SELF-ORGANIZATION OF SPEECH SOUND INVENTORIES

IN THE FRAMEWORK OF COMPLEX NETWORKS

Animesh Mukherjee

SELF-ORGANIZATION OF SPEECH SOUND INVENTORIES

IN THE FRAMEWORK OF COMPLEX NETWORKS

A dissertation submitted to the
Indian Institute of Technology, Kharagpur
in partial fulfillment of the requirements of the degree

of

Doctor of Philosophy

by

Animesh Mukherjee

Under the supervision of

Dr. Niloy Ganguly
and
Prof. Anupam Basu

Department of Computer Science and Engineering

Indian Institute of Technology, Kharagpur

December 2009

APPROVAL OF THE VIVA-VOCE BOARD

Certified that the thesis entitled ***"Self-Organization of Speech Sound Inventories in the Framework of Complex Networks"*** submitted by Animesh Mukherjee to the Indian Institute of Technology, Kharagpur, for the award of the degree of Doctor of Philosophy has been accepted by the external examiners and that the student has successfully defended the thesis in the viva-voce examination held today.

Members of the DSC

Supervisor

Supervisor

External Examiner

Chairman

Date:

CERTIFICATE

This is to certify that the thesis entitled ***"Self-Organization of Speech Sound Inventories in the Framework of Complex Networks"****, submitted by Animesh Mukherjee to the Indian Institute of Technology, Kharagpur, for the partial fulfillment of the award of the degree of Doctor of Philosophy, is a record of bona fide research work carried out by him under our supervision and guidance.*

The thesis in our opinion, is worthy of consideration for the award of the degree of Doctor of Philosophy in accordance with the regulations of the Institute. To the best of our knowledge, the results embodied in this thesis have not been submitted to any other University or Institute for the award of any other Degree or Diploma.

Niloy Ganguly	Anupam Basu
Associate Professor	Professor
CSE, IIT Kharagpur	CSE, IIT Kharagpur

Date:

DECLARATION

I certify that

(a) The work contained in the thesis is original and has been done by myself under the general supervision of my supervisors.

(b) The work has not been submitted to any other Institute for any degree or diploma.

(c) I have followed the guidelines provided by the Institute in writing the thesis.

(d) I have conformed to the norms and guidelines given in the Ethical Code of Conduct of the Institute.

(e) Whenever I have used materials (data, theoretical analysis, and text) from other sources, I have given due credit to them by citing them in the text of the thesis and giving their details in the references.

(f) Whenever I have quoted written materials from other sources, I have put them under quotation marks and given due credit to the sources by citing them and giving required details in the references.

Animesh Mukherjee

Date:

ACKNOWLEDGMENTS

For every thesis, as a "technical" story builds up over the years in the foreground, there is a "social" story that emerges at the background. While the rest of the thesis is meant to elucidate the technical story, this is the only place where one takes the liberty to narrate the social story. Here is how I view the social story of this thesis . . .

The M.Tech curriculum of the Department of Computer Science and Engineering, IIT Kharagpur requires that a candidate spends the final year of the course in research work for preparing the master's thesis. In order to meet this requirement as a registrant of this course, in 2004, I joined the Communication Empowerment Laboratory (CEL) run by Prof. Anupam Basu and primarily funded by the Media Lab Asia (MLA). It was during this time that I got introduced to the complexity of human languages and readily fell in love with this topic although my M.Tech thesis had almost nothing to do with it. The whole credit for churning this interest in me goes to one of the most charming personalities I have ever encountered in life – Dr. Monojit Choudhury whom I call "Monojit da" ("da" refers to elder brother in Bengali). This interest had turned into a passion by the time I graduated as an M.Tech in summer 2005.

