, such as community detection. For instance, the main drawback of the network models of Ke and Yao (2008) is that they neglect the statistical properties of the input data, such as frequency of co-occurrence, which are very important in language acquisition. This is a good example of a thorough study, which, however, uses an over-simplified network model disregarding link weights. In addition, linguistic network models are either semantic or syntactic (word co-occurrence networks also model syntax), but not both, thereby failing to fuse together these two important properties of language.

In addition to the network modelling and analysis problems outlined above, a second key issue in language is the validation of models of language acquisition, such as MOSAIC, which is important for understanding mechanisms of learning, and evaluating theories of learning, such as UG and distributional analysis. Based on MOSAIC's previous success in modelling numerous phenomena of language acquisition, it is expected that the hypothesis that there will be significant similarity between its output and children' utterances will be accepted. Hence, if MOSAIC performs well then there is reason to believe that the nativist theory may be inconsistent with empirical data reflecting children's linguistic abilities.

2.6 Research Solutions

This thesis addresses the problems outlined in the previous section by proposing a common methodology to both applied case studies that is based on network theory (Fig. 2.8). It is important to note that although the case studies (air transportation and language acquisition) appear very different with little in common, they can in fact be described and analysed using the same tools from network theory, which provide a logical abstraction of a complex system and serve to mediate the generality of this research. The following section discusses the proposed research solutions in more detail.

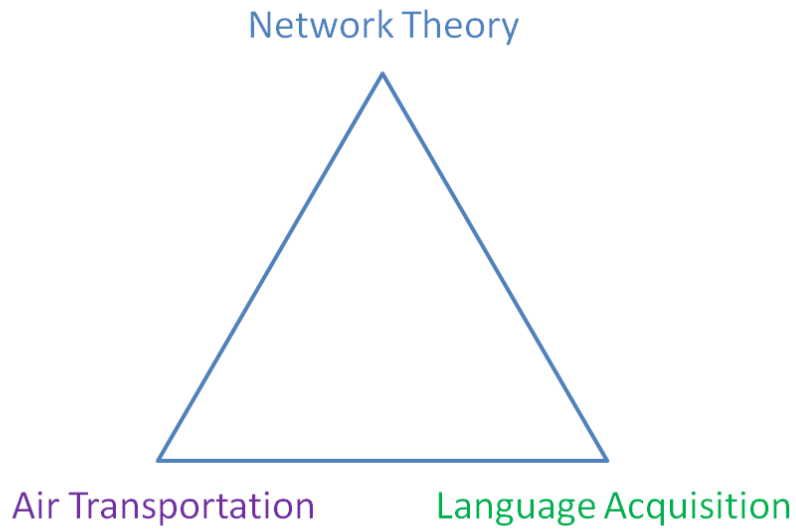

Fig. 2.8. Common methodology for air transportation and language acquisition based on network theory.

2.7 Research Methodology

This section begins by describing the cluster damage method that addresses the robustness problem in complex networks. Furthermore, it describes in more detail the solutions to the research problems that are common to both domains: air transportation and language acquisition. Specifically, detailed network modelling, space-independent community structure, and dynamics and evolution have been identified as key solutions to the limitations and drawbacks of existing approaches that model and analyse complex networks.

2.7.1 Cluster Damage

Cluster damage is a novel method to structurally disturb a network in order to measure another dimension of its robustness. A cluster is defined as a group of nodes interconnected with links. Note that this is not necessarily a community, since community structure requires particularly strong internal and weak external interactions maximising modularity, which is not the case for the type of clusters defined here.

2.7.2 Detailed Network Modelling

The models developed in this thesis are more detailed than existing models in both domains in terms of network dimensions. In other words, the networks have five dimensions: topology, link weights, link directionality, space and time. Topology refers to the structure of the network in terms of the connections. Link weights refer to the strength of some type of interaction among nodes. Link directionality refers to the directed nature of some relationships, e.g. A - B but not B - A. Space refers to the Euclidean distance between airports in the airport networks, and to the semantic distance between words in the language acquisition networks. Time refers to the dynamics of the networks that occur over a time period.

Network theory offers numerous statistical parameters that usually measure some structural property of the underlying network, so the most prominent parameters are selected for analysis. Since they are quite general, they are often used across many disciplines that exploit the potential of network modelling and analysis. Six individual parameters are investigated: number of nodes (N); number of links (E); size of Giant Connected Component (GCC); average degree ($<k>$); characteristic path length (L); and clustering coefficient (C). In addition, three functions are computed: the in-degree distribution $P(k_{in})$; the out-degree distribution $P(k_{out})$; and the ranked weight distribution $W(r)$ (Gegov *et al.*, 2011) [P3], (Gegov *et al.*, 2012) [P2]. It is worth mentioning that $W(r)$ is an indicator of dynamics *on* the network, as opposed to the other indicators that measure some property of the network structure. It was chosen because it contains information

about the absolute value of the link weights and every link is explicitly present in the distribution. In the case study of air transportation, *W(r)* is used as a measure of the volume of passengers travelling between all connected airports instead of the commonly used (cumulative) probability distribution of link weights. However, *W(r)* has been neglected by the linguistic research community in recent years, but given its effectiveness in other research domains, such as air transportation, it is also expected to reveal useful information about linguistic networks, and therefore, it is also used in the case study on language acquisition networks. Since networks are a direct reflection of the properties and characteristics of the underlying data sets, analysing and comparing these networks (using the statistical parameters above) will reveal important information regarding the nature of these data. It is worth mentioning that this information (patterns or trends, for example) may remain hidden when analysing data using standard techniques at the microscopic level, but emerges only through the application of network theory at the macroscopic level.

2.7.3 Space-Independent Community Structure

In order to find more realistic community structure, Expert's (2011) null model is coupled with the spatial dimension of the networks, producing very detailed and accurate information on the hidden community structure within (Gegov *et al.*, 2012; Gegov *et al.*, to be published) [P4, P1]. The model is able to uncover space-independent community structure (as shown in Expert's paper for the Belgian mobile network, compared to the NG null model), and hence, it is applied to both of the applied case studies – air transportation and language acquisition – each consisting of 18 network snapshots (representing topology and link weights). The inputs are the adjacency matrix A_{ij} (encoding the snapshot), the distance matrix D_{ij} (containing the Euclidean distance between all pairs of airports, or the semantic distance between pairs of words), the importance vector N_i (holding the passenger flow at each airport, or the occurrence frequency of a word), and the bin size, which is used to bin the data from the distance matrix. In order to obtain fair results, it is necessary to select a bin size such that the bins are sufficiently populated, without losing too much spatial resolution.

The bin sizes used are described in the respective chapters. The output of Expert's (2011) null model is the modularity matrix:

$$Q_{ij} = (A_{ij} - P_{ij}^{Spa}) \tag{2-5}$$

which is then fed into a community detection algorithm (Jutla and Mucha, n.d.) that searches for a network partition, maximising modularity. It is worth mentioning that this is the same algorithm used by Expert for the Belgian mobile network. The output of the community detection algorithm is a vector, assigning each airport/word to a specific community in which all members have particularly strong interactions in terms of passenger flows/co-occurrences, given their spatial separation. It is worth mentioning that since community structure relates to the cohesiveness of certain groups of nodes in terms of their strong internal relationships, link directionality does not play a major role in the identification of such groups (communities), as the main emphasis is on the strength and not the directionality of the relationships. Therefore, most community structure methods are designed to work with undirected networks, and for this purpose, all the networks presented in this thesis are converted from directed to undirected, for the detection of community structure. In other words, unidirectional and bidirectional links are replaced by undirected links that are weighted with the sum of the weights on the directed links.

2.7.4 Dynamics and Evolution

To investigate the time dimension, the model proposes eighteen network snapshots for each applied case study (air transportation and language acquisition), partitioned into three discrete developmental stages consisting of six networks (Gegov *et al.*, 2011) [P5]. In other words, for air transportation, long-term network evolution is captured by the stages and short-term network dynamics are captured by the snapshots within a stage. Note that in the language acquisition networks the snapshots within a stage are used to study the characteristics of the individual children, and aggregated stage networks are built to model the average child's linguistic development over time. Aggregated

annual networks are not built for the USAN since they would only provide a more coarse-grained resolution on the seasonal flows within the network.

2.8 Summary

This chapter reviewed recent literature on complex networks, highlighting research problems in the field, and proposing solutions in the form of methodologies. The main idea is to use more advanced network modelling and analysis techniques to understand complex systems at an abstract level.

Chapter 3
Robustness to Cluster Damage

This chapter describes a novel approach to investigate network robustness by damaging entire clusters of nodes. A partial network model of the Internet at the Autonomous System level is presented. The robustness of the network to node and cluster damage is discussed.

This chapter is motivated by the fact that there is a large body of research on the error and attack tolerance of complex networks to node or link damage but not cluster damage (Gegov, 2009) [P6]. For example, in the Internet it may be the case that instead of a single server or communication channel, an entire local area network is damaged, either accidentally (error) or by intention (attack).

3.1 Methodology

The standard method to test robustness is to iteratively remove network components according to various strategies, such as errors and attacks, which simulate random failure and targeted damage to components, respectively. This chapter proposes the use of clusters for the components being removed. The attacking strategy targets the most central clusters, i.e. those which are responsible for keeping the network interconnected. Here, it is necessary to define Cluster Betweenness Centrality (CBC), which is equivalent to node betweenness centrality in a network where each cluster is represented by a single node and all inter-cluster links are represented as normal links. Robustness is defined by the network's ability to maintain its function, which is measured at each iterative step.

In order to properly assess the robustness of a network it is essential to define first the type of damage being considered and the function of the network that is of primary interest. In this chapter, the severity of the damage caused by the proposed cluster failure type is measured by the number of nodes removed from the network. Since more important clusters are often larger (in terms of nodes), a

cluster attacking strategy will decrease the network size very rapidly compared to a random strategy. Hence, it is essential to compare results consistently by using the number of damaged nodes as a measure of the level of damage caused. The general level of tolerance (robustness) is measured by three parameters: number of links E; size of Giant Connected Component (GCC); and average geodesic length L.

By definition, robustness is the ability of a system to maintain its function in the presence of disturbances. Hence, in the context of a network such as the Internet, the most fundamental function is to remain efficiently connected so robustness can be defined as the level of connectivity in the presence of structural damage (simulated by removing network components). The average geodesic length L is a good indicator of efficient connectivity as it increases to a maximum (the breaking point) as the network is progressively damaged, and then decreases when the GCC breaks up into smaller disconnected components. Therefore, the breaking point can be considered as the minimum connectivity for satisfactory network function. Based on this, specific threshold robustness is defined by the level of damage necessary to cause unsatisfactory network function, i.e. the percentage of nodes removed at the breaking point.

The cluster damage (errors and attacks) methodology is summarised in the following algorithm:

1. Partition network into clusters.
2. Build network of clusters (for cluster attacks only).
3. For cluster errors: select a cluster at random.

   For cluster attacks: identify node with highest betweenness centrality in network of clusters.
4. Remove corresponding cluster in original network.
5. Measure network parameters.
6. Go to step 3 and repeat until the network is fragmented.

The network is first divided into similarly-sized modules (step 1) using spectral partitioning (Newman, 2006). This technique recursively bisects the network into equally-sized sub-networks by minimising the number of links that are cut. Therefore, the end result is a network partition into clusters that tend to have few links between them. Since the objective of this chapter is to present an application of the cluster damage methodology, it is not necessary to find the optimum network partition into meaningful communities, and therefore this spectral partitioning approach is sufficient in this context.

3.2 The Internet

The Internet is a real-world technological network. Specifically, it is a global network of interconnected computer networks called *Autonomous Systems* (ASs) (Fig. 3.1). Since the entire network is very large, it is usually studied at the AS level, as opposed to the individual computer level. For this theoretical case study, the network model was constructed using Internet traffic data that are publicly available from CAIDA (http://www.caida.org/home/). Only a small part of the Internet (1390 ASs) is captured by the network model due to the specific and partial traffic data that were used. However, this small snapshot represents the structure of the Internet well, due to its fractal nature. In the network model, a node represents an AS: a network under a single admin and routing policy; and a link represents a direct channel for traffic exchange.

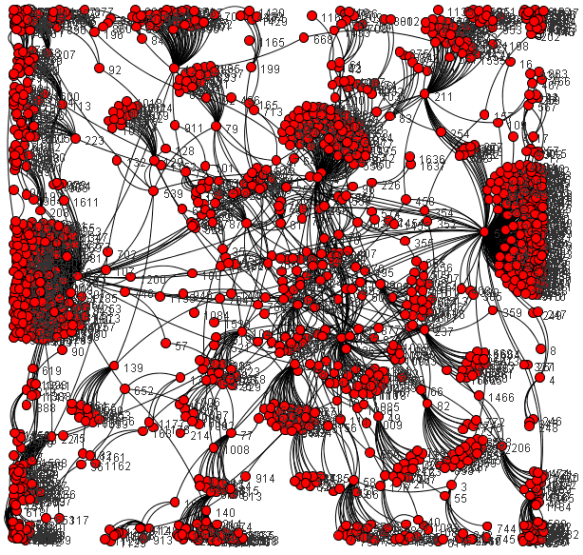

Fig. 3.1. Network of Autonomous Systems.

3.3 Results

By tuning the recursion depth and the quality of partitions in the spectral partitioning algorithm, it is possible to obtain many different partitions into clusters using the same network model. Here, the model network is split into seven different partitions, each composed of a different number of clusters that have modularity values presented in Table 3.1. Note that the highest modularity of 0.7747 (very high for a real-world network) was obtained for 74 clusters. The lowest possible value of 0 was obtained for 1390 clusters, i.e. when every node is a cluster. For each partition (into a different number of clusters), a cluster is represented by a single node and inter-cluster links are represented by ordinary links, in a new network of clusters (step 2). This network is used to identify the most central modules in the original configuration (step 3) in order to target them first when simulating cluster attacks (step 4).

Table 3.1. Summary of different network partitions into clusters.

Clusters	Modularity
6	0.4925
33	0.7522
74	0.7747
110	0.4844
152	0.3773
182	0.3248
229	0.2648
1390	0

The highest modularity, attained for a partition into 74 clusters, implies that this is the most meaningful of the obtained partitions. Therefore, this particular partition is used to test the general robustness of the network to cluster damage. Figs. 3.2-3.4 show how the number of links E, the average geodesic L, and the Giant Connected Component (*GCC*) behave as a function of damage in terms of nodes removed from the network, i.e. the remaining nodes in the x-axes decrease from left to right as the network is progressively damaged. *Node errors* refers to the random removal of nodes; *node attacks* refers to the targeted removal of nodes; *cluster errors* refers to the random removal of clusters; and *cluster attacks* refers to the targeted removal of clusters. In addition, modularity (Table 3.1) and

specific threshold robustness (defined earlier) are presented in Figs. 3.5 & 3.6 as a function of the number of clusters obtained from the seven partitions.

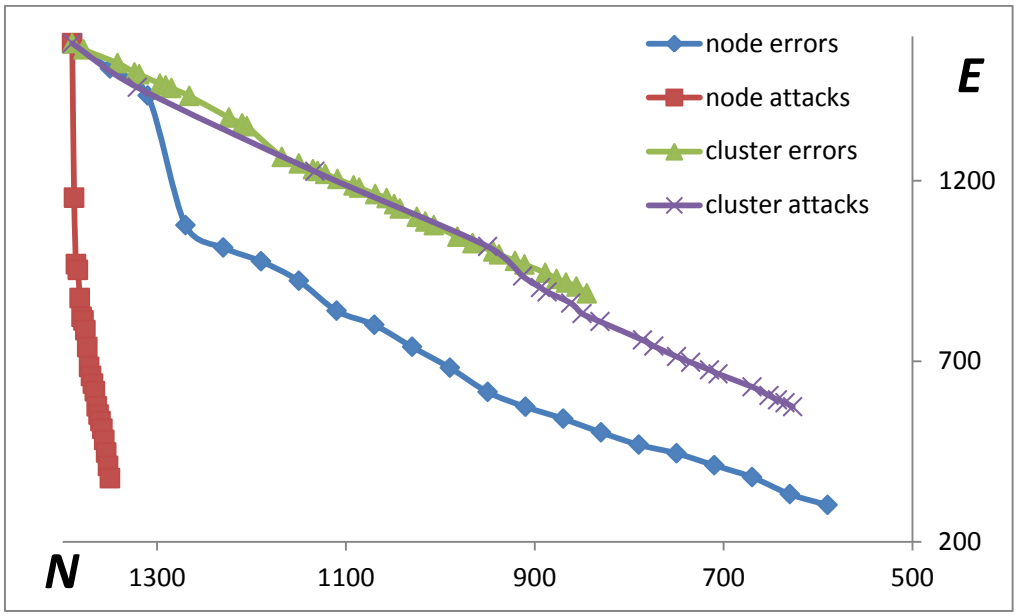

Fig. 3.2. Links *E* as a function of nodes *N* for 74 clusters.

Fig. 3.2 shows how the number of links (AS connections) diminishes as the network becomes increasingly damaged. Clearly, the network best maintains its connections under cluster errors and attacks (equally well), which has two implications. Firstly, AS connections appear to be more robust to cluster damage (especially attacks). Secondly, AS connections appear to be equally robust to both cluster errors and attacks, suggesting that the betweenness of clusters is independent of the number of their internal links. In other words, there is no correlation between cluster importance and size (in terms of connections). It is assumed that the node errors strategy knocked out a high-degree node at an early stage, which is represented by the sudden dip in the blue curve in the top left. The node attacking strategy causes the most severe damage to the AS connections.

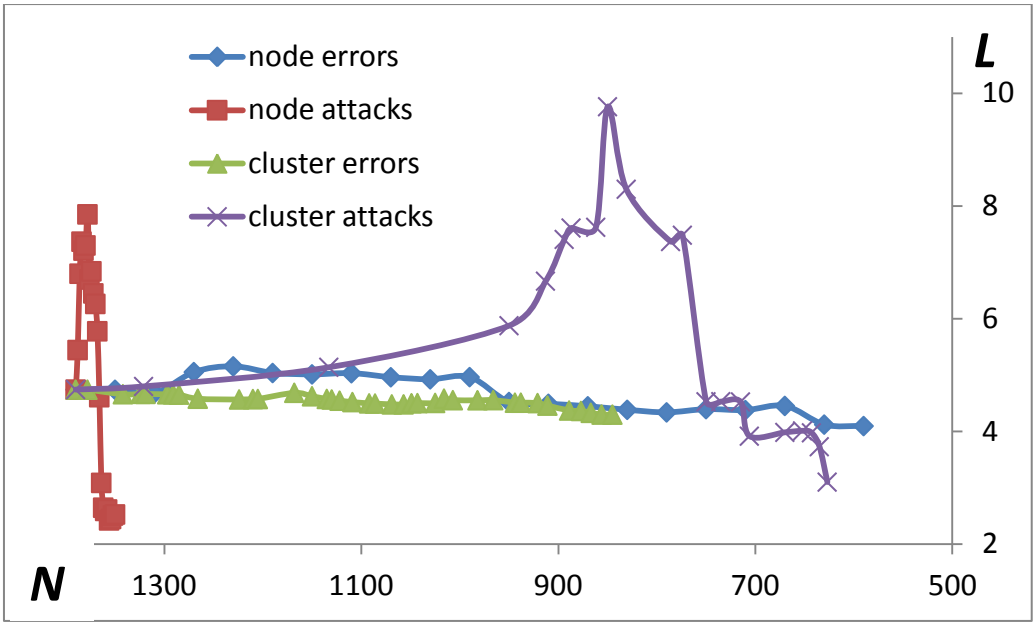

Fig. 3.3. Average geodesic *L* as a function of nodes *N* for 74 clusters.

Fig. 3.3 displays how the average geodesic *L* behaves as the network becomes increasingly damaged. Node and cluster errors do not appear to affect *L* significantly, which means that random failures are unlikely to affect the efficiency and function of the Internet. This is in line with a large body of research in the field of Internet robustness. However, target attacks (by hackers for example) cause severe damage to the Internet, especially when they target specific critical network components. This is demonstrated by the two peaks in the attack curves, which represent the point where the network becomes fragmented into multiple small disconnected components. Note how the focused node attacks break-up the network very early on compared to the more distributed cluster attacks that disconnect the network at a much later stage after many more nodes have been removed.

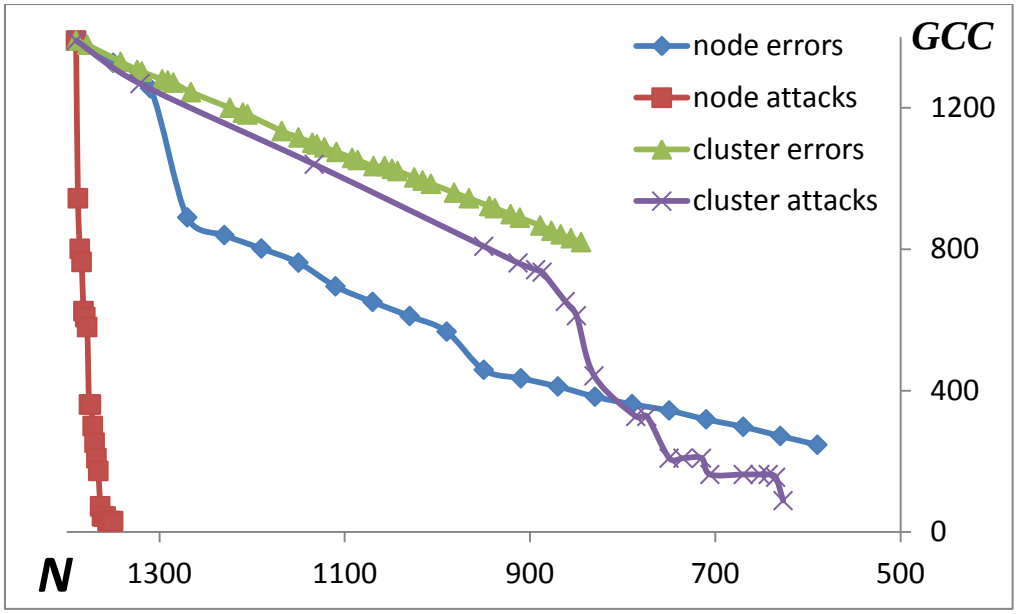

Fig. 3.4. Giant Connected Component *GCC* as a function of nodes *N* for 74 clusters.

Fig. 3.4 shows how the *GCC* decreases as the network becomes increasingly damaged. The trends are similar to those for the number of links *E* and therefore suggest a relationship between *GCC* and *E*. In other words, the number of links is positively correlated with the number of nodes as the network is damaged. This means that node/cluster errors/attacks knock out components with a similar ratio of nodes to links.

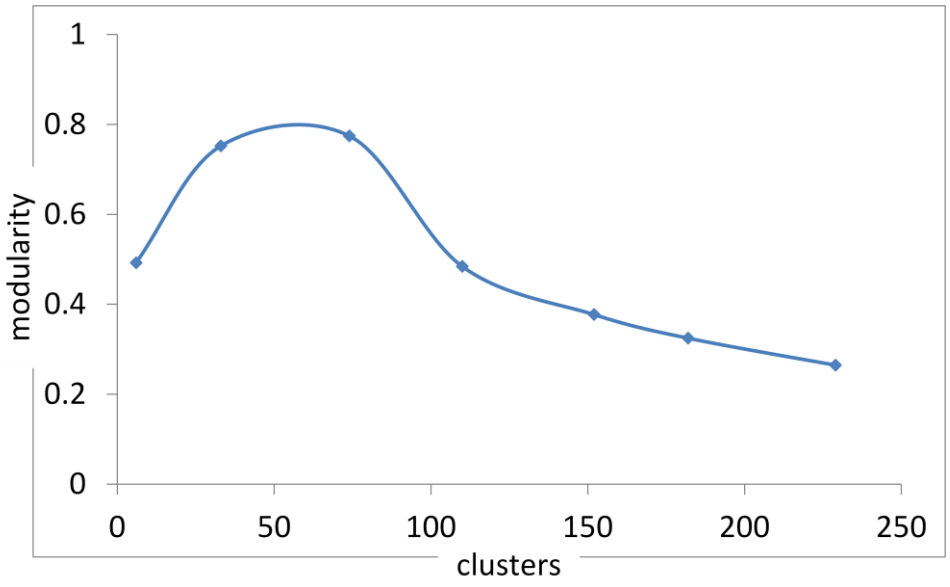

Fig. 3.5. Modularity as a function of the number of clusters.

Fig. 3.5 presents the modularity Q of the seven network partitions into clusters (the exact values are shown in Table 3.1). Since spectral partitioning is not intended for optimising Q, its maximum value and corresponding partition is not discovered, but this is irrelevant for the purpose of this chapter. Fig. 3.5 is intended to show the range of the discovered partitions in order to identify a possible relationship between cluster modularity and robustness to cluster damage. The highest Q is obtained for a network partition into 74 clusters, after which Q predictably decays for partitions into more clusters.

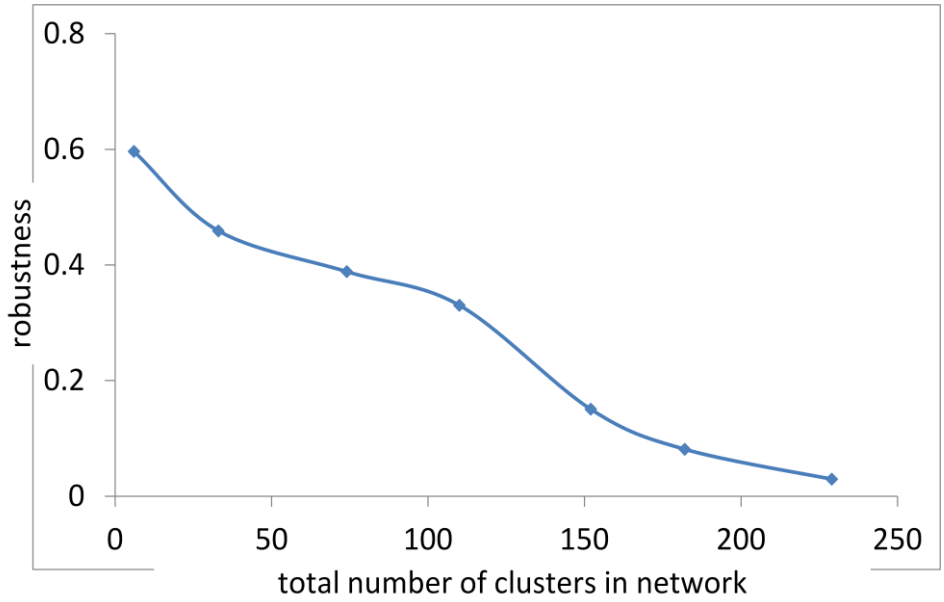

Fig. 3.6. Robustness to cluster attacks as a function of the number of clusters.

Fig. 3.6 depicts the specific threshold robustness of the Internet as a function of the seven network partitions into clusters. This plot refers to the cluster attacking strategy, since the peak in L is reached after a considerable amount of damage, and hence, a significant robustness is observed. The node attacking strategy is not shown as the network is fragmented too early, resulting in 1% robustness, i.e. if 1% of nodes are attacked then the network is fragmented. Moreover, there is a negative near-linear relationship between robustness and the number of clusters, as shown by the smooth curve in Fig. 3.6. In other words, robustness to cluster damage increases linearly with a decreasing number of clusters. Basically, fewer clusters are relatively larger so then the cluster damage is less focused, resulting in better robustness.

It is important to highlight that there is no meaningful observed relationship between modularity and robustness (see Fig. 3.7). This implies that cluster modularity does not affect robustness to cluster attacks. However, Kitano (2004) has shown that modularity does in fact facilitate network robustness to other types of damage, such as node damage. Therefore, this result suggests that the current theory cannot be extended to new types of damage, such as cluster attacks. This is an interesting finding but more results are necessary in order to draw firm conclusions. For example, detecting more meaningful clusters by applying more advanced community detection methods may reveal new trends in the analysis of robustness to cluster damage presented in the current chapter. In this context, chapters 4 and 5 present the application of one such novel community detection method that is able to identify communities based on the level of their internal interactions and the spatial distances between nodes.

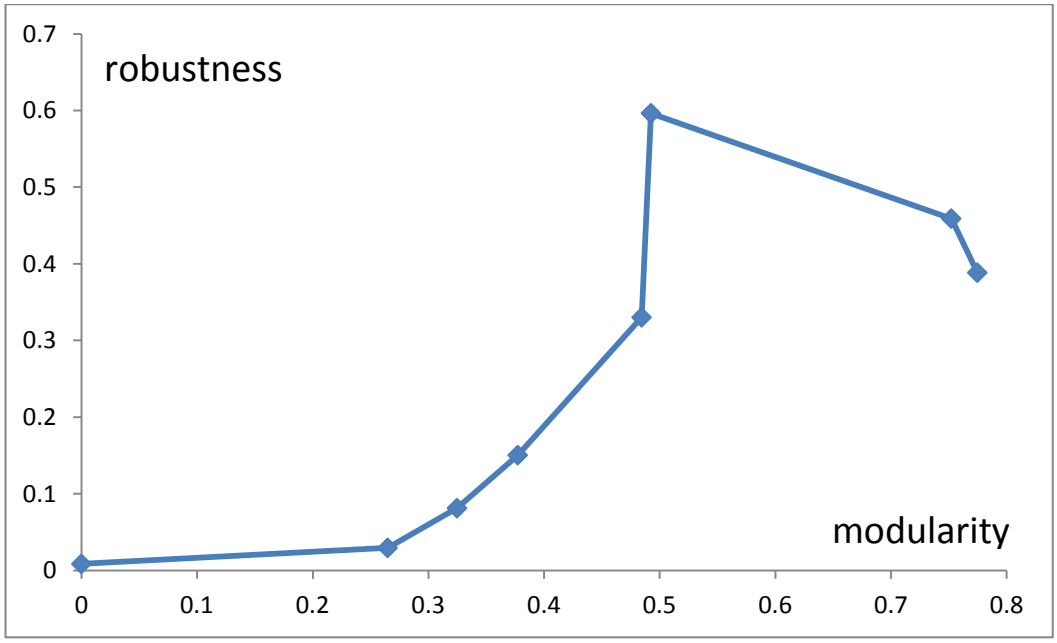

Fig. 3.7. Robustness to cluster attacks as a function of modularity.

3.4 Summary

This chapter presented a novel methodology for damaging clusters in order to measure a new type of network robustness. The methodology was applied to a simple model of the Internet at the Autonomous System level to serve as an example that validates the theory. The results suggest that the Internet, and

perhaps all networks, are less robust to more targeted attacks that seek to maximise their damage.

Chapter 4
Air Transportation Networks

This chapter presents the first applied case study in this thesis on the US Airport Network. US air travel and migration are introduced. An evolution-based model of the network is proposed. The general properties and the community structure of the network are presented.

4.1 Domain Description

This chapter presents a case study of a continuously developing air transportation network that is vital for the mobility of millions of passengers per day. The USAN was chosen for several reasons. Firstly, it is large and growing, so it is clearly a good candidate for studying network evolution. Secondly, there are few detailed models that trace the network for more than a few years. Thirdly, there is a large quantity of available data, dating back to 1990, when the network looked very different to what it is today.

Over the past few decades air travel in the US has changed considerably. Apart from the obvious increase in the number of airports, connections and passengers, the structure (topology) of the USAN has been transformed, thereby affecting all aspects of air travel. Up to the 1970s the USAN had mainly a hub-and-spoke architecture: flights coming from many origins (spokes) converge to the airport (hub) from which new flights start toward other destinations (spokes). The hub-and-spoke architecture is characterised by a high spatial network concentration, and a time co-ordination of flights at the hub according to a *flight wave* concept (Burghouwt and de Wit, 2005). The ideal wave is the set of arriving and departing flights such that for each arriving flight there is a departing one allowing travellers to get an easy transfer to the final destination, and the integration of air services at the hub (e.g. baggage transfer). The main disadvantage of the hub-and-spoke architecture for passengers is that they would have to change flights at the hub, taking more time to reach their final

destination. Furthermore, passengers travelling between other destinations may experience poor service, including infrequent flights and many changes (Hsu and Wen, 2003). As a result, a number of low-cost airlines emerged in the 1980s, providing point-to-point direct services between poorly connected destinations. One example is JetBlue, which is still considered very successful even when compared against larger airlines, such as American Airlines and United Airlines (Bounova, 2009). Consequently, the resulting USAN topology is a combination of both hub-and-spoke and point-to-point architectures.

Migration can be thought of as population redistribution within a country or between countries. It is often linked to an asymmetric distribution of employment and affluence: people are attracted to areas with better job markets, services and quality of life. These aspects relate to the concept of *city competitiveness*, in other words, attractive cities (or regions) are efficient, accessible and offer economic opportunities to both investors and workers (Bulu, 2012; Chorianopoulos *et al.*, 2010; Camagni, 2002; Cervero, 2001). In terms of accessibility, attractive areas have efficient transport systems mainly in terms of external connections linking those areas to other parts of a large territory. In this context, air services can play an important role because they provide fast links among distant locations, even though there may be alternative forms of transportation. In large countries, such as the US, the domestic airport network is a key factor in facilitating domestic migration, i.e. the movement of people within the United States. Particularly, migrants are defined as people moving among states (inter-state migration). Incoming migration (in-migration) is defined as movements into an area during a given period, while outgoing migration (out-migration) is defined as movements out of an area during the same period. The fusion of air transportation with migration is an important novelty in the field of transportation since migration is a driving factor behind many passenger flows in an airport network.

4.2 Data Set

Firstly, it is necessary to decide which specific interactions in the airport network are of particular interest. For example, these can be the number of passengers

flying between airports, the number of aircraft flying between airports, or quite possibly any other metric describing the link between a pair of airports. Then, a long enough time interval is chosen, such that there are available data to be modelled, and the scale of the observed evolution is maximised. The chosen interval is partitioned into equal time slices, depending on the required level of granularity. In the case where a long interval and high granularity result in an unfeasible number of time slices, a sample of those can be selected for the actual modelling.

The number of passengers flying from an origin to a destination airport was chosen as the variable for this study, because it is the common choice in the literature, and it is perhaps the most influential factor in the expansion and organisation of the network. The longest possible time period – from 1990 to 2010 – was selected, based on the availability of data for this period. To investigate seasonal variation within a given year and to build more precise models of the network, time slices of length two months offer a good balance, so a year is divided in six equal parts. To reduce the huge amount of modelling (120 networks) without losing too much information, only three years are modelled in this study: 1990, 2000, and 2010. These years capture the oldest, the intermediate, and the newest, open source states of the network. Data were obtained from the Bureau of Transportation Statistics (BTS) (Bureau of Transportation Statistics, n.d.), of the US Department of Transport. More specifically, data contained monthly records of origin-destination pairs of domestic airports and the number of passengers carried.

4.3 Methodology

The methodology consists of two parts: Network Modelling and Network Analysis. The former describes how the USAN time-series model is developed and the latter presents the network analysis techniques that are used to quantify various properties of the networks.

4.3.1 Network Modelling

To investigate the evolution of the USAN from 1990 to 2010 the network is modelled in a discrete time-series consisting of three stages: 1990, 2000 and 2010. Each of those is further split into six bi-monthly intervals, in order to capture finer temporal detail and to explore seasonal variations in the network. Hence, the network model consists of 18 network snapshots depicting topology and traffic for a two-month time-slice. Each network is defined by a set of nodes (the airports) and a set of directed, weighted links (the flight connections) representing topology. Link directionality reflects the difference in passengers flying from A to B and vice versa. Links weight represents the total number of passengers that flew on that connection within the specified time-slice. Each network includes a number of isolated nodes and self-loops. Isolated nodes denote airports that handled aeroplane departures and/or arrivals, but no actual passengers. Self-loops occur when an aeroplane takes off and lands at the same airport for some reason, such as an emergency.

Using network modelling, both dynamics *on* the network in terms of traffic fluctuations and dynamics *of* the network in terms of topology fluctuations are studied. The more recent structure of the network (reference year: 2010) is compared with migration patterns among the four US macro-regions (West, Midwest, Northeast and South), in order to identify possible relationships. Fig. 4.1 shows a map of the US regions and states, including the locations of the main airports in terms of 2010 passenger flows.

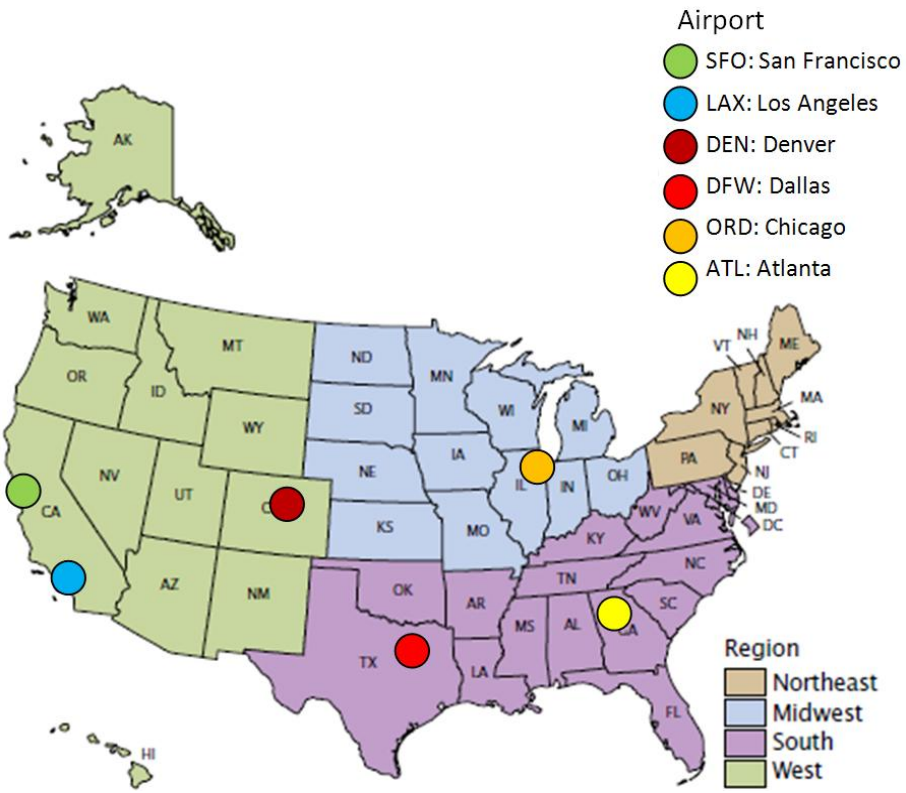

Fig. 4.1. US macro-regions and major airports in 2010 (Mackun *et al.*, 2011).

4.3.2 Network Analysis

The analysis of the USAN involves simple statistical parameter analysis and more complex community structure analysis. The idea of the former is to identify general network properties of the USAN as a whole, such as the average number of airport connections $<k>$. The latter exposes specific traffic patterns at the airport level, thereby revealing deeper individual characteristics. For example, if New York and Los Angeles happen to be members of the same community, then this implies that there is significantly more air traffic between them than expected, given their distance apart.

- **Network Parameters**

In the USAN model, N is the total number of US airports; E is the total number of one-way domestic connections; GCC is the number of airports in the largest connected subnetwork; $<k>$ is the average number of domestic connections per airport; L is the average number of flights that need to be taken to get from A to B; and C is the expected proportion of airport neighbours (all connected to the

airport) that are connected themselves. The latter two of those are calculated for a simple (unweighted and undirected) version of the network due to computational complexity but most connections are bidirectional anyway so the results should be fairly accurate. $P(k_{in})$ and $P(k_{out})$ are the probability distributions of a randomly chosen airport having k_{in} incoming and k_{out} outgoing connections, respectively. By extracting the first two data points (0 and 1 connection) and taking them as separate parameters p and q, the degree distributions are well-approximated by a power-law fitting function of the form:

$$P(k) = ak^n \tag{2-6}$$

where a is the scaling factor, k is in/out-degree, and n is the exponent. Fig. 4.2 shows an example in-degree distribution for the Nov-Dec 2010 snapshot of the USAN.

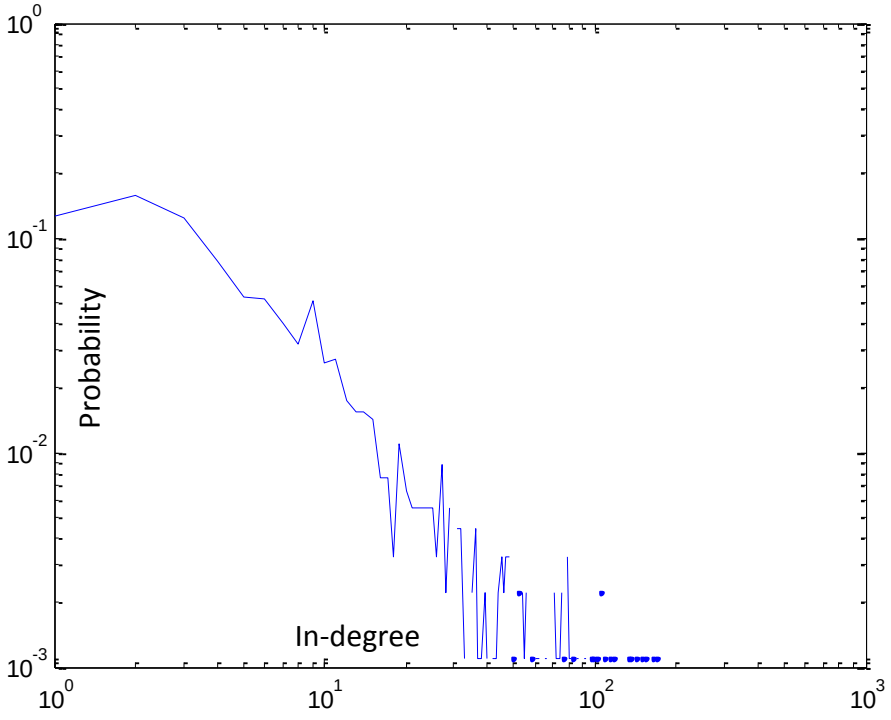

Fig. 4.2. In-degree probability distribution for Nov-Dec 2010 snapshot of the USAN.

W(r) is the rank-ordered passenger distribution on all network connections. For systematic analysis across all networks, *W(r)* is normalised to be in the range (0, 1]. This function is well-approximated by a logarithmic fit of the form:

$$W(r) = bLn(r) + c \tag{2-7}$$

where *b* is the scaling factor, *Ln* is the natural logarithm, *r* is the rank, and *c* is the intercept. In this context, *b* and *c* are the parameters that define the linear transformation needed to map the standard natural logarithm function onto the observed data. Therefore, they have no practical meaning but they are studied here in order to measure the change in the passenger distributions on different network snapshots. Fig. 4.3 shows an example *W(r)* plot for the Nov-Dec 2010 snapshot of the USAN.

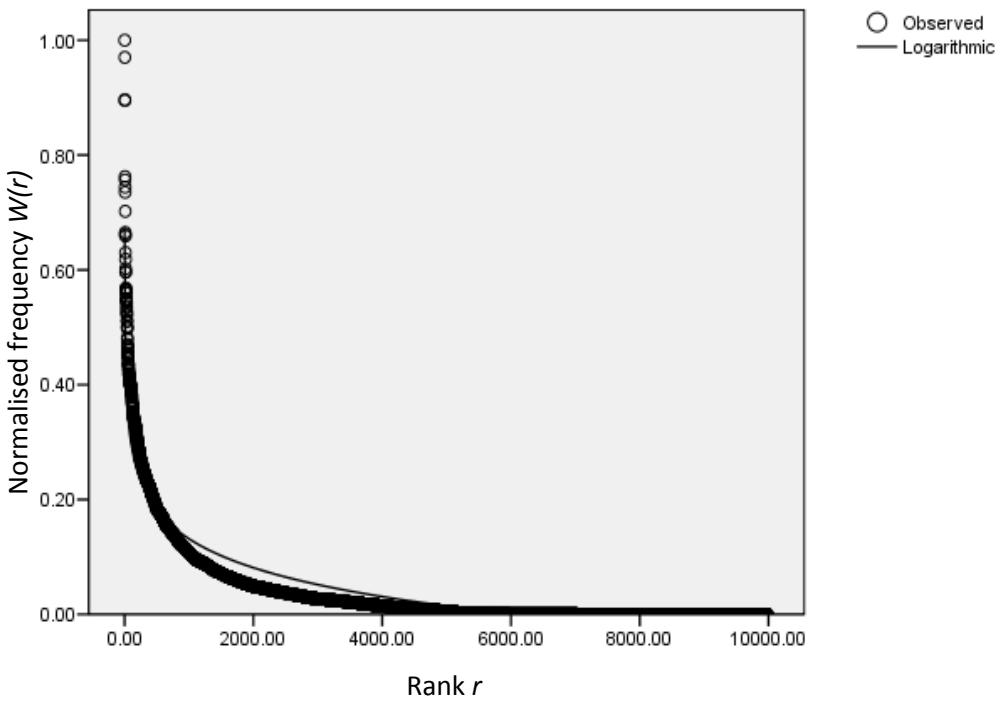

Fig. 4.3. Ranked weight (normalised frequency of passengers) *W(r)* for Nov-Dec 2010 snapshot of the USAN.

Hence, the functions are described by their parameters: p_{in}, q_{in}, a_{in}, and n_{in} of $P(k_{in})$; p_{out}, q_{out}, a_{out}, and n_{out} of $P(k_{out})$; and, *b* and *c* of *W(r)*. To sum up, the networks are analysed in terms of six individual parameters (denoted by capital letters and $<k>$), and ten function parameters (denoted by lower case letters). In

addition, the correlation coefficient R – which measures how well the best-fit approximates the real data – is calculated. The individual parameters are calculated using Network Workbench; the degree distributions are fitted using the EzyFit toolbox for Matlab; and the ranked weight distributions are fitted in SPSS. For each parameter and for each of the three years (1990, 2000, and 2010), the mean parameter value and the Standard Error of the Mean (SEM) of all six network snapshots were calculated. The SEM indicates the amount of bi-monthly variation.

- **Community Structure**

Since distance is expressed in terms of degrees of arc length where one degree is approximately 60 miles, the largest distance in the distance matrix is 149. The bin populations and the deterrence function were checked for bin sizes 0.1, 1, 2 and 3, and 1 was chosen as it provided balanced bin populations and a smooth deterrence function.

There is potentially a large number of nearly-optimal partitions (Good, De Montjoye and Clauset, 2010), and therefore, a non-deterministic implementation of the algorithm is applied twice to each USAN network snapshot, in order to discover better partitions and to check their stability (similar partitions for the same snapshot). This is achieved using Normalised Variation of Information (NVI) (Meila, 2003), which measures the distance between two partitions in the range 0-1 (0 if they are identical, approaching 1 if they are very different). The average NVI values across the six snapshots for the years 1990, 2000 and 2010 are 0.40, 0.34 and 0.26, respectively. These values indicate that the community detection is considerably stable.

4.4 Results

Over the past twenty years, the USAN experiences dramatic growth: airports triple from about 350 to over 1,100, and direct connections double from 5,000 to 10,000.

4.4.1 Network Parameters

Figs. 4.4-4.19 illustrate the trend of each parameter average over the twenty-year period, and the vertical error bars (where visible, due to higher variance) indicate the SEM. Figs. 4.4-4.9 present the six individual network parameters in green. Figs. 4.10-4.17 show the eight degree distribution parameters in blue for in-degree and orange for out-degree. Figs. 4.18 and 4.19 report the ranked weight distribution parameters, *b* and *c*, in red. The results are discussed in section 6.1.1.

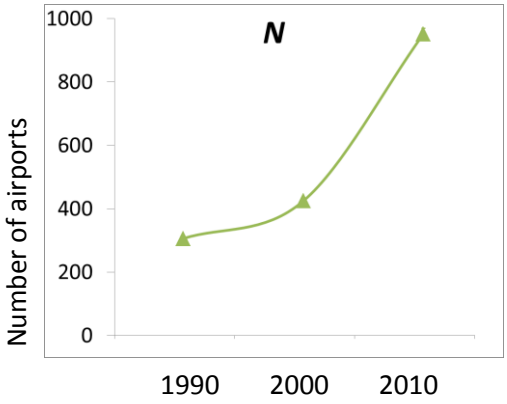

Fig. 4.4. Number of airports as a function of time.

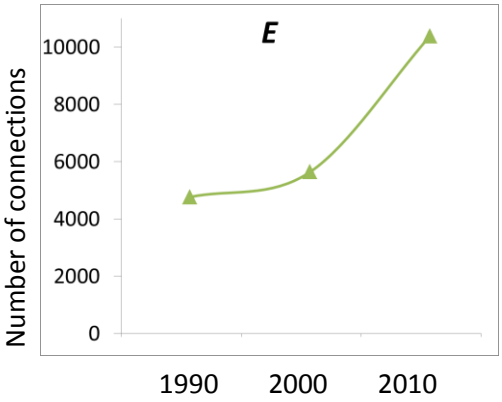

Fig. 4.5. Number of connections as a function of time.

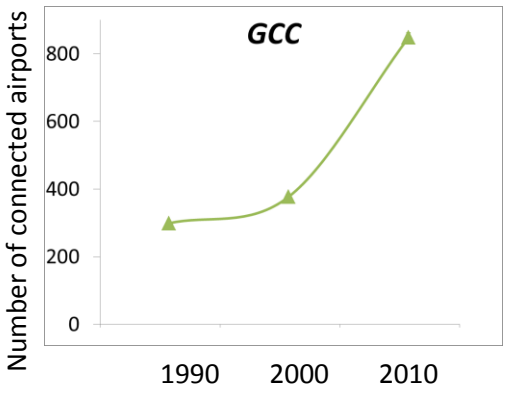

Fig. 4.6. Number of connected airports as a function of time.

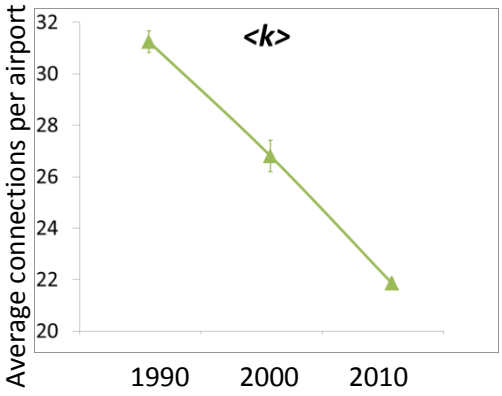

Fig. 4.7. Average connections per airport as a function of time.

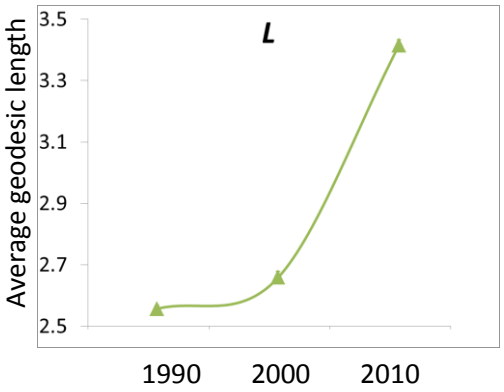

Fig. 4.8. Average geodesic length as a function of time.

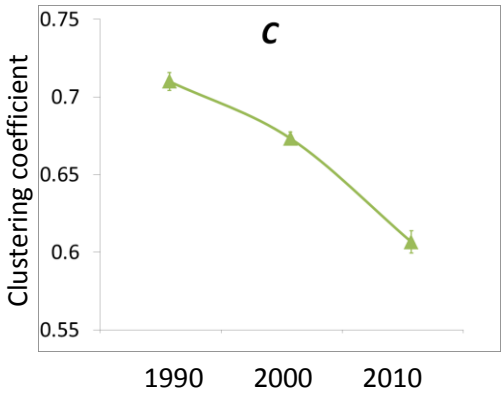

Fig. 4.9. Clustering coefficient as a function of time.

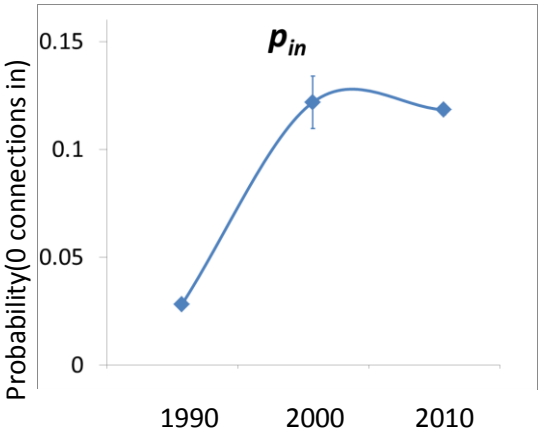

Fig. 4.10. Probability(0 connections in) as a function of time.

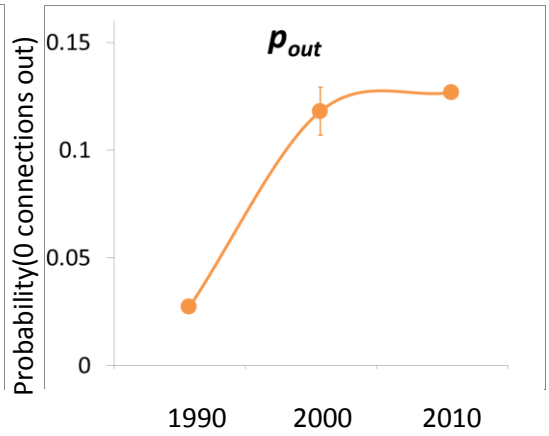

Fig. 4.11. Probability(0 connections out) as a function of time.

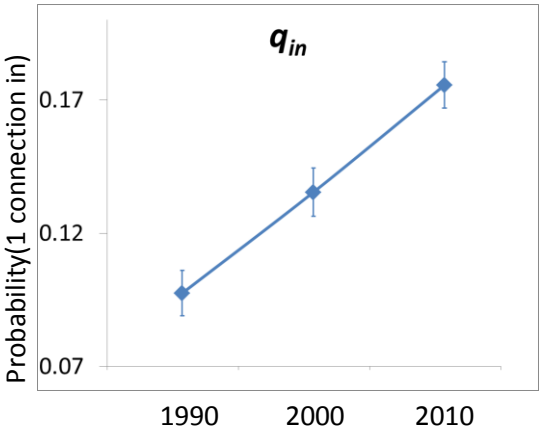

Fig. 4.12. Probability(1 connection in) as a function of time.

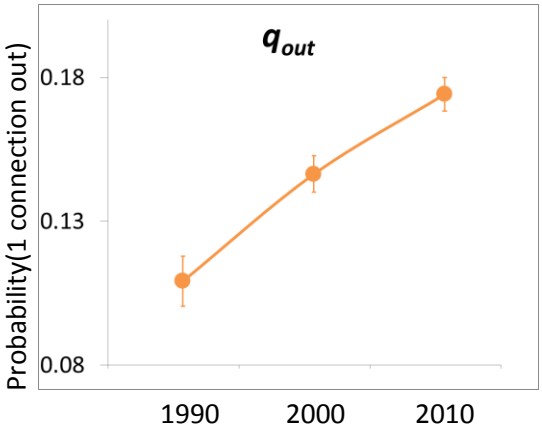

Fig. 4.13. Probability(1 connection out) as a function of time.

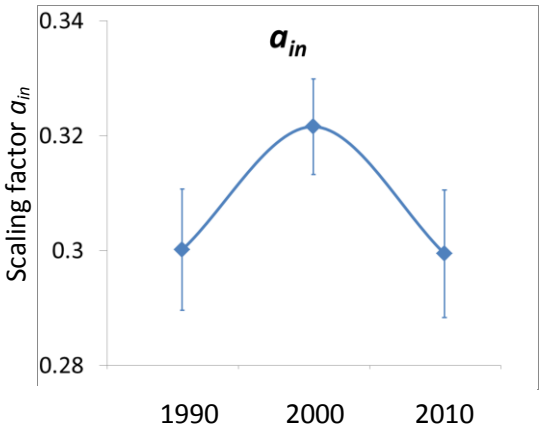

Fig. 4.14. Scaling factor a_{in} as a function of time.

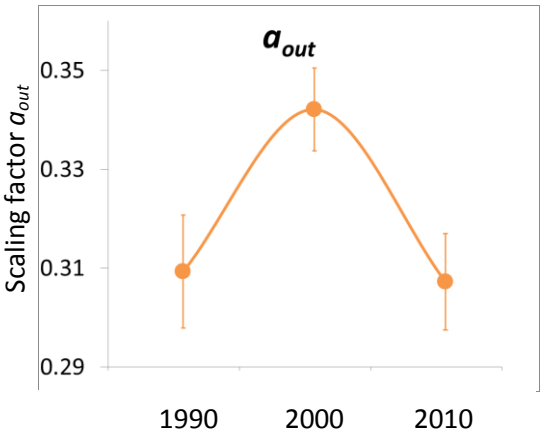

Fig. 4.15. Scaling factor a_{out} as a function of time.

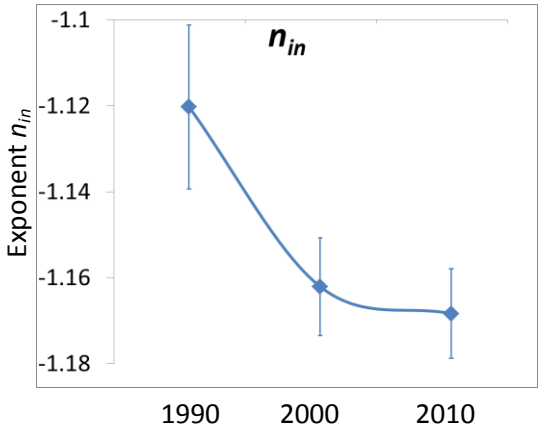

Fig. 4.16. Exponent n_{in} as a function of time.

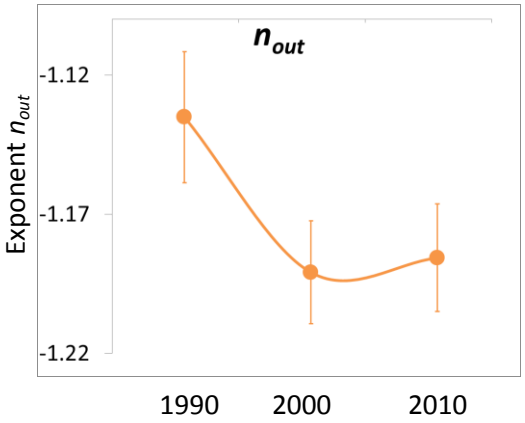

Fig. 4.17. Exponent n_{out} as a function of time.

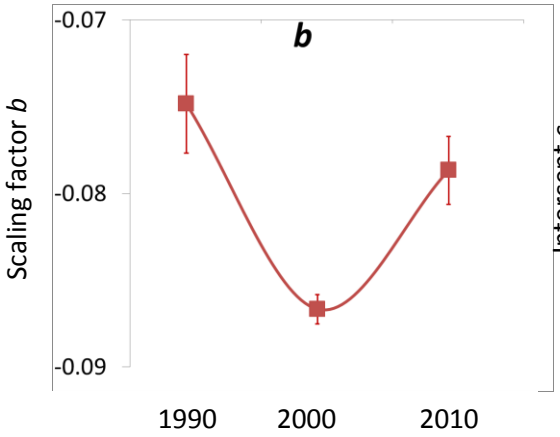

Fig. 4.18. Scaling factor *b* as a function of time.

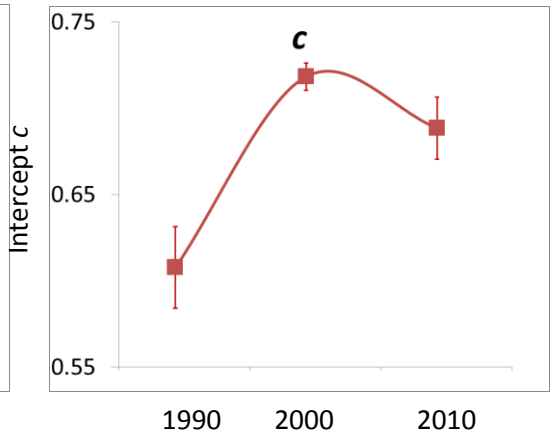

Fig. 4.19. Intercept *c* as a function of time.

4.4.2 Community Structure

Figs. A.1-A.18 (see Appendix) represent the USAN at various stages over time, where each airport is denoted by a circle with a surface area that is directly proportional to the passenger flow (inbound and outbound passengers), and the colour represents the community. Airport connections and airport-to-airport flows are not shown for clarity, and colour is not consistent across the networks as it is only used to differentiate between different communities in a single network (the software used does not allow the user to consistently assign colours

to communities). In other words, the figures depict the size of airports by passengers handled, and the groups of identically coloured airports that have particularly strong passenger flows between them. Alaska, Hawaii and the Mariana Islands are not shown here but they represent a very small fraction of the network. The airport in the bottom right is for the Virgin Islands. In the following analysis of results, the term *hub* is used to describe an airport that handles a high volume of passengers, and the terms *community* and *cluster* are used interchangeably.

- **Year 1990**

In Jan-Feb (Fig. A.1) there is a well-defined cyan community of west-coast airports, such as Los Angeles (LA) and San Francisco, together with Chicago, indicating high passenger mobility between those locations. In Fig. A.2 the network for Mar-Apr implies a particularly large community (light-green) of the main US airports. This means that there were particularly active interactions between all the light-green locations during this time, in contrast to the previous image for Jan-Feb. May-Jun in Fig. A.3 displays a geographically clustered set of communities in the east, together with the largest community in red which spans almost the entire US. In other words, the geographically clustered communities represent the regions where passengers mainly flew locally, and the red community refers to long-distance passengers. Jul-Aug (Fig. A.4) shows a very inter-mixed network, with significant long-distance travel suggested by the spatial spanning of the communities. However, the cyan Dallas cluster is an exception, as it covers only Dallas and small nearby airports. Sep-Oct (Fig. A.5) sees an overall decline in air travel flagged by the noticeable reduction in general size of circles, matching the end of the tourist season, and two large communities in blue and green. In Fig. A.6 Nov-Dec has no major change in traffic patterns apart from the fact that Chicago (a key US hub) is taken over by the spanning blue community, implying that it was used extensively for air travel, particularly among these blue regions.

- **Year 2000**

Jan-Feb in Fig. A.7 displays a prevailing cyan community of most major airports dominating the west and a large part of the rest of the US. In Fig. A.8, Mar-Apr displays a very similar pattern but the number of passengers has increased, which is reflected by the larger circles. In particular, yellow Atlanta (ATL) is clearly the leading US airport in terms of passengers handled during this period. May-Jun in Fig. A.9 suggests that Dallas and Chicago have separated from the largest community in the previous image, forming their own community (in blue) with a few more airports in the north-east. Again, Atlanta is nearly the only member of its yellow cluster, but its size implies that it plays the role of the main hub in the US, connecting many of the other regions. This is explored in more detail in the discussion section. Jul-Aug (Fig. A.10) appears similar to the networks for Jan-Apr, with a main green cluster covering most of the US and Atlanta still on its own. In Fig. A.11 Sep-Oct the number of passengers has predictably decreased. The east appears to be mixed while the west, Dallas and Chicago are all part of the same red cluster. Nov-Dec in Fig. A.12 is similar to the previous network for Sep-Oct.

- **Year 2010**

Fig. A.13 Jan-Feb has two large clusters in red and green covering the west and a large part of the US, respectively. Atlanta (blue) is still the largest hub but passenger demand is low due to the low season. Mar-Apr in Fig. A.14 shows an increase in passengers and a clearly dominating red community in the west. The south is covered by the pink Dallas cluster, and yellow Atlanta and light-green Chicago are the first and second largest hubs, respectively. May-Jun in Fig. A.15 is different in two respects. Firstly, Chicago has formed a yellow cluster covering the south-west and the east, and secondly, orange Dallas has separated from the south cluster, so it has become more of a long-distance travel airport than in the previous two months. Atlanta is still the largest airport by far, providing the connections for many more passengers than any other airport in the US. Jul-Aug (Fig. A.16) is very similar to May-Jun. This means that there is a particularly high volume of travellers among the east coast, the west coast and Chicago. Sep-

Oct (Fig. A.17) has a good mix of many clusters, suggesting that during these months there has been more long-distance travel within the US. The green, yellow and blue communities are particularly well spread out, highlighting the extent of long-range travel. Nov-Dec (Fig. A.18) is similar to the previous two months but now the Chicago and LA clusters have merged again (see May-Jun and Jul-Aug), forming one of the two largest clusters (red and green).

4.5 Summary

This chapter presented the first applied case study on the US Airport Network. The key contribution of this chapter is the first application of space-independent community detection in air transportation. Specifically, Expert's method found high-resolution non-trivial communities of airports with particularly high-traffic internal connections. In addition, a comprehensive study of US air travel in the last two decades revealed detailed trends and relationships among US airports.

Chapter 5
Language Acquisition Networks

This chapter presents the second applied case study on language acquisition networks. The main types of linguistic networks are introduced. A development-based model of the networks is proposed. The general properties and the community structure of the networks are presented.

5.1 Domain Description

As discussed in the literature review, there are three main types of linguistic networks: co-occurrence, syntactic and semantic. Since the co-occurrence network is more general than the latter two in the sense that it can be used to extract both syntactic and semantic content, it is more suitable for modelling language at a high level. This is in line with the main goal of this chapter, which is to provide evidence that MOSAIC is a good model of language acquisition. The main novelty here is the formal validation of MOSAIC that is based on co-occurrence network analysis. In addition, the relationship between mothers and their children is also investigated. Since English is a word-based language, the most meaningful linguistic *chunks* are words, and therefore, word co-occurrence networks are used to model language in this thesis. These networks are easy to build, simple to understand, and contain a lot of encoded information that can be obtained with suitable analysis techniques. Specifically, a word co-occurrence network is defined by a set of nodes representing words (the vocabulary), and a set of directed links representing the flow of words within utterances, i.e. these networks show how words are linked in sentences in terms of the order of occurrence. The key idea is that when large linguistic data sets are modelled, the statistical properties of the data emerge in the model, which aggregates all pairs of adjacent words into a network.

Network analysis is expected to reveal interesting new insights such as trends and patterns in children's distributional analysis of language, which is reflected by

their own linguistic production. In addition, the detection of community structure has not yet been applied to linguistic networks and should expose specific linguistic properties in terms of the clustering of frequently co-occurring words. Another key point is the fact that standard community detection methods are only suitable for non-spatial networks and although co-occurrence networks fit this requirement it is possible to extend them into a non-physical space by introducing an additional parameter describing a non-physical distance between words. For example, apart from the frequency of co-occurrence, the links can also be weighted by the semantic distance between words, thereby incorporating a second dimension of information within the network. In other words, the co-occurrence network describing syntax can be extended to also describe semantics. Then, the semantic distance plays the role of a spatial distance for the purposes of the community structure model by (Expert *et al.*, 2011), which finds communities based on three characteristics of the networks: topology (co-occurrence structure), link weight (frequency of co-occurrence), and spatial separation (semantic distance).

5.2 Data Sets

This section describes four sources of data: mothers, children, MOSAIC and the baseline.

5.2.1 Mothers and Children

The mothers' and the children's data come from the Manchester corpus of the CHILDES database (Theakston *et al.*, 2001; MacWhinney, 2000), which holds large files of logged conversations between mothers and their children, produced while they are interacting at home. Over a significant developmental time period, the children are regularly visited by an experimenter that records all the interactions for a fixed time period. The utterances are recorded on audio tapes that are transcribed to text files by keeping all clearly audible utterances and ignoring anything inaudible. The children's files are partitioned into three discrete, non-overlapping stages of development, and all the files for a given stage are combined to produce three data sets: stage 1, stage 2 and stage 3. The

ages of the children at the start of the stages are presented in Table 5.1. For each mother, the files at the three stages are combined to produce just one data set, as their language should remain fairly stable.

Table 5.1. Age of children at start of stages in years;months.days.

	Stage 1	Stage 2	Stage 3
Ann	1;10.7	2;3.20	2;8.24
Ara	1;11.12	2;1.28	2;4.20
Bec	2;0.7	2;2.22	2;5.8
Car	1;8.22	1;11.12	2;1.25
Dom	2;1.11	2;4.4	2;9.19
Gai	1;11.27	2;2.12	2;4.28

5.2.2 Model Of Syntax Acquisition In Children

Model Of Syntax Acquisition In Children (MOSAIC) is a computer model of language acquisition that simulates the development in children's linguistic capabilities (Freudenthal, Pine and Gobet, 2006). MOSAIC uses distributional analysis to capture precise statistical properties in child-directed speech, such as the location of specific word classes within a sentence. It takes as input transcribed utterances and learns to output progressively longer utterances that can be directly compared to children's utterances over their early linguistic development. MOSAIC is based on a discrimination network consisting of nodes connected by test links. The network always has an empty root node but the other nodes hold words or phrases. Links define the difference between the contents of two nodes. The model encodes utterances by parsing them left-to-right. As the network receives input it creates new nodes below the root. Level 1 nodes (just below the root) are *primitive* nodes. As more input is received, new nodes are added at deeper levels. The model has two learning mechanisms. The first, based on discrimination, adds new nodes and links to the network probabilistically. The second, based on similarity, adds *generative* links between nodes holding phrases encountered in similar contexts. MOSAIC is an extension of CHREST (Chunk Hierarchy and REtrieval STructure), which is in the EPAM (Elementary Perceiver And Memoriser) group (Feigenbaum and Simon, 1984). CHREST models have been successful in simulating novice–expert differences in chess

(Gobet and Simon, 2000), memory for computer programs, and language acquisition (Freudenthal, Pine and Gobet, 2001; Croker, Pine and Gobet, 2000).

- **Example**

An empty network receives the utterance *did he go* so the first input is the word *did.* As the network is empty there are no test links and the model therefore creates a test link and a node under the root node (the model is *learning* the word). Now the new node and test link both hold *did*. The next input is *he* and the model checks the links from the root but since *did* and *he* are different the model now creates a second link and node below the root that encode the word *he* (similarly for *go*). Fig. 5.1 shows the network at this stage.

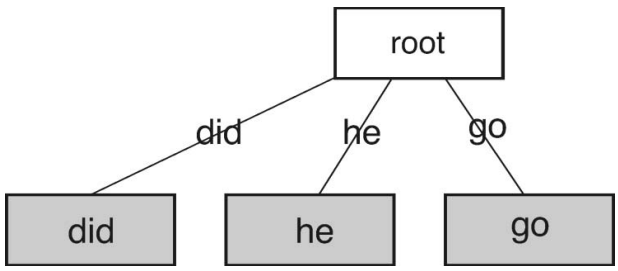

Fig. 5.1. MOSAIC network after one appearance of *did he go* (Freudenthal, Pine and Gobet, 2006)**.**

If the network receives the same utterance once again, it finds a link for *did* (the model *recognises* the word), follows it down, and moves on to the next input *he*. The network now considers test links originating from the *did* node but as there are none and as *he* has already been learnt as a primitive, a new test link and node is created below the *did* node. The link holds the word *he* and the node holds the phrase *did he.* Next, the network recognises *go* but does not learn it since there is no input remaining. When the same utterance is presented a third time, the model parses it until reaching the *did he* node and finds that there is no *go* link, so it creates one under the *did he* node, thereby recording the fact that it has seen the utterance *did he go*. In addition, it also records that *he* has been followed by *go* into the primitive node *he*. Hence, on this pass the model has encoded that *did he* has been followed by the word *go,* as well as the fact that *he* has been followed by the word *go.* Fig. 5.2 shows the network at this stage.

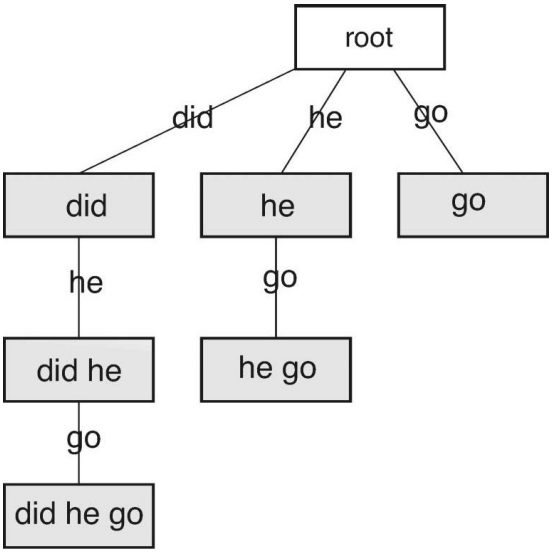

Fig. 5.2. MOSAIC network after three appearances of *did he go* (Freudenthal, Pine and Gobet, 2006)**.**

Suppose the model now sees the phrase *he walks.* It first recognises the word *he.* When it reaches *walks* it tries to create a new test link under *he.* However, there is no primitive *walks* node so the model creates one. When seeing the phrase *he walks* again, it creates the test link *walks* (and node *he walks*) below the *he* node. Now the *he* node has two links encoding that *he* has been followed by *go* and *walks*. Fig. 5.3 shows the network at this stage.

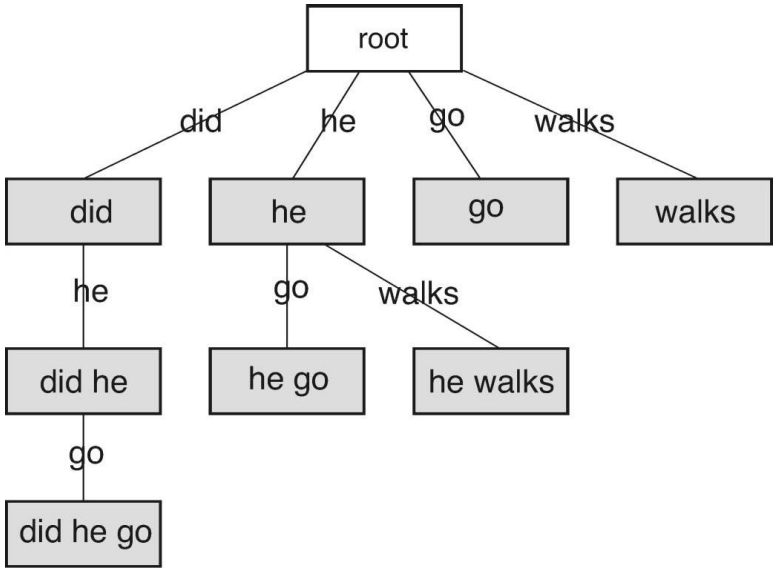

Fig. 5.3. MOSAIC network after three appearances of *did he go* and two appearances of *he walks* (Freudenthal, Pine and Gobet, 2006)**.**

Even though in the example so far consecutive nodes differ by only one word, the model can also represent larger phrases as a single unit. If the model in Fig. 5.3 were to learn the word *does* and then sees *does he go*, it can create a *does he go* node below the *does* node. Since the model already has a node encoding *he go*, it can recognise this phrase as one unit. This *chunking* mechanism enables the model to learn frequent phrases quickly.

In the above example nodes are always added when possible, but in the actual model the addition of nodes is determined by a *Node Creation Probability* (NCP) (Freudenthal, Pine and Gobet, 2006). When NCP = 1, a node is always created (as in the above example), but when NCP < 1 a node may or may not be created. In other words, the lower the NCP, the less likely it is that a node is created. This probabilistic node creation has two advantages. Firstly, a lower NCP value reduces the learning rate of the model, which prevents it from learning long utterances too quickly. Secondly, a lower NCP value makes the model more frequency sensitive. To simulate the range of *MLU*s of young children and to generate enough output, the NCP is set to monotonically increasing values as the network grows. This is in line with empirical observations confirming that children learn new words faster as their vocabulary size increases (Bates and

Carnavale, 1993). In addition, nodes for longer phrases have a lower creation probability. Eq. 2-8 defines the NCP:

$$NCP = \left(\frac{N}{M}\right)^{W} \tag{2-8}$$

Where M is a constant (70,000), N is the number of nodes ($N \leq M$), and W is the length of the phrase in words. Hence, in a small network learning is slow but as the network grows the learning rate increases. The exponent W simply reduces the probability of adding nodes for longer phrases.

- **Generative Links**

MOSAIC's more advanced learning mechanism is based on the creation and removal of *generative* links. These are created between phrases that share a context overlap in terms of the preceding and the following words within the utterance. Since new nodes are constantly added, the percentage overlap between two phrases may drop below a threshold (typically 10%), resulting in the link being removed. If two words belong to the same *word class*, they are likely to be in the same position in a sentence, so they are preceded and followed by similar words.

- **Producing Utterances**

Utterances are produced by traversing the network from the top and reading the contents of the links. By following only test links, the model only produces *rote-learnt* utterances that were present in the input. By also following generative links, it also produces novel *generated* utterances (Fig. 5.4). For example, since *she* and *he* have a generative link, the model can output the novel utterance *she sings*.

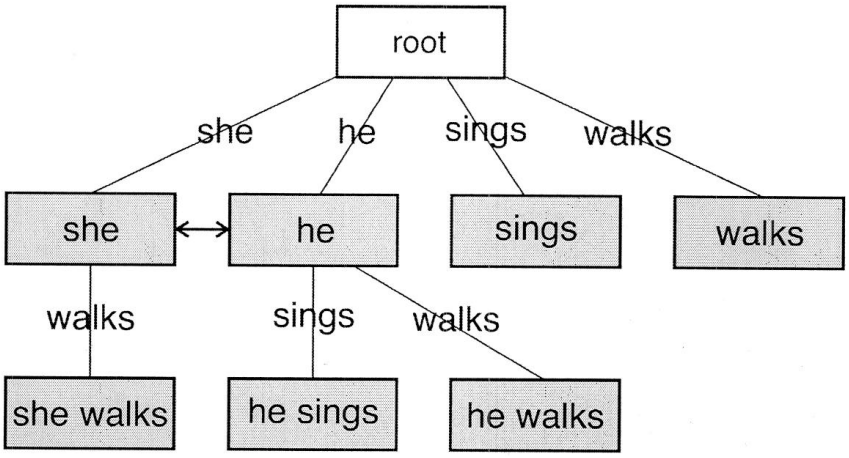

Fig. 5.4. MOSAIC network with a generative link (arrow) (Freudenthal, Pine and Gobet, 2006)**.**

- **The Utterance-Final Constraint**

MOSAIC's output is restricted in order to be more realistic. Specifically, an utterance is produced only if the final word in the utterance was the final word in an input utterance. This is encoded in the model by adding an *end marker* to utterance-final phrases, ensuring that the output consists only of utterance-final phrases and utterances produced by substituting a word into an utterance-final phrase though a generative link. Based on empirical data, research suggests that children are particularly sensitive to utterance-final phrases in terms of learning and understanding (Naigles and Hoff-Ginsberg, 1998).

- **Training and Output**

MOSAIC was trained by iteratively (due to relatively small corpora) presenting the entire corpus of the mother until the model reached an MLU that is close to the MLU of the corresponding child, for a given stage. The output generated by exhaustively traversing MOSAIC's discrimination network was recorded on file, and was later processed as described in 5.3 Methodology.

5.2.3 Baseline

The baseline model is a simple model based on the maternal data. It is developed in four steps, as shown in Fig. 5.5.

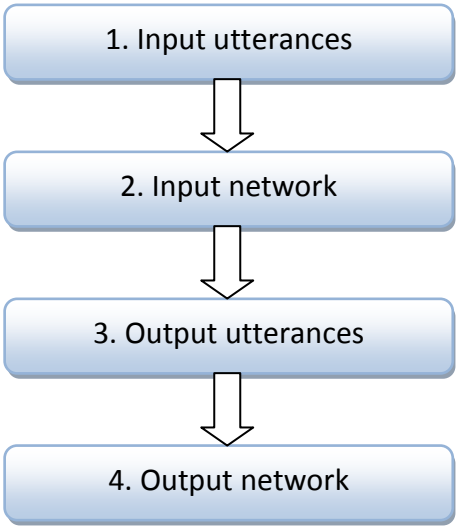

Fig. 5.5. Steps of building the baseline co-occurrence networks.

- **Input Utterances**

To begin with, the maternal data sets are transformed in two ways. Firstly, each data set is reduced to 1000 utterances that are randomly selected from the entire set. The reason for this is twofold: to ensure that all inputs to the baseline model are of equal length for consistency; and to enforce a mechanism of selective reproduction, since the baseline reproduces utterances with every possible co-occurrence that was present in the input. Secondly, each maternal utterance is marked by a beginning marker (a marker here is just an ordinary word and the $ sign is used to differentiate it from the normal words) $BEG and an end marker $END, to denote the start and finish of an utterance. The purpose of this is to enable the baseline to produce utterances that only start with a word that was itself a start word in the maternal utterances, and end on a word that was an end word in the maternal utterances. This simple rule has two benefits. The primary one is that it forces the baseline to mimic the language acquisition of young children, who tend to focus on, remember, and reproduce the beginning and end words of utterances they hear. The secondary one is that it restricts the number of possible utterances that the baseline can produce, which filters out a lot of syntactically incorrect utterances. The transformed maternal utterances are henceforth referred to as input utterances for the baseline model (step 1 in Fig. 5.5).

- **Input Network**

In step 2, the input utterances are converted to a word co-occurrence network – an input network to the baseline model – as described in section 5.3.2. Construction of Networks. Note that the input network is a little different to the ordinary word co-occurrence networks, since it contains two special nodes – a $BEG node and an $END node – that each have a unique property. The former has no incoming links, but it has outgoing links equal to the number of unique starting words that appear in the maternal utterances. Similarly, the latter has no outgoing links, but it has incoming links equal to the number of unique ending words present in the maternal data. The network appears in step 2 of Fig. 5.5.

- **Output Utterances**

In step 3, the input network is used to generate the output utterances of the baseline model using Depth First Search (DFS). A Matlab m-file is written to search for all possible paths (utterances) between the $BEG node and the $END node, below a given length. Note that the $BEG and $END nodes are not part of the path itself. Since all the sources' MLU for a given stage should be more or less the same for consistency in the analysis, the maximum recursion depth (path length) of the DFS is set to either 3 or 4, yielding utterances with MLU just below 3 or 4, respectively. The reason for this is that there are many more paths with length x+1 than with length x that outweigh the shorter paths. The stage 1 scenario is not modelled using the baseline because the children's MLU is 2.27 so the maximum recursion depth must be set to 2, but in this case the entire output consists of unique utterances of length 1 or 2, resulting in no repeated word co-occurrence, which is a trivial scenario. The output utterances generated by the baseline model are represented by step 3 in Fig. 5.5.

- **Output Network**

In the final step 4, the output utterances are converted to an output word co-occurrence network –an output network of the baseline model (step 4 in Fig. 5.5) – as described in 5.3.2 Construction of networks.

5.3 Methodology

In order to test MOSAIC, and hence, children's distributional analysis ability, language acquisition networks built from data sets of the real children's utterances can be directly compared with networks built from utterances produced by MOSAIC. However, it would be difficult to determine the statistical significance of these results alone, as there is no scale to define the level of overlap between MOSAIC and children. Therefore, it is also necessary to test the second linguistic simulation model (the baseline) in order to be able to quantify the quality of MOSAIC's output in relation to the baseline. Then, it would be possible and fair to say exactly how well MOSAIC performs, and therefore, how well it replicates real children's linguistic development (based solely on mothers' child-directed speech). In addition, it is expected that the more basic baseline model will display a much poorer resemblance to the children.

The methodology is composed of three parts: Filtering and reduction, Construction of networks, and Analysis. Filtering and reduction explains the various filtering techniques used to ensure that the data are consistent for network modelling and analysis. Construction of networks presents the steps involved in the creation of word co-occurrence networks from the data. Analysis describes the statistical analysis techniques that are used to measure and to compare the networks.

5.3.1 Filtering and Reduction

Since the raw output of a data set consists of a long list of utterances, some of which have duplicates, those duplicates are removed in order to obtain consistent networks. The reason behind this is that the focus of this work is on the language acquisition of children, and more precisely, the pattern of combining pairs of words together to form sentences. If duplicate utterances – which are mainly caused by the highly repetitive nature of mother-child interactions – are allowed, they would introduce a lot of noise in the data. The baseline model, however, does not produce any duplicate utterances so it needs no filtering. After they are filtered, all the data files' lengths – in terms of the number of utterances they contain – are recorded. For a given stage, it is noted that the children's data files

are always the shortest, except for MOSAIC's stage 1 data file for Carl. This result is a little surprising but it is reasonable to assume that it is due to the fact that for stage 1, MOSAIC is lacking output due to a lack of input data, while Carl is surprisingly talkative at such an early stage. Again, to make the data files as consistent as possible for later systematic analysis, all longer files (except the maternal, since they do not correspond to any of the 3 stages) are randomly reduced to the length of the shortest file for a particular stage. The reduction is done using Matlab's random permutation function, which assigns each utterance a unique natural number between 1 and the total number of utterances. Then, to complete the reduction, all utterances that were assigned a number that is larger than the size of the required data set are discarded. To check the quality of the reduction, the MLU of the reduced data sets is re-calculated and the obtained differences are negligible.

5.3.2 Construction of Networks

Word co-occurrence networks are built using the data in a process consisting of four stages. Fig. 5.6 illustrates a simple version of such a network for the following two sentences: *The cat sleeps*. *The dog wakes the cat*.

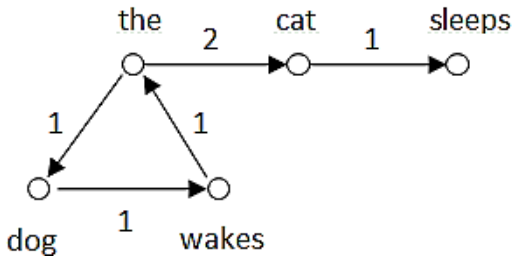

Fig. 5.6. Simple word co-occurrence network.

In the first stage, the utterances are split into overlapping pairs of adjacent words, such that each co-occurrence is represented by a single pair. For example, the first two pairs in this sentence are "For example" and "example the", when ignoring any punctuation marks. Also, single word utterances are kept unchanged because they become isolated nodes in the final network, if they do not appear in any pair elsewhere. In stage two, the pairs are transformed into a corresponding network of nodes and links using a simple mapping: for every pair, insert a

directed link from the first word to the second word. The network is represented as a list of nodes and links, where the nodes are trivially obtained by taking all words from the pairs, and removing duplicates. However, this network is a multidigraph, i.e. there exist multiple links from a source to a target node, which can be simplified to an equivalent network representation by using link weights. Therefore, in stage three, the network is converted to a weighted digraph – with no parallel links – where the weight (frequency) of a link denotes the original number of links from the source to the target node in the multidigraph. This digraph is used in all of the analysis except for the calculation of two network parameters – the average geodesic length and the clustering coefficient – which require a simple graph, i.e. an unweighted, undirected graph with no self-loops. Hence, for the final stage four, the weighted digraph is converted to a simple graph by erasing all link weights, converting all links from directed to undirected (and removing duplicates), and removing all self-loops.

5.3.3 Network Analysis

The goal of this analysis is to carry out a thorough comparison of all word co-occurrence networks by employing a number of established statistical analysis techniques. Furthermore, the results of this analysis will highlight specific features of the networks, such as their power-law ranked frequency distribution. These features will be used to compare MOSAIC with the simple baseline model, in terms of their ability to simulate language acquisition in children.

- **Network Parameters**

Parameter analysis involves the calculation and analysis of the selected network parameters that are described in more detail below (most of them were briefly described in section 2.4):

1. *MLU*

The Mean Length of Utterance (*MLU*) is the average number of words in a sentence within a data set. Therefore, this parameter measures the length of the produced utterances. The *MLU* is perhaps the most basic parameter since it is a measure on the data set, not on the network. It is frequently used to define the

level of language development, and therefore, it is used in this research to classify the data sets into three developmental levels, in order to be able to analyse and compare data sets belonging to the same discrete level.

2. *N*

The number of nodes *N* in a word co-occurrence network simply reflects the number of unique words, which were produced within a data set. It is a typical measure of vocabulary size.

3. *E*

The number of links *E* in the network measures the number of unique word co-occurrences that were produced within a data set. More links indicate that more diverse utterances were created.

4. *GCC*

The Giant Connected Component (*GCC*) is the largest connected subnetwork of the original network. Therefore, the size of the GCC represents the number of core words, which are commonly used to form sentences.

5. <*k*>

The average degree <*k*> is the average number of links adjacent to a node. This parameter represents the complexity of the vocabulary, since a higher <*k*> means that more unique co-occurrences are produced. The average degree is a function of the number of nodes *and* links within the network, and therefore it provides a good relative measure of the diversity of the utterances.

6. *L*

The average geodesic length *L* is the average number of links on the shortest paths between all pairs of nodes. This parameter is calculated for a simple graph and gives an indication of how well-interconnected the graph is. The shorter the length, the quicker you can get from word A to word B. The typical average geodesic in co-occurrence networks for natural language is low.

7. *C*

Roughly speaking, the clustering coefficient *C* is a measure of the redundant links in the network. More precisely, it is a measure of how many neighbours of a node are directly connected themselves. This parameter is also calculated for a simple graph and gives an indication of how clustered the graph is. High clustering implies that diverse utterances are produced using a small number of words, suggesting more advanced language skills. Note that there is an important distinction between *C* and community structure as *C* measures general clustering but community structure reflects the modular nature of certain groups of nodes within the network.

8. *P(k)*

Since the networks are directed the degree distribution consists of the in-degree $P(k_{in})$ and out-degree $P(k_{out})$ distributions. $P(k_{in})$ is the probability distribution of a given node having some number of links pointing to it. A high in-degree node has many other nodes pointing to it and hence, the word represented by the node has been preceded by many other words in the respective utterances. $P(k_{out})$ is the probability distribution of a given node having some number of links pointing away from it. A high out-degree node points to many other nodes and hence, the word represented by the node has been followed by many other words in the respective utterances. Fig. 5.7 shows an example in-degree distribution for the Ann stage 3 network.