I always wanted to be in research and due to the encouragements that I received from my M.Tech supervisor Prof. Basu and Monojit da, I decided to join the PhD programme of the department in the fall of 2005. The topic of the thesis was set to be – no not something associated with the complexity of human languages, but "cognitive models of human computer interaction"! Here is where another prime mover of this story enters – Prof. Niloy Ganguly who after our very first meeting agreed to supervise this work along with Prof. Basu. It was he who introduced us to the new and emerging field of complex network theory. Immediately, Monojit da and I realized that the problem of sound inventories which we had been discussing for quite some time could be nicely modeled in this framework. I decided to carry on this work as a "side business" apart from the work pertaining to the topic of the thesis. Soon I landed up to various interesting results in this side business that made me glad. Nevertheless, the fact that I was not able to do much on the thesis front used to keep me low most of the times. Therefore, I decided to share my excitement about the side business with Prof. Ganguly and further confessed to him that there was almost no progress pertaining to the thesis. It was in this meeting that by showing some initial results Monojit da and I tried to convince Prof. Ganguly to allow me to devote my full time in the side business for at least the next three months. He kindly agreed to our proposal and that marked the beginning of a journey meant to convert my passion about the intricacies of natural languages into something noteworthy, that is, this thesis!

There are a large number of people and organizations to be thanked for collectively making this journey comfortable and gratifying. First of all, I must acknowledge all those who worked with me on certain parts of this thesis – Fernando Peruani (parts of the analytical solution for the growth model of the bipartite network), Shamik Roy Chowdhury (experiments related to the community analysis of vowel inventories), Vibhu Ramani (sorting the consonant inventories according to their language families), Ashish Garg and Vaibhav Jalan (experiments related to the dynamics across the language families) and above all Monojit da for his innumerable comments and suggestions as and when I needed them. I am also thankful to all my co-authors with whom I had the opportunity to work on various interesting and intriguing problems (the list is quite long and it is difficult to enumerate).

I received a lot of encouragements as well as many valuable feedbacks while presenting parts of this work at various conferences, workshops and institutions. In this context, I would like to express my sincere thanks to Prof. John Nerbonne, Prof. T. Mark Ellison and Prof. Janet B. Pierrehumbert for all the highly informative discussions that I had with them at ACL 2007 (Prague). I am grateful to Prof. Juliette Blevins for giving me an opportunity to present my work at the Max Planck Institute for Evolutionary Anthropology (Leipzig) as well as for her invaluable suggestions. Let me also say "thank you" to Dr. Andreas Deutsch for inviting me at the Technical University of Dresden, Germany and, thereby, giving me a chance to interact with his team of researchers which later turned out to be extremely productive. Thanks to Prof. Ravi Kannan and the entire Multilingual Systems (MLS) Research Group for their co-operation, support and guidance during my internship at Microsoft Research India (MSRI).

I have been fortunate to come across many good friends, without whom life would be bleak. Special thanks to the members of CEL and CNeRG (Complex Networks Research Group), who apart from extending personal and professional helps from time to time, have reminded me that there are other important things in life than a PhD thesis (namely chats, movies, dinners, concerts and plays). I am also indebted, in particular, to Mousumi di and Tirthankar for painstakingly proof-reading parts of this thesis.

I express my gratitude to the staffs and the faculty members of the department. I would specially like to thank Prof. S. Sarkar, Prof. S. Ghose and Prof. P. Mitra for their suggestions about the work at various occasions. I am indebted to many organizations including MSRI, ACL, IARCS, ISI Foundation, Italy for their financial assistance towards either a monthly fellowship or a foreign travel or both. I must also thank the International Phonetic Association (Department of Theoretical and Applied Linguistics, School of English, Aristotle University of Thessaloniki, Greece) for allowing me to reproduce the International Phonetic Alphabet in Appendix D.

Finally, I must admit that a “thank you” is something too small for the role played by my parents and my supervisors in shaping this thesis. In their absence, this thesis would have been very different or perhaps would not have existed at all.

Animesh Mukherjee

Date:

ABSTRACT

The sound inventories of the world's languages show a considerable extent of symmetry. It has been postulated that this symmetry is a reflection of the human physiological, cognitive and societal factors. There have been a large number of linguistically motivated studies in order to explain the self-organization of these inventories that arguably leads to the emergence of this symmetry. A few computational models in order to explain especially the structure of the smaller vowel inventories have also been proposed in the literature. However, there is a need for a single unified computational framework for studying the self-organization of the vowel as well as other inventories of complex utterances like consonants and syllables.