9. *W(r)*

W(r) is the rank-ordered frequency distribution on all network connections. It shows how the magnitude of link frequencies decreases when the frequencies are sorted in descending order. To compute this function, all the links of the given network are ranked in order of frequency. This frequency is then normalised for consistency to a value between 0 and 1 by dividing all frequencies by the highest frequency in the network. The distribution is defined as the normalised frequency as a function of the rank. This distribution is particularly interesting for studying language because previous research (Corominas-Murtra, Valverde and Solé, 2010) has shown that *P(f)* – the probability distribution of a node with frequency *f* – in children's syntactic networks follows a power-law. Fig. 5.8 shows an

example *W(r)* from our maternal results, which also appears to follow a power-law, i.e. most co-occurrences have low frequency of repetition whereas some particular co-occurrences have exceptionally high frequency of repetition. A power law is usually described by two parameters: a scaling factor *a* and an exponent *n*:

$$f(x) = ax^n \tag{2-9}$$

The scaling factor simply determines how much the function is shifted along the y-axis. The exponent controls the slope of the function; thus, a higher *n* (in absolute sense) results in a more skewed distribution. In a co-occurrence network, the presence of a power law means that language productivity is very biased towards some word co-occurrences, which are produced much more often than other co-occurrences. Based on this, *n* is expected to increase over incremental stages of linguistic development.

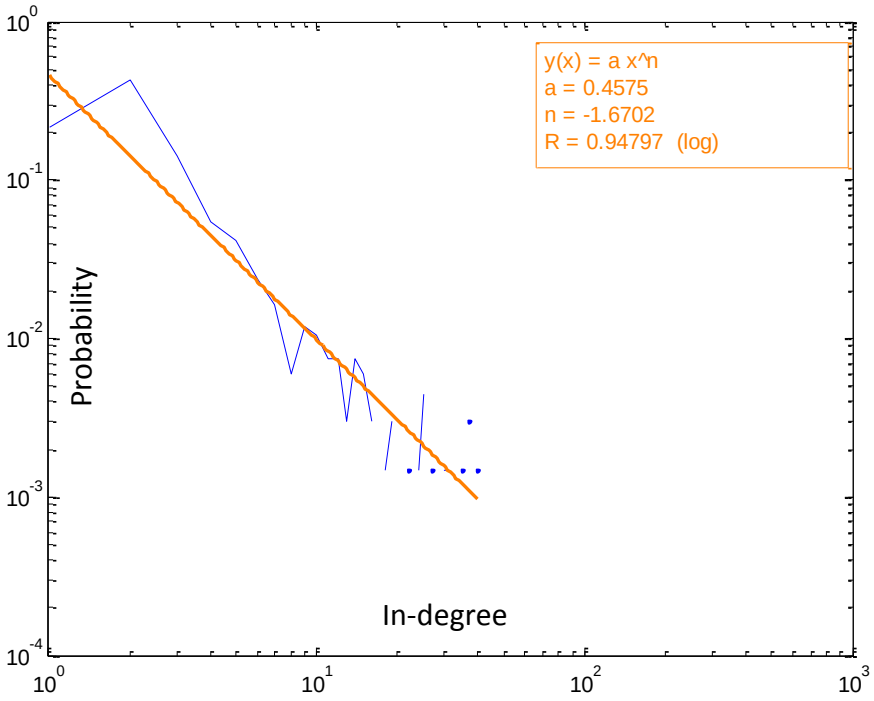

Fig. 5.7. In-degree probability distribution for Ann stage 3 network.

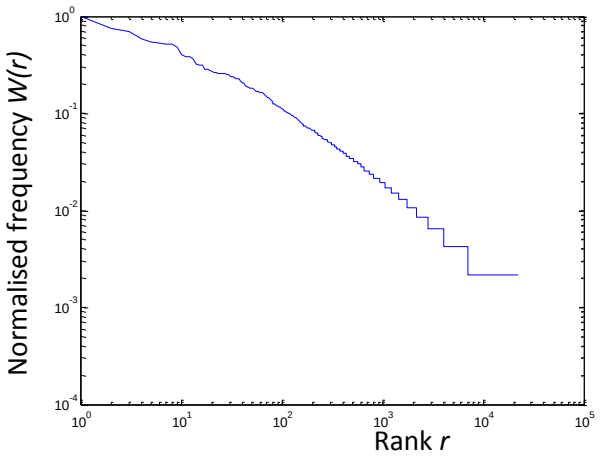

Fig. 5.8. Ranked weight (normalised frequency) distribution *W(r)*.

The flowchart in Fig. 5.9 describes the process of analysing the word co-occurrence networks. The analysis begins with a decision. If a single parameter is being analysed, the value of this parameter is used and the fitting step (described next) is by-passed. By contrast, if a function (i.e. the frequency distribution or the degree distribution) is being analysed, it is necessary to fit a best-fit curve to the data so that the parameters of the function can be used for further analysis. Both functions are fitted using a power-law fit of the form of Eq. 2-9 where *a* is the scaling factor and *n* is the exponent. In the frequency distributions, *x* represents the rank and *f(x)* represents the normalised frequency for that rank. In the degree distributions, *x* represents the degree and *f(x)* represents the probability of a randomly chosen node with degree *x*. The correlation coefficient R – which measures how well the best-fit approximates the real data – is also calculated. Then, the analysis process follows each of two branches. The first branch is concerned with the average for a given source and stage, thereby ignoring the specifics of the individual children. For each parameter, the average value across the stage is calculated and a summary plot is produced. The second branch of the analysis involves the correlations between pairs of sources for a given child and stage thereby focusing on the details of the individual children. Therefore, all the parameters are tabulated and pair-wise correlations are calculated for all possible pairs of data: children-MOSAIC, children-baseline, children-mothers, MOSAIC-baseline, MOSAIC-mothers, and baseline-mothers. The purpose of the

correlations is to identify common properties and characteristics among networks from different sources.

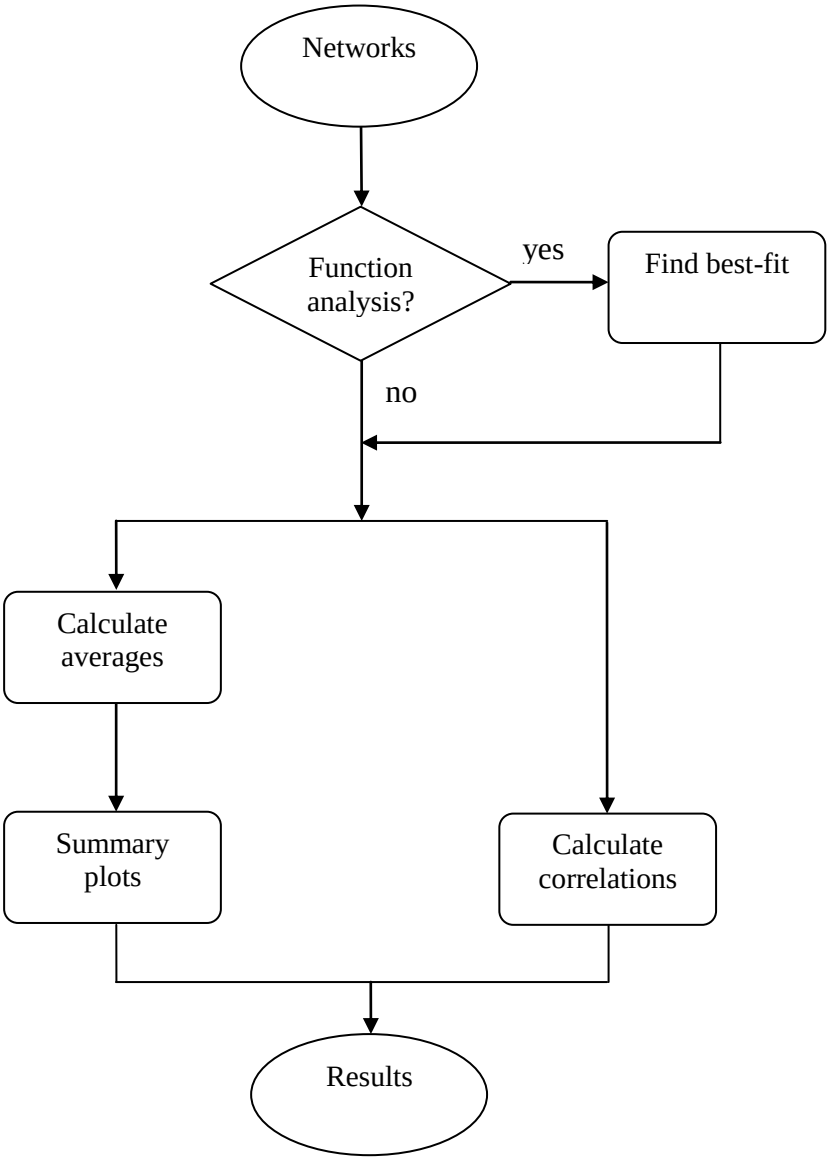

Fig. 5.9. Analysis flowchart.

- **Community Structure**

Semantic distances between words were obtained using Exrelan (Russell *et al.*, to be published), a Python-based tool that takes as input data in the form of text files, and returns semantic distances. These were then normalised in the range 1-100 for community detection. The bin populations and the deterrence function

were checked for bin sizes 0.5, 1, 2, 3, 4, 5, 10, 20, 25 and 34. Bin size 2 provided balanced bin populations and a smooth deterrence function so it was used in all applications of the spatial null model.

5.4 Results

5.4.1 Network Parameters

The results obtained from the parameter analysis are discussed in three parts. Part A addresses the individual parameters, which are based on the global structure of the entire network. Part B covers the degree distribution parameters, which describe the structure of the word co-occurrence networks in terms of the number of unique words that co-occur with each word. Part C focuses on the weight distribution parameters, which highlight the high heterogeneity in the words' frequency of co-occurrence with other words. All of the parameters obtained are presented in tables. The summary plots that follow present the averages for a given source and stage. Each parameter is presented on a single plot showing the change of the parameter value over the three discrete stages. Each of the four sources is associated with a specific colour and symbol. The data points represent the mean parameter value over all six children, and the vertical error bars represent the Standard Error of the Mean (SEM). The mothers' results are shown next to the other sources' stage 3 results for comparison. Finally, correlations between all pairs of sources are reported in order to identify which parameters, if any, from one source resemble those from another source. Again, the mothers' parameters are classified as stage 3. Overall, some of the correlations are moderate ($0.30 \leq r < 0.50$) or even strong ($r \geq 0.50$), to use Cohen's criteria (Cohen, 1988). However, given the small number of observations (6) and hence degree of freedom ($n = 4$), none of the correlations are statistically significant. The quality of all best-fit functions is checked by computing the correlation coefficient R between the real data and the fitted function, in order to check the accuracy of the fit. For the degree distribution fits all correlations are above 0.98, indicating an almost perfect power-law relationship in the data. For the weight distribution fits all correlations are above 0.81, which is good.

- **Individual Network Parameters**

The individual network parameters of the word co-occurrence networks are summarised in Table 5.2.

Table 5.2. Individual network parameters.

		MLU	*N*	*E*	*GCC*	*<k>*	*L*	*C*
Mothers	ann	5.13	3286	23375	3093	14.23	2.74	0.43
	ara	6.11	4669	33964	4604	14.55	2.77	0.45
	bec	5.01	2696	16765	2588	12.44	2.79	0.41
	car	4.56	2428	15340	2297	12.64	2.80	0.38
	dom	5.26	2845	19314	2702	13.58	2.80	0.42
	gai	4.86	4179	22123	3913	10.59	2.87	0.43
Children	ann 1	2.14	764	1429	556	3.74	3.39	0.19
	ann 2	3.32	1384	4942	1205	7.14	2.97	0.33
	ann 3	3.44	672	1678	594	4.99	3.32	0.17
	ara 1	2.43	643	1660	487	5.16	2.95	0.34
	ara 2	3.16	725	2119	651	5.85	3.00	0.33
	ara 3	3.84	1408	5254	1309	7.46	2.96	0.32
	bec 1	1.92	754	1020	498	2.71	3.76	0.15
	bec 2	2.74	956	2330	824	4.87	3.22	0.20
	bec 3	3.59	1385	5407	1245	7.81	2.96	0.31
	car 1	2.45	467	1188	385	5.09	3.07	0.29
	car 2	2.66	714	2578	628	7.22	2.93	0.33
	car 3	3.58	1203	5948	1138	9.89	2.76	0.38
	dom 1	2.43	650	1738	558	5.35	3.13	0.25
	dom 2	3.29	1227	5039	1087	8.21	2.98	0.30
	dom 3	3.38	510	1296	449	5.08	3.21	0.15
	gai 1	2.27	889	1381	622	3.11	3.53	0.14
	gai 2	3.05	1005	2415	865	4.81	3.26	0.19
	gai 3	3.41	1397	4200	1230	6.01	3.11	0.29
MOSAIC	ann 1	2.11	602	1285	376	4.27	3.28	0.20
	ann 2	3.14	1142	4436	984	7.77	3.01	0.28
	ann 3	3.97	735	2146	715	5.84	3.06	0.21
	ara 1	2.13	718	1382	419	3.85	3.40	0.24
	ara 2	3.12	1013	2695	892	5.32	3.24	0.26
	ara 3	3.88	1541	5754	1463	7.47	2.95	0.35
	bec 1	1.96	499	816	249	3.27	3.49	0.26
	bec 2	3.02	732	2255	614	6.16	3.05	0.25
	bec 3	3.71	1173	5178	1084	8.83	2.84	0.33
	car 1	1.82	503	726	245	2.89	3.60	0.25
	car 2	2.82	767	2634	644	6.87	3.02	0.24
	car 3	3.57	1189	6190	1092	10.41	2.79	0.30
	dom 1	2.12	585	1303	349	4.45	3.13	0.26
	dom 2	3.22	964	4318	833	8.96	2.88	0.29
	dom 3	3.84	668	1780	637	5.33	3.23	0.18
	gai 1	2.07	567	1006	281	3.55	3.49	0.25

	gai 2	2.65	916	2104	698	4.59	3.22	0.20
	gai 3	3.32	1308	4364	1168	6.67	3.10	0.27
Baseline	ann 2	2.93	613	1759	611	5.74	3.21	0.15
	ann 3	3.90	484	1352	481	5.59	3.02	0.21
	ara 2	2.94	538	1446	538	5.38	3.07	0.22
	ara 3	3.92	698	2227	698	6.38	2.94	0.31
	bec 2	2.93	531	1364	529	5.14	3.14	0.20
	bec 3	3.90	689	2066	689	6.00	3.09	0.23
	car 2	2.92	524	1418	519	5.41	3.15	0.18
	car 3	3.90	670	2072	669	6.19	3.08	0.22
	dom 2	2.92	570	1770	570	6.21	3.13	0.18
	dom 3	3.91	436	1256	436	5.76	2.95	0.26
	gai 2	2.93	556	1385	556	4.98	3.15	0.19
	gai 3	3.90	682	1912	682	5.61	3.08	0.24

The table of network parameters is large, and therefore, this section highlights the most notable observations before each of the four sources is individually described. An interesting observation of Table 5.2 is that (for all multi-stage sources) the stage 3 networks for Anne and Dominique (represented as ann 3 and dom 3 in the table, respectively) are smaller than the corresponding stage 2 networks, which is unexpected for two reasons. Firstly, the stage 3 networks are based on a later stage of linguistic development, so it is highly likely that the vocabulary (number of nodes) grows and linguistic complexity (<*k*>) increases. Secondly, stage 3 is defined by a higher *MLU*, which means that the data files have longer utterances – containing more words and co-occurrences – so the number of unique words (nodes) and distinct co-occurrences (links) should increase. In particular, those stage 3 networks are smaller in terms of the number of nodes and links, the *GCC* and <*k*>. Furthermore, *L* increases and *C* decreases, except for the baseline model, which displays the opposite trend, possibly because its <*k*> does not drop that much compared to MOSAIC and the children.

The following summary briefly describes the individual sources. All the maternal networks are relatively similar in terms of network parameters, but some are larger than others in terms of nodes and links, and hence, in *GCC*. Compared to the other sources' networks, the maternal have the highest *MLU*, the most nodes and links, the largest *GCC*, the highest <*k*>, low *L* and high *C*. This is rooted in the mothers' experienced use of language. The children's networks differ mainly across the three stages. In general, for higher stages the *MLU* is higher (this is in

fact always true for all sources, since the *MLU* defines the stage of development), the network is bigger (more nodes and links), the *GCC* is larger, and <*k*> is higher. MOSAIC's networks also generally differ across the stages, with higher stages having larger networks, larger *GCC*, and higher <*k*>. Similarly to the previous two sources, by comparing the two stages of the baseline, it is easy to see that the stage 3 networks are generally larger, with a bigger *GCC*, higher <*k*>, lower *L*, and higher *C*. Note that the baseline is the only source from the three multi-stage sources to display a consistent drop in *L* and rise in *C*, for all six children.

Figs. 5.10-5.16 present the summary plots for *MLU*, *N*, *E*, *GCC*, <*k*>, *L* and *C*, respectively.

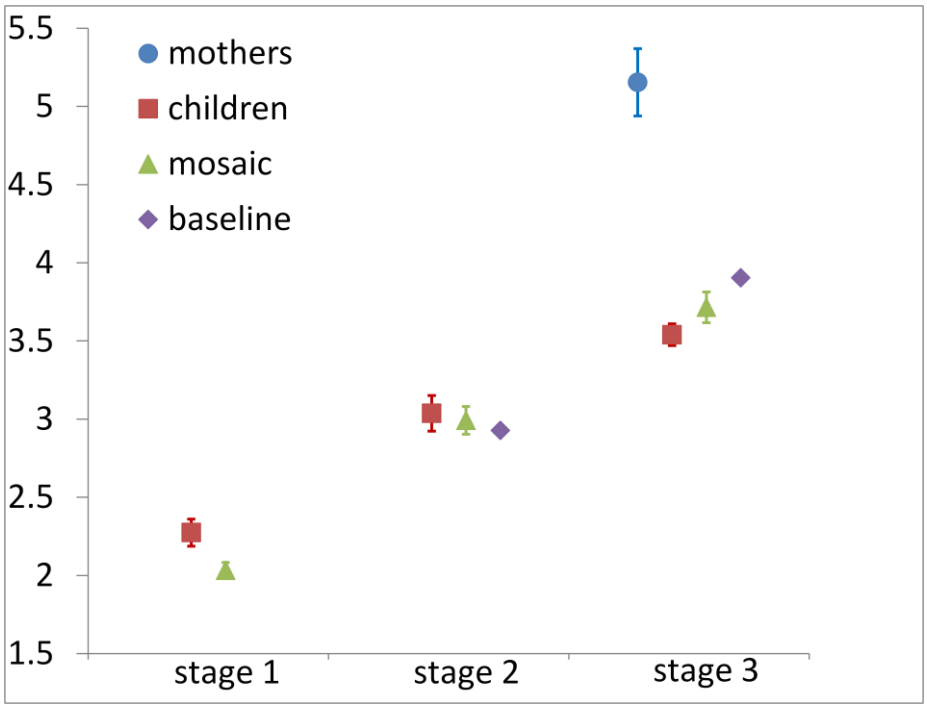

Fig. 5.10. Summary plot for *MLU*.

Fig. 5.10 shows the steady increase in *MLU* for the three stages of linguistic development. Note that, except for the mothers, the sources are close to each other for a given stage, highlighting the fact that they are closely matched according to *MLU*, in order to provide as accurate results as possible. The maternal *MLU* is clearly much higher, due to the longer sentences produced by the mothers.

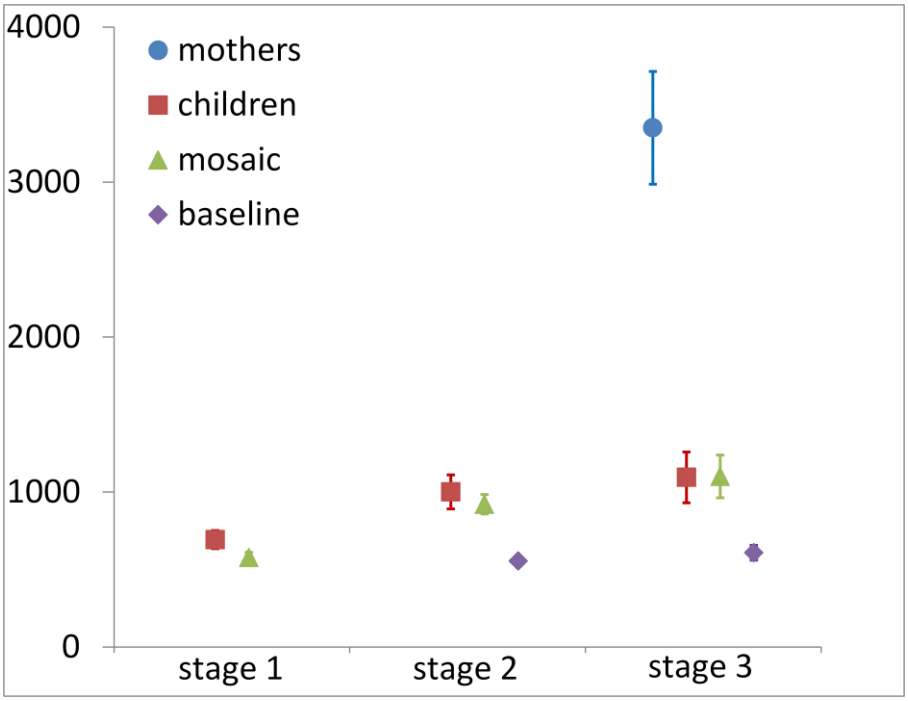

Fig. 5.11. Summary plot for *N*.

MOSAIC and the children clearly correlate very well on the number of nodes (unique words), whereas the baseline seems to underperform, and the mothers are way above the others (Fig. 5.11). Also, the vocabulary seems to grow more between stages 1 and 2 than between stages 2 and 3.

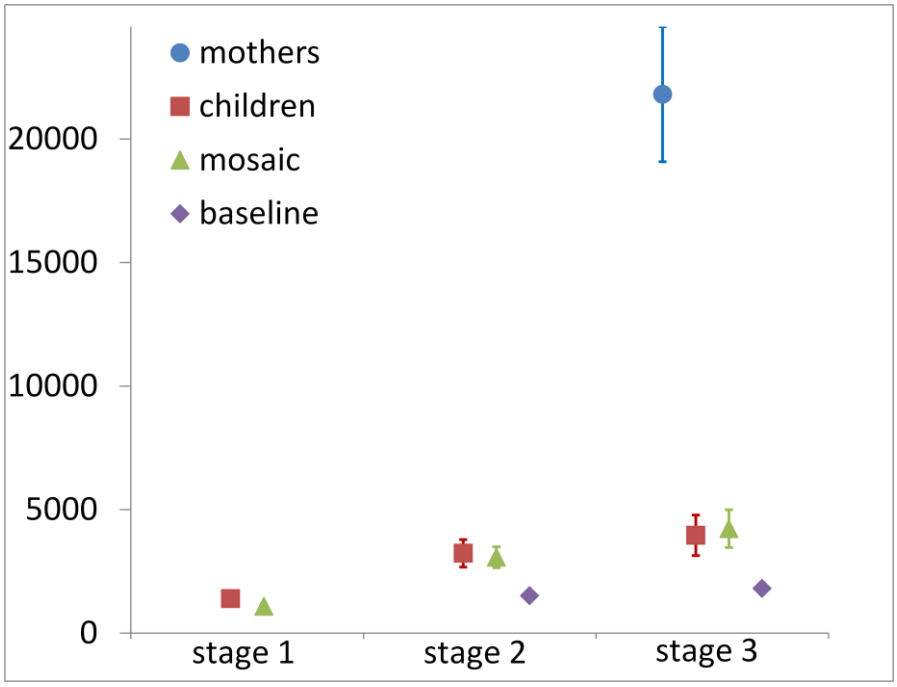

Fig. 5.12. Summary plot for *E*.

The number of links grows in a similar manner to the number of nodes, but this time the baseline model performs more like the children and MOSAIC (Fig. 5.12). This trend implies that as children grow, they not only increase their vocabulary, but also produce new word co-occurrences. The extent to which these new co-occurrences are a direct effect of the increasing vocabulary can be determined by the average node degree, $<k>$, which is summarised in Fig. 5.14.

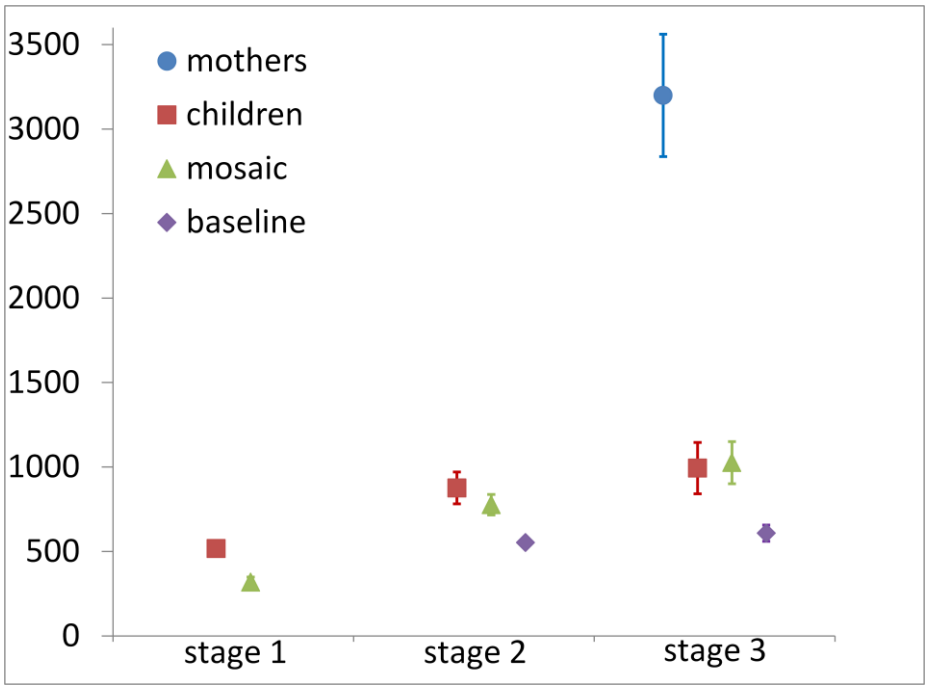

Fig 5.13. Summary plot for *GCC*.

The size of the giant connected component appears to be behaving in a similar fashion to the number of nodes and links, but in fact, for the children and MOSAIC, the fraction of the network that is connected (GCC/nodes) increases over time (Fig 5.13). In addition, the mothers, and especially the baseline, have a particularly high GGC/nodes ratio approaching 1.

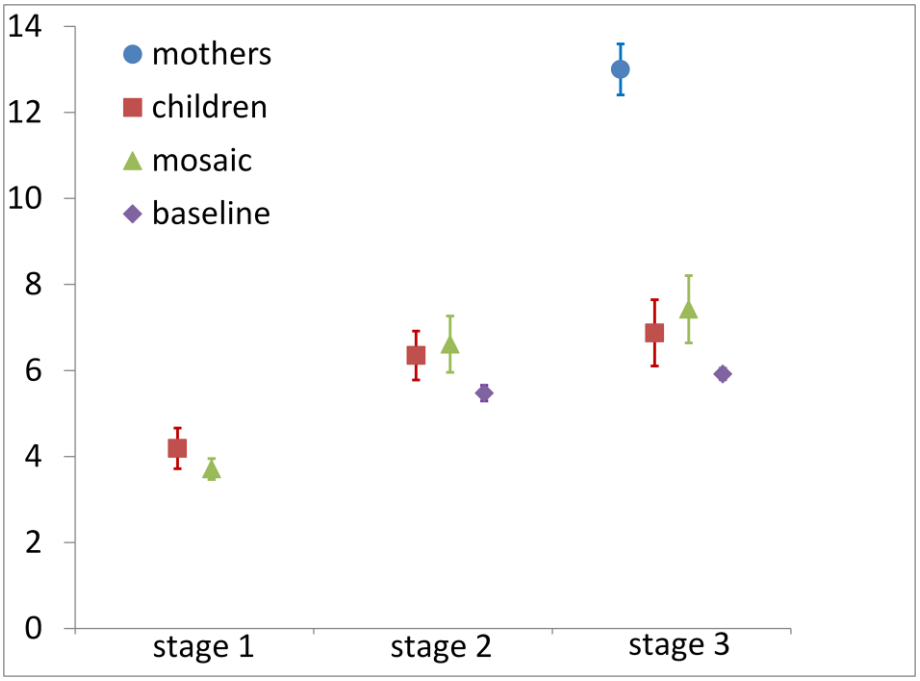

Fig. 5.14. Summary plot for <*k*>.

The average degree <*k*> is clearly increasing over time for all multi-stage sources, which means that there are more links per node (Fig. 5.14). This implies that new co-occurrences are being produced *not* just as a result of the increasing vocabulary, but also as a result of the developing ability to produce linguistic diversity in the form of unique co-occurrences, and hence, novel utterances. Again, note how well MOSAIC mimics the children in this respect.

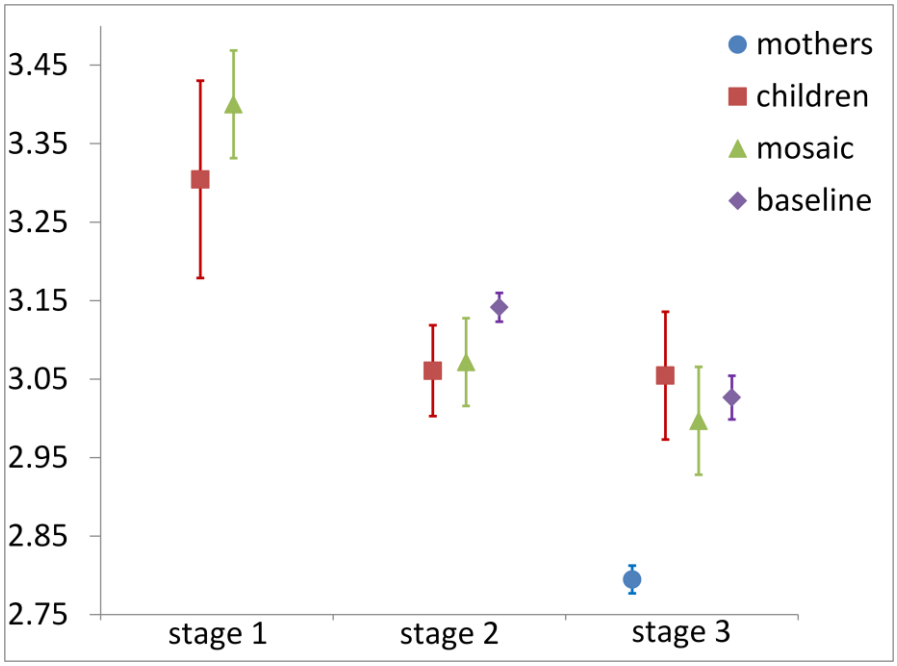

Fig. 5.15. Summary plot for *L*.

In the children's and MOSAIC's networks, the average geodesic, *L*, drops quite significantly from stage 1 to stage 2, but then relatively little, from stage 2 to 3 (Fig. 5.15). Naturally, since <*k*> is increasing, *L* should decrease at a similar rate if the proportion of shortcut links (that connect otherwise distant nodes) remains steady. Therefore, since the drop in *L* is much larger compared to the rise in <*k*> from stage 1 to 2, it must be due to an increased proportion of shortcut links. In other words, in stage 1 there is a lower fraction of co-occurrences between words that are otherwise far apart, i.e., connected by a long chain of co-occurrences. Something interesting to note here is that if *L* continued to drop at the same rate between stages 2 and 3, then it would be very close to the maternal *L*, meaning that the children and MOSAIC produce utterances by combining words from different contexts more frequently than the mothers, and thus greatly reducing *L*, without a significant increase in <*k*>.

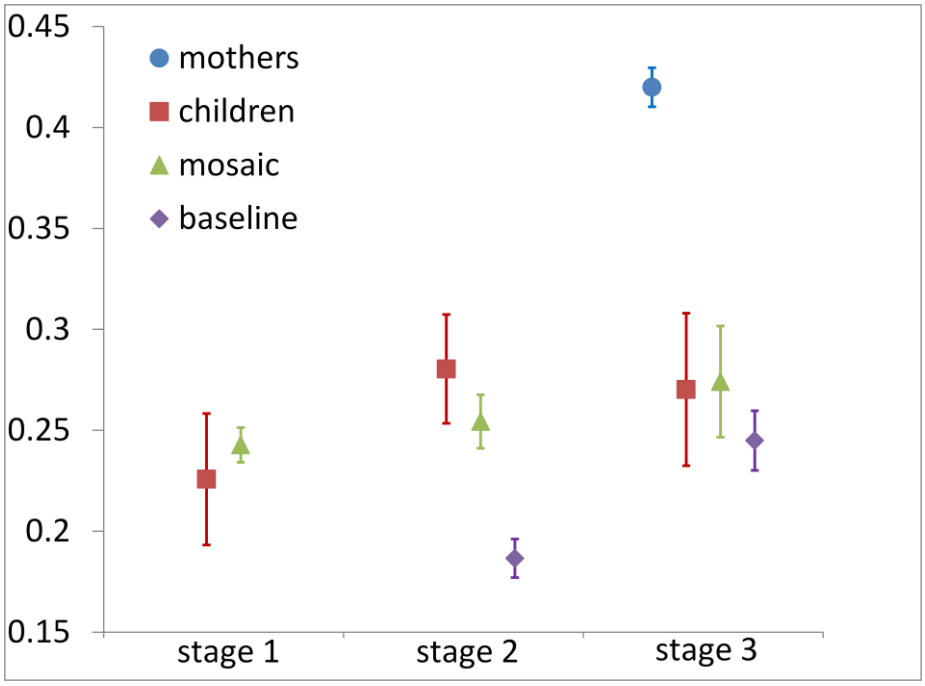

Fig. 5.16. Summary plot for *C*.

The clustering coefficient, *C*, of the children's networks does not behave in an expected fashion, since it peaks for stage 2, after which it drops slightly for stage 3 (Fig. 5.16). By inspecting the other network parameter summary plots and the other sources' *C* values on this plot, it is clear that *C should* be monotonously increasing. This implies that there are three possible cases: either, the children's clustering for stage 2 is too high, or, the children's clustering for stage 3 is too low, or both. In any case, the clustering is relatively high in general, meaning that all networks are small-world networks, since their average geodesic is low.

In summary, the children's networks suggest that *MLU, N, E, GCC,* <*k*> and *L* are behaving predictably with respect to the developmental stage of the children – the former five are increasing and *L* is decreasing because the networks are becoming more connected. On the other hand, *C* drops slightly in stage 3, suggesting that the children probably experimented with a wider variety of utterances, resulting in fewer word co-occurrence loops. This curvilinear function is interesting and is in line with what was found for the *n* parameter of the ranked frequency distribution.

The correlations between all pairs of the four sources are calculated and presented in Table 5.3. In the table – and all subsequent tables of correlations

between the sources – the values above 0.7 are highlighted since they are particularly high, and of primary interest for this analysis.

Table 5.3. Correlations of individual network parameters.

	MLU	*N*	*E*	*GCC*	*<k>*	*L*	*C*
Child-MOSAIC							
stage 1	.06	.08	.89	.26	.24	.21	-.04
stage 2	.55	.59	.96	.56	.89	.45	.60
stage 3	.24	.94	.99	.93	.98	.83	.88
Child-Baseline							
stage 2	.23	.91	.98	.92	.92	-.02	-.23
stage 3	.61	.99	.97	1.00	.77	-.35	.08
Child-Mother							
stage 3	.62	.39	-.02	.39	-.14	-.19	-.39
MOSAIC-Baseline							
stage 2	.09	.84	.99	.78	.93	-.43	-.34
stage 3	.46	.94	.97	.92	.66	-.52	.34
MOSAIC-Mother							
stage 3	.58	.58	.06	.63	-.23	.18	-.02
Baseline-Mother							
stage 3	.91	.32	.20	.33	.35	.41	.63

From the table, it is possible to identify the sources that correlate well by scanning down the rows and focusing on those with the highest correlation parameters. It is clear that stage 3 MOSAIC has excellent correlation with the children on all network parameters except for the *MLU*, which is, in fact, a data set parameter. Nevertheless, the network similarities for this particular pair of sources are astonishing, and none of the other pairs display such high positive correlation on six network parameters. For example, the second highest correlating pair is the stage 2 child-baseline pair, which correlates well on four of the parameters, but not on the average geodesic and the clustering coefficient.

- **Degree Distribution Parameters**

After the node in-degree and out-degree distributions were calculated, it was found that the first data point (representing in/out-degree 0) is far from a power-law fit due to its unique nature, and therefore it is regarded as an additional parameter *p***.** The reason for this behaviour is that the in/out-degree 0 nodes

decrease as the network becomes increasingly connected. The parameters of the in-degree and out-degree distributions are summarised in Table 5.4. The first parameter, *p*, is the actual value of the first data point, i.e. the probability of a node having in/out-degree 0. The second and third parameters (*a* and *n*), are the parameters of the best-fit of the degree distribution without the first data point. The fourth parameter R is the correlation between the data and the fit.

Table 5.4. Parameters of the best-fit of the in-degree and out-degree distributions.

		In-degree				Out-degree			
		p	**a**	**n**	**R**	**p**	**a**	**n**	**R**
Mothers	**ann**	0.09	0.36	-1.40	1.00	0.25	0.29	-1.39	1.00
	ara	0.03	0.42	-1.50	1.00	0.17	0.36	-1.51	1.00
	bec	0.07	0.38	-1.41	1.00	0.24	0.32	-1.47	1.00
	car	0.08	0.38	-1.47	1.00	0.27	0.29	-1.39	1.00
	dom	0.08	0.40	-1.53	1.00	0.23	0.33	-1.53	1.00
	gai	0.10	0.43	-1.58	1.00	0.30	0.33	-1.61	1.00
Children	**ann 1**	0.45	0.29	-1.68	1.00	0.47	0.29	-1.72	1.00
	ann 2	0.21	0.36	-1.49	1.00	0.42	0.29	-1.65	1.00
	ann 3	0.21	0.43	-1.72	1.00	0.40	0.31	-1.75	1.00
	ara 1	0.35	0.31	-1.59	1.00	0.44	0.28	-1.60	1.00
	ara 2	0.16	0.43	-1.58	1.00	0.38	0.32	-1.69	1.00
	ara 3	0.13	0.42	-1.57	1.00	0.33	0.33	-1.61	1.00
	bec 1	0.47	0.30	-1.71	1.00	0.55	0.25	-1.71	1.00
	bec 2	0.24	0.40	-1.67	1.00	0.48	0.28	-1.87	1.00
	bec 3	0.15	0.38	-1.45	1.00	0.42	0.28	-1.67	1.00
	car 1	0.30	0.32	-1.49	1.00	0.45	0.27	-1.63	1.00
	car 2	0.23	0.30	-1.32	1.00	0.39	0.27	-1.51	1.00
	car 3	0.09	0.37	-1.38	1.00	0.33	0.28	-1.42	1.00
	dom 1	0.24	0.34	-1.42	.99	0.46	0.25	-1.53	1.00
	dom 2	0.18	0.34	-1.41	1.00	0.39	0.28	-1.57	1.00
	dom 3	0.18	0.44	-1.70	1.00	0.39	0.29	-1.59	1.00
	gai 1	0.45	0.32	-1.85	1.00	0.50	0.29	-1.85	1.00
	gai 2	0.23	0.44	-1.78	1.00	0.45	0.30	-1.86	1.00
	gai 3	0.19	0.43	-1.67	1.00	0.44	0.30	-1.76	1.00
MOSAIC	**ann 1**	0.50	0.22	-1.48	1.00	0.59	0.20	-1.89	.99
	ann 2	0.26	0.28	-1.28	.99	0.40	0.25	-1.43	1.00
	ann 3	0.12	0.44	-1.55	1.00	0.35	0.33	-1.65	1.00
	ara 1	0.55	0.21	-1.55	1.00	0.59	0.22	-1.94	1.00
	ara 2	0.25	0.37	-1.55	1.00	0.39	0.32	-1.69	1.00

	ara 3	0.14	0.38	-1.41	.99	0.31	0.34	-1.61	1.00
	bec 1	0.62	0.18	-1.54	1.00	0.63	0.20	-2.07	.99
	bec 2	0.28	0.31	-1.40	.99	0.41	0.26	-1.49	1.00
	bec 3	0.16	0.32	-1.29	.99	0.34	0.25	-1.29	.99
	car 1	0.66	0.15	-1.48	1.00	0.65	0.17	-1.68	.99
	car 2	0.29	0.28	-1.32	1.00	0.42	0.24	-1.48	1.00
	car 3	0.18	0.26	-1.14	.98	0.32	0.24	-1.23	.99
	dom 1	0.54	0.18	-1.36	.99	0.57	0.20	-1.63	.99
	dom 2	0.25	0.26	-1.20	.99	0.37	0.22	-1.28	1.00
	dom 3	0.18	0.43	-1.63	1.00	0.31	0.36	-1.71	1.00
	gai 1	0.60	0.18	-1.51	1.00	0.64	0.19	-2.04	.99
	gai 2	0.37	0.32	-1.58	1.00	0.51	0.22	-1.50	1.00
	gai 3	0.22	0.36	-1.47	1.00	0.40	0.27	-1.47	.99
Baseline	**ann 2**	0.15	0.45	-1.65	1.00	0.41	0.27	-1.56	1.00
	ann 3	0.14	0.47	-1.70	.99	0.38	0.29	-1.58	1.00
	ara 2	0.11	0.49	-1.69	1.00	0.47	0.23	-1.47	1.00
	ara 3	0.07	0.51	-1.72	1.00	0.33	0.34	-1.71	1.00
	bec 2	0.13	0.49	-1.77	1.00	0.48	0.22	-1.50	1.00
	bec 3	0.08	0.52	-1.80	1.00	0.36	0.31	-1.64	1.00
	car 2	0.11	0.47	-1.65	1.00	0.48	0.22	-1.47	1.00
	car 3	0.07	0.50	-1.71	1.00	0.33	0.33	-1.75	1.00
	dom 2	0.08	0.45	-1.57	1.00	0.40	0.24	-1.38	1.00
	dom 3	0.09	0.53	-1.93	1.00	0.38	0.27	-1.43	1.00
	gai 2	0.14	0.51	-1.95	1.00	0.49	0.23	-1.54	1.00
	gai 3	0.11	0.53	-1.89	1.00	0.36	0.33	-1.74	1.00

Figs. 5.17 and 5.18 report the summary plots for parameter p of the in-degree and out-degree distributions, respectively. Likewise, Figs. 5.19 and 5.20 present the summary plots for parameter a of the in-degree and out-degree distributions, respectively. Finally, Figs. 5.21 and 5.22 show the summary plots for parameter n of the in-degree and out-degree distributions, respectively.

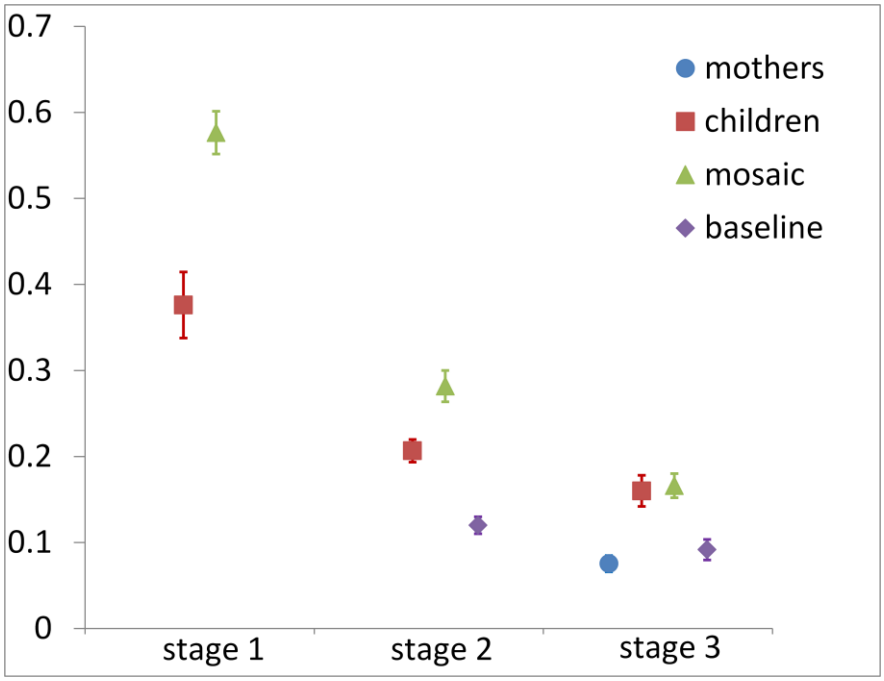

Fig. 5.17. Summary plot for parameter *p* of the in-degree distribution.

Evidently, *p* is decreasing with time, indicating a reduction in the proportion of starting words, which do not appear anywhere else but at the very beginning of an utterance (Fig. 5.17). Furthermore, MOSAIC is clearly approaching the children over time, whereas the baseline is lower and closer to the mothers, possibly because it is so heavily based on the maternal data. Note that for stage 3 the difference between MOSAIC and the children is negligible, highlighting the accuracy of the model in this case.

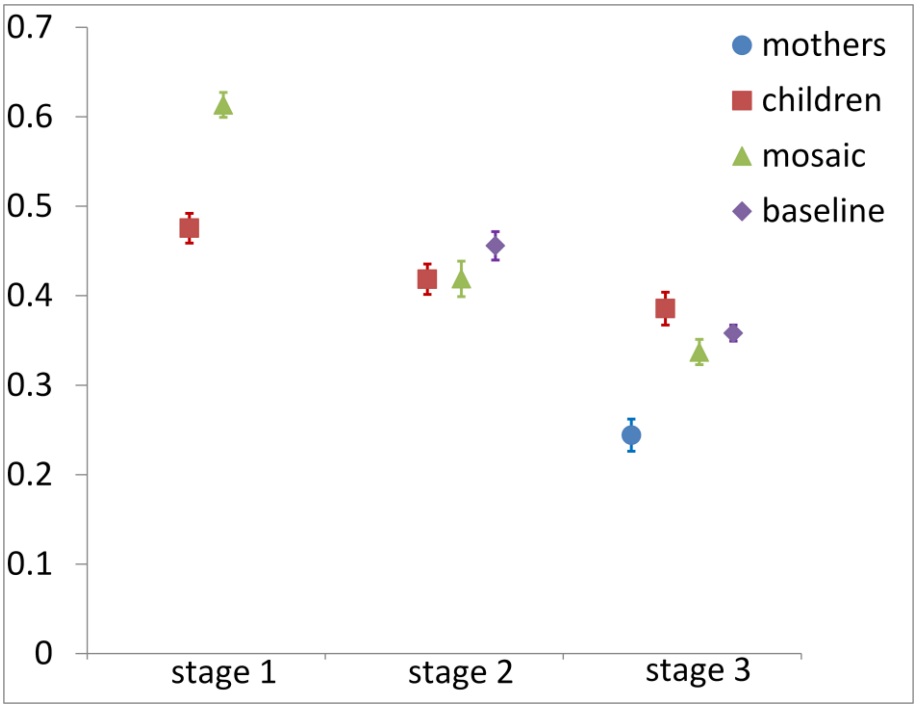

Fig. 5.18. Summary plot for parameter *p* of the out-degree distribution.

Again, *p* is decreasing with time, but there are some notable differences (Fig. 5.18). Firstly, the parameter values are generally higher than those for the in-degree, meaning that there are more end words than start words. Secondly, the children follow a more linear decrease in *p* than before. Also, MOSAIC is closest to the children at stage 2 instead of 3, and the baseline is closer to MOSAIC instead of the mothers.

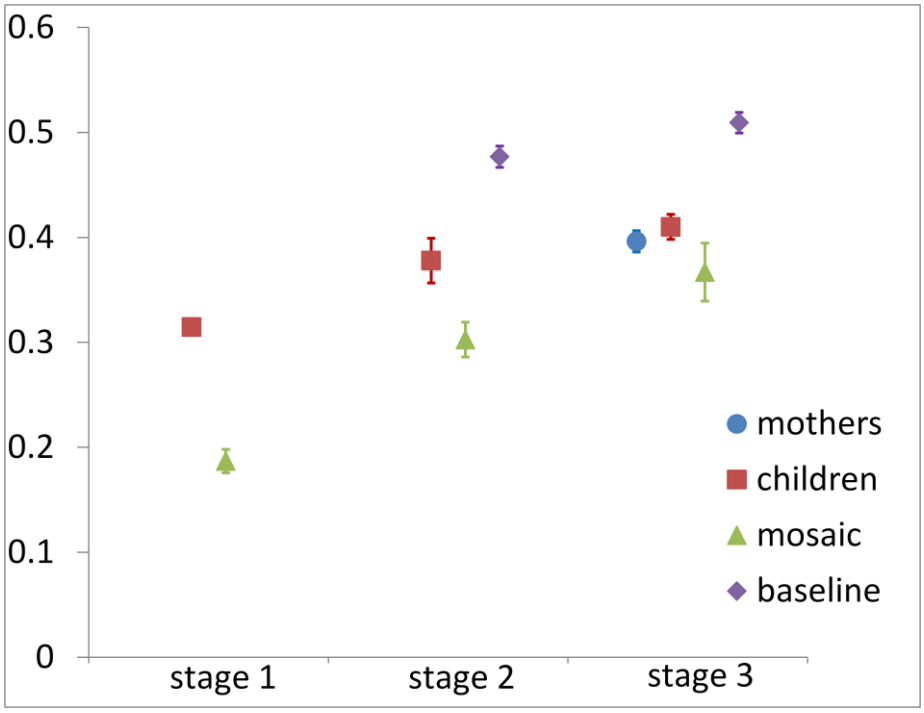

Fig. 5.19. Summary plot for parameter *a* of the in-degree distribution.

Parameter *a* is the probability of degree 1, obtained from the best-fit curve (Fig. 5.19). Clearly, it is increasing with time, and MOSAIC is steadily converging on the children, which approach the mothers in the final stage 3. This is another good example of how MOSAIC outperforms the baseline model in terms of simulating syntax acquisition in children.

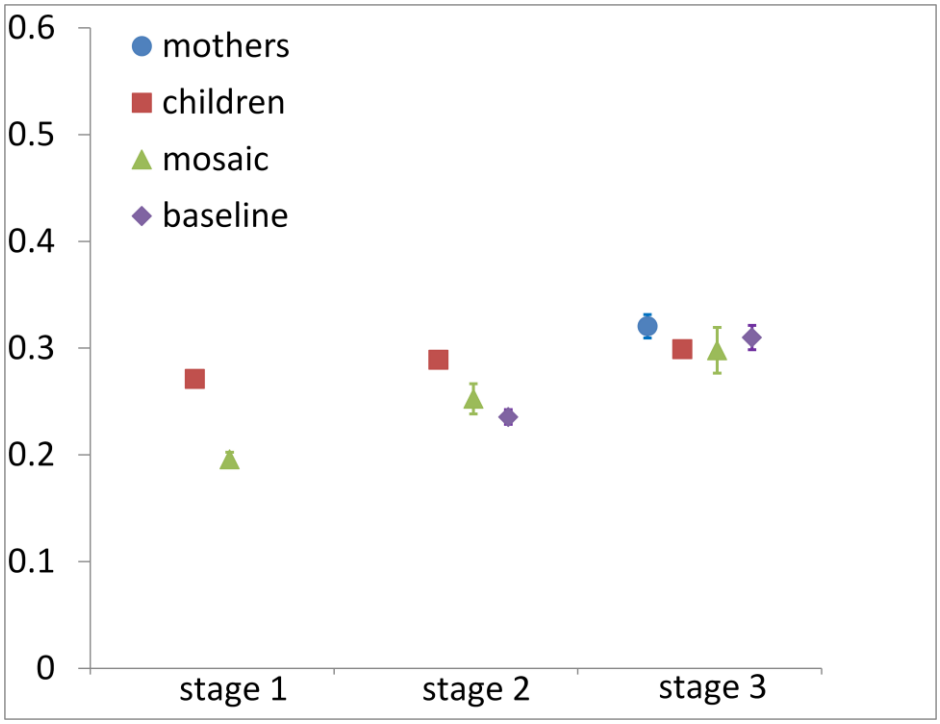

Fig. 5.20. Summary plot for parameter *a* of the out-degree distribution.

Again, *a* is increasing with time (Fig. 5.20), but here, all four sources are much closer to each other and the probabilities are slightly lower than for the in-degree.

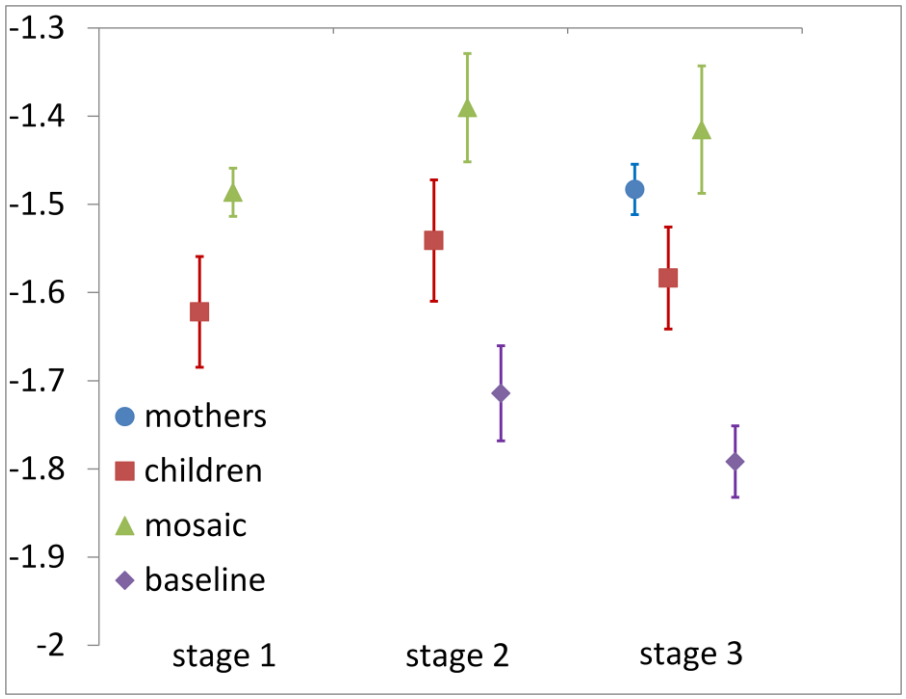

Fig. 5.21. Summary plot for parameter *n* of the in-degree distribution.

The exponent *n* is the most important of the degree distribution parameters, since it specifies the slope of the best-fit curve, i.e., how extremely the probabilities decay towards 0 (Fig. 5.21). Note that the error bars in Fig. 5.21 – which represent the standard error of the mean – are relatively long, implying that the variance within each data set is high. Nevertheless, MOSAIC is following the same trend over time as the children, but since the trend is non-linear, it is not certain whether the baseline does too, as there are only two data points for it.

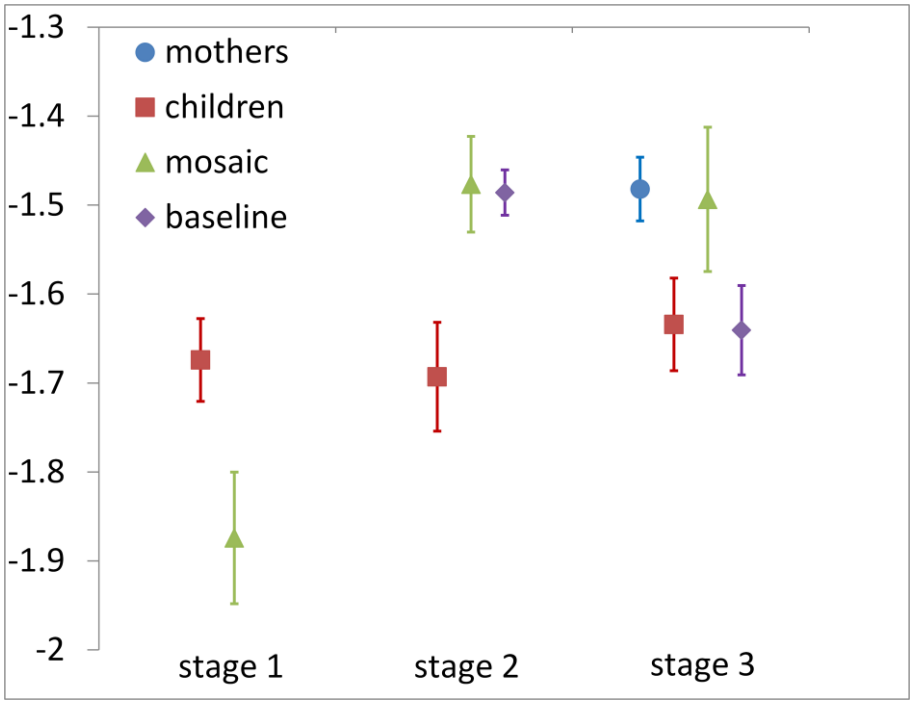

Fig. 5.22. Summary plot for parameter *n* of the out-degree distribution.

Unexpectedly, for the out-degree the children follow quite the opposite trend over time (Fig. 5.22). The exponent first drops and then rises slightly, whereas MOSAIC and the mothers are far from this trend for all stages. There are two interesting observations, however. The baseline is almost identical to the children for stage 3 and to MOSAIC for stage 2.

In summary, by examining the in-degree and out-degree distributions, the children's networks suggest a clear decrease in *p* and a small increase in *a*, but non-linear behaviour in *n*. However, the magnitude of these changes in parameter *n* is small. Also, the children appear to be close to the mothers with respect to all three degree distribution parameters. This leads us to believe that children in

general produce utterances that are statistically similar to adults, in terms of the bias to particular word usage. In addition, the standard deviation of both mothers and children is significantly different from 0, for all 15 parameters under study. For 9 of the parameters, the SEM, and therefore, the variability of the children, is greater than that of the mothers. Specifically, for the more complex parameters (*L*, *C*, *n* (in-degree), *n* (out-degree), and *n* (frequency)), the children have a significantly higher variability compared to the mothers, and vice versa, for the more simple parameters (*MLU*, *N*, *E*, *GCC*), the mothers have higher variability. For the moderately complex parameters (<*k*>, *p* (in-degree), *p* (out-degree), *a* (in-degree), *a* (out-degree), and *a* (frequency)), the children and the mothers have relatively similar variability.

The correlations between all pairs of the four sources are calculated and presented in Table 5.5.

Table 5.5. Correlations of best-fit parameters of the in-degree and out-degree distributions.

	In-degree			Out-degree		
	p	**a**	**n**	**p**	**a**	**n**
Child-MOSAIC						
stage 1	-.03	-.58	.63	.37	-.11	.74
stage 2	.66	.74	.76	.52	.65	.34
stage 3	-.08	.91	.96	.74	.69	.40
Child-Baseline						
stage 2	.56	.73	.88	.39	.17	.53
stage 3	.89	.08	.42	.75	.09	-.12
Child-Mother						
stage 3	.43	.41	.32	.54	.54	.41
MOSAIC-Baseline						
stage 2	.56	.66	.80	.58	-.07	.40
stage 3	-.26	-.19	.52	.31	-.57	-.65
MOSAIC-Mother						
stage 3	.49	.02	.32	.79	.50	.25
Baseline-Mother						
stage 3	.52	.66	.67	.31	.09	-.01

MOSAIC and the children appear to correlate well in terms of their stage 3 degree distributions. The parameters of the in-degree best-fit have particularly

high coefficients (both above 0.9), suggesting that MOSAIC is imitating the children quite well in terms of in-degree. This is not the case for the out-degree, since the exponent *n* has a low correlation of 0.4. It is interesting that for stage 1, this correlation is in fact 0.74. The parameters of the in-degree best-fit for stage 2 are also well correlated for the Child-MOSAIC pair, as well as the Child-Baseline pair. The fact that the simple baseline model based on DFS is able to reproduce such a good in-degree distribution is somewhat of a mystery, and the most reasonable explanation for it is that perhaps children at stage 2 produce utterances similarly to the baseline, at least in some sense. Note that for stage 3, the Child-Baseline pair is also correlating well in parameter *p*, for both the in-degree and out-degree distributions, meaning that the baseline is producing beginning and end words with a similar frequency to the children.

- **Weight Distribution Parameters**

The parameters of the best-fit of the ranked weight distribution are summarised in Table 5.6.

Table 5.6. Parameters of the best-fit of the ranked weight distribution.

		a	n	R
Mothers	ann	1.37	-0.62	.97
	ara	1.01	-0.65	.98
	bec	1.23	-0.64	.98
	car	1.04	-0.63	.99
	dom	1.80	-0.58	.92
	gai	1.35	-0.60	.96
Children	ann 1	0.86	-0.50	.98
	ann 2	1.49	-0.54	.96
	ann 3	1.33	-0.47	.96
	ara 1	1.23	-0.56	.98
	ara 2	1.33	-0.59	.96
	ara 3	1.13	-0.60	.99
	bec 1	0.71	-0.42	.94
	bec 2	1.38	-0.51	.97
	bec 3	1.55	-0.56	.95
	car 1	1.25	-0.55	.98
	car 2	0.90	-0.57	.99

	car 3	1.42	-0.58	.95
	dom 1	0.95	-0.62	.98
	dom 2	0.84	-0.65	.98
	dom 3	0.90	-0.56	.99
	gai 1	1.12	-0.52	.98
	gai 2	0.77	-0.62	.96
	gai 3	0.91	-0.61	.98
MOSAIC	ann 1	1.51	-0.44	.95
	ann 2	1.42	-0.61	.97
	ann 3	1.23	-0.53	.98
	ara 1	1.06	-0.49	.98
	ara 2	1.05	-0.56	.99
	ara 3	0.93	-0.58	.99
	bec 1	0.97	-0.55	.98
	bec 2	1.17	-0.65	.97
	bec 3	1.21	-0.61	.99
	car 1	1.03	-0.45	.97
	car 2	1.37	-0.57	.97
	car 3	1.00	-0.63	.99
	dom 1	1.11	-0.48	.97
	dom 2	1.57	-0.53	.94
	dom 3	1.04	-0.48	.98
	gai 1	1.33	-0.51	.96
	gai 2	1.12	-0.55	.99
	gai 3	1.20	-0.54	.98
Baseline	ann 2	2.08	-0.40	.86
	ann 3	1.60	-0.38	.91
	ara 2	1.88	-0.37	.87
	ara 3	1.92	-0.41	.87
	bec 2	1.78	-0.37	.86
	bec 3	1.31	-0.43	.91
	car 2	1.57	-0.39	.88
	car 3	1.30	-0.43	.89
	dom 2	2.01	-0.38	.81
	dom 3	1.33	-0.39	.93
	gai 2	1.71	-0.37	.86
	gai 3	1.51	-0.42	.89

It is worth mentioning that the baseline model's ranked weight distributions do not follow a perfect power-law, since the lowest correlation with the best-fit is 0.81, which is significantly lower than the other sources' correlations. This is

possibly due to fact that the baseline model is a very simple model based on Depth First Search (DFS), with no built-in mechanisms to simulate the natural acquisition of language in young children. Nevertheless, 0.81 is still a relatively good correlation, meaning that the baseline model is able to reproduce words with a power-law-like distribution of co-occurrences. This is achieved in two steps. Firstly, the input networks (built from the maternal utterances) have a set of *beginning* nodes and a set of *end* nodes, and output is generated by finding all unique paths between all pairs of beginning and end nodes. Secondly, the DFS algorithm is exhaustive so it finds all unique paths. Hence, it is inevitable that the search visits some key links very frequently, but most links are visited infrequently (when searching down a new branch, for example), resulting in a power-law-like frequency of co-occurrences.

Figs. 5.23 and 5.24 present the summary plots for parameters *a* and *n*, respectively. These plots demonstrate the effectiveness of the ranked frequency distribution in capturing patterns in linguistic development.

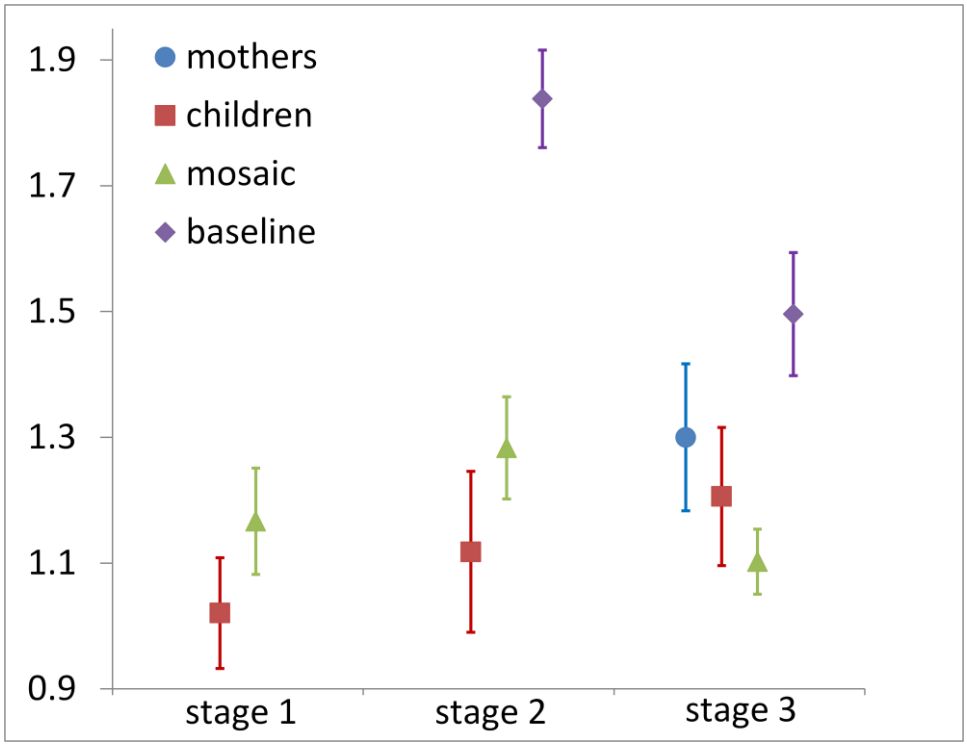

Fig. 5.23. Summary plot for parameter *a*.

From Fig. 5.23 it is apparent that MOSAIC and the children follow a similar trend, approaching the mothers in their final stage 3. The baseline, however, does not display this property.

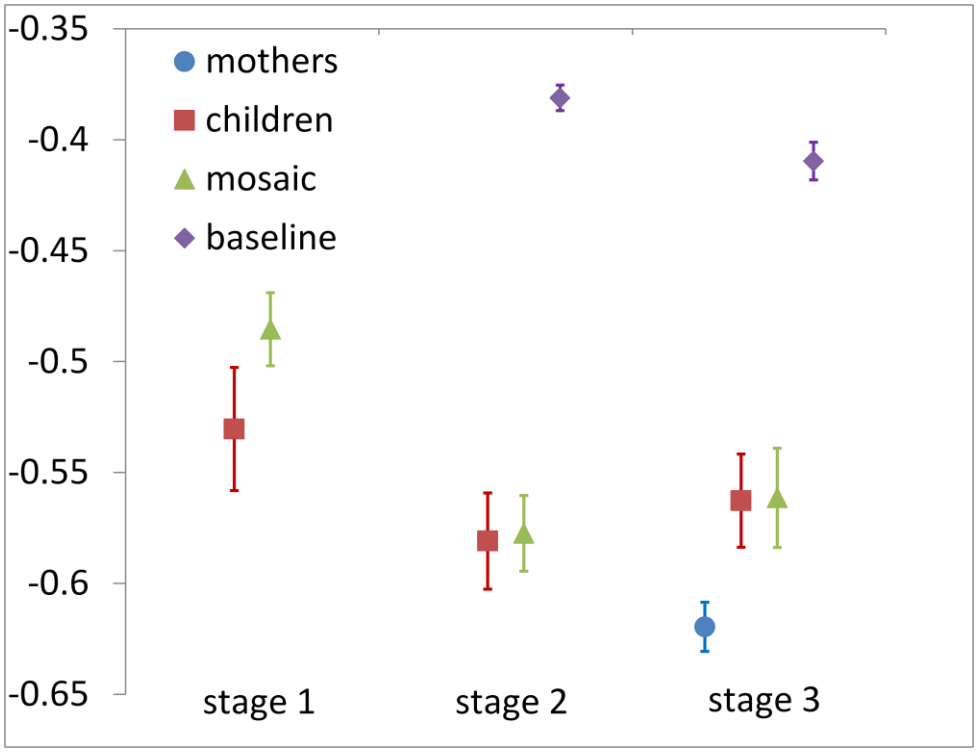

Fig. 5.24. Summary plot for parameter *n*.

Again, MOSAIC and the children are close, with virtually identical parameter values and variations for stages 2 and 3 (Fig. 5.24). The baseline, however, is much further away.

The correlations between all pairs of the four sources are calculated and presented in Table 5.7.

Table 5.7. Correlations of best-fit parameters of the ranked weight distribution.

	a	**n**
Child-MOSAIC		
stage 1	-.13	-.54
stage 2	-.23	-.96
stage 3	.21	.29
Child-Baseline		
stage 2	.47	-.18
stage 3	-.23	.69
Child-Mother		

stage 3	-.56	.06
MOSAIC-Baseline		
stage 2	.41	.05
stage 3	-.30	.82
MOSAIC-Mother		
stage 3	.28	.82
Baseline-Mother		
stage 3	-.39	.49

From the table it is clear that for stage 3 MOSAIC has high correlation (0.82) in parameter *n* with the baseline model. This relationship is interesting because it means that a simple baseline model performs similarly to MOSAIC for later stages of language acquisition. Also, MOSAIC has high correlation (0.82) in parameter *n* with the mothers. This suggests that MOSAIC produced realistic output that is similar to grown-ups' output. In addition, for stage 3 there is good correlation (0.69) in parameter *n* between the children and the baseline. This suggests that for later developmental stages, the baseline model produces output with a rank-frequency curve exponent *n* that is similar to the children's, which is remarkable given the simple nature of the model. This may be explained by the fact that the baseline only repeats co-occurrences that were present in the mothers' input data (without generating any new ones), leading to a more robust but less realistic output.

5.4.2 Community Structure

Community structure is investigated for two types of network: the *aggregated* type contains all six children's data and the *individual* type contains each child's data. Hence, the aggregated networks give a general overview of the average linguistic development whereas the individual networks provide specific detail for the given child.

Figs. 5.25-5.27 present the aggregated children's community structure at each of the three stages, where the font size is proportional to the square root of the occurrence frequency (since the range of frequencies is large this ensures more balanced font sizes that are more clearly visible). The axes dimensions represent

two semantic categories: goodness and size (in absolute terms, in the sense that good/bad and big/small are not distinguished) on the y and x-axis, respectively. The semantic distance (calculated by DISCO (Kolb, 2008)) between each word in the networks and each of the two dimensions is used to plot the data. Only the words *good* and *big* were used in the calculations as the semantic distance does not distinguish between synonyms and antonyms (hence the absolute values).

The colour represents the community assignment. Words located at the origin (bottom-left) are not shown since they lack semantics according to the axes. For better clarity, only the largest 50 words (in terms of font size) are shown in each respective plot, and the axes are in the range 0-50% semantic similarity since there are very few words that appear outside this range. Furthermore, word connections and co-occurrence frequencies are not shown, and colour is not consistent across the networks as it is only used to differentiate between different communities in a single network. By observing the distribution of words relative to the dashed diagonal line it is possible to determine the bias towards goodness or size related words in the utterances produced. To summarise, the figures depict the word occurrence frequency by their font size, and groups of identically coloured words have particularly high frequencies of co-occurrence, given their semantic distance. Figs. B.19-B.36 (see Appendix) present the individual children's community structure at each of the three stages.

- **Stage 1 Aggregated Children**

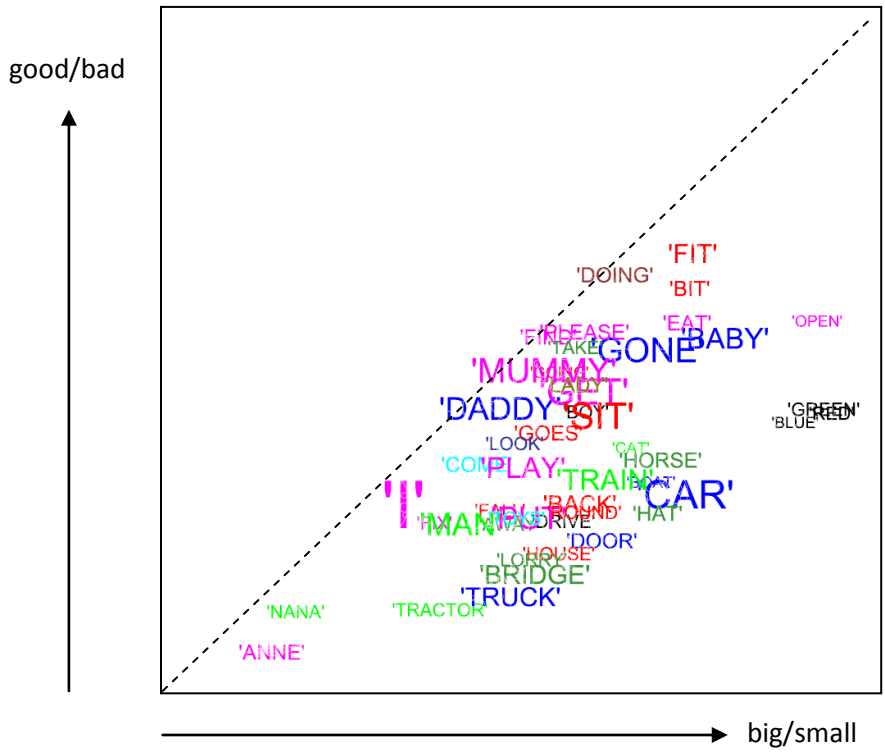

Fig. 5.25. Community structure in aggregated children in stage 1.

- **Stage 2 Aggregated Children**

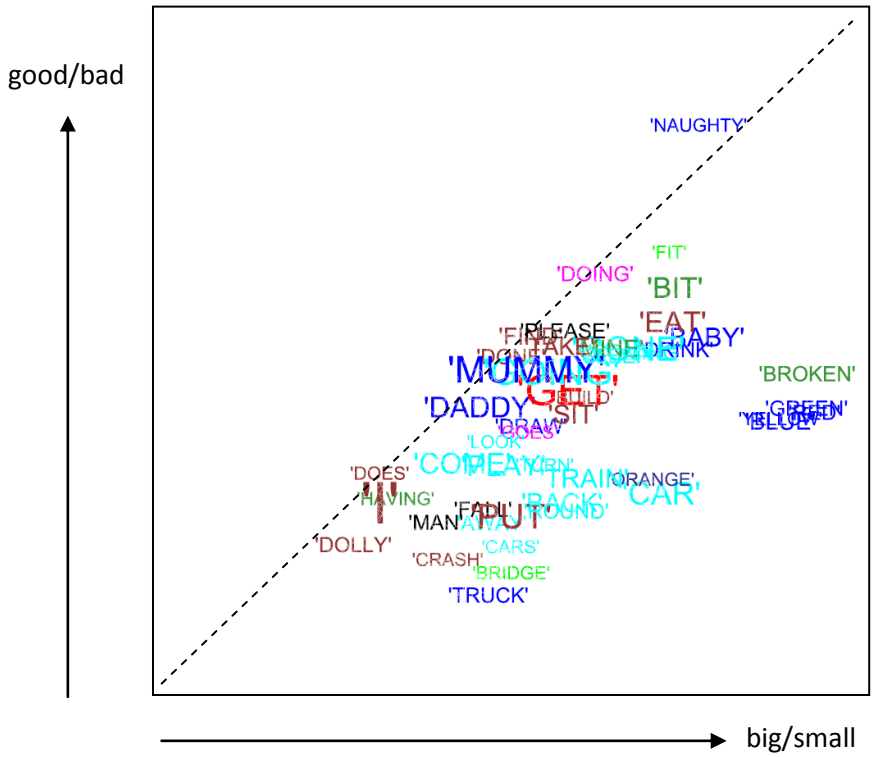

Fig. 5.26. Community structure in aggregated children in stage 2.

- **Stage 3 Aggregated Children**

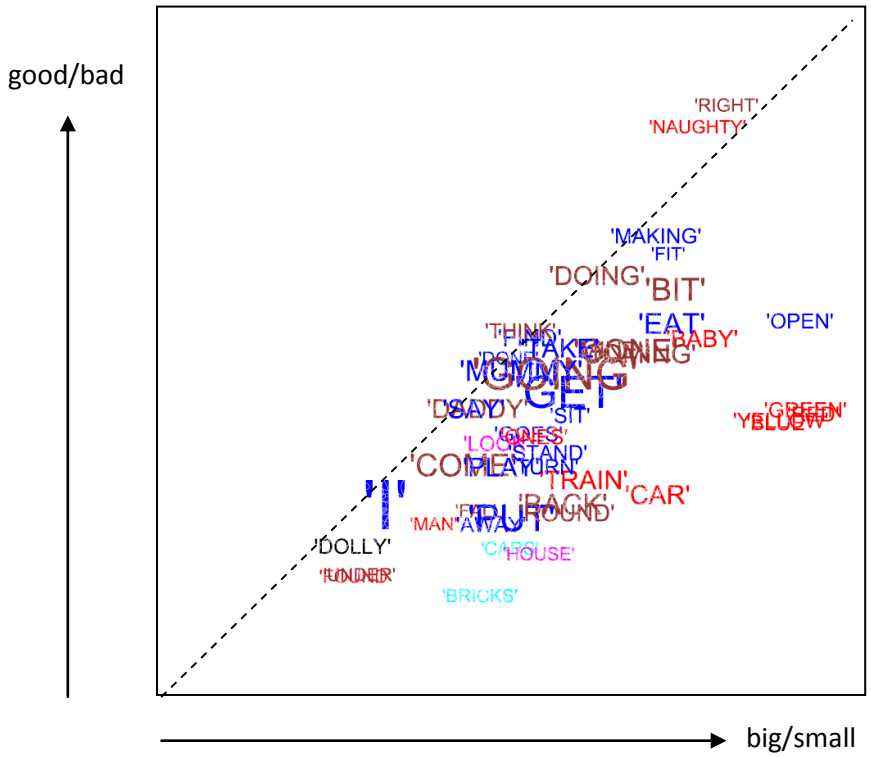

Fig. 5.27. Community structure in aggregated children in stage 3.

5.5 Summary

This chapter presented the second applied case study on language acquisition networks. The key contribution of this chapter is the validation that MOSAIC is a good model of children's syntax acquisition, which supports the Distributional Analysis theory in language. In addition, the proposed semantics-independent community structure revealed interesting patterns in children's word co-occurrence networks, opening up a new horizon for further study.

Chapter 6
Discussion

This chapter discusses the general properties and the community structure of the air transportation networks and language acquisition networks. The results of Expert's spatial and Newman's non-spatial community detection methods are compared. The two applied case studies are also compared and generalities are presented.

6.1 Air Transportation Networks

6.1.1 Network Parameters

The results obtained from the parameter analysis are discussed in three parts. The first part addresses the individual parameters, which are based on the global structure of the entire network. The second part covers the degree distribution parameters, which describe the structure of the air transportation networks in terms of the airports' number of incoming and outgoing connections from/to other airports. The final part focuses on the weight distribution parameters, which highlight the high heterogeneity in the number of passengers on different connections.

- **Individual Network Parameters**

Figs. 4.4-4.6 show the growth of the network in terms of airports, connections, and connected airports. Clearly, the expansion is much larger from 2000 to 2010, indicating a non-linear growth process. This observed behaviour is not unusual, as any transportation network is constantly affected by economic decisions, supply and demand, and many other factors. What is rather unusual is the fact that the average number of airport connections, Fig 4.7, displays a linear decline in time, due to the faster increase in number of airports compared to the number of airport connections. This means that many (probably small size) airports were introduced but they were not interconnected that well, unless already established

airports lost some connections. Because of this rapid growth, the average geodesic length (Fig. 4.8) between any two airports in the US jumped from 2.5 to 3.5, within the past ten years. However, this does not imply that the average journey would need more changes; to the contrary, the network was optimised over time to reduce the changes of the average passenger by interconnecting airports with higher passenger demands, and disconnecting those less profitable. This is evident from the recent boom in low-cost airlines, providing many point-to-point flights between poorly connected destinations. Based on these facts, it is natural to assume that the clustering in the network increases, but Fig. 4.9 contradicts this; again, this must be due to the huge number of new airports. All these parameters have confirmed the immense development of the USAN, particularly in the first decade of the 21st century, and the next section explains this phenomenon in more detail.

- **Degree Distribution Parameters**

Figs. 4.10 and 4.11 show the probability of an airport having zero incoming and outgoing connections, respectively. In other words, this parameter measures the proportion of very remote airports that only have some arrivals, or departures, per two months. Clearly, the fraction rises from 1990 to 2000, indicating a significant increase in such poorly connected airports, but more interesting is the 2000 to 2010 period, which experienced no major change. Figs. 4.12 and 4.13 present the fraction of airports with just one incoming and outgoing connection, respectively. Again, these trends quantify the presence of minor airports, which increases linearly over the two decades. Figs. 4.14 and 4.15 report the fitting functions' estimates for the parameters from the previous two figures. Basically, they confirm that the fits are not able to approximate (especially for the year 2000) the first two data points that were extracted as p and q, since they do not obey the power-law relationship that the rest of the data do. The key parameter in a power-law is the exponent, as it controls the skew of the distribution. Therefore, between 1990 and 2000, Figs. 4.16 and 4.17 suggest an increasing exponent in absolute terms, since the scale of the figures is negative. This implies stronger preferential attachment, which means that already highly connected airports

obtained more connections, while poorly connected airports received few new, or even lost existing, connections. The fact that the change between 2000 and 2010 is small, suggests that although there was a lack of point-to-point flights in the 90s, it may have been resolved in the 00s.

- **Weight Distribution Parameters**

The ranked passenger distribution is the only characteristic of the dynamics on the network that is considered in this thesis, and as such, cannot be taken as a complete description of the function of the network. Nevertheless, the results are interesting, and can be used as a basis for further analysis. Figs. 4.18 and 4.19 depict the two parameters of the logarithmic fit, and although further work is necessary to arrive at more precise conclusions, one thing is certain: the USAN exhibits considerable passenger variability over the course of a year. This is demonstrated by the error bars in the figures.

6.1.2 Community Structure

First of all, it is important to highlight the fact that some communities have airports that are very far apart, suggesting that spatial community detection discovers more meaningful communities that are not occupying a single region on the map. The seasonal variation within each of the three years and the long-term evolution of the network between those years are explored in the following two sections. In addition, the obtained space-independent community structure is validated through comparison to the standard space-dependent community structure.

- **Seasonal Variation**

The seasonal variation in passenger flows within each year is investigated qualitatively by visually examining the obtained community structure, and quantitatively, using Normalised Mutual Information (NMI) (Danon *et al.*, 2005).

In terms of qualitative analysis, there appear to be significant changes in the community structure of the USAN in 1990. In other words, there were considerable seasonal variations in the volume of passengers on network

connections. Specifically, Jan-Feb had a very mixed structure, Mar-Apr had a large (green) super-cluster, and the rest of the year was mixed again, with some similarities between May-Jun and Sep-Oct. In the last two months of the year, Chicago joined the blue LA cluster, forming a similar structure to Jan-Feb, which indicates the presence of an annual cycle of passenger demand. Throughout 2000 (apart from May-Jun), the community structure remained fairly stable, implying low seasonal variation. In particular, the network had a large super-cluster covering most of the US, and Atlanta was the super-hub. May-Jun, however, was different as Dallas and Chicago were in a separate cluster of their own, so there was a particularly strong passenger flow between them and other smaller airports in the north-east during these months. In 2010, similarly to 1990, there were notable fluctuations in the community structure of the network. Jan-Feb was mixed, Mar-Apr had a dominant red cluster, and in the rest of the year there were two dominant clusters (Denver and Chicago). LA and San Francisco formed their own community in green in Sep-Oct.

Quantitative analysis of network snapshots involves NMI, which measures the similarity between two network partitions (in this case two consecutive snapshots), returning 1 if they are identical and 0 if they are completely independent. It is typically used to quantify the stability of community structure over time, but it is also used in tests of community detection algorithms (Lancichinetti and Fortunato, 2009). In order to calculate NMI, it was necessary to filter airports that do not appear in all snapshots for a given year. These few, small airports are rarely used and their traffic is very low, so their effect on the network is insignificant.

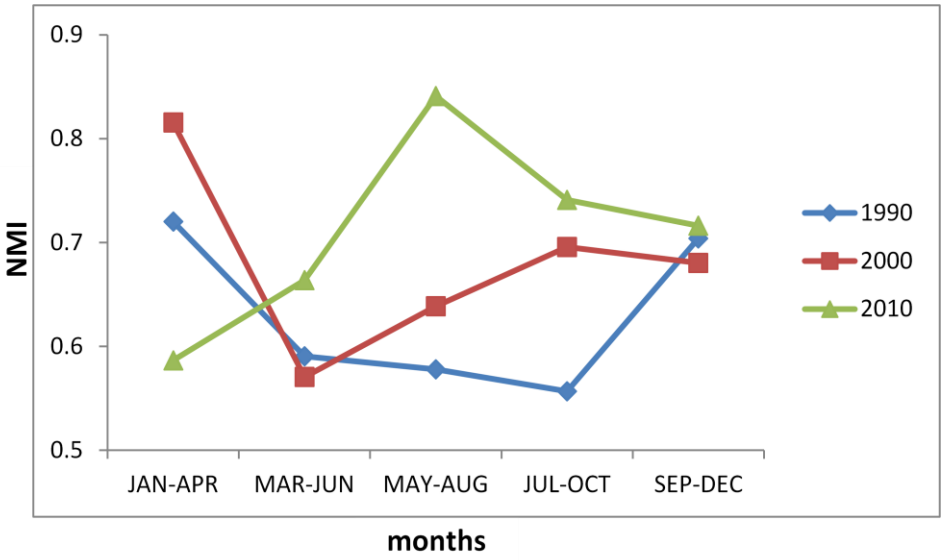

Fig. 6.1. Normalised Mutual Information (NMI) of consecutive network snapshots.

Fig. 6.1 presents NMI over time. For example, JAN-APR refers to the stability of the community structure in the period January to April, using the NMI of the partitions for Jan-Feb and Mar-Apr. The connecting lines do not indicate continuity, but are there to facilitate interpretation of the graph. Fig. 6.1 suggests that in general, the community structure is fairly stable over the course of a year as the NMI is always above 0.5. In addition, annual stability has increased over the three years investigated as the average NMIs for 1990, 2000 and 2010 are 0.63, 0.68 and 0.71, respectively. Specifically, for May-Aug and Jul-Oct the network has shown consistent improvement in stability over its evolution, whereas for Jan-Apr it has become more unstable. The intervals Mar-Jun and Sep-Dec are virtually unchanged over the two decades. In particular, Jan-Apr 2000 and May-Aug 2010 were highly stable (NMI > 0.8), while 1990 was a relatively unstable year. The existence of an annual cycle is confirmed and quantified by calculating the NMI of the pair Jan-Feb and Nov-Dec, which is 0.69, 0.79 and 0.52 for the years 1990, 2000 and 2010, respectively. In other words, in terms of community structure, Jan-Feb resembles Nov-Dec (not so much in 2010), indicating the presence of an annual cycle of passenger demand.

- **Evolution**

This part describes the evolution of the USAN from 1990 to 2010 by focusing on three key issues: volume of air travel, bi-monthly snapshots, and the main hub Atlanta. In addition, the migration levels between, and within, the four US macro-regions are discussed.

The quantity of domestic air traffic can be described by the total number of passengers carried across the USAN. Since the surface area of the airport nodes in Figs. 4.20-4.37 is proportional to the number of passengers, it is easy to determine the volume of air travel by observing the size of the nodes. The volume of air travel grew significantly from 1990 to 2000, with a particularly strong concentration of travellers via Atlanta. The first decade of the 21st century, however, did not see a significant increase in air travel, which, to a certain extent, may have been caused by key events, such as the September 11 terrorist attacks in 2001, and the start of the global economic recession in 2008. It is interesting that although most airports did not grow much from 2000 to 2010, there are some, such as Denver, that did experience a steady growth in terms of passengers. The specific changes in passenger distribution among airports are highlighted by the changes in the size of circles in Figs. 4.20-4.37.

In addition to the analysis of seasonal variation, it is also necessary to study long-term evolution, by focusing on individual bi-monthly snapshots and observing the changes in the network from 1990 to 2000, and from 2000 to 2010. Therefore, each of the six bi-monthly periods is analysed separately in order to illustrate the precise changes in passenger flows and community structure for the specified period, that have occurred in each of the two decades.

January-February: In terms of community structure, 1990 has a mixed pattern of clusters apart from the south (Fig. A.1), 2000 has a large cyan super-cluster covering all of the US (Fig. A.7), and 2010 again has a mixed structure (Fig. A.13). This indicates that in 1990 and 2010 there were numerous popular connections that saw a large number of air passengers, but in 2000 the passengers were more evenly distributed among the possible connections, resulting in a

single super-community. In addition, Atlanta and some airports in the south and north-east had their own specific traffic patterns, as shown in Fig. A.7.

March-April: Generally, the community structure for this period is stable, but from 2000 to 2010 there is a clear transition of two hubs – Chicago and Dallas – from the main cluster to their own local-scale clusters (Figs. 4.27 & 4.33). In other words, these two airports became regional hubs in the first decade of the 21st century, at least for the months of March and April.

May-June: Community structure changes significantly for the period 1990-2010, highlighting the specific changes in passenger trends over the years. In particular, 1990 is composed of one large red cluster covering all but the south, one medium-sized pink cluster in the south, and several regional clusters (Fig. A.3). This structure indicates that the red airports are the national long-range hubs, the south is somewhat more isolated, and the rest of the airports provide more local services. On the other hand, in 2000 Chicago and Dallas belong to the same cluster, and Atlanta is by far the top airport in the US (Fig. A.9). In 2010, there are two main clusters – the Chicago cluster in yellow, and the Denver cluster in blue – that cover the US together with Atlanta and Dallas, acting as national super-hubs (Fig. A.15).

July-August: Community structure in July-August suggests that in 1990 passengers preferred specific long-range connections (Fig. A.4). Most clusters cover large areas of the US, so many people travelled all over the US, specifically among airports of the same colour. On the other hand, in 2000 passengers were more evenly distributed within the green cluster, and more intricately concentrated on certain routes only in the north-east (Fig. A.10); while in 2010 the picture is, again, completely different, with two large clusters in red and pink, and two key hubs – Atlanta and Dallas – in blue and green, respectively (Fig. A.16).

September-October: The network in 1990 (Fig. A.5) is mainly composed of the blue LA cluster and the green Dallas cluster, with Chicago and Atlanta as hubs, and the usual mix of clusters in the densely populated north-east. In 2000, however, there is one red super-cluster, Atlanta is the main hub, and there is also

a lot of activity in the Chicago region, as illustrated by the many colours that indicate the specific passenger trends in September-October (Fig. A.11). 2010 has a mix of multiple large clusters revealing new passenger flows (Fig. A.17). This is a sign of long-range travel among community members that are far apart.

November-December: In 1990 (Fig. A.6) the USAN is split into a large blue cluster and a yellow Dallas cluster in the south, but in 2000 (Fig. A.12) they have converged to a single yellow super-cluster, covering all but some regions in the north-east and the main hub Atlanta. In 2010 (Fig. A.18), the super-cluster has broken down, leaving Dallas as a national hub, and two red and green clusters spanning a large part of the US.

The role of Atlanta (ATL) as a leading US airport depends on factors, such as air services and their locations, as well as investments into growth and development. In 1967, the city of Atlanta and the airlines began to work on a master plan for the future development of the airport. Many investments were made in the following years, leading to new passenger terminals, runways, and facilities both inside (such as the people mover system linking parts of the terminal), and outside (such as the Red/Gold rail line, operated by the Metropolitan Atlanta Rapid Transit Authority, linking the airport to the counties of Fulton and DeKalb, in addition to Atlanta itself). ATL is also the primary base of many airlines, such as Delta Air Lines, who built one of the world's largest airline bases in 1930. Delta was an early adopter of the hub-and-spoke system, with Atlanta as its primary hub between the Midwest and Florida. This gave it an early competitive advantage, as Florida has been an attractive destination within the US for many decades. Although there is a decrease in the volume of migration in recent years, Florida and the South are still very popular destinations. In 1990, Atlanta was one of the three leading US airports for domestic flights. By 2000 it became the top airport (Figs. 4.20-4.37). Atlanta is also the only significant member in its community for all three years. This implies that it is equally well connected to other airports, thereby possibly serving as a national hub. Since ATL handles so many passengers but there are no other major airports of the same colour, it follows that all ATL connections have relatively similar traffic loads, with longer connections having less traffic due to the effect of spatial separation. Therefore,

ATL has no strong preferential attachment to any other major airport. To verify that Atlanta is a national hub, it is also necessary to check its number of direct connections. Table 6.1 summarises ATL's number of connections and the highest number of connections for the months Jan-Feb, in each of the years studied.

Table 6.1. Atlanta's connections.

	1990	2000	2010
Atlanta	101	142	167
Max	139	142	172

Clearly, Atlanta ranks very high in terms of connections, so it has a direct influence on a large part of the US territory. For example, in Jan-Feb 2010, ATL handled 10.7 million passengers (top in the US) on 167 connections, with an average of 64,000 passengers per connection, compared with the US highest figures of 172 and 73,000, respectively. In summary, ATL became the top US hub for domestic flights by the year 2000.

According to recent figures and US Census data (United States Census Bureau, n.d.), American people move many times during their adult lives, mainly in their twenties. Preferred destinations of domestic migration were Southern states, mainly Florida, possibly because they are considered attractive places to live and work. Although US domestic migration has fallen noticeably since the 1980s, it is still higher than that within most other developed countries and during the period 2000-2004 it continued to redistribute the country's population (Perry, 2006). Nevertheless, the current slowdown in domestic migration due to the impact of the economic situation has changed the picture of movements within the US. In-migration towards states like Arizona, Florida and Nevada has slowed down, while Massachusetts, New York and California now have considerably less out-migration (Internal Revenue Service (IRS), n.d.; United States Census Bureau, n.d.). In the years 2009 and 2010 mobility among states slowed nationwide and only a small percentage difference was observed during the two-year period (Table 6.2). Migration is considered only for 2010 as this is the most recent year in the airport network model but a comprehensive investigation into

the long-term relationships between migration and air travel is beyond the scope of this work.

Table 6.2. In-migration, representing the number of people migrating to specific US states in 2009-2010 (United States Census Bureau, n.d.)**.**

State	Year 2010	Year 2009	Diff %	State	Year 2010	Year 2009	Diff %
Alabama	108,951	124,658	-0.14	Montana	35,641	31,015	0.13
Alaska	36,345	40,474	-0.11	Nebraska	51,290	53,214	-0.04
Arizona	223,324	226,457	-0.01	Nevada	103,179	109,257	-0.06
Arkansas	79,214	85,857	-0.08	New Hampshire	39,423	37,940	0.04
California	445,972	460,161	-0.03	New Jersey	130,101	136,212	-0.05
Colorado	187,240	182,854	0.02	New Mexico	74,237	64,797	0.13
Connecticut	79,360	81,546	-0.03	New York	276,167	277,482	0.00
Delaware	31,713	35,085	-0.11	North Carolina	265,206	284,171	-0.07
District of Columbia	1,244	38,907	0.24	North Dakota	30,100	29,970	0.00
Florida	495,857	475,871	0.04	Ohio	174,773	171,894	0.02
Georgia	250,469	280,221	-0.12	Oklahoma	106,720	117,850	-0.10
Hawaii	53,581	53,270	0.01	Oregon	117,521	127,489	-0.08
Idaho	55,871	57,790	-0.03	Pennsylvania	241,855	232,316	0.04
Illinois	206,014	206,151	0.00	Rhode Island	32,335	32,108	0.01
Indiana	127,925	132,755	-0.04	South Carolina	152,710	33,616	0.78
Iowa	72,706	74,704	-0.03	South Dakota	25,777	145,873	-4.66
Kansas	95,127	102,695	-0.08	Tennessee	159,778	29,632	0.81
Kentucky	118,622	122,184	-0.03	Texas	490,738	168,174	0.66
Louisiana	98,291	90,957	0.07	Utah	78,163	511,166	-5.54
Maine	27,962	24,672	0.12	Vermont	22,529	90,375	-3.01
Maryland	165,096	174,958	-0.06	Virginia	260,813	19,390	0.93
Massachusetts	143,247	148,500	-0.04	Washington	191,784	271,600	-0.42
Michigan	117,581	118,054	0.00	West Virginia	39,791	192,654	-3.84
Minnesota	89,911	90,944	-0.01	Wisconsin	93,586	50,155	0.46
Mississippi	73,135	67,245	0.08	Wyoming	28,046	95,475	-2.40
Missouri	146,093	150,271	-0.03	Puerto Rico	31,732	30,889	0.03

Despite the current tendency to stagnancy, the role of the airport network in the context of US domestic migration is important. Since an airport network is continuously evolving depending on passenger demand, it is increasingly well-optimised for a number of functions, such as carrying more passengers, minimising flight changes for the average passenger, and making profit. As the USAN has evolved to attract passengers that are typically travelling to popular destinations, it is directly facilitating migration. Although most passengers fly for short-term business or leisure, there is evidence that a significant fraction of passengers are in fact migrating with a migration probability inversely

proportional to the distance (Levy, 2010; Schwartz, 1973). According to Census data, Figs. 6.2-6.4 show the migration patterns for the years 1990, 2000 and 2010. Data refer to people that are moving to a given macro-region or within it. The scale is relative to the maximum value and therefore not consistent across the three years, but they are comparable, in order to identify any potential variations in migration patterns over the two decades. Migration within the macro-regions is higher than that among them (decay of interaction as distance increases), and migration within the South is the highest, suggesting strong dynamics among the member States. Furthermore, the South region attracts the most people from outside for all three years. This is in line with the fact that Atlanta airport (located in the South) has the highest passenger flow, as discussed above, but it does not necessarily follow that the entire flow is related to the South, as many of the passengers change flights in Atlanta en route to other regions. Nevertheless, the migration patterns do have a clear overlap with the community structure discovered in the USAN.

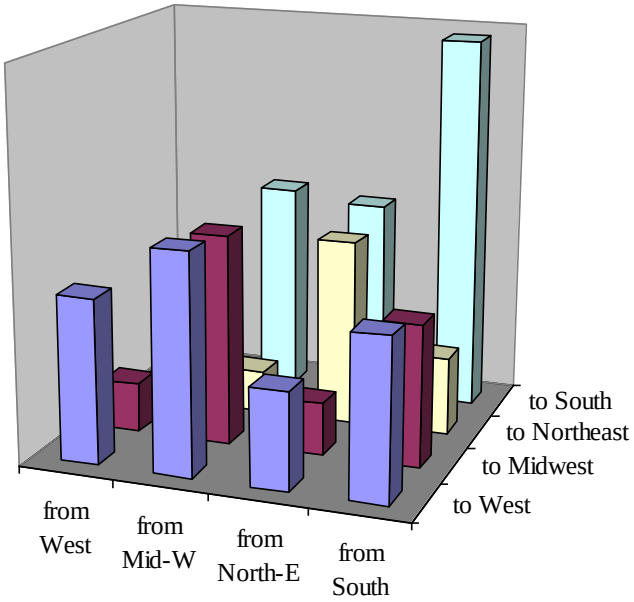

Fig. 6.2. 1990 migration patterns among the four macro-regions: West, Midwest, Northeast and South.

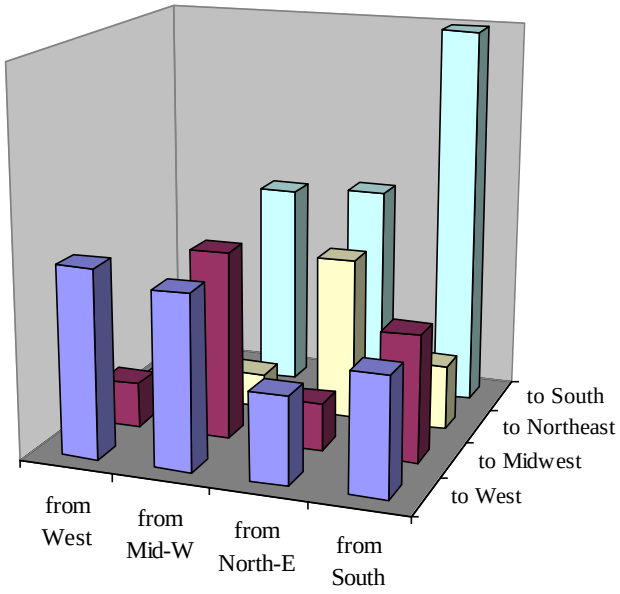

Fig. 6.3. 2000 migration patterns among the four macro-regions: West, Midwest, Northeast and South.

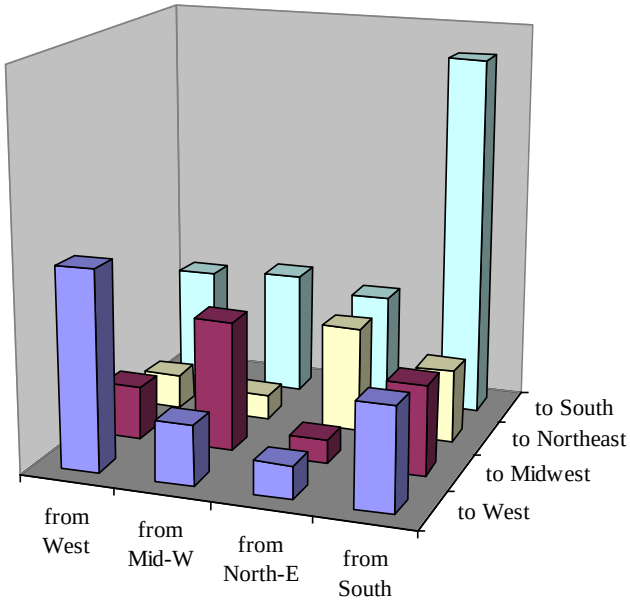

Fig. 6.4. 2010 migration patterns among the four macro-regions: West, Midwest, Northeast and South.

- **Comparative Validation**

The purpose of this section is to demonstrate and to evaluate the effectiveness of Expert's (2011) space-independent community structure detection in comparison to Newman and Girvan's (2004) general community detection. To this end, Newman's method (referring to the null model proposed by Newman and

Girvan) was applied to all eighteen network snapshots of the USAN in order to compare the communities obtained. In addition, Ball, Karrer and Newman (2011) have identified an overlapping community structure in the USAN but the resolution of the communities is low since only two major communities are identified (splitting the US into east and west). This, however, does not provide any detailed information regarding particularly high-traffic connections in the US, as identified by Expert's null model in this thesis. Regarding the application of Newman's method to the networks presented in this thesis, Fig. 6.5 shows the NOV-DEC 2010 snapshot, which is representative of all eighteen snapshots (see Figs. C.37-C.54 in Appendix). In comparison, Fig. 6.6 shows the same snapshot for Expert's method. Here, the communities are not region-based but cover a large area of the US, exposing particularly high-traffic connections, given their distance.

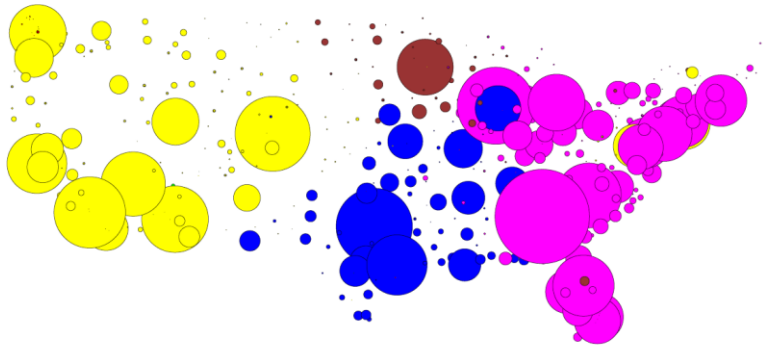

Fig. 6.5. Community structure in USAN in NOV-DEC 2010 identified using Newman's method (same as Fig. C.54 in Appendix).

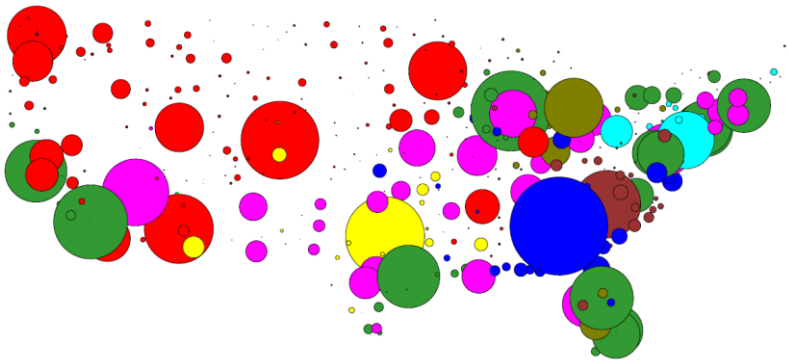

Fig. 6.6. Community structure in USAN in NOV-DEC 2010 identified using Expert's method (same as Fig. A.18 in Appendix).

All partitions identified using Newman's non-spatial NG null model reveal an identical and trivial community structure within the USAN. Specifically, there are four main regional communities that always cover the same region within the US: the east, west, north or south. These isolated communities of airports only provide a very low-resolution picture of the major flows within the US, based solely on the passenger volumes among airports, and disregard the non-linear spatial influence on passenger flows. In summary, Expert's spatial null model has revealed many particularly high flows among distant airports within the US, but Newman's general null model only reveals four regions of high internal traffic, which results from the spatial networks bias towards stronger short-range interactions in terms of passenger flows.

In order to provide a quantitative comparison of the results obtained using Expert's (Expert *et al.*, 2011) and Newman and Girvan's (2004) null models, it is possible to use a community structure comparison measure, such as Normalised Mutual Information (NMI) or Normalised Variation of Information (NVI). Since the purpose of this section on comparative validation is to highlight the contribution of Expert's spatial null model, NVI is a better candidate since it measures the difference between two partitions (in this case Expert's and Newman's), thus quantifying the significance of using spatial information in the detection of communities.

NVI for Expert's (2011) and Newman and Girvan's (2004) community structure is shown in Fig. 6.7 where each trend represents the NVI over the course of a given year. Basically, the plot shows by how much Expert's and Newman's results differ over time, in each of the three years studied. High NVI means high variation (large difference) between the two partitions. Hence, since NVI ranges from 0 to 1, values above 0.5 are large and therefore the plot suggests that there are large differences in the community structure obtained by Newman's method and Expert's method. Specifically, 1990 generally has the highest NVI (especially JUL-AUG); 2000 is almost identical apart from a much lower NVI in JUL-AUG; and 2010 generally has the lowest and steadiest NVI. In summary, quantitative comparison of the communities obtained using Expert's and Newman's null models suggests that there is a significant difference between the

partitions obtained. This means that there is indeed a need to use tailored spatial community detection techniques for spatial networks, as the results obtained would be very different. Assuming that Expert's model accurately captures the bias in the spatial networks, it follows that the higher the NVI, the better Expert's model performs in comparison to Newman's model.

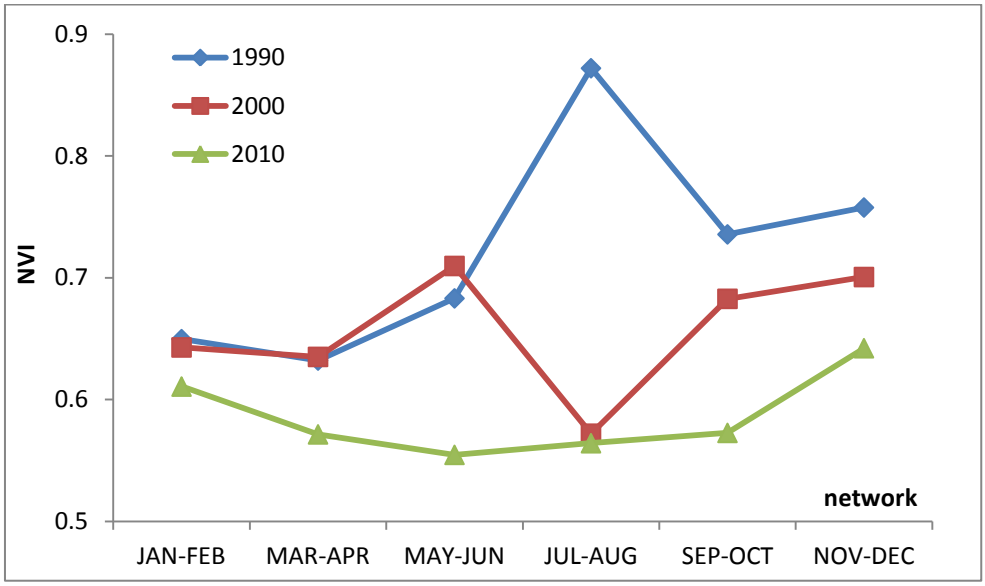

Fig. 6.7. Normalised Variation of Information (NVI) among community structure identified using Expert's and Newman's null models.

6.2 Language Acquisition Networks

6.2.1 Network Parameters

This section discusses the ranked weight distribution parameters in Figs. 5.23 and 5.24 by comparing two pairs of sources: mothers and children; and MOSAIC and children. The reason for the former is that the relationship between mothers' child-directed speech and their children's speech is a key issue in linguistics. Specifically, the linguistic research community is very interested in the level of dependence of children's linguistic abilities and characteristics on their mothers' speech. The reason for the latter is that in order to test the hypothesis of section 2.5 it is necessary to evaluate MOSAIC's performance in relation to the baseline model. This is achieved by comparing how well the two models simulate the children's speech in terms of similar network parameters. Note that whereas the individual network parameters and the degree distributions were unable to

significantly differentiate MOSAIC and the baseline model, the ranked weight distribution did. In other words, it was demonstrated that *W(r)* of a word co-occurrence network is a powerful tool that can provide deeper insights into the language acquisition of young children. Therefore, its use in future research in this field is encouraged and the language acquisition community should exploit its full potential.

Regarding the correlations that were obtained between the sources, the coefficients are difficult to interpret without strong *a priori* theoretical hypotheses, which are beyond the scope of this thesis. In addition, the small number of degrees of freedom means that it is very difficult to reach the significance level, which none of the correlations did. Thus, research with larger samples must be awaited before stronger conclusions can be made about the meaning of these correlations.

- **Mothers and Children**

Fig. 5.19 suggests that the scaling factor *a* of the children is increasing in time in a linear fashion, converging to the mothers'. Here, *a* is the best-fit estimate of the top-ranked normalised frequency. Therefore, it should be close to 1, and the fact that it is increasing above 1 over the three stages implies that the best-fit is diverging from the data for the top-ranked frequency, in order to provide a better fit to the rest of the data. This suggests that over the stages, the top frequency is falling below the expected power-law frequency. In other words, the most common pair of words occurs a little less frequently than a power-law would predict. However, the exponent *n* is in fact the key parameter in a power-law, as it controls how skewed the function is. Therefore, an increasing *n* in absolute terms (decreasing in Fig. 5.20 due to negative sign), implies more bias towards certain pairs of words in language productivity. Note that the mothers are well-aligned with the children of stages 1 and 2, but the stage 3 children have in fact diverged completely from the expected linear trend. This is a clear indicator that in terms of word combinations, children are still developing their linguistic skills at this stage 3. It is suspected that the children probably experimented with a variety of new, or relatively new, word combinations when producing utterances,

thereby reducing the frequency of the more common word pairs, which are already well-known. Even though the other network parameters (except *C*) and the previous two stages for this parameter suggest otherwise, the exponent of the ranked frequency distribution has uncovered something very surprising. This result is also supported by the trend in *C*, but more importantly, the frequency distribution is a function parameter of the link weights, so it is telling us a lot more about the dynamics on the network.

- **MOSAIC and Children**

Fig. 5.19 clearly implies that MOSAIC and the children are close in terms of parameter *a* of the ranked weight distribution. The baseline, on the other hand, is much further away. This suggests that MOSAIC is significantly outperforming the baseline in this respect. In terms of parameter *n* (Fig. 5.20), MOSAIC and the children are very close, with virtually identical parameter values and variations for stages 2 and 3. Note that *n* is the exponent in the power-law of the fit to the ranked frequency distribution, which means that it determines the slope of the curve, i.e. the decay of the frequencies. This result is impressive because it implies that MOSAIC produces co-occurrences with frequencies that are very similar to the children's. The baseline, however, does not. In addition, both the children and MOSAIC follow an interesting but unexpected trend over time for *n*: whereas they steadily approach the mothers for *a* (Fig. 5.19), they diverge for *n* (Fig. 5.20), but the mothers are exactly where the children are expected to be by linear extrapolation (Fig. 5.20).

It would be useful to compare directly the main findings for MOSAIC and the children, and hence, Table 6.3 summarises the main qualitative and quantitative properties of the networks.

Table 6.3. Main findings in MOSAIC and children.

Property	MOSAIC	Children
Degree distribution	Power-law	Power-law
Weight distribution	Power-law	Power-law
Average <*k*>	5.92	5.81
Average *L*	3.16	3.14
Average *C*	0.26	0.26

The degree and weight distributions of MOSAIC and the children all follow a power-law, confirming the ability of MOSAIC to simulate real children's syntactic acquisition. In addition, the average $<k>$, L and C across all eighteen networks are also very similar for MOSAIC and the children. The average $<k>$ validates the ability of MOSAIC to simulate the connectivity of children's word co-occurrence network. The average geodesic length L is low for all networks and again MOSAIC's L is very close to that of the children. The average clustering coefficient C is identical for MOSAIC and children. Since a thorough comparison to an ensemble of random networks in terms of L and C has not been carried out, it is not possible to state with certainty whether any of the networks obey the small-world property of low L and high C, but the values clearly show that the networks at least resemble a small-world.

In summary, the hypothesis in section 2.5 has been verified, i.e. MOSAIC significantly outperforms the baseline at simulating children's linguistic development. Therefore, it is a good model of language acquisition, which draws two conclusions. Firstly, MOSAIC should be developed and exploited further in order to reveal deeper insights into children's language acquisition. Secondly, MOSAIC has illustrated the mechanisms involved in children's distributional analysis of language. Therefore, it supports the distributional analysis theory where children's linguistic abilities are shaped by their external environment. It would be interesting to see whether supporters of the nativist theory are able to explain these phenomena. In general, then, network analysis provides powerful constrains for theories of language acquisition.

6.2.2 Community Structure

Similarly to the discovered community structure in the air transportation networks, it is important to note that in the children's networks some communities have words that are very far apart, suggesting that spatial community detection discovers more meaningful communities that are not occupying a single region on the map. In the following two sections the obtained community structure is discussed. For each of the three developmental stages, the individual children are analysed using the individual networks. General linguistic

development in the children is addressed by analysing the growing complexity of the aggregated networks. It is important to note that no linguistic pre-processing has been applied and so all of the obtained results are extracted directly from the networks using the community detection. In addition, the space-independent community structure detection is validated through comparison to the standard space-dependent community structure detection.

- **Dynamics and Evolution of Individual Children's Networks**

This section discusses the community structure obtained for the individual children (see Appendix). Firstly, each of the three individual stages is qualitatively described. Then, linguistic development is quantitatively assessed by calculating the NMI of the community structure of each of the children over the three stages.

Fig. B.19 (Ann 1) shows a large pink community of words including *Anne, I, find, get, put, take, brush* and others, suggesting a particularly high use of active verbs together with *Anne* and *I*, which means that she was most probably doing something herself during the data collection. Fig. B.20 (Ara 1) contains a large cyan community of *dump, truck, door, boat*, and other nouns and verbs, which, again, point to the focus of the child's interaction with their mother. Fig. B.21 (Bec 1) has a number of communities but they are relatively small in terms of frequency. Fig. B.22 (Car 1) has a large pink community including *nana, man, bridge, apple*, and *train*, which makes sense since some of them are clearly related, thereby confirming the quality of the obtained partitions. Fig. B.23 (Dom 1) shows very high frequency words in various colours, such as *I* and *play* in pink, *mummy* and *get* in blue, *car* and a number of colours in red, and *lorry* and *bridge* in cyan. These are some very interesting word pairs that perfectly illustrate the emergence of statistically significant and meaningful patterns in children's utterances. Fig. B.24 (Gai 1) does not appear to have a dominating community of words but there are individual communities that include words that are clearly related, which is an encouraging result. In summary, in stage 1 Bec did not produce any word with particularly high frequency, whereas Ara and Dom repeated certain words a lot, as illustrated by their dense plots.

Fig. B.25 (Ann 2) has words such as *I, come* and *sleep* in dark green, *put* and *get* in light green, and *bit* and *baby* in red. Fig. B.26 (Ara 2) has *I* and *get* in cyan, and a large red community of *tractor, train, bus, car, come, back, sit* and others. Fig. B.27 (Bec 2) has *I* and *find* in brown and multiple other communities of low frequency words. Fig. B.28 (Car 2) has *I, draw, get, find* and *mummy* in black, and *daddy, baby, car* and *truck* in pink. Fig. B.29 (Dom 2) has very high frequency words, such as *I, play* and *eat* in red, *going, gone* and *car* in brown, and *get, come* and *back* in pink. Fig. B.30 (Gai 2) has *I, put, take, wear* and *mummy* in the largest community in black. In summary, in stage 2 Bec continued with low frequency words, as well as Gai, whereas Dom (but not Ara) continued with high frequency words.

Fig. B.31 (Ann 3) has unexpectedly low frequency words with no significantly large communities. Fig. B.32 (Ara 3) has *I, put, get, going* and *eat* in the top black community, and *bricks, toys, back, train, car, horse, stuck* and *sit* in the second largest light green community. Fig. B.33 (Bec 3) has *I* as a particularly high frequency word, *get, going,* and *doing* in pink, and *put, back, stand* and *eat* in black. Fig. B.34 (Car 3) has many high frequency words, such as *I* and *get* in light green, *going* in blue, and *train, car, back, water, bridge, under* and *trucks* in black. Fig. B.35 (Dom 3) has particularly low frequency words and no dominating community. Fig. B.36 (Gai 3) has *I, going, find* and *take* in blue, *come, put* and *open* in light green, *get, right* and *look* in black, and *bit* in red. In summary, in stage 3 Ann and Dom (Dom previously produced very high frequency words) began to produce low frequency words, whereas Car began to produce high frequency words. This is a very interesting phase transition in Dom's language that suggests that the bias towards the repetition of specific words has been lost.

By observing the distribution of words relative to the diagonal, it is clear that generally, children's linguistic production focused more on size related words (since the majority of words are concentrated in the bottom right half of the plots) apart from Dom 2, Car 2, and Car 3 who produced more balanced utterances. This makes sense since young children have a better understanding of the simple notion of size, as opposed to the more complex concept of good and bad.

To calculate NMI it is necessary to compare community structure of the same input network, i.e. having the same nodes or words in this case. Therefore, for each child only the words present in all three stages are extracted and their community membership is used to measure NMI. Table 6.4 summarises the number of words in each of the networks as well as the number of common words that are used for the NMI. Although the number of common words is low compared to some of the bigger networks, the number is stable across the six children and the words are stable over the three stages, so these are the most significant words to study using NMI.

Table 6.4. Number of words in children's networks and number of common words present in all three stages.

	ann	**ara**	**bec**	**car**	**dom**	**gai**
Stage 1	556	487	498	385	558	622
Stage 2	1205	651	824	628	1087	865
Stage 3	594	1309	1245	1138	449	1230
Common	*246*	*276*	*291*	*236*	*253*	*309*

Fig. 6.8 shows the NMI of the community structure of consecutive stage networks of individual children. The two data points per child represent the NMI of community structure of stage 1 compared to stage 2 and stage 2 compared to stage 3. Generally, the values are low (NMI ranges from 0 to 1), especially when compared to NMI for the seasonal variation in the air transportation networks, suggesting that there is little common in the community structures obtained for consecutive stages. The most contrasting communities are those found in Ara's stage 1 and stage 2 networks (NMI = 0.1 in bottom left), suggesting that she significantly changed her word co-occurrence frequencies during this time. The other children have relatively similar and low NMI values. Specifically, Gai appears to develop steadily over the three stages; Ann and Dom slightly decrease their rate of change over time (NMI increases); and Bec and Car slightly increase their rate of change over time (NMI decreases).

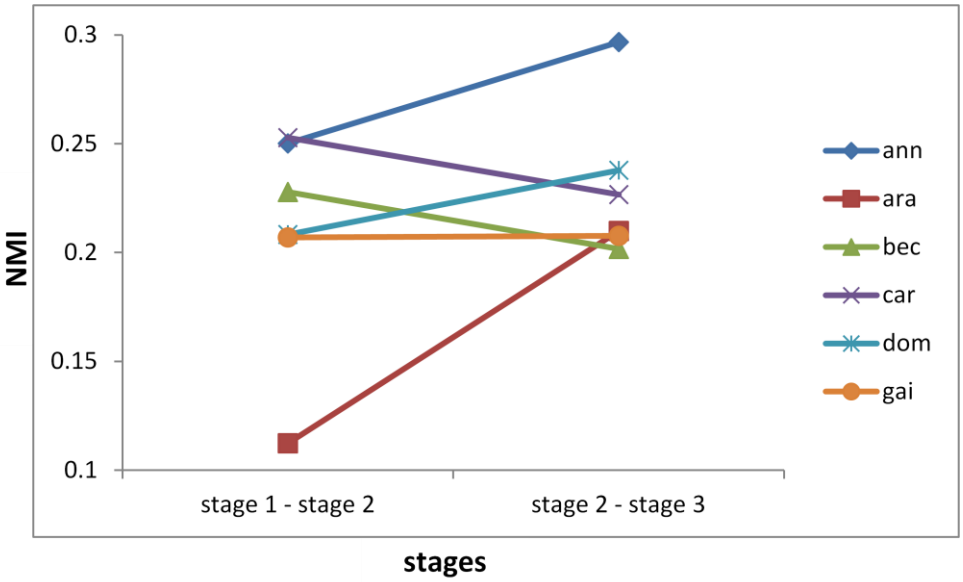

Fig. 6.8. Normalised Mutual Information (NMI) of consecutive stage networks of individual children.

The NMI values presented in Fig. 6.8 are also particularly important because they do not agree with Chomsky's theory of Universal Grammar. Assuming that UG parameters are set at an early age before the beginning of stage 1, UG would predict a high NMI since the theory suggests that once the parameters for language are set, linguistic production should not be affected by the environment and therefore the community structure of the networks should remain stable over time, resulting in high NMI. However, the fact that the networks represent observed interactions means that they are biased towards the statistical irregularities of the empirical data, so the communities are not a pure function of knowledge, but also of the nature of the specific data set obtained. Therefore, it is necessary to normalise for this effect, for example by ignoring the frequency of co-occurrence (i.e. the link weights), resulting in an unweighted network. In this case, however, the quantity of empirical data would determine the presence or absence of links since the more data is modelled the more likely it is that more unique links exist. Hence, it would be difficult to determine how much data to use and therefore it is not possible to make any solid conclusions regarding the agreement of UG with the NMI values obtained.

In summary, Fig 6.8 confirms the dynamic nature of children's language acquisition by exposing the long-term evolutionary changes in the community

structure of word co-occurrence networks incorporating both syntax and semantics.

- **Dynamics and Evolution of Aggregated Children's Networks**

This section discusses the community structure obtained for the average child (see aggregated networks in Figs. 5.25-5.27) in each of the three stages, and the main changes in the networks that took place during this developmental period. NMI has not been calculated for the aggregated networks as it is simply the average of the NMI of the individual networks.

Fig. 5.25 (stage 1) shows a pink community of the words *Anne, I, put, play, get, mummy, find, please, eat* and *open*, which are mostly verbs describing the actions performed by the children and/or the mothers. The blue community of *car, door, truck, daddy, baby* and *gone* (nouns and *gone*) clearly illustrates the child's observation of something *going* (*car* is the largest so *car gone* is definitely the main pair in this community). There is also a light green community of *man, train, nana* and *tractor*, and a dark green community of *bridge, lorry, take, hat* and *horse*. For stage 1 the phrase *I get* is particularly popular.

Fig. 5.26 (stage 2) shows a dark blue community *of mummy, daddy, baby, drink, draw, truck,* and a number of colours. *Drink* and *draw* are clearly the main family activities at this stage since they are in the same community as *baby, mummy* and *daddy*. *Truck* is probably the most common toy and the colours probably relate to the *truck* or to *draw*. There is a light blue community of *going, gone, car, train, come, back,* and *play,* which suggests that the child talked about a *car* or a *train going* or *com*ing *back* or being *play*ed with. The red word *get* is particularly popular (second highest frequency after *I*) but there are no other red words visible in the figure which means that it was widely used in combinations with many other less frequently occurring words. The brown community of *I, put, eat, take, sit, find, done, dolly, does* and *crash* indicate the main actions of *I* and *dolly*. For stage 2 the pair *I put* is particularly popular.

Fig. 5.27 (stage 3) shows a dark blue community of *I, get, put, mummy, eat, take, play* and many other verbs, so the children were able to describe more of the

activities that they became involved in, thanks to their continuous acquisition of language skills. The brown community of *going, gone, come, back, bit, doing, daddy, think, round* and *fall* mainly describes something or someone in motion since the top frequency word in the community is *going*. In other words, community detection has discovered two main communities that represent two very important but distinct types of verb: the blue community refers to actions carried out by the child (*I*), whereas the brown community refers to actions carried out by others (*going*) since *going* is probably the simplest verb to learn and the easiest action to observe. The red community of *train, car, baby, naughty, man, found, ones* and a number of colours is also interesting. In stage 3 the phrase *I get* has returned to be the most popular pair as it was in stage 1.

The total frequency of the top 50 words remains fairly stable over the three stages, as illustrated by the colour density in the plots, but clearly certain words' frequency changes over time, as depicted by the font sizes and the presence or absence of given words (each word has a single, unique location in the semantic space plots based on its semantic distance to each of the two dimension's semantic categories). As observed for the individual children's networks, the aggregated children's networks clearly show a bias towards words that are semantically closer to size, as opposed to goodness (most words are below the diagonal). It is interesting how for stage 1 the most popular pair of words (given their semantic distance) is *I get*, then for stage 2 it is *I put*, and then again for stage 3 it is back to *I get*.

- **Comparative Validation**

Similarly to the comparative validation section within the air transportation case study, the goal of this section is to demonstrate and to evaluate the effectiveness of Expert's (2011) space-independent community structure detection (Spa) in comparison to Newman and Girvan's (2004) general community detection (NG) for the language acquisition case study. Again, Newman's method was applied to all eighteen of the children's networks to compare the communities obtained. It is worth mentioning that at the time of writing this thesis community structure had not yet been identified for any linguistic network and therefore there was no

previous work to compare against. However, the partitions obtained using the two null models are significantly different to allow a comparison among the two, and to reveal the advantages of Expert's model that also uses semantics in the community detection to identify more coherent communities. Fig. 6.9 shows the network of Gai at stage 3 and all eighteen networks are presented in the Appendix (see Figs. D.55-D.72).

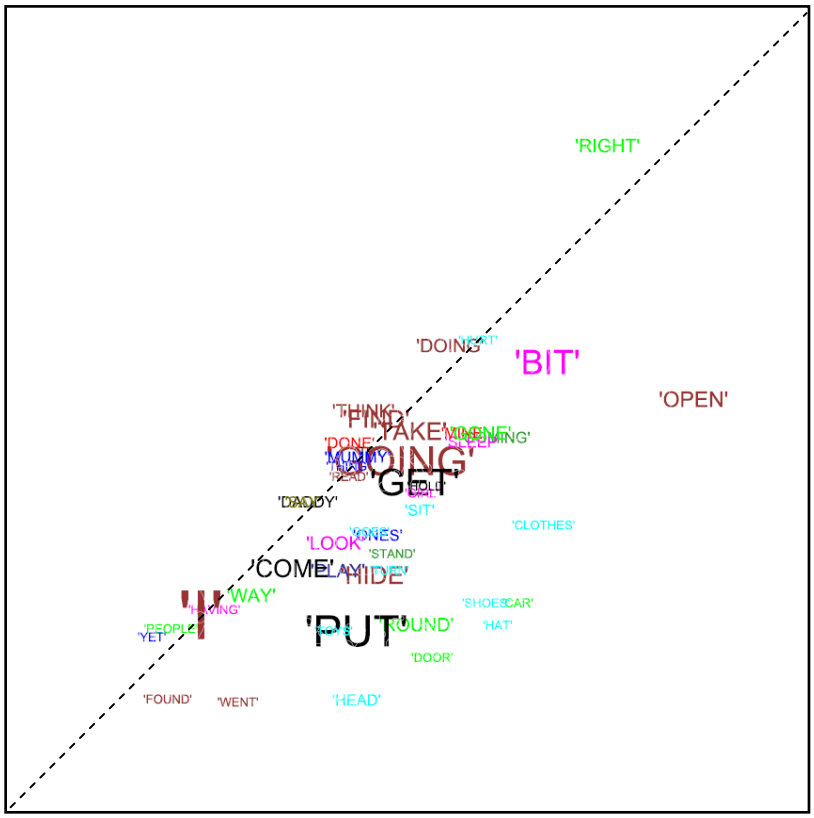

Fig. 6.9. Community structure in Gai 3 identified using Newman's method (same as Fig. D.72 in Appendix).

By visually comparing the community structure obtained using Expert's and Newman's null models it is clear that there is significant overlap in the partitions, even though the colours denoting communities may be inconsistent. Specifically, Ara 1's partitions are particularly similar, suggesting that the semantic distance between words did not affect their co-occurrence frequency significantly enough to be picked up by Expert's null model.

In order to demonstrate the benefits of using semantics in Expert's (2011) null model it is necessary to focus on the differences between partitions. For example, Bec 3's NG partition has *I, going* and *get* in pink, whereas the Spa partition has

going and *get* in pink but *I* in brown. This means that although *I, going* and *get* interacted strongly on a syntactic level, when also considering the linguistic bias towards stronger interactions for lower semantic distances (i.e. higher co-occurrence frequency if words are more similar semantically), then Spa places *I* in another community of particularly highly interacting words. Car 2's NG partition assigns all colours to a light green community whereas Spa assigns *black* and *green* to two different communities. This means that Spa differentiates these two colours from the rest thereby revealing new relationships among these two individual colours and other co-occurring words. Car 3's NG structure has *coming, train* and *car* in pink, whereas Spa has *coming* and *round* in cyan but *train* and *car* in black. In other words Spa has disentangled *coming* from the vehicles and has revealed an important connection *coming round* that is especially popular in this child's linguistic production. Finally, to provide the last example of some of the main differences between NG and Spa, Dom 2's NG plot has *going* and *gone* in cyan but *car* in pink, while Dom 2's Spa plot has *going*, *gone* and *car* in brown. This example shows how Spa has picked up on the fact that *car going* and *car gone* are particularly regular when semantics is also considered. In summary, Expert's model has revealed some specific particularly high-frequency co-occurrences among words, given their semantic similarity. On the other hand, Newman's model only reveals the pure communities based solely on the co-occurrence frequencies, without considering the semantic similarity between the co-occurring words.

To quantify the exact difference between the results of Expert's (2011) and Newman and Girvan's (2004) models, the NVI in the language acquisition networks in presented in Fig. 6.10. Here, each trend represents the NVI of a child over the three stages of development. Since the values calculated are between 0.3 and 0.6, they are lower than those for air transportation, but they are nevertheless significant. Therefore, the plot suggests that, again, there are considerable differences in the community structure obtained by Newman's method and Expert's method. Interestingly, NVI variation among the children is lower at stage 2, meaning that the difference between Expert and Newman is more stable across the children at this stage. Specifically, two of the trends go up, two go

down, and two are inverted, over time. Ann and Car go down, Ara and Bec go up, and the other two criss-cross. Dom first goes up then down, but Gai first goes down then up. These interesting patterns in the trends reveal the performance of Expert's model in comparison to Newman's, as the higher the NVI, the higher the impact of semantics in the community structure detection.

In summary, quantitative comparison of the communities obtained using Expert's (2011) and Newman and Girvan's (2004) null models suggests that there is a considerable difference between the partitions obtained. In other words, it is necessary to use semantics for community detection in language acquisition networks. Again, assuming that Expert's model accurately captures the bias towards more co-occurrences of semantically similar words, it follows that the higher the NVI, the better Expert's model performs in comparison to Newman's model.

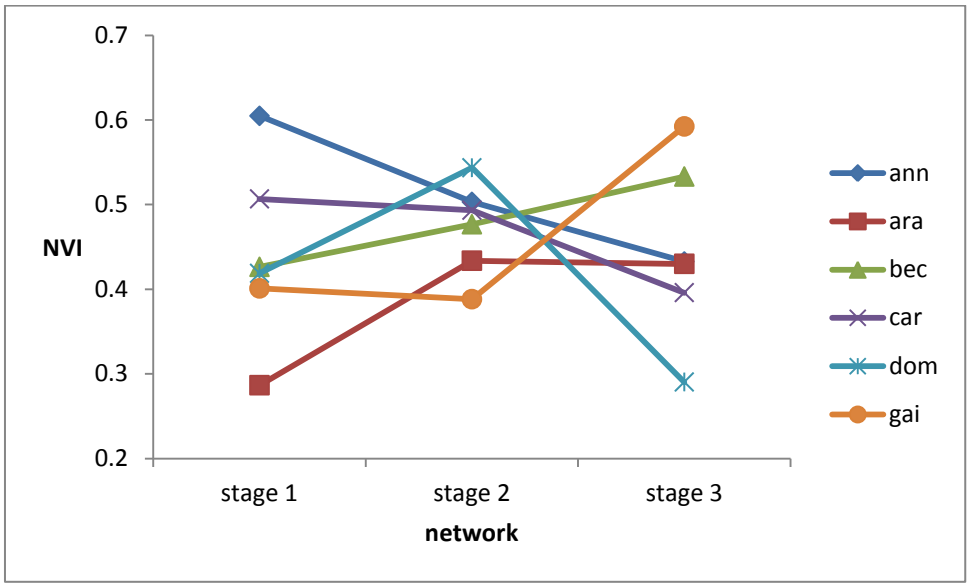

Fig. 6.10. Normalised Variation of Information (NVI) among community structure identified using Expert's and Newman's null models.

6.3 Comparison and Generalities

Now that the results of the two applied case studies have been individually discussed, it is appropriate to address them together in order to reveal their specifics, similarities, differences, and generalities. Fig. 6.11 (same as Fig. 2.8) shows the basic structure of this research: case studies in air transportation and language acquisition are both modeled using network theory, and this section

discusses the overlap among the two domains using the results of the analyses of the network models, as well as their generalities, such as common findings, trends, patterns, and relationships. It is important to mention that the common ground among the two applied case studies has been intentionally emphasized where possible, to facilitate comparison and to identify potential similarities. For example, apart from the model structure (3 stages, 6 networks per stage) and common network modeling methodology, the network analysis techniques are also kept as consistent as possible. A good example of this is the spatial community structure detection, which is applied to a specifically augmented language acquisition network that contains both syntax and semantics (semantics playing the role of spatial distance for the community structure identification).

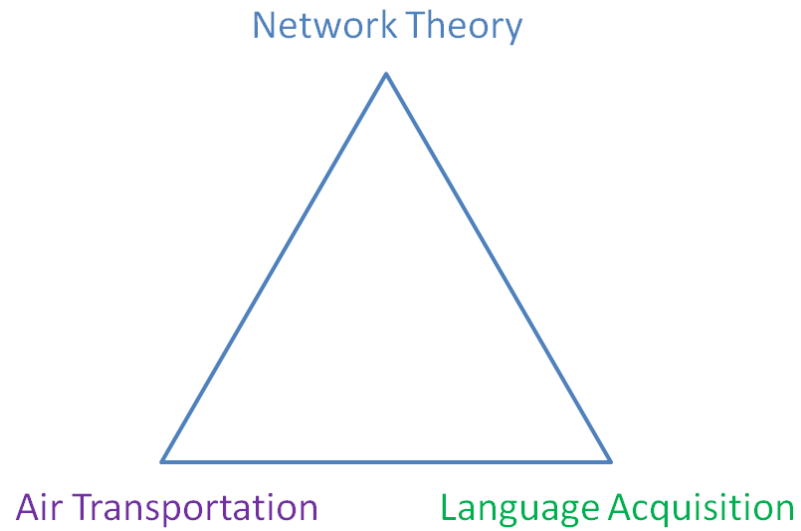

Fig. 6.11. Comparison and generalities of air transportation and language acquisition using network theory.

Although air transportation and language acquisition appear to have little in common at first, this thesis shows that in fact they do have a lot in common at a high abstract level. In addition to the common network modelling methodology that was used to analyse complex systems within each field, numerous common network analysis tools (such as statistical parameters and community structure detection) are also employed to study these systems. In terms of results of these analyses, one cannot expect any level of common trends or patterns in the properties of two networks that are so distinct. However, what both types of network definitely have in common is continuous dynamics of the network. In

other words, they may be growing, changing, developing, or even shrinking, but they are always evolving, driven by the environment that they interact with. For example, children's linguistic networks are constantly shaped by the children's social interactions, and air transportation networks are shaped by the interactions between airlines' flights supply and the passenger demand. Another interesting commonality between these two networks (and many other types of network) is that they are both formed as a result of the flows on the networks. For example, in language the flow is the order of the words in utterances and in air transportation the flow is the passengers flying from A to B. However, the real complexity emerges only when the flows begin to interact on the network, and this complexity has been revealed and discussed in this thesis.

It would be useful to compare directly the results obtained in each of the two applied case studies, and hence, Table 6.5 summarises the main qualitative and quantitative findings in children's language acquisition and air transportation. Here, only the actual children are compared, as the mothers, the baseline and MOSAIC are all related to the children, who are the primary source of interest in the language acquisition case study.

Table 6.5. Main findings in air transportation and children's language acquisition.

Property	Air transportation	Language acquisition
Degree distribution	Power-law	Power-law
Weight distribution	Logarithmic	Power-law
Community structure	Distributed	Distributed
Average $<k>$	26.64	5.81
Average L	2.88	3.14
Average C	0.66	0.26

Networks from both domains are scale-free since all in-degree and out-degree distributions follow a power-law of the form $f(x) = ax^n$. This means that connections are distributed in a highly non-uniform manner with preferential attachment to a small number of key nodes. However, the weight distributions are different since the passengers follow a logarithmic trend but the co-occurrence frequencies follow a power-law. This means that the passenger distribution is less skewed than the co-occurrence distribution, suggesting stronger preferential interactions in the language acquisition networks. The

identified space-independent community structure in both domains is distributed as opposed to regional, in the sense that community members are dispersed over large distances as opposed to being concentrated within a single region in space. This suggests that the communities found have particularly high internal interactions given the distance between them (which is the main aim of community detection). This is in contrast to previous community structure identified in the USAN (Ball, Karrer and Newman, 2011), which has been largely affected by the short-range interaction bias in spatial networks.

The average $<k>$ over all airport and children's networks reveals the high connectivity of US airports compared to words in children's word co-occurrence networks. The average geodesic length L is low and similar for networks of both domains. This means that it is very easy to get from any one node to another node in terms of the number of links that need to be traversed. However, the average clustering coefficient C is much higher in the USAN compared to the language acquisition networks, meaning that there are much more closed triangles of connections. This is most probably due to the higher density of connections in the USAN, which is reflected by the higher average $<k>$. As stated earlier in the case for language, since a thorough comparison to an ensemble of random networks in terms of L and C has not been carried out, it is not possible to state with certainty whether any of the networks obey the small-world property of low L and high C, but the values clearly show that the networks at least resemble a small-world.

6.4 Summary

This chapter discussed the main results in air transportation and language acquisition. A comparison of the two domains revealed some emerging generalities that highlight the common ground among them. In summary, this chapter has shown that network theory can provide a useful abstraction of a complex system that reveals valuable domain-specific knowledge, while bridging the gap among distant disciplines.

Chapter 7
Conclusion

This chapter concludes the thesis. The major contributions to research in the field of complex networks are summarised. Theoretical implications for each individual case study and for complex networks in general are drawn based on the analysis of the results. Recommendations for future work are provided at the end.

7.1 Research Contributions

This section summarises four theoretical contributions of this thesis to general research in the field of complex networks: *cluster damage*; *detailed network modelling*; *space-independent community structure*; and *dynamics and evolution*. Cluster damage is a novel approach to test the robustness of complex networks to cluster failure. Detailed network modelling refers to the models developed in this thesis that are more detailed than existing models in terms of the number of network dimensions (topology, link weights, link directionality, space and time). The latter two more complex dimensions are especially neglected by the complex systems research community but they are important so they have been addressed by the two contributions that follow. Space-independent community structure refers to community detection methods that are tailored specifically for spatial networks. Dynamics and evolution refers to the topological and flow changes of the USAN both on short (dynamics) and long-term (evolution) time scales, and also the changes in children's language in the short (dynamics) and long-term (evolution).

In addition, there are two applied contributions to the fields of air transportation and language acquisition. They highlight the benefits to science of multi-disciplinary research which facilitates the cross-fertilisation of ideas and methods among distant subjects.

- **Cluster Damage**

This theoretical contribution presents the robustness of the Internet to cluster damage at the Autonomous System level. Since existing approaches only investigate simple node or link damage, they do not consider scenarios where nodes within a cluster are so entangled that the failure of a small number of nodes can knock out the entire cluster. For example, if the cluster hubs within a cluster of the USAN are congested or not available, a cascade of congestion and flight cancellations may sweep across the entire cluster, or even across the entire network depending on the severity of the affected traffic. Therefore, when investigating robustness it is important to consider the possible failure of entire clusters, especially when the network is highly modular. The benefit of this contribution to science is that it presents a generalisation on the standard strategy for robustness testing.

- **Detailed Network Modelling**

Since some research still does not consider even the first three basic dimensions (topology, link weights and link directionality) even though they may be important for the given context, the fact that they are incorporated in this work is a contribution in its own right within each of the two applied case studies presented in this thesis. The benefit of this contribution to science is that it has shown how more comprehensive models are able to reveal much more information about the object being modelled.

- **Space-Independent Community Structure**

The detection of meaningful community structure is of great significance for the understanding and analysis of complex systems of any type. Therefore, this thesis has confirmed that spatial networks need tailored community detection methods, such as the null model proposed by Expert (2011). The main contribution here is the application of this novel null model to two very different case studies that were developed in this thesis. It is important to stress that there is no alternative method for uncovering space-independent community structure and Expert's method has only been applied to the Belgian mobile network. The benefit of this

contribution to science is that it has shown how the same method can be applied to very different problems as long as they are represented in the same way (in this case in the form of networks). This suggests that not only should methods be developed to be as generic as possible, but also problems should be encoded in a format that is as generic as possible (e.g. network model), in order to facilitate the application of methods in order to solve problems.

- **Dynamics and Evolution**

The short and long-term dynamics of an evolving complex network are important since they reveal trends in the network and allow the forecasting of future behaviour based on historical data. However, most research in complex networks does not address this essential issue. The benefit of this contribution to science is that the time dimension has been shown to be of significant importance in the study of systems that change over time.

- **Seasonal Variation and Evolution in Air Transportation**

The fact that the networks from language acquisition are partitioned into three stages of six children suggested an analogous treatment of the data from air transportation. In other words, the language acquisition case study was developed first and the structure of the model was applied for the first time in the field of air transportation, and specifically to the USAN case study, since it offered a much deeper insight into the evolutionary dynamics of the network. The key contribution here is the analysis of both seasonal variation within a calendar year and long-term evolution over two decades. The benefit of this contribution to science is that it shows how different time scales reveal different insights about the changes that occur in evolving complex systems, such as air transportation networks.

- **Syntax and Semantics in Language Acquisition**

Since the spatial community structure detection was first investigated for the case study on air transportation, it sparked an interesting idea for the language acquisition networks. Even though the word co-occurrence networks are not

embedded in space, they were extended to a virtual semantic space (i.e. first application of semantics in a syntax network), thereby incorporating an additional dimension of information in the form of semantics. Then, this fits perfectly with the spatial community detection, allowing the discovery of more meaningful communities based on both syntax in the form of word co-occurrence frequencies, and semantics in the form of semantic distances between words. The benefit of this contribution to science is that it reveals the significance of modelling mutually dependent variables, as opposed to neglecting one of them, such as semantics in word co-occurrence networks.

7.2 Theoretical Implications

The general robustness of part of the Internet has been shown to be very dependent on the kind of perturbation being considered. Specifically, the removal of two types of components (nodes and clusters) according to two strategies (errors and attacks) has revealed high heterogeneity in the robustness of the network. In other words, the *robust yet fragile* property common to many different complex networks has been exposed in the Internet. In addition, the obtained specific threshold robustness suggests that when larger clusters are attacked, the network is more robust since the attacks are less focused and more distributed, and hence, less effective since more nodes need to be damaged in order to damage the critical nodes. Therefore, in order to protect against targeted component attacks, the results suggest that by coupling nodes into dense clusters that need to be taken down as one whole (i.e. where individual internal nodes cannot fail without the entire cluster failing) would result in a more robust network configuration. In summary, the proposed network perturbation strategy based on the removal of entire clusters of nodes has been successfully applied to a partial model of the Internet at the AS level. The results have revealed an intuitive relationship between robustness and perturbation focus, which is also expected to hold for other kinds of real-world and computer-generated complex networks.

The USAN is a complex system that is continuously evolving to meet the growing demands for air travel. Investigating the community structure within has

illuminated important hidden characteristics of the network's topology and dynamics. Specifically, the findings reveal high heterogeneity in both space and time. In other words, the network is non-uniform (in space) and non-linear (in time) in terms of its connections and traffic. In addition, spatial community detection has identified a more realistic picture of the intricate structure within the network, which is invaluable for understanding this critical transportation system. Furthermore, this network model may be used for forecasting future trends in the USAN. For example, the identification of reliable communities can be the first step to study how external factors, such as natural disasters (e.g. tornados, which are common in large parts of the US), affect the function of the network. Moreover, the communities emerging from socio-economic interactions, as in the case of migration, reflect both the social influence radii and the activity system configuration (the distribution of activities in terms of location). Variations in the activity system will possibly modify such relationships and the resulting community structure. Finally, there is a clear relationship between domestic US air travel and migration. In particular, the identified community structures map well onto the migration patterns among the four macro-regions and within the region.

Young children's first language acquisition is a complex process that is continuously driven by the immediate environment, and specifically, by the social interactions with parents or other children. This has been confirmed by the language acquisition network analyses, such as the evolving community structure identified, which reveals the heterogeneity in children's word combinations captured by the word co-occurrence frequencies in the networks. In addition, the extension of the syntactic language acquisition networks to enable the modelling of semantics for the application of spatial community detection has allowed a deeper and more accurate understanding of the particularly strong word interactions in the language produced. In addition, the network model can be used for the development of growth models that simulate the development of language acquisition networks, and also for predicting future properties and trends in the language of children that are currently learning (and hence there are no data of future networks yet). The identification of reliable communities can

also be used to study how external factors, such as parent's language, the social environment, or even linguistic disabilities, affect the structure and dynamics of language acquisition networks, thereby revealing the intricate effects on linguistic production. The key conclusion of this case study, however, relates to the hypothesis that MOSAIC is a good simulation tool for children's syntactic acquisition. Specifically, the results of the network analyses suggest that MOSAIC performs very well indeed, implying that language acquisition is definitely affected by the environment and specifically child-directed speech. This, in turn, questions the validity of the mainstream Universal Grammar theory proposed by Chomsky. In summary, based on the findings of this research it is fair to say that perhaps the best explanation for children's extraordinary ability to learn their first language is that they both have an innate predisposition for language *and* built-in mechanisms for analysing (and learning from) the statistical properties of the language they are subjected to.

Evolution-based modelling of networks promises to be a useful tool for extracting detailed information about the complex interactions in networks that are typically growing, as demonstrated by the applied case studies presented in this thesis. Specifically, where possible, it is recommended to study complex networks in line with the contributions outlined above: cluster damage has to be considered; models need to be sufficiently detailed; spatial networks deserve special attention; and network evolution also needs to be considered. In addition, both seasonal variation and evolution should be investigated in air transportation models; and both syntax and semantics should be combined in language acquisition models. It is worth mentioning that the approach described in this thesis is simple and straightforward, and may be applied to the study of any transportation or language acquisition network, and more generally, to any evolving complex network.

7.3 Future Work

Future work in the general field of complex networks needs to propose new theoretical methods for the analysis and understanding of complex systems, and new simulation models that display the behaviour of empirical models of

complex systems. Specifically, it is necessary to develop and evaluate alternative methods for the discovery of space-independent community structure. One important question regarding the application of Expert's (2011) method is exactly how much bias there is towards stronger short-range interactions in the network being studied. Also, it is important to identify new network parameters that describe sufficiently well both dynamics on and of the network. Finally, given the proposed evolution-based network models, it would be useful to forecast potential future trends in the networks.

Future work in the more specific fields of robustness, air transportation and language acquisition has to address a number of issues. First of all, the proposed cluster damage strategy should be tested for clusters that are identified using a community detection method (thus becoming community damage) in order to generalise the current findings to the failure of meaningful communities (as opposed to the more trivial clusters), which is more appropriate for simulating real world failures. Regarding air transportation, the relationship between migration and air travel in the US needs to be explored further as the preliminary results in this thesis suggest that inter-state migration may be a significant contributing factor to US air travel in general. Regarding language acquisition, clearly it is necessary to increase the number of observations (only 6 children in this thesis) in order to identify statistically significant correlations among multiple subjects. In addition, there are currently no network growth models that aim to simulate the development of children's language acquisition networks (e.g. word co-occurrence), which would be useful for the study of language acquisition (there exist growth models for air transportation). Another particularly useful extension is to identify the community structure in MOSAIC's networks in order to compare the extent of overlap among MOSAIC and the children in terms of communities of densely interacting words.

The author looks forward to exciting new developments in the multi-disciplinary field of complex networks and hopes that in time they will help to solve ever-more challenging problems.

Bibliography

Adamo, M. and Boylan, S. (2008) *A network approach to lexical growth and syntactic evolution in child language acquistion*, manuscript edn.

Albert, R., Jeong, H. and Barabási, A.-. (2000) "Error and attack tolerance of complex networks", *Nature,* vol. 406, no. 6794, pp. 378-382.

Albert, R., Jeong, H. and Barabási, A.-. (1999) "Diameter of the world-wide web", *Nature,* vol. 401, no. 6749, pp. 130-131.

Alderson, D. and Willinger, W. (2005) "A contrasting look at self-organization in the Internet and next-generation communication networks", *IEEE Communications Magazine,* vol. 43, no. 7, pp. 94-100.

Almendral, J.A., Leyva, I., Li, D., Sendiña-Nadal, I., Havlin, S. and Boccaletti, S. (2010) "Dynamics of overlapping structures in modular networks", *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics,* vol. 82, no. 1.

Amaral, L.A.N., Scala, A., Barthélémy, M. and Stanley, H.E. (2000) "Classes of small-world networks", *Proceedings of the National Academy of Sciences of the United States of America,* vol. 97, no. 21, pp. 11149-11152.

Atherton, M. and Bates, R. (2005) "Robustness and Complexity" in *Nature and Design,* eds. M. Collins, M. Atherton and J. Bryant, WIT Press, Southampton.

Australian National University (n.d.) *chap5Newth-final-5*. Available at: http://epress.anu.edu.au/cs/chap5Newth-final-5.jpg 2008).

Bagler, G. (2008) "Analysis of the airport network of India as a complex weighted network", *Physica A: Statistical Mechanics and its Applications,* vol. 387, no. 12, pp. 2972-2980.

Ball, B., Karrer, B. and Newman, M.E.J. (2011) "Efficient and principled method for detecting communities in networks", *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics,* vol. 84, pp. 036103.

Barabási, A.-. and Albert, R. (1999) "Emergence of scaling in random networks", *Science,* vol. 286, no. 5439, pp. 509-512.

Barrat, A., Barthélemy, M., Pastor-Satorras, R. and Vespignani, A. (2004) "The architecture of complex weighted networks", *Proceedings of the National Academy of Sciences of the United States of America,* vol. 101, no. 11, pp. 3747-3752.

Barthélemy, M., Barrat, A., Pastor-Satorras, R. and Vespignani, A. (2005) "Characterization and modeling of weighted networks", *Physica A: Statistical Mechanics and its Applications,* vol. 346, no. 1-2 SPEC. ISS., pp. 34-43.

Barthélémy, M. and Nunes Amaral, L.A. (1999) "Small-world networks: Evidence for a crossover picture", *Physical Review Letters,* vol. 82, no. 15, pp. 3180-3183.

Bates, E. and Carnavale, G.F. (1993) "New directions in research on child development.", *Developmental Review,* vol. 13, pp. 436-470.

Blondel, V.D., Guillaume, J.-., Lambiotte, R. and Lefebvre, E. (2008) "Fast unfolding of communities in large networks", *Journal of Statistical Mechanics: Theory and Experiment,* vol. 2008, no. 10.

Bounova, G. (2009) *Topological evolution of networks: case studies in the US airlines and language Wikipedias,* MIT.

Bulu, M. (2012) *City Competitiveness and Improving Urban Subsystems: Technologies and Applications,* IGI Global.

Bureau of Transportation Statistics (n.d.) *Bureau of Transportation Statistics* . Available at: http://www.bts.gov/ 2011).

Burghouwt, G. and de Wit, J. (2005) "Temporal configurations of European airline networks", *Journal of Air Transport Management,* vol. 11, no. 3, pp. 185-198.

Button, K.J. (2010) "Economic aspects of regional airport development" in *Developments of Regional Airports,* ed. M.N. Postorino, WIT Press, Southampton, UK.

Calabrese, F., Dahlem, D., Gerber, A., Paul, D., Chen, X., Rowland, J., Rath, C. and Ratti, C. (2011) "The connected states of America: Quantifying social radii of influence", *Proceedings - 2011 IEEE International Conference on Privacy, Security, Risk and Trust and IEEE International Conference on Social Computing, PASSAT/SocialCom 2011*, pp. 223.

Camagni, R. (2002) "On the concept of territorial competitiveness: Sound or misleading?", *Urban Studies,* vol. 39, no. 13, pp. 2395-2411.

Cancho, R.F.I. and Solé, R.V. (2001) "The small world of human language", *Proceedings of the Royal Society B: Biological Sciences,* vol. 268, no. 1482, pp. 2261-2265.

Carreras, B.A., Lynch, V.E., Dobson, I. and Newman, D.E. (2002) "Critical points and transitions in an electric power transmission model for cascading failure blackouts", *Chaos,* vol. 12, no. 4, pp. 985-994.

Carter, S.L., Brechbühler, C.M., Griffin, M. and Bond, A.T. (2004) "Gene co-expression network topology provides a framework for molecular characterization of cellular state", *Bioinformatics,* vol. 20, no. 14, pp. 2242-2250.

Cartwright, T.A. and Brent, M.R. (1997) "Syntactic categorization in early language acquisition: formalizing the role of distributional analysis", *Cognition,* vol. 63, no. 2, pp. 121-170.

Cervero, R. (2001) "Efficient urbanisation: Economic performance and the shape of the metropolis", *Urban Studies,* vol. 38, no. 10, pp. 1651-1671.

Chomsky, N. (1981) *Lectures on government and binding,* Foris, Dordrecht, The Netherlands.

Chomsky, N. (1957) *Syntactic structures,* Mouton and Co., The Hague.

Chorianopoulos, I., Pagonis, T., Koukoulas, S. and Drymoniti, S. (2010) "Planning, competitiveness and sprawl in the Mediterranean city: The case of Athens", *Cities,* vol. 27, no. 4, pp. 249-259.

Christiansen, M.H. and Chater, N. (2001) *Connectionist psycholinguistics: capturing the empirical data,* Elsevier Science.

Cohen, J. (1988) *Statistical Power Analysis for the Behavioral Sciences,* 2nd edn, Erlbaum, Mahwah, NJ.

Corominas-Murtra, B., Valverde, S. and Solé, R.V. (2010) "Emergence of scale-free syntax networks", *Evolution of Communication and Language in Embodied Agents,* vol. 83.

Criado, R., Flores, J., HernÃ¡ndez-Bermejo, B., Pello, J. and Romance, M. (2005) "Effective measurement of network vulnerability under random and intentional attacks", *Journal of Mathematical Modelling and Algorithms,* vol. 4, no. 3, pp. 307-316.

Croker, S., Pine, J.M. and Gobet, F. (2000) "Modelling optional infinitive phenomena: A computational account", *Proceedings of the Third International Conference on Cognitive Modeling,* eds. N. Taatgen and J. Aasman, Universal Press, Veenendaal, The Netherlands.

Crucitti, P., Latora, V. and Marchiori, M. (2004) "Model for cascading failures in complex networks", *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics,* vol. 69, no. 4 2, pp. 045104-1-045104-4.

Danon, L., Díaz-Guilera, A., Duch, J. and Arenas, A. (2005) "Comparing community structure identification", *Journal of Statistical Mechanics: Theory and Experiment,* , no. 9, pp. 219-228.

De Montis, A., Barthélemy, M., Chessa, A. and Vespignani, A. (2007) "The structure of interurban traffic: A weighted network analysis", *Environment and Planning B: Planning and Design,* vol. 34, no. 5, pp. 905-924.

Elman, J.L. (1993) "Learning and development in neural networks: The importance of starting small", *Cognition,* vol. 48, pp. 71-99.

Erdös, P. and Rényi, A. (1959) "On the evolution of random graphs", *Publ. Math. Inst. Hung. Acad. Sci.,* vol. 5.

Estrada, E. and Hatano, N. (2009) "Communicability graph and community structures in complex networks", *Applied Mathematics and Computation,* vol. 214, no. 2, pp. 500-511.

Expert, P., Evans, T.S., Blondel, V.D. and Lambiotte, R. (2011) "Uncovering space-independent communities in spatial networks", *Proceedings of the National Academy of Sciences of the United States of America,* vol. 108, no. 19, pp. 7663-7668.

Faloutsos, M., Faloutsos, P. and Faioutsos, C. (1999) "On power-law relationships of the internet topology", *Computer Communication Review,* vol. 29, no. 4, pp. 251-261.

Feigenbaum, E.A. and Simon, H.A. (1984) "EPAM-like models of recognition and learning", *Cognitive Science,* vol. 8, no. 4, pp. 305-336.

Freudenthal, D., Pine, J. and Gobet, F. (2001) "Modeling the optional infinitive stage in MOSAIC: A generalisation to Dutch", *Proceedings of the Fourth International Conference on Cognitive Modeling,* eds. E.M. Altmann, A. Cleeremans, C.D. Schunn and W.D. Gray, Lawrence Erlbaum Associates, Inc., Mahwah, NJ.

Freudenthal, D., Pine, J.M., Aguado-Orea, J. and Gobet, F. (2007) "Modeling the developmental patterning of finiteness marking in English, Dutch, German, and Spanish using MOSAIC", *Cognitive Science,* vol. 31, no. 2, pp. 311-341.

Freudenthal, D., Pine, J.M. and Gobet, F. (2006) "Modeling the Development of Children's Use of Optional Infinitives in Dutch and English Using MOSAIC", *Cognitive Science,* vol. 30, no. 2, pp. 277-310.

Girvan, M. and Newman, M.E.J. (2002) "Community structure in social and biological networks", *Proceedings of the National Academy of Sciences of the United States of America,* vol. 99, no. 12, pp. 7821-7826.

Gobet, F. and Simon, H.A. (2000) "Five Seconds or Sixty? Presentation Time in Expert Memory", *Cognitive Science*, vol. 24, no. 4, pp. 651-682.

Gomez, R.L. and Gerken, L. (1999) "Artificial grammar learning by 1-year-olds leads to specific and abstract knowledge", *Cognition*, vol. 70, pp. 109-135.

Good, B.H., De Montjoye, Y.-. and Clauset, A. (2010) "Performance of modularity maximization in practical contexts", *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, vol. 81, no. 4.

Guimerà, R., Mossa, S., Turtschi, A. and Amaral, L.A.N. (2005) "The worldwide air transportation network: Anomalous centrality, community structure, and cities' global roles", *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 22, pp. 7794-7799.

Haitao, L., and Fengguo, H. (2008) "What role does syntax play in a language network?", *Europhysics Letters*, vol. 83, no. 1, pp. 18002.

Holme, P., Kim, B.J., Yoon, C.N. and Han, S.K. (2002) "Attack vulnerability of complex networks", *Phys.Rev.E*, vol. 65, no. 5, pp. 056109.

Hsu, C.-. and Wen, Y.-. (2003) "Determining flight frequencies on an airline network with demand-supply interactions", *Transportation Research Part E: Logistics and Transportation Review*, vol. 39, no. 6, pp. 417-441.

Huberman, B.A. and Adamic, L.A. (1999) "Growth dynamics of the world-wide web", *Nature*, vol. 401, no. 6749, pp. 131.

ibiblio (n.d.) *peer_to_peer1*. Available at: http://www.ibiblio.org/team/intro/search/peer_to_peer1.gif 2008).

Internal Revenue Service (IRS) (n.d.) *migration data 2010*. Available at: http://www.irs.gov/taxstats/ 2012).

Jen, E. (2003) "Stable or robust? What's the difference?", *Complexity*, vol. 8, no. 3, pp. 12-18.

Jeong, H., Tombor, B., Albert, R., Oltval, Z.N. and Barabásl, A.-. (2000) "The large-scale organization of metabolic networks", *Nature*, vol. 407, no. 6804, pp. 651-654.

Jutla, I.S. and Mucha, P.J. (n.d.) *A generalized Louvain method for community detection implemented in MATLAB*. Available at: http://netwiki.amath.unc.edu/GenLouvain 2012).

Kaiser, M. and Hilgetag, C.C. (2004) "Edge vulnerability in neural and metabolic networks", *Biological cybernetics*, vol. 90, no. 5, pp. 311-317.

Ke, J. and Yao, Y. (2008) "Analysing language development from a network approach", *Journal of Quantitative Linguistics,* vol. 15, no. 1, pp. 70-99.

Kitano, H. (2004) "Biological robustness", *Nature Reviews Genetics,* vol. 5, no. 11, pp. 826-837.

Kitano, H. (2002) "Systems biology: A brief overview", *Science,* vol. 295, no. 5560, pp. 1662-1664.

Kolb, P. (2008) "DISCO: A Multilingual Database of Distributionally Similar Words", *Proceedings of KONVENS-2008*Berlin.

Krakauer, D.C. and Plotkin, J.B. (2005) "Principles and parameters of molecular robustness", *In Robust Design: A Repertoire for Biology, Ecology and Engineering*Oxford University Press, , pp. 71.

Kumar, R., Raghavan, P., Rajagopalan, S. and Tomkins, A. (1999) "Trawling the Web for emerging cyber-communities", *Computer Networks,* vol. 31, no. 11, pp. 1481-1493.

Lancichinetti, A. and Fortunato, S. (2009) "Community detection algorithms: A comparative analysis", *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics,* vol. 80, no. 5.

Lancichinetti, A., Radicchi, F., Ramasco, J.J. and Fortunato, S. (2011) "Finding statistically significant communities in networks", *PLoS ONE,* vol. 6, no. 4.

Leigh, J.R. (1992) *Control Theory: A guided Tour,* IEE Publishing, London.

Levy, M. (2010) "Scale-free human migration and the geography of social networks", *Physica A: Statistical Mechanics and its Applications,* vol. 389, no. 21, pp. 4913-4917.

Li, J. and Zhou, J. (2007) "Chinese character structure analysis based on complex networks", *Physica A: Statistical Mechanics and its Applications,* vol. 380, no. 1-2, pp. 629-638.

Li, W. and Cai, X. (2004) "Statistical analysis of airport network of China", *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics,* vol. 69, no. 4 2, pp. 046106-1-046106-6.

Liang, W., Shi, Y., Tse, C.K., Liu, J., Wang, Y. and Cui, X. (2009) "Comparison of co-occurrence networks of the Chinese and English languages", *Physica A: Statistical Mechanics and its Applications,* vol. 388, no. 23, pp. 4901-4909.

Mackun, P., Wilson, S., Fischetti, T. and Goworowska, J. (2011) *2010 Census Brief* .

MacWhinney, B. (2000) *The CHILDES project: Tools for analysing talk* , 3rd edn, Erlbaum, Mahwah, NJ.

Meila, M. (2003) "Comparing clusterings by the variation of information", *Lecture Notes in Artificial Intelligence (Subseries of Lecture Notes in Computer Science)*, pp. 173.

Meunier, D., Lambiotte, R. and Bullmore, E. (2010) "Modular and hierarchically modular organization of brain networks", *Frontiers in Neuroscience*, vol. 4.

Milgram, S. (1967) " The small-world problem", *Psychology Today*, vol. 2.

Morine, M.J., Gu, H., Myers, R.A. and Bielawski, J.P. (2009) "Trade-offs between efficiency and robustness in bacterial metabolic networks are associated with niche breadth", *Journal of Molecular Evolution*, vol. 68, no. 5, pp. 506-515.

Motter, A.E., De Moura, A.P.S., Lai, Y.-. and Dasgupta, P. (2002) "Topology of the conceptual network of language", *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, vol. 65, no. 6, pp. 065102/1-065102/4.

Motter, A.E. (2004) "Cascade Control and Defense in Complex Networks", *Phys.Rev.Lett.*, vol. 93, no. 9, pp. 098701.

Naigles, L. and Hoff-Ginsberg, E. (1998) "Why are some verbs learned before other verbs? Effects of input frequency
and structure on children's early verb use.", *Journal of Child Language*, vol. 25, pp. 95-120.

Newman, M.E.J. (2003) "The Structure and Function of Complex Networks", *SIAM Review*, vol. 45, no. 2, pp. 167-256.

Newman, M.E.J. (2006) "Finding community structure in networks using the eigenvectors of matrices", *Phys.Rev.E*, vol. 74, no. 3, pp. 036104.

Newman, M.E.J. and Girvan, M. (2004) "Finding and evaluating community structure in networks", *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, vol. 69, no. 2 2, pp. 026113-1-026113-15.

Patuelli, R., Reggiani, A., Gorman, S.P., Nijkamp, P. and Bade, F.-. (2007) "Network analysis of commuting flows: A comparative static approach to German data", *Networks and Spatial Economics*, vol. 7, no. 4, pp. 315-331.

Perry, M.J. (2006) *Domestic Net Migration in the United States: 2000 to 2004*, U.S. Department of Commerce, Economics and Statistics Administration.

Pinker, S. (1984) *Language learnability and language development.* Harvard University Press, Cambridge, MA.

Prehofer, C. and Bettstetter, C. (2005) "Self-organization in communication networks: Principles and design paradigms", *IEEE Communications Magazine,* vol. 43, no. 7, pp. 78-85.

Redington, M., Chater, N. and Finch, S. (1998) "Distributional information: A powerful cue for acquiring syntactic categories", *Cognitive Science,* vol. 22, no. 4, pp. 425-469.

Rouwendal, J. (2004) "Search theory and commuting behavior", *Growth and Change,* vol. 35, no. 3, pp. 391-418.

Russell, Y.I., Murzac, A., Gobet, F. and Whitehouse, H. (to be published) *Semantic network analysis of religious pamphlets*.

Saffran, J.R., Aslin, R.N. and Newport, E.L. (1996) "Statistical Learning by 8-Month-Old Infants", *Science,* vol. 274, no. 5294, pp. 1926-1928.

Schwartz, A. (1973) "Interpreting the effect of distance on migration", *Journal of Political Economy,* vol. 81.

Shi, Y., Liang, W., Liu, J. and Tse, C.K. (2008) "Structural equivalence between co-occurrences of characters and words in the Chinese language", *1. International Symposium on Nonlinear Theory and its Applications*.

Solé, R.V., Corominas-Murtra, B., Valverde, S. and Steels, L. (2010) "Language networks: Their structure, function, and evolution", *Complexity,* vol. 15, no. 6, pp. 20-26.

Staab, S., Heylighen, F., Gershenson, C., Flake, G.W., Pennock, D.M., Fain, D.C., Roure, D.D., Aberer, K., Shen, W., Dousse, O. and Thiran, P. (2003) "Neurons, Viscose Fluids, Freshwater Polyp Hydra-and Self-Organizing Information Systems", *IEEE Intelligent Systems,* , pp. 72-86.

Sumpter, D.J.T. (2006) "The principles of collective animal behaviour", *Philosophical Transactions of the Royal Society B: Biological Sciences,* vol. 361, no. 1465, pp. 5-22.

Theakston, A.L., Lieven, E.V.M., Pine, J.M. and Rowland, C.F. (2001) "The role of performance limitations in the acquisition of verb-argument structure: An alternative account", *Journal of child language,* vol. 28, no. 1, pp. 127-152.

United States Census Bureau (n.d.) *US census data 2011* . Available at: http://www.census.gov/ 2012).

Wagner, A. (2005a) *Robustness and Evolvability in Living Systems* Princeton University Press, Princeton, NJ.

Wagner, A. (2005b) "Distributed robustness versus redundancy as causes of mutational robustness", *BioEssays,* vol. 27, no. 2, pp. 176-188.

Wang, X.F. and Chen, G. (2003) "Complex networks: Small-world, scale-free and beyond", *IEEE Circuits and Systems Magazine,* vol. 3, no. 1, pp. 6-20.

Wasserman, S. and Faust, K. (1994) *Social Network Analysis* , Cambridge University Press, Cambridge.

Watts, D.J. (2002) "A simple model of global cascades on random networks", *Proceedings of the National Academy of Sciences of the United States of America,* vol. 99, no. 9, pp. 5766-5771.

Watts, D.J. and Strogatz, S.H. (1998) "Collective dynamics of 'small-world9 networks", *Nature,* vol. 393, no. 6684, pp. 440-442.

Wuellner, D.R., Roy, S. and D'Souza, R.M. (2010) "Resilience and rewiring of the passenger airline networks in the United States", *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics,* vol. 82, no. 5.

Xu, Z. and Harriss, R. (2008) "Exploring the structure of the U.S. intercity passenger air transportation network: A weighted complex network approach", *GeoJournal,* vol. 73, no. 2, pp. 87-102.

Zhou, S., Hu, G., Zhang, Z. and Guan, J. (2008) "An empirical study of Chinese language networks", *Physica A: Statistical Mechanics and its Applications,* vol. 387, no. 12, pp. 3039-3047.

Publications

P1. Gegov, E., Postorino, M.N., Atherton, M. and Gobet, F. Community structure detection in the evolution of the United States airport network. *Journal of Advances in Complex Systems* (2013).

P2. Gegov, E., Gegov, A., Gobet, F., Atherton, M., Freudenthal, D. and Pine, J. Cognitive modelling of language acquisition with complex networks. In: *Computational Intelligence* (A. Floares (Ed)). Nova Science, Hauppauge NY (2012).

P3. Gegov, E., Gobet, F., Atherton, M., Freudenthal, D. and Pine, J. Modelling language acquisition in children using network theory. In: *European Perspectives on Cognitive Science* (B. Kokinov, A. Karmiloff-Smith and N. Nersessian (Ed)). New Bulgarian University Press, Sofia (2011).

P4. Gegov, E., Gegov, A., Postorino, M.N., Atherton, M. and Gobet, F. Space-independent community structure detection in United States air transportation. *Proceedings of IFAC 2012: International Symposium on Control in Transportation Systems* (2012).

P5. Gegov, E., Gegov, A., Atherton, M. and Gobet, F. Evolution-based modelling of complex airport networks. *Proceedings of COSY 2011: International Conference on Complex Systems* (2011).

P6. Gegov, E. Robustness of complex networks to node and cluster damage. *Proceedings of ECCS 2009: European Conference on Complex Systems* (2009).

P7. Gegov, A., Petrov, N., Gegov, E. Rule base identification in fuzzy networks by boolean matrix equations. *Journal of Intelligent and Fuzzy Systems* (2013).

Appendix A
USAN Community Structure by Expert's Method

The following figures depict the community structure found in the USAN using Expert's method. In each figure, all community members are assigned the same colour.

- **Year 1990**

Figs. A.1-A.6 depict bi-monthly snapshots of the USAN for the year 1990.

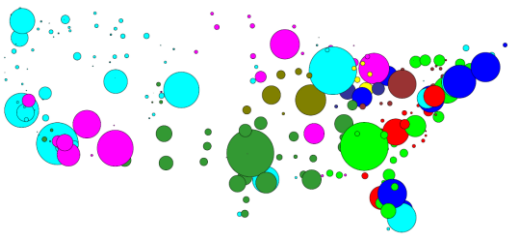

Fig. A.1. JAN-FEB 1990 community structure with Expert.

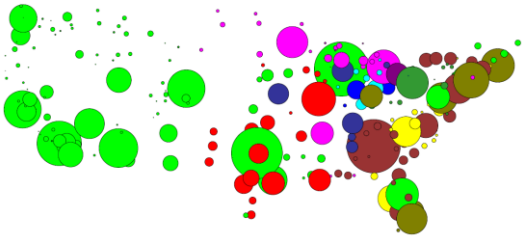

Fig. A.2. MAR-APR 1990 community structure with Expert.

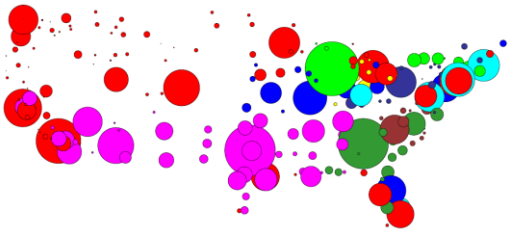

Fig. A.3. MAY-JUN 1990 community structure with Expert.

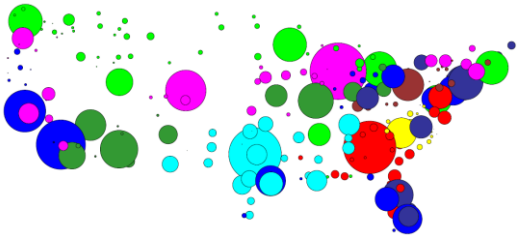

Fig. A.4. JUL-AUG 1990 community structure with Expert.

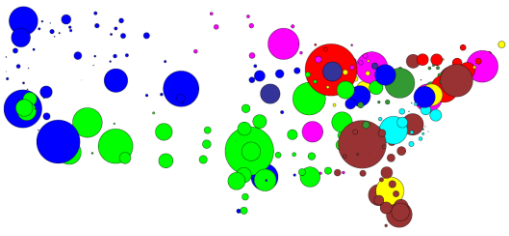

Fig. A.5. SEP-OCT 1990 community structure with Expert.

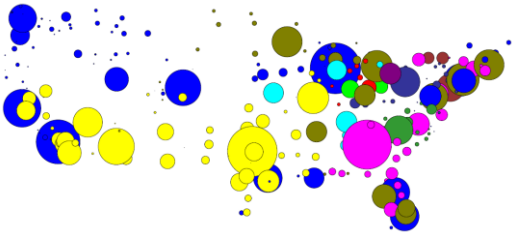

Fig. A.6. NOV-DEC 1990 community structure with Expert.

- **Year 2000**

Figs. A.7-A.12 depict bi-monthly snapshots of the USAN for the year 2000.

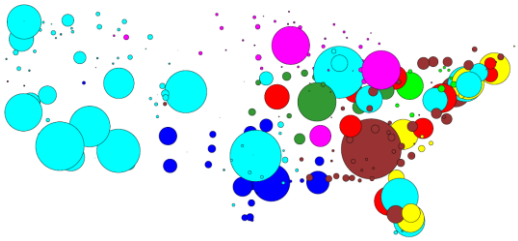

Fig. A.7. JAN-FEB 2000 community structure with Expert.

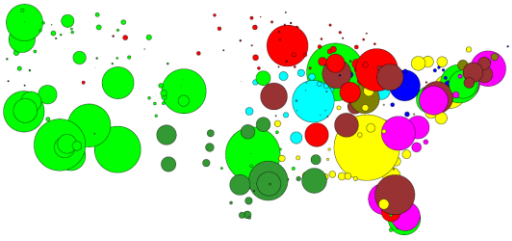

Fig. A.8. MAR-APR 2000 community structure with Expert.

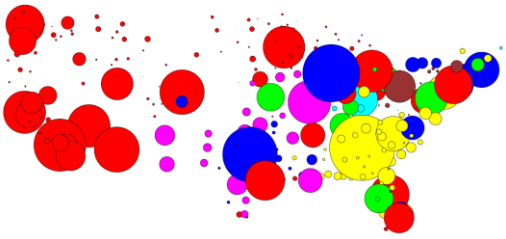

Fig. A.9. MAY-JUN 2000 community structure with Expert.

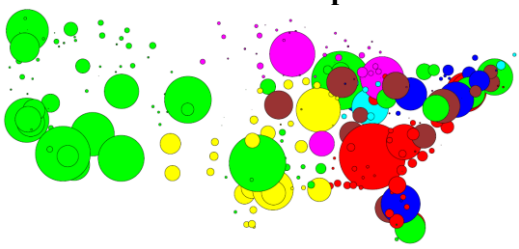

Fig. A.10. JUL-AUG 2000 community structure with Expert.

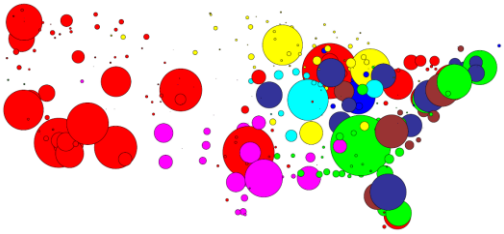

Fig. A.11. SEP-OCT 2000 community structure with Expert.

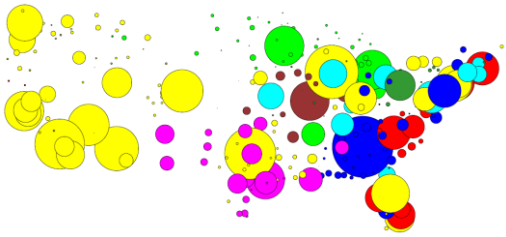

Fig. A.12. NOV-DEC 2000 community structure with Expert.

- **Year 2010**

Figs. A.13-A.18 depict bi-monthly snapshots of the USAN for the year 2010.

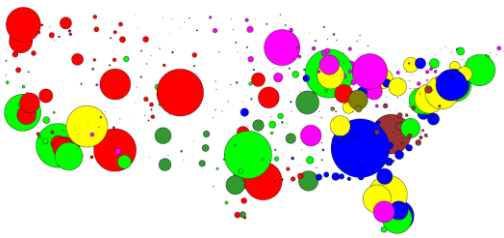

Fig. A.13. JAN-FEB 2010 community structure with Expert.

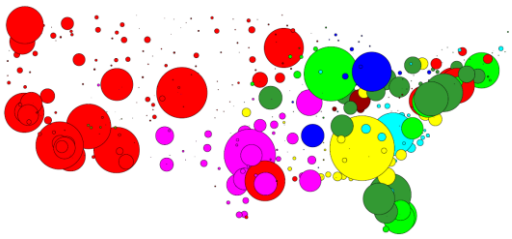

Fig. A.14. MAR-APR 2010 community structure with Expert.

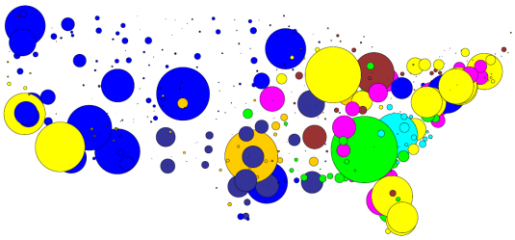

Fig. A.15. MAY-JUN 2010 community structure with Expert.

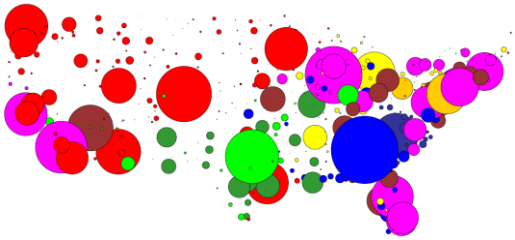

Fig. A.16. JUL-AUG 2010 community structure with Expert.

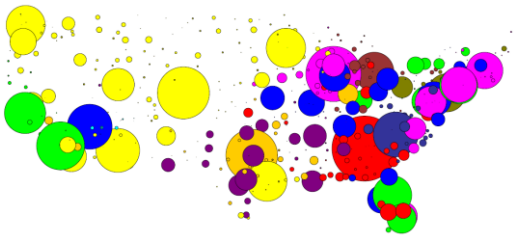

Fig. A.17. SEP-OCT 2010 community structure with Expert.

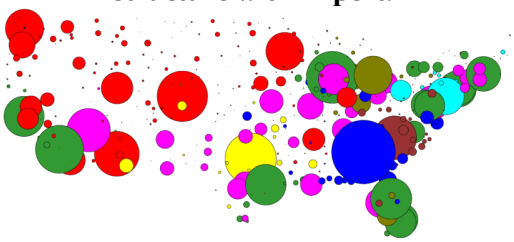

Fig. A.18. NOV-DEC 2010 community structure with Expert.

Appendix B
Children's Community Structure by Expert's Method

The following figures depict the community structure found in the children using Expert's method. In each figure, all community members are assigned the same colour.

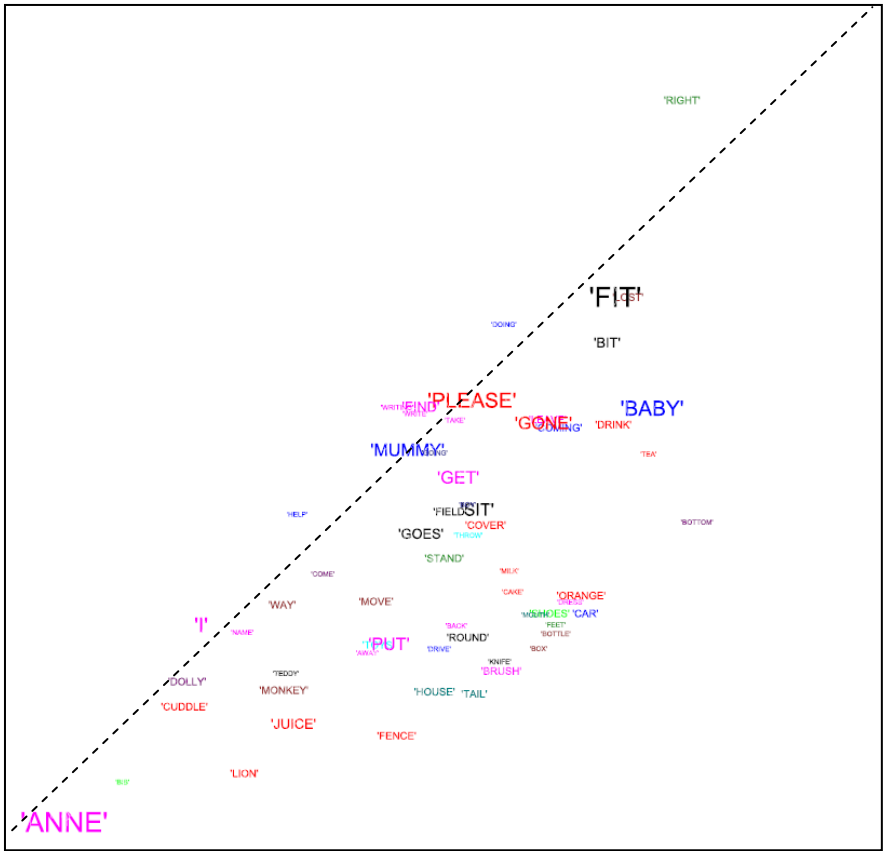

Fig. B.19. Ann 1 community structure with Expert.

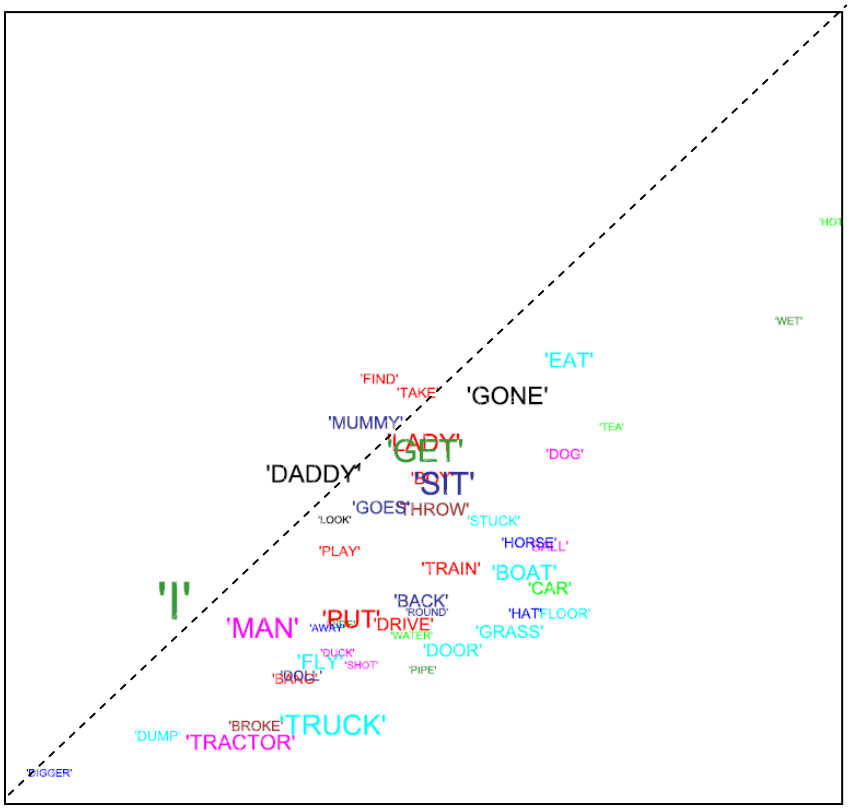

Fig. B.20. Ara 1 community structure with Expert.

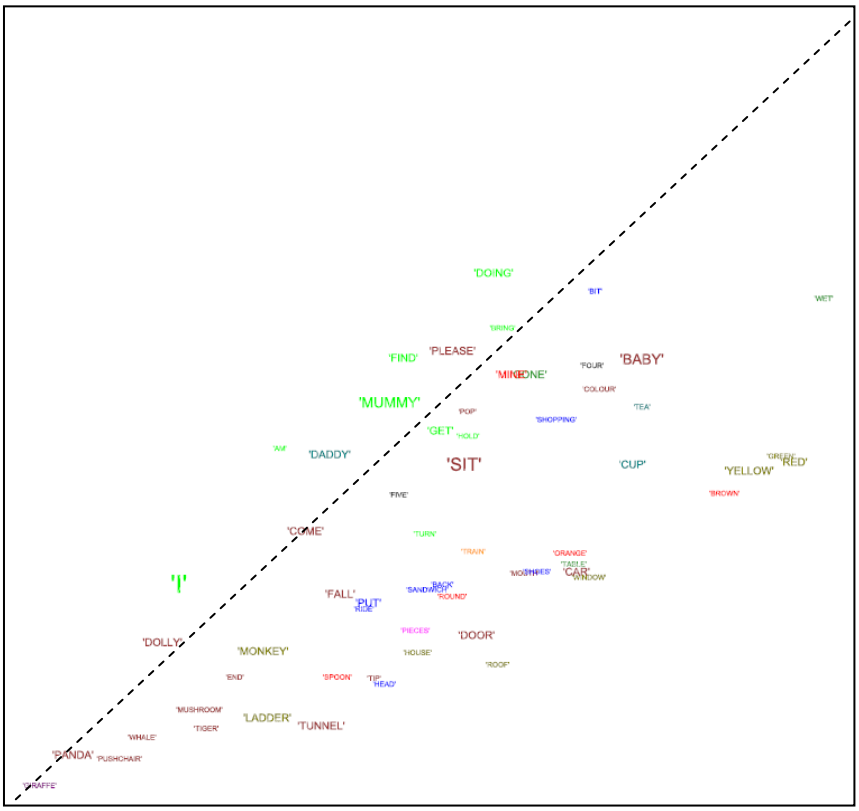

Fig. B.21. Bec 1 community structure with Expert.

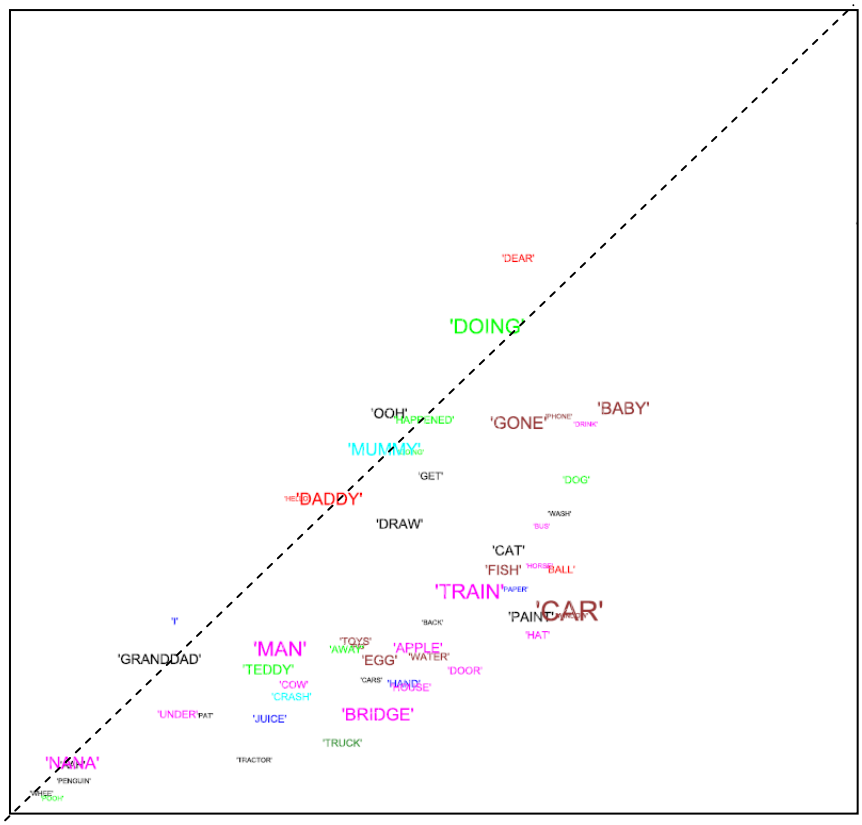

Fig. B.22. Car 1 community structure with Expert.

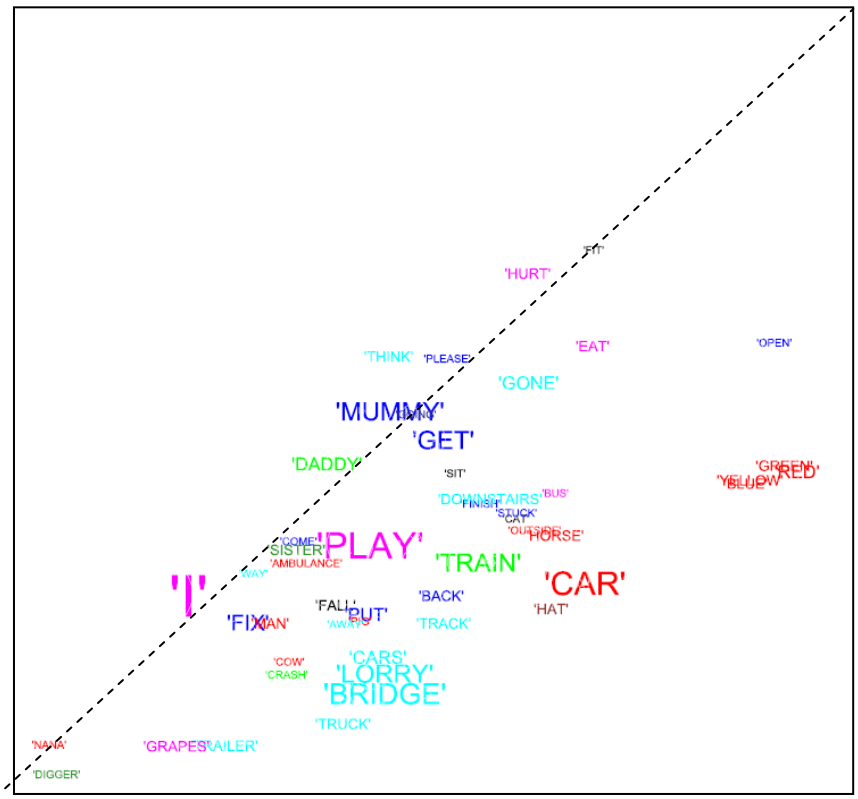

Fig. B.23. Dom 1 community structure with Expert.

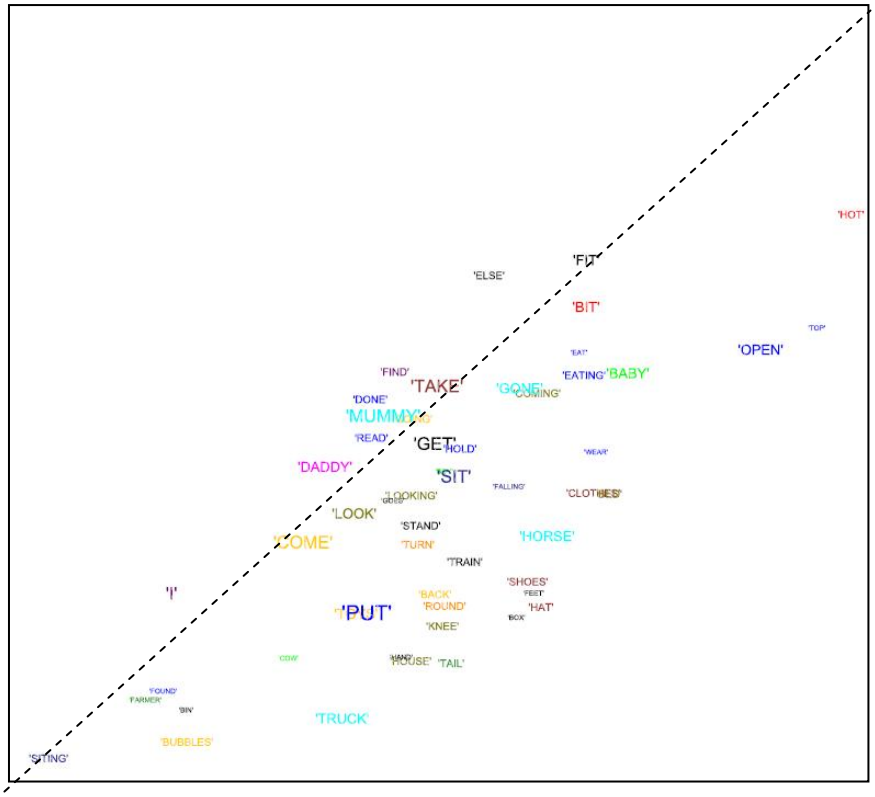

Fig. B.24. Gai 1 community structure with Expert.

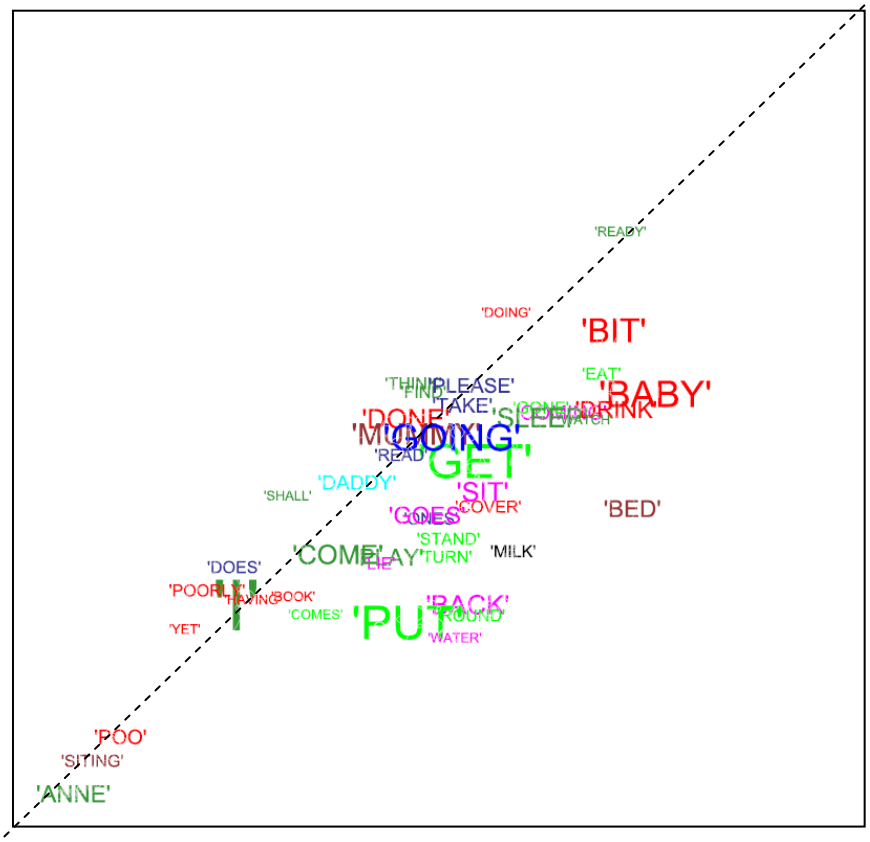

Fig. B.25. Ann 2 community structure with Expert.

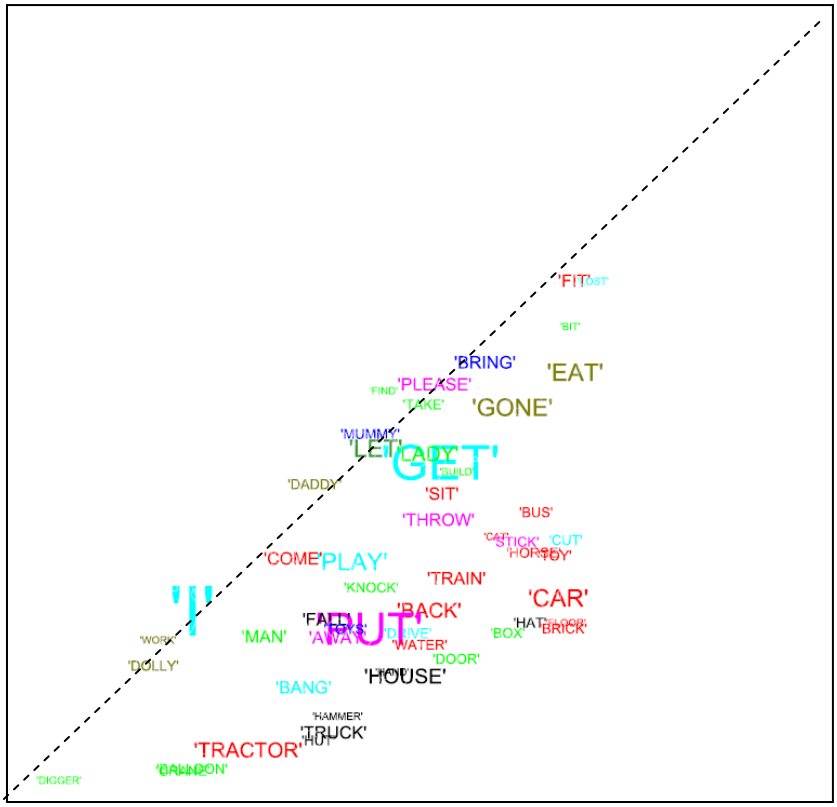

Fig. B.26. Ara 2 community structure with Expert.

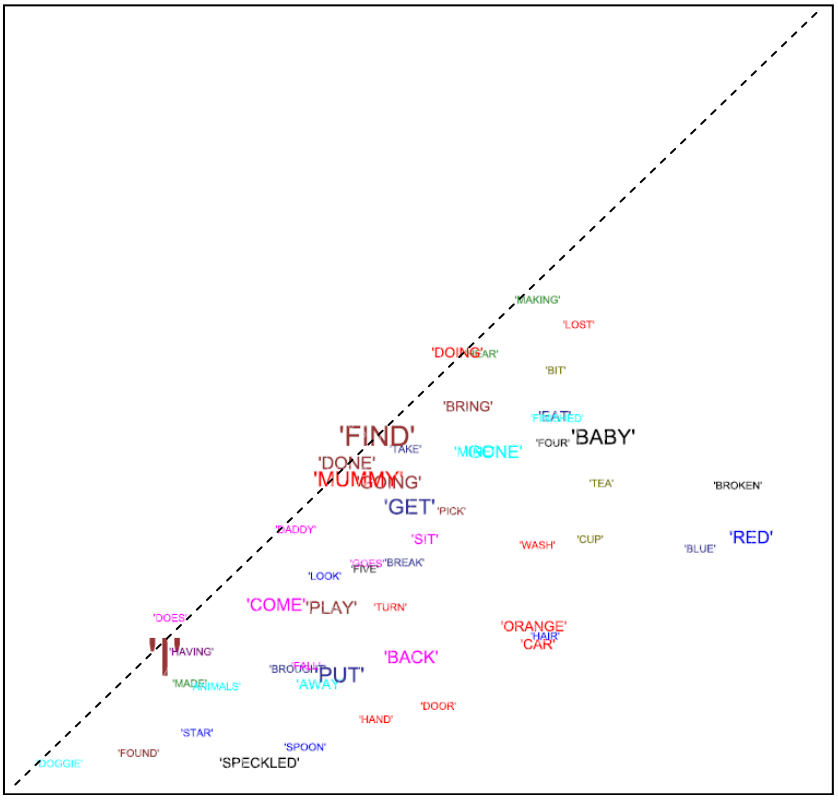

Fig. B.27. Bec 2 community structure with Expert.

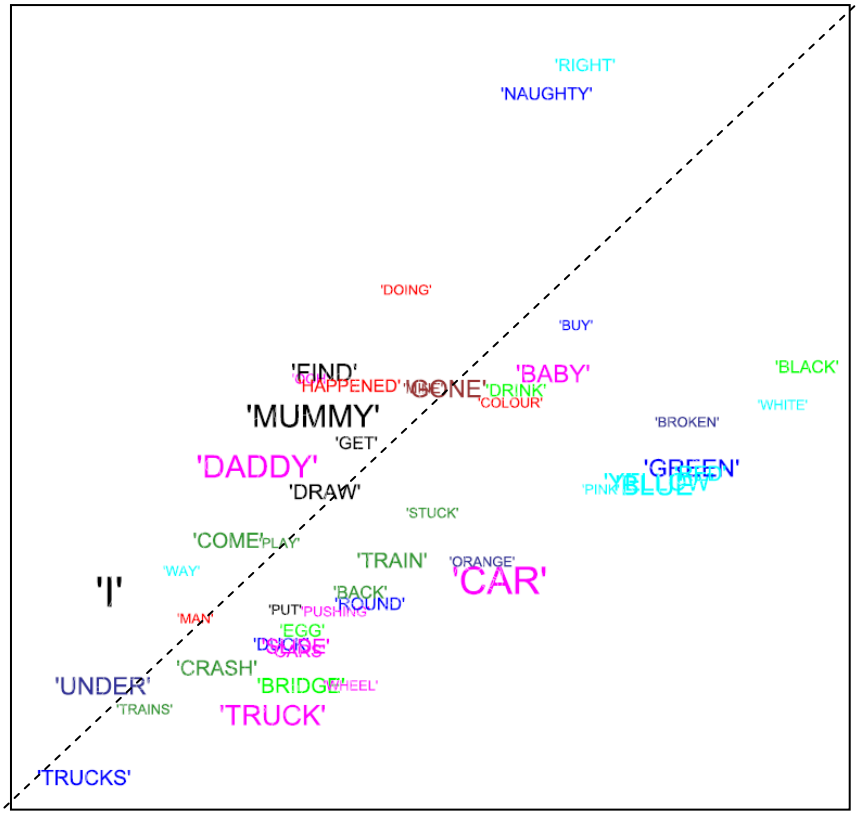

Fig. B.28. Car 2 community structure with Expert.

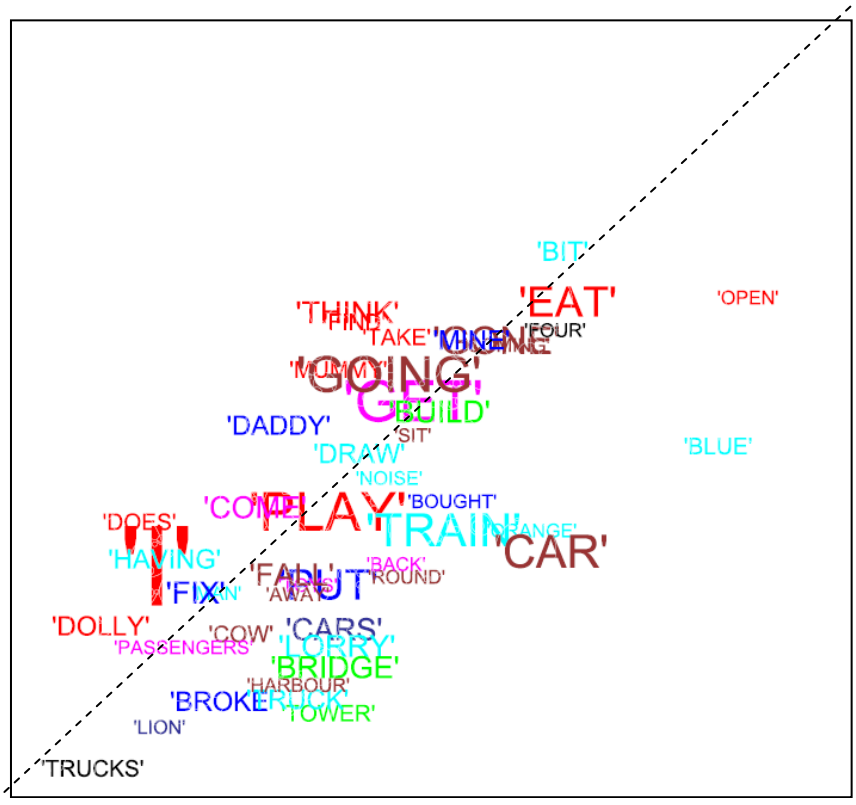

Fig. B.29. Dom 2 community structure with Expert.

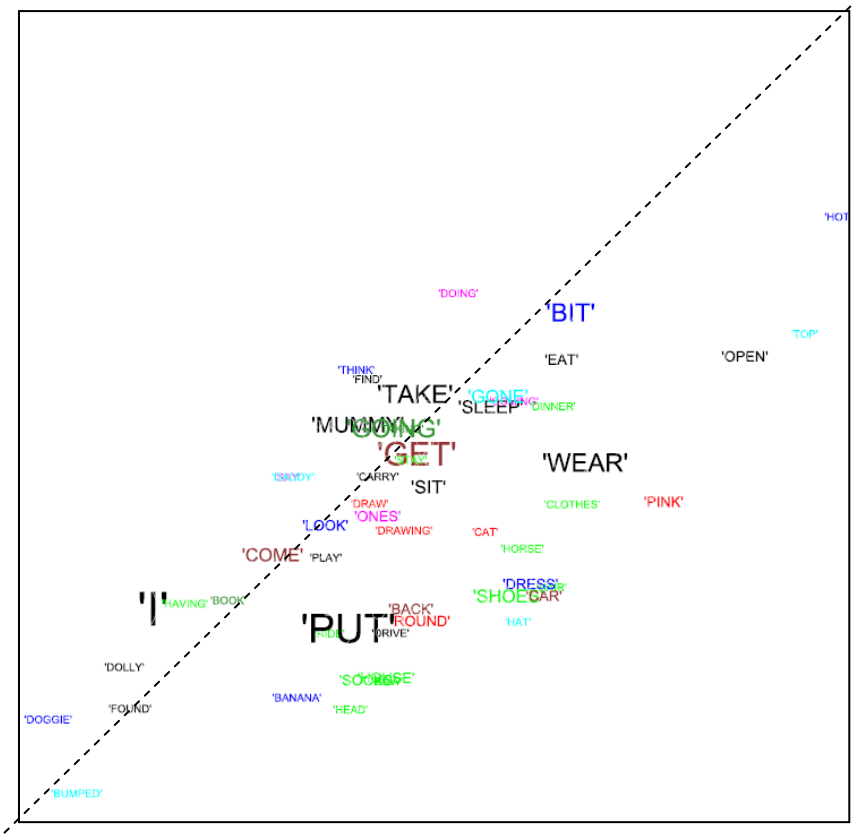

Fig. B.30. Gai 2 community structure with Expert.

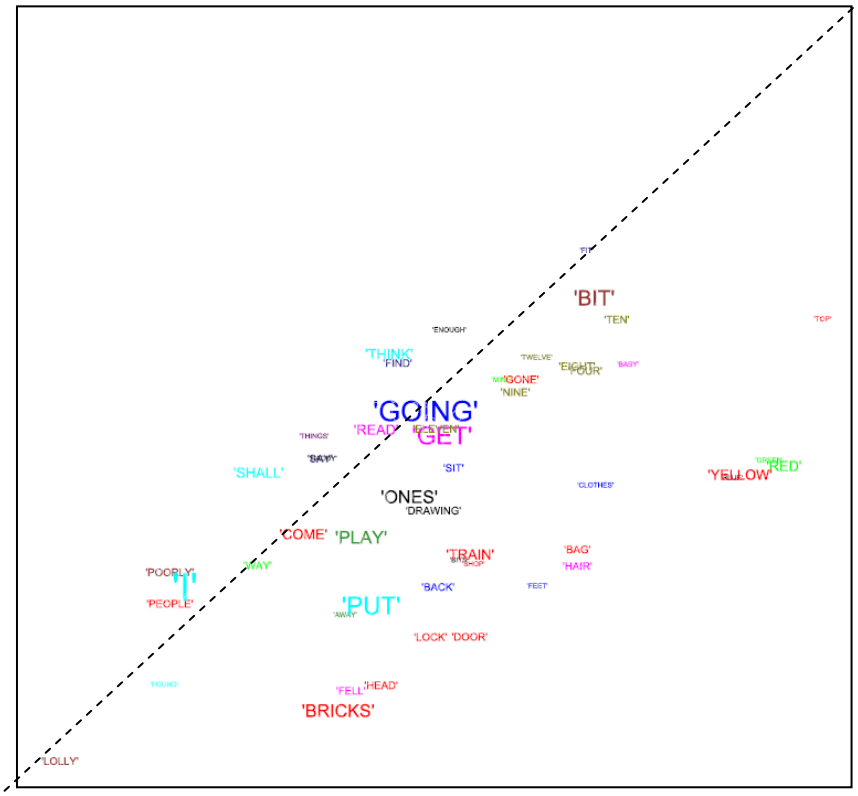

Fig. B.31. Ann 3 community structure with Expert.

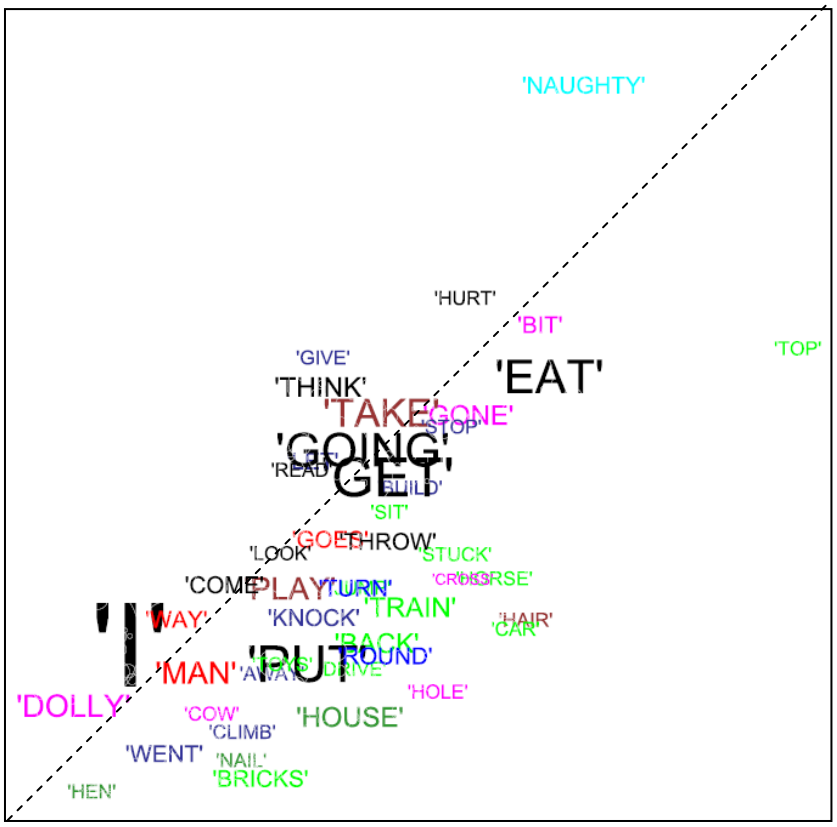

Fig. B.32. Ara 3 community structure with Expert.

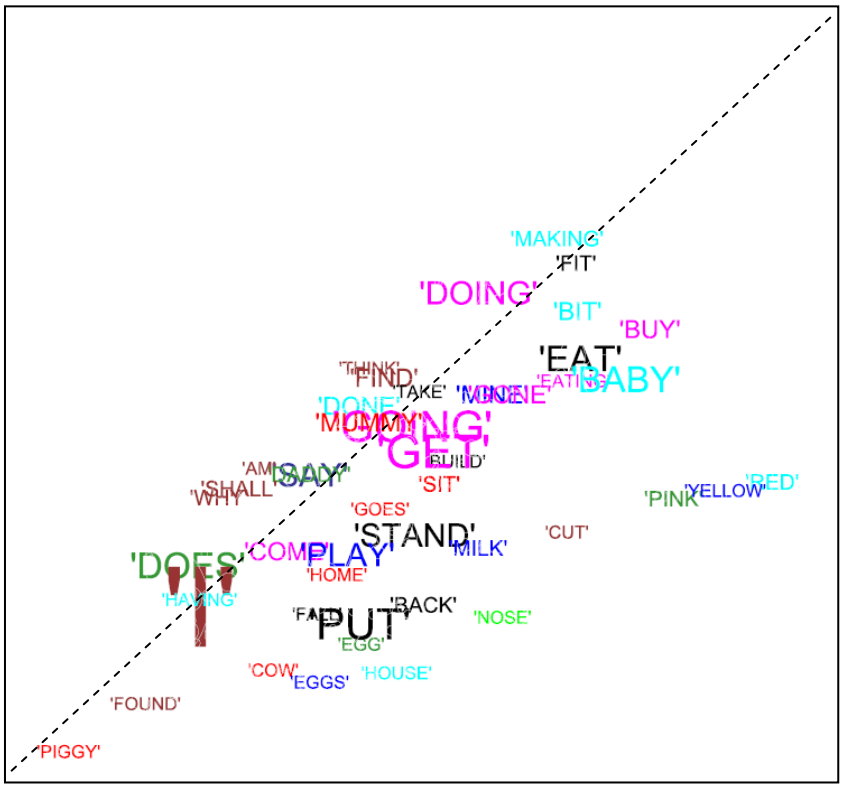

Fig. B.33. Bec 3 community structure with Expert.

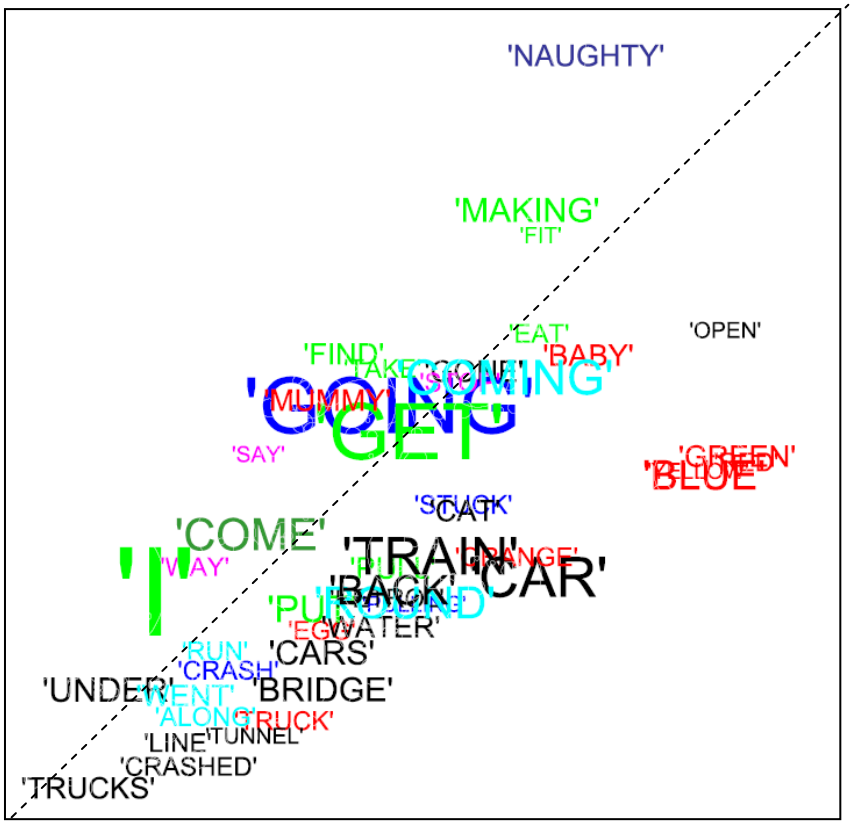

Fig. B.34. Car 3 community structure with Expert.

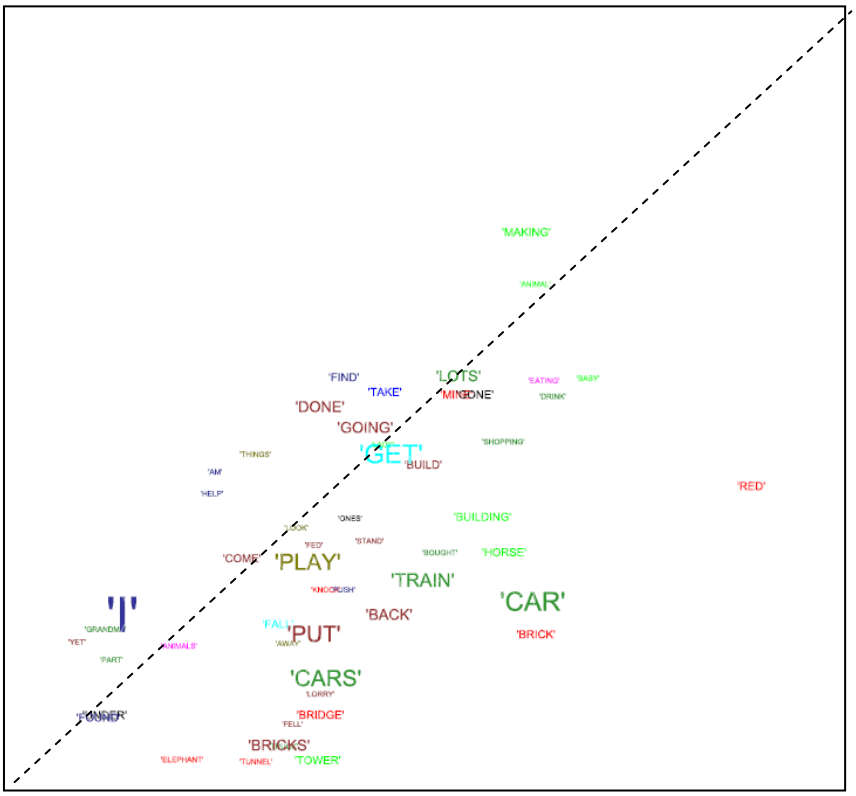

Fig. B.35. Dom 3 community structure with Expert.

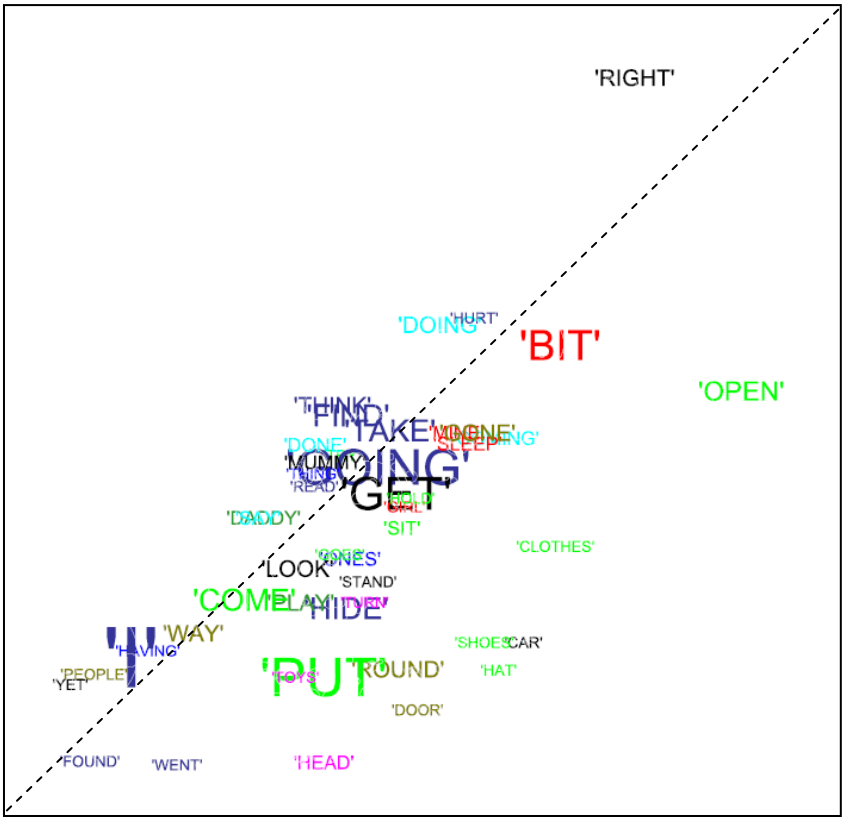

Fig. B.36. Gai 3 community structure with Expert.

Appendix C
USAN Community Structure by Newman's Method

The following figures depict the community structure found in the USAN using Newman's method. In each figure, all community members are assigned the same colour.

- **Year 1990**

Figs. C.37-C.42 depict bi-monthly snapshots of the USAN for the year 1990.

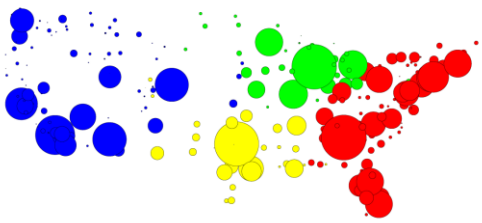

Fig. C.37. JAN-FEB 1990 community structure with Newman.

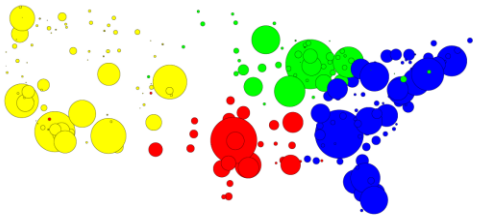

Fig. C.38. MAR-APR 1990 community structure with Newman.

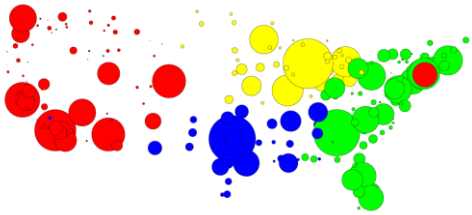

Fig. C.39. MAY-JUN 1990 community structure with Newman.

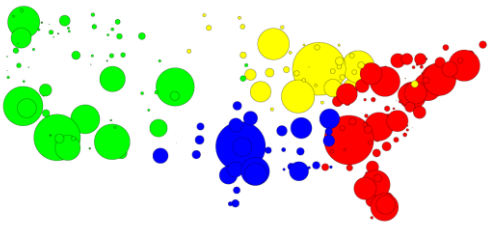

Fig. C.40. JUL-AUG 1990 community structure with Newman.

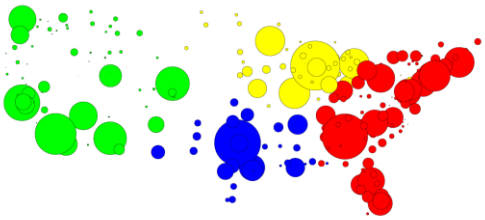

Fig. C.41. SEP-OCT 1990 community structure with Newman.

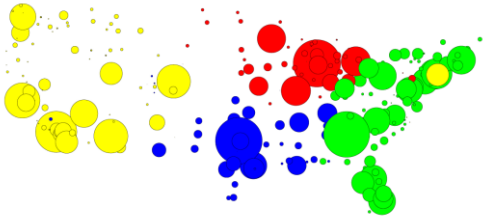

Fig. C.42. NOV-DEC 1990 community structure with Newman.

- **Year 2000**

Figs. C.43-C.48 depict bi-monthly snapshots of the USAN for the year 2000.

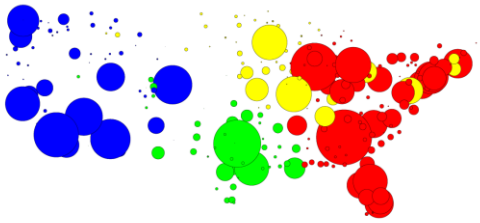

Fig. C.43. JAN-FEB 2000 community structure with Newman.

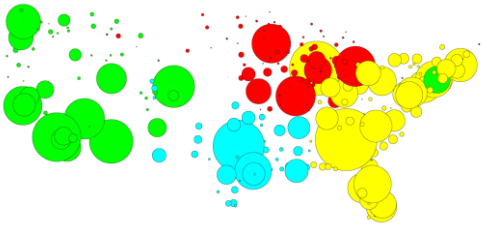

Fig. C.44. MAR-APR 2000 community structure with Newman.

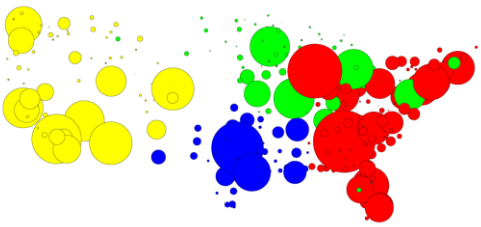

Fig. C.45. MAY-JUN 2000 community structure with Newman.

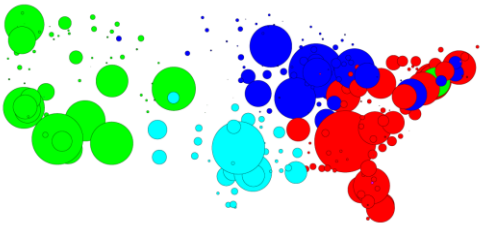

Fig. C.46. JUL-AUG 2000 community structure with Newman.

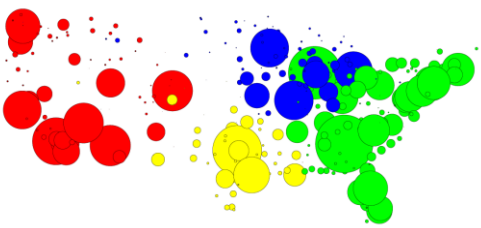

Fig. C.47. SEP-OCT 2000 community structure with Newman.

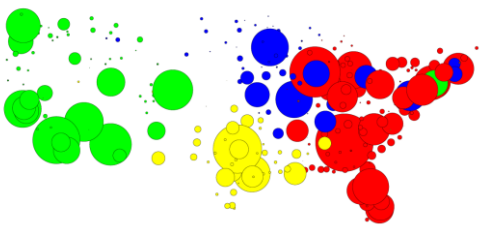

Fig. C.48. NOV-DEC 2000 community structure with Newman.

- **Year 2010**

Figs. C.49-C.54 depict bi-monthly snapshots of the USAN for the year 2010.

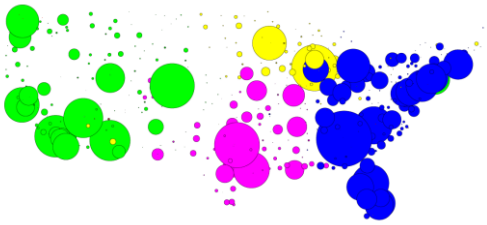

Fig. C.49. JAN-FEB 2010 community structure with Newman.

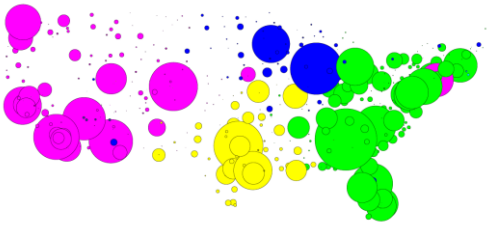

Fig. C.50. MAR-APR 2010 community structure with Newman.

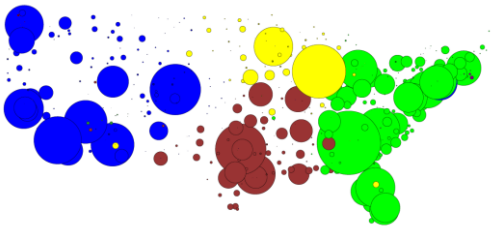

Fig. C.51. MAY-JUN 2010 community structure with Newman.

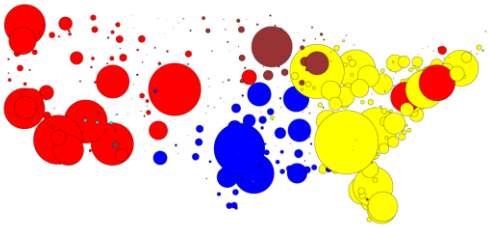

Fig. C.52. JUL-AUG 2010 community structure with Newman.

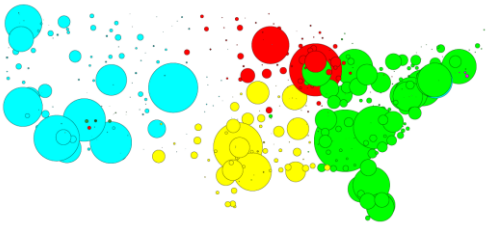

Fig. C.53. SEP-OCT 2010 community structure with Newman.

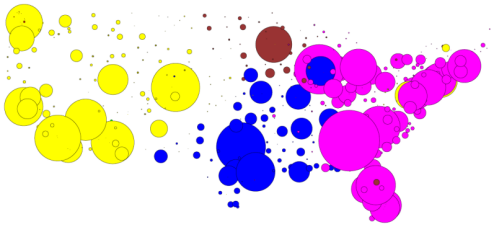

Fig. C.54. NOV-DEC 2010 community structure with Newman.

Appendix D
Children's Community Structure by Newman's Method

The following figures depict the community structure found in the children using Newman's method. In each figure, all community members are assigned the same colour.

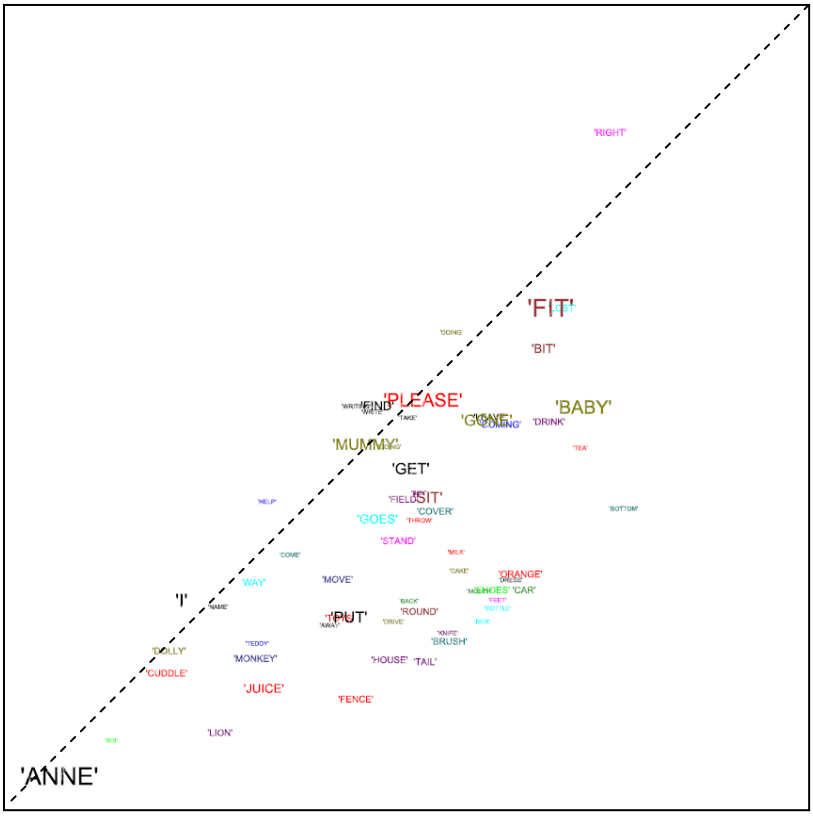

Fig. D.55. Ann 1 community structure with Newman.

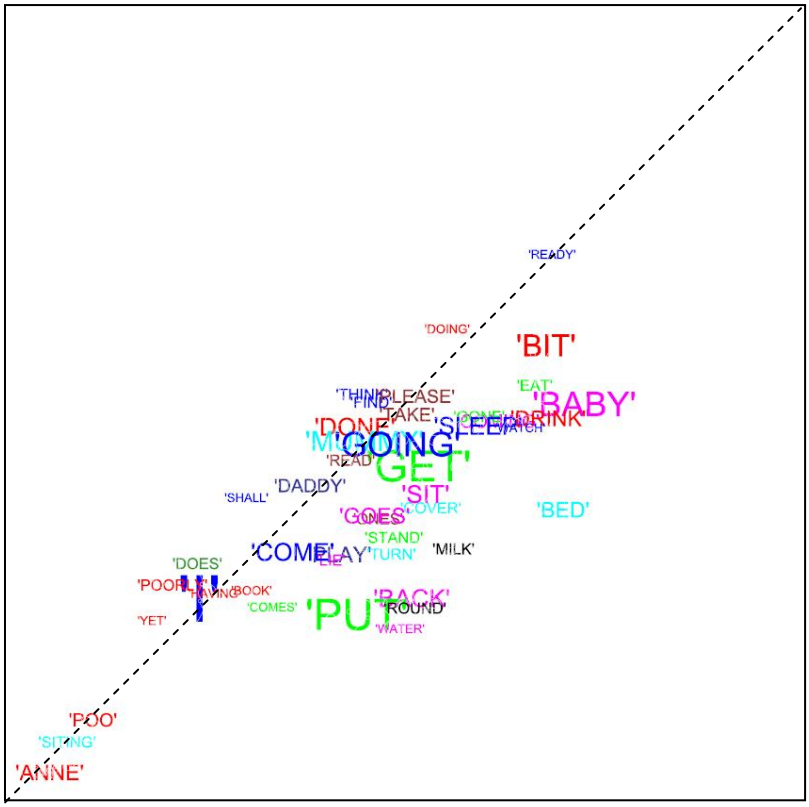

Fig. D.56. Ann 2 community structure with Newman.

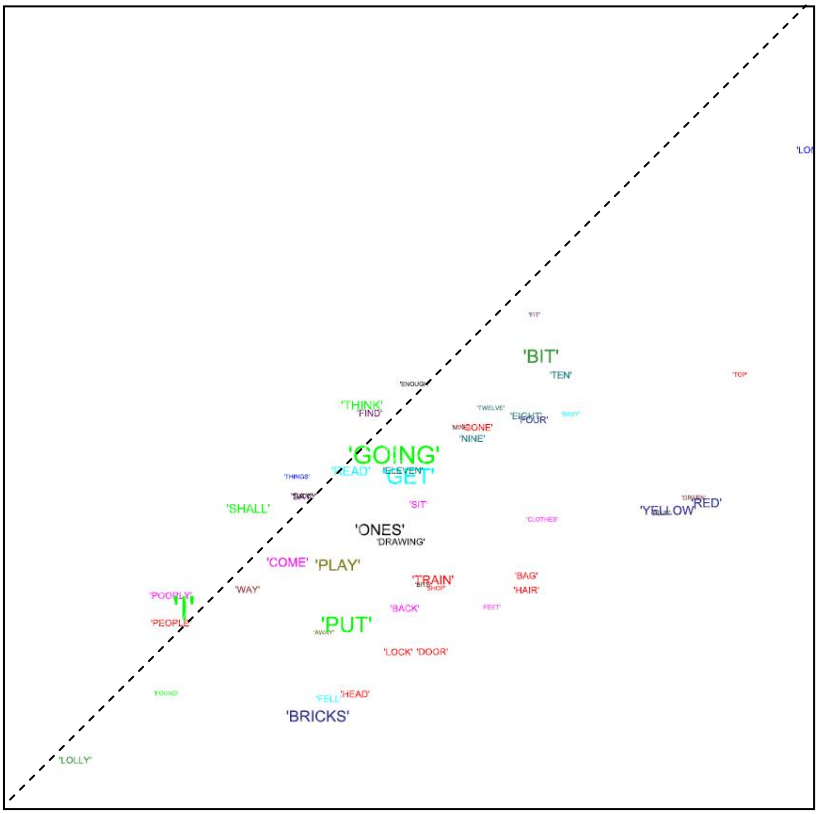

Fig. D.57. Ann 3 community structure with Newman.

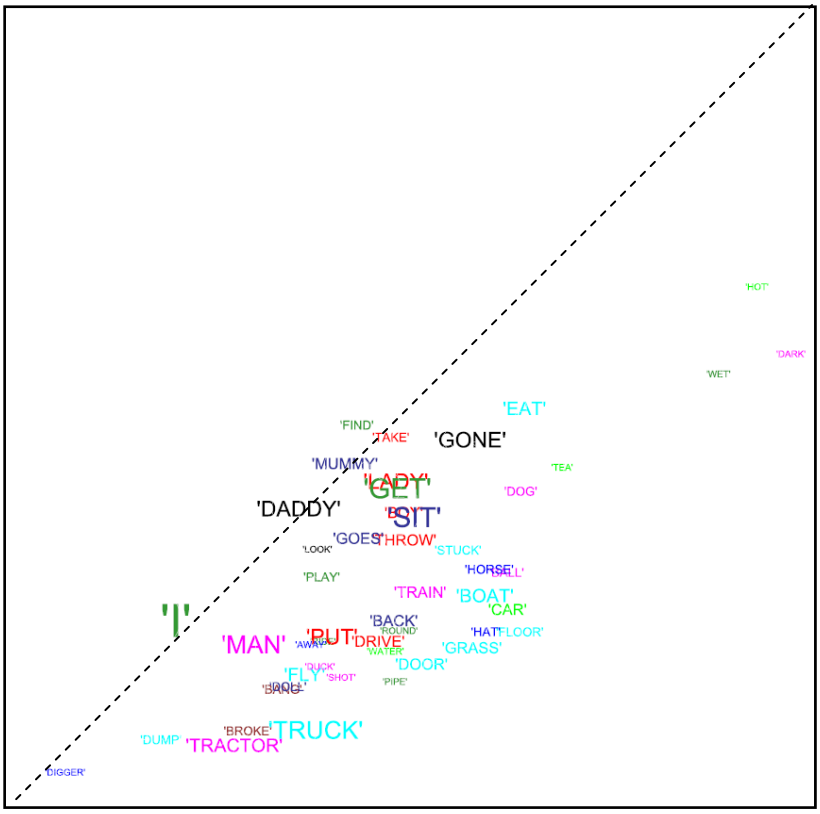

Fig. D.58. Ara 1 community structue with Newman.

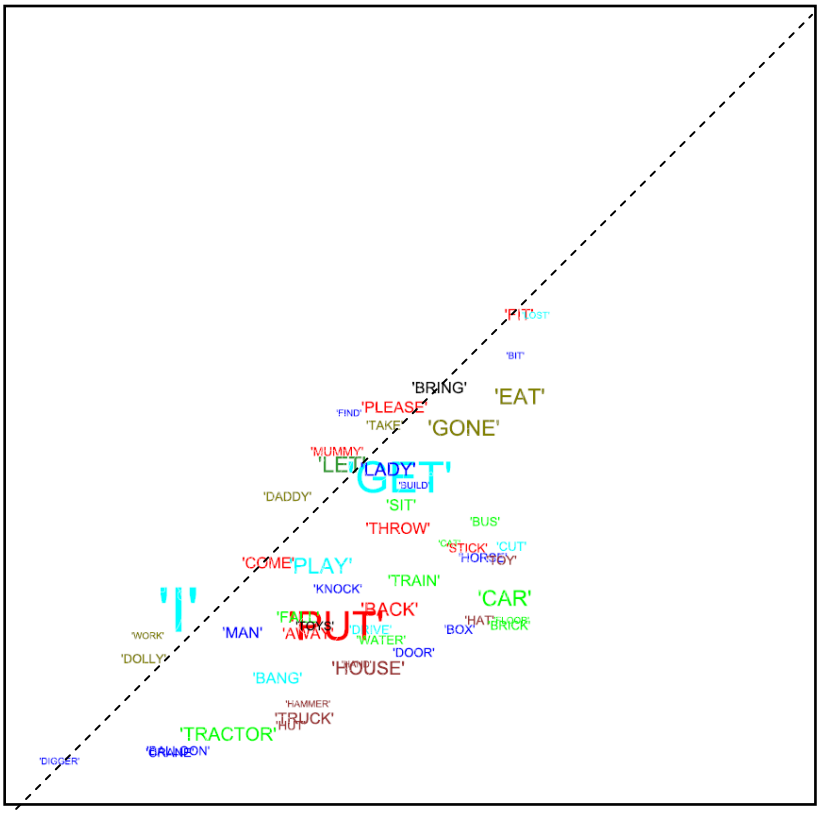

Fig. D.59. Ara 2 community structure with Newman.

Fig. D.60. Ara 3 community structure with Newman.

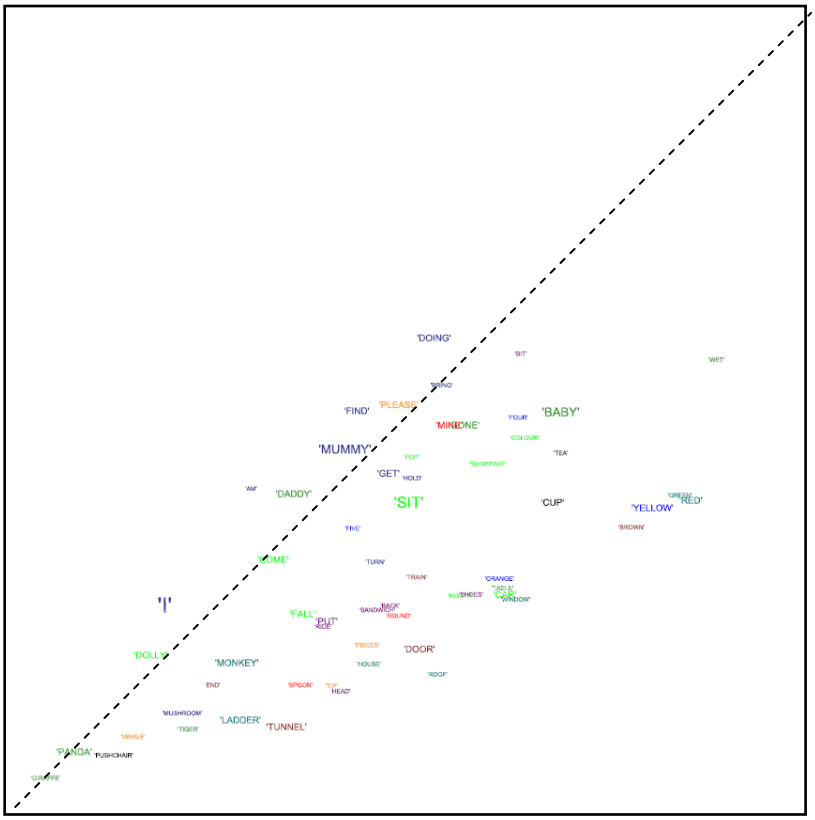

Fig. D.61. Bec 1 community structure with Newman.

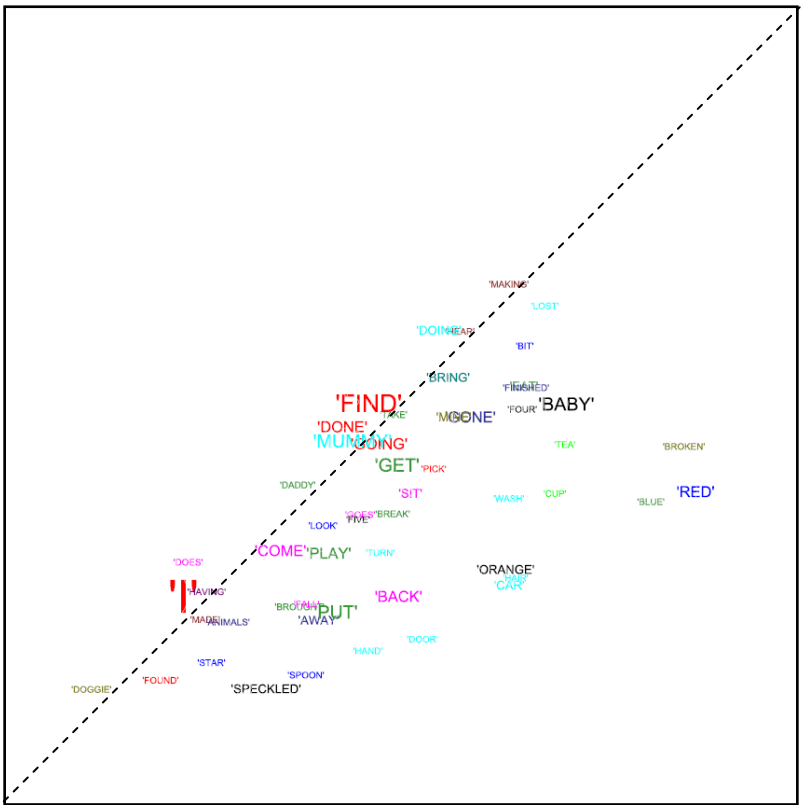

Fig. D.62. Bec 2 community structure with Newman.

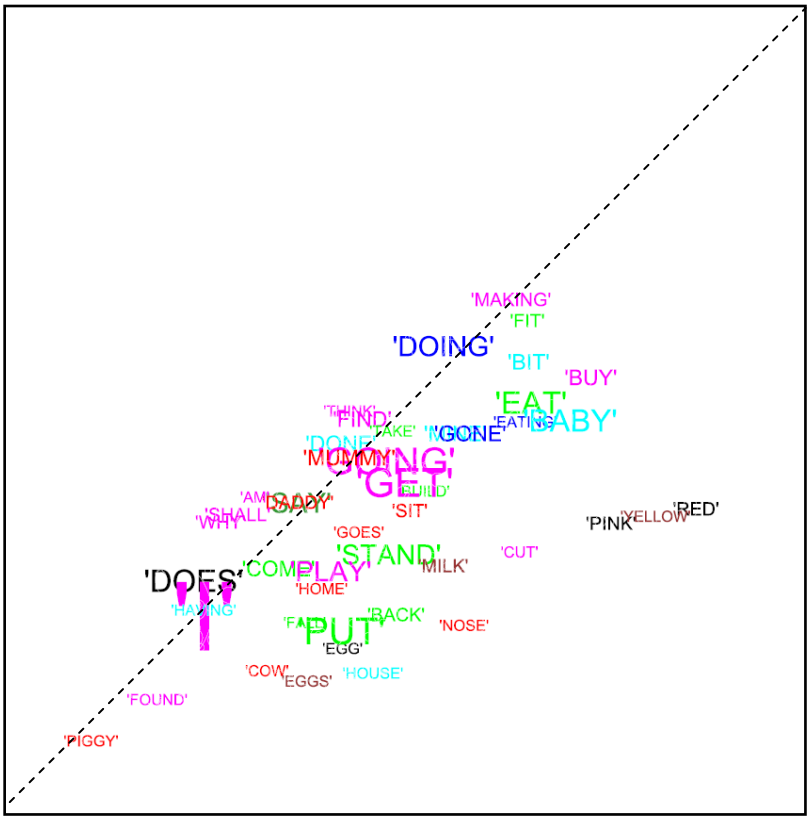

Fig. D.63. Bec 3 community structure with Newman.

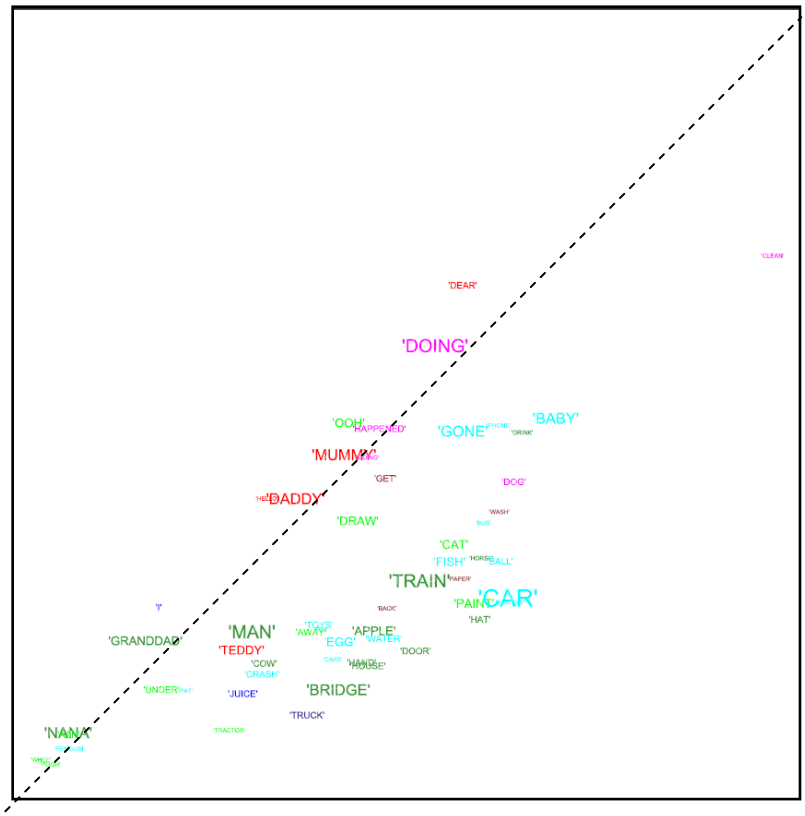

Fig. D.64. Car 1 community structure with Newman.

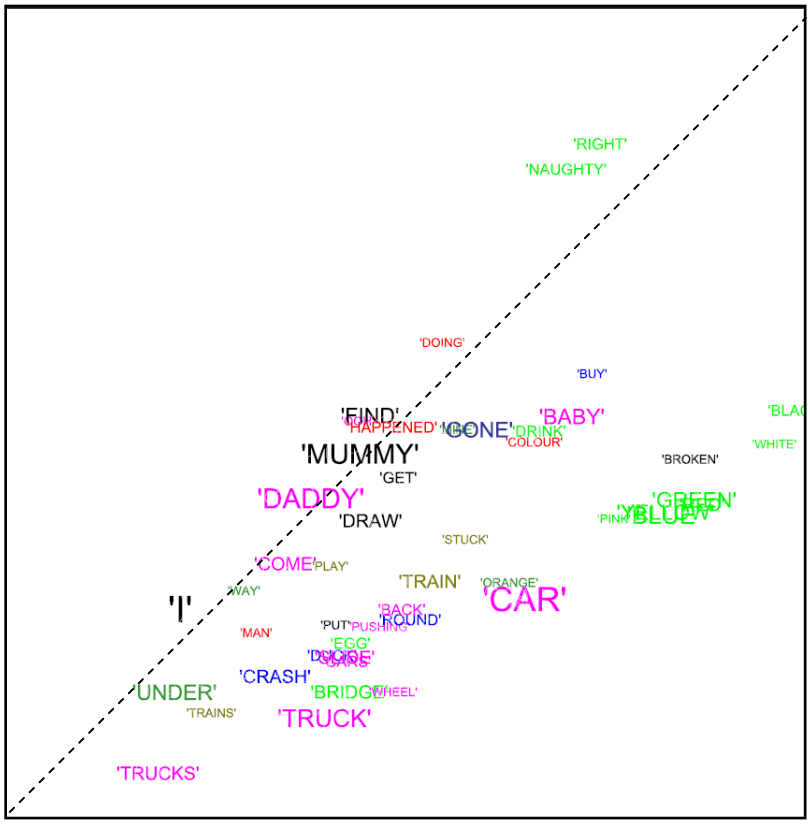

Fig. D.65. Car 2 community structure with Newman.

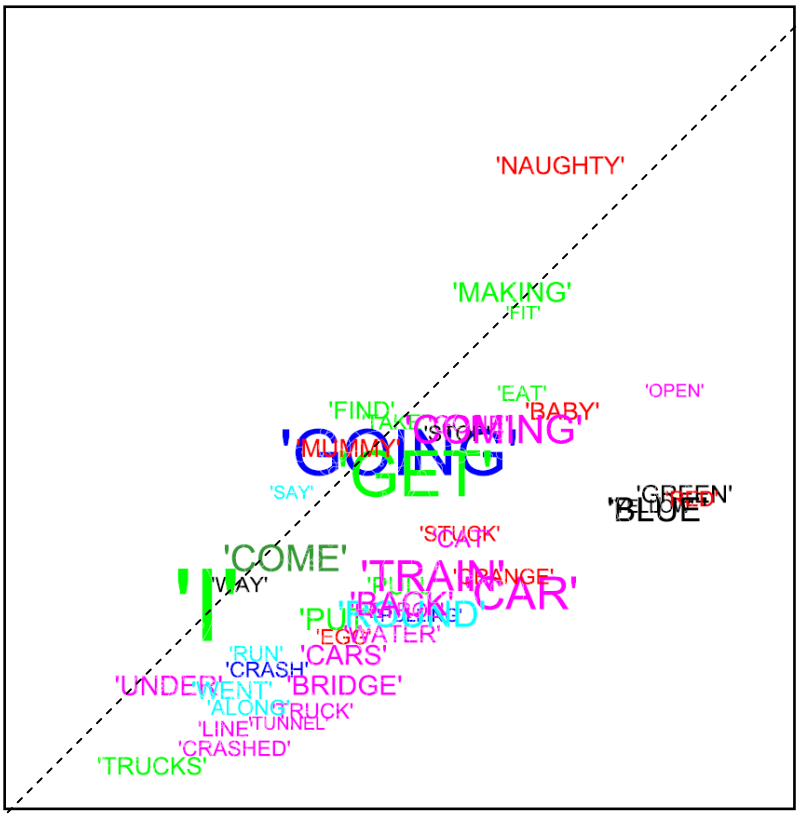

Fig. D.66. Car 3 community structure with Newman.

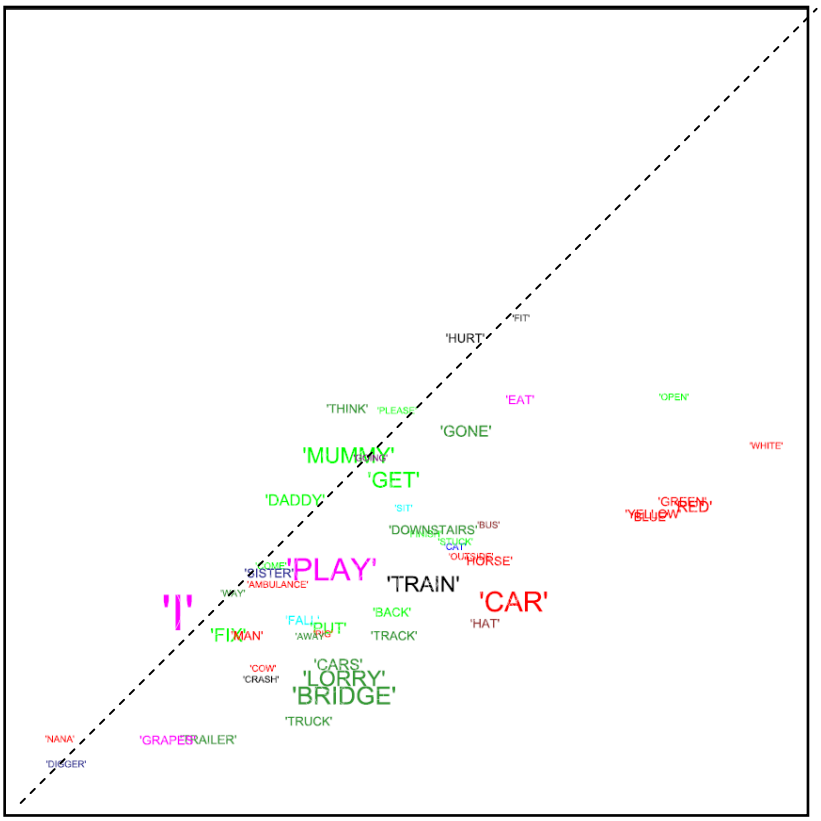

Fig. D.67. Dom 1 community structure with Newman.

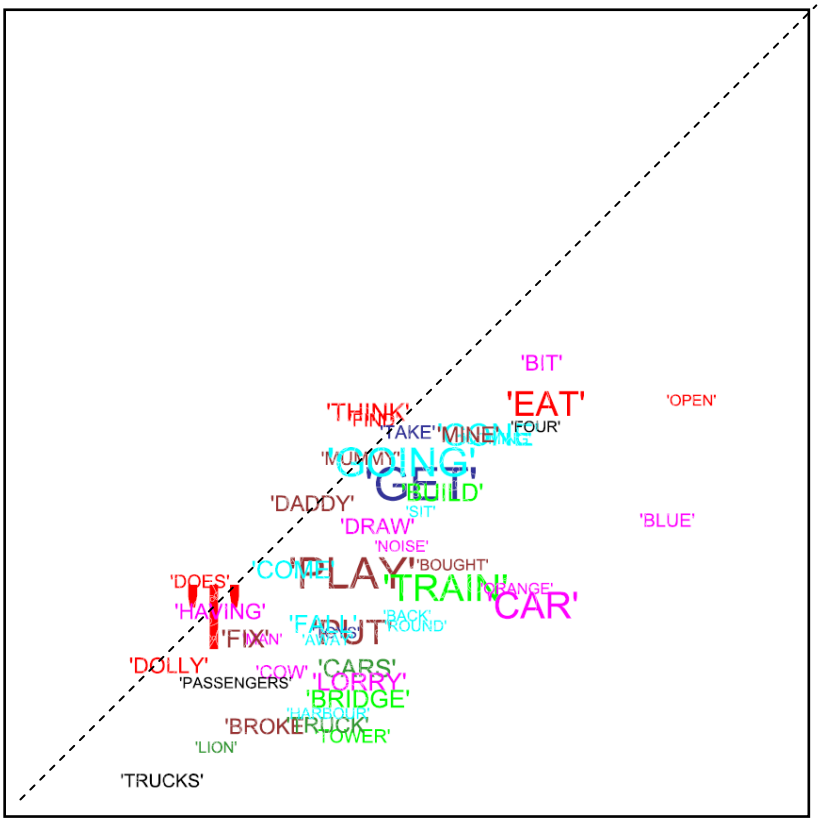

Fig. D.68. Dom 2 community structure with Newman.

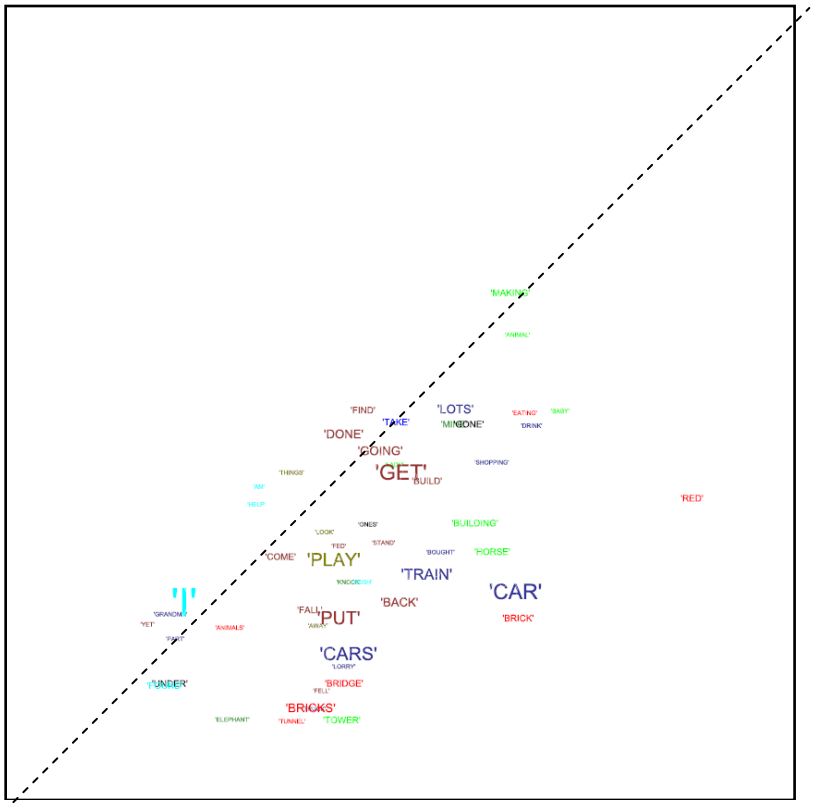

Fig. D.69. Dom 3 community structure with Newman.

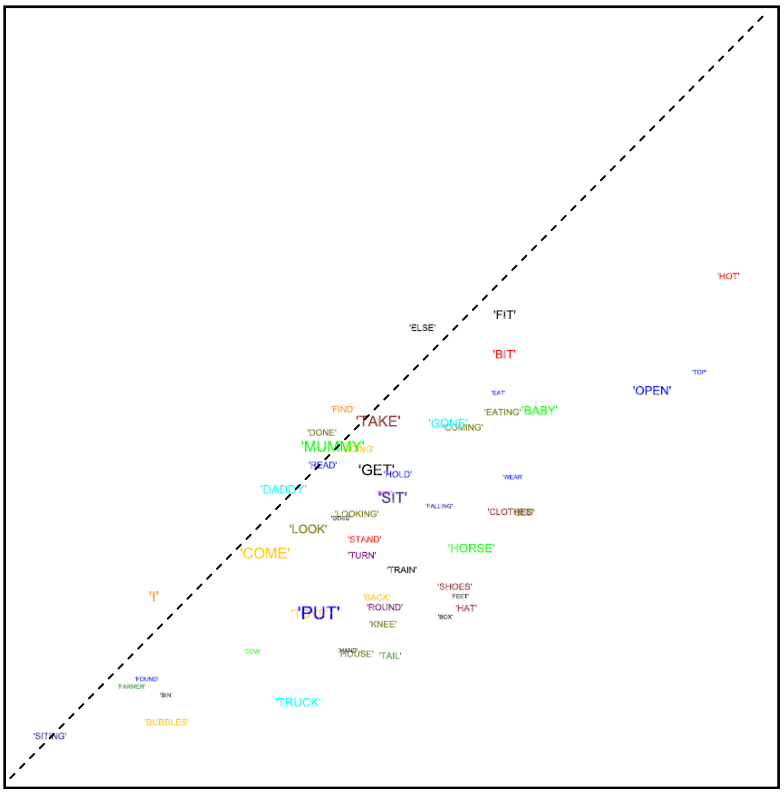

Fig. D.70. Gai 1 community structure with Newman.

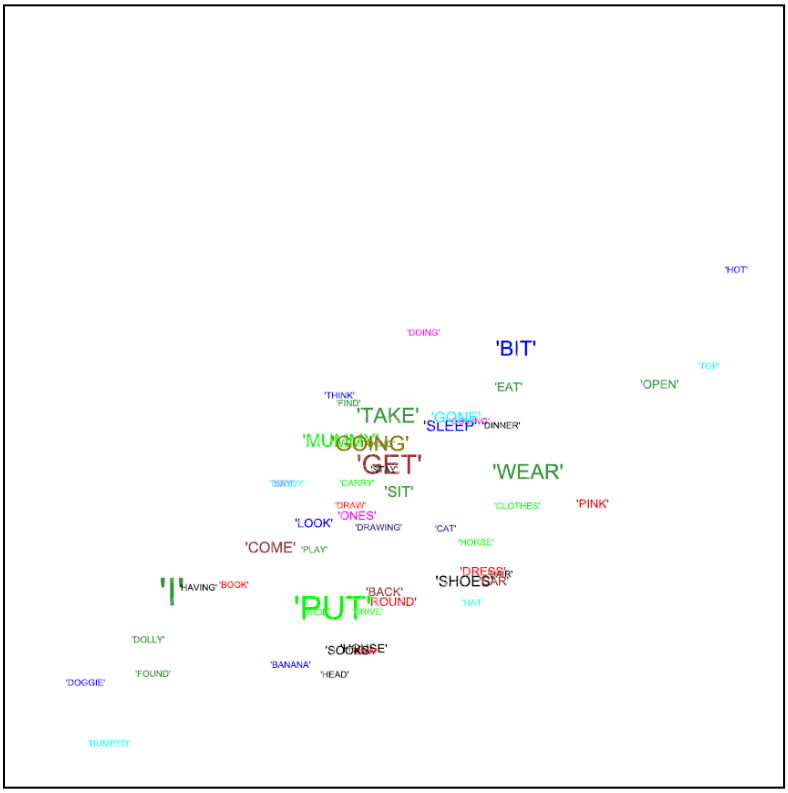

Fig. D.71. Gai 2 community structure with Newman.

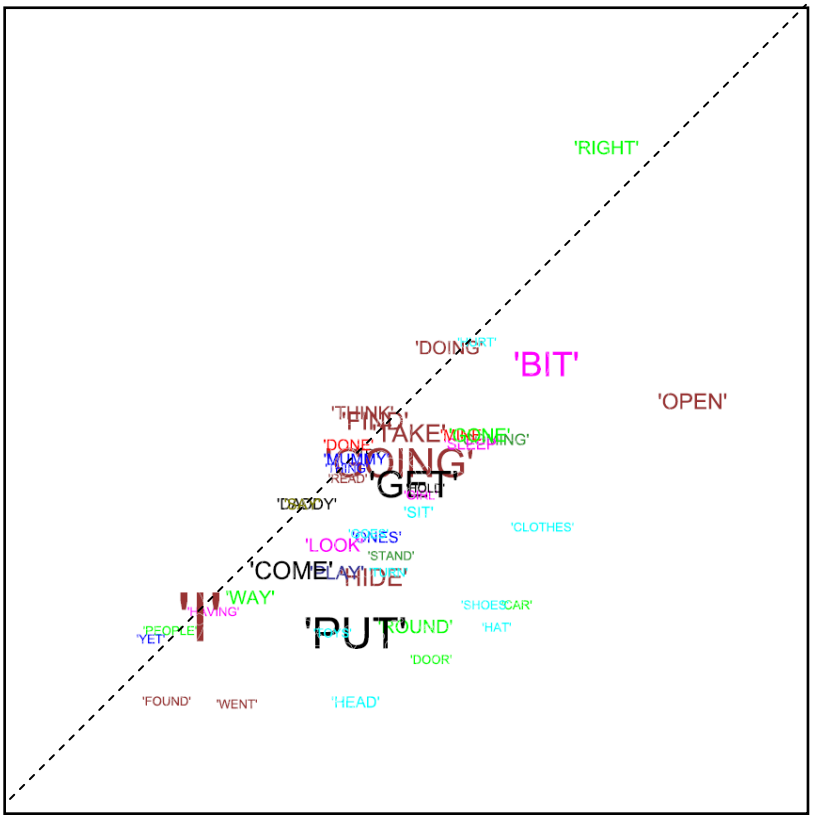

Fig. D.72. Gai 3 community structure with Newman.