In this thesis, we reformulate this problem in the light of statistical mechanics and present complex network representations of these inventories. The central objective of the thesis is to study and explain the self-organization and emergence of the consonant inventories. Nevertheless, in order to demonstrate the versatility of our modeling methodology, we further apply it to investigate and detect certain interesting properties of the vowel inventories.

Two types of networks are considered - a language-consonant bipartite network and a consonant-consonant co-occurrence network. The networks are constructed from the UCLA Phonological Segment Inventory Database (UPSID). From the systematic analysis of these networks we find that the occurrence and co-occurrence of the consonants over languages follow a well-behaved probability distribution. The co-occurrence network also exhibits a high clustering coefficient. We propose different synthetic models of network growth based on preferential attachment so as to successively match with higher accuracy the different statistical properties of the networks. Furthermore, in order to have a deeper understanding of the growth dynamics we analytically solve the models to derive expressions for the emergent degree distribution and clustering coefficient. The co-occurrence network also exhibits strong community structures and a careful inspection indicates that the driving force behind the community formation is grounded in the human articulatory and perceptual factors. In order to quantitatively validate the above principle, we introduce an information theoretic definition of this factor – feature entropy – and show that the natural language inventories are significantly different in terms of this quantity from the randomly generated ones. We further construct similar networks for the vowel inventories and study various interesting similarities as well as differences between them and the consonant inventories.

To summarize, this thesis shows that complex networks can be suitably used to study the self-organization of the human speech sound inventories. In this light, we deem this computational framework as a highly powerful tool in future for modeling and explaining the emergence of many other complex linguistic phenomena.

Keywords: consonants, vowels, distinctive features, occurrence network, co-occurrence network, self-organization, emergence, preferential attachment, community structure, feature entropy.

Contents

List of Figures

List of Tables

List of Symbols

V_L	Set of nodes in the language partition of PlaNet
V_C	Set of nodes in the consonant partition of PlaNet
E_{pl}	Set of edges that run between V_L and V_C
k	Degree of a node
p_k	Fraction of nodes having degree equal to k
P_k	Fraction of nodes having degree greater than or equal to k
γ	Tunable randomness component of the synthesis model
t	Time step
$p_{k,t}$	Fraction of nodes having degree equal to k at time t
μ	Average degree of the nodes in V_L
N	Number of nodes in V_C
$\delta_{k,0}$	Kronecker delta function
E_{ph}	Set of edges in PhoNet
w_{ij}	Weight of the edge (i, j) in PhoNet
c_{av}	Average clustering coefficient of PhoNet
q	Degree of a node in the one-mode projection
p_u(q)	Degree distribution of the nodes in one-mode projection
f_d	Sampling distribution
$F_k(q)$	Probability that a node of degree k in the bipartite network ends up as a node of degree q in the one-mode projection
$f(x)$	Generating function for the distribution of node degrees in V_L
$p(x)$	Generating function for degree distribution of the V_C nodes in the bipartite network
$g(x)$	Generating function for degree distribution $p_u(q)$ of the one-mode projection
λ	Power-law exponent
$g_{app}(x)$	Generating function for the approximate expression of $p_u(q)$
p_t	Triad formation probability

S	Strength of an edge
η	Threshold for the MRad algorithm
O_L	Occurrence ratio
O_{av}	Average occurrence ratio
DC	Discriminative capacity
F_E	Feature entropy
RR	Redundancy ratio
MSE	Mean square error
V_V	Set of nodes in the vowel partition of VlaNet
E_{vl}	Set of edges that run between V_L and V_V

Chapter 1

Introduction

If you are reading the thesis for the first time, then every new sentence that you come across tends to make you feel increasingly curious about – “what follows?”. Undoubtedly, curiosity about the world as well as ourselves is one of the most important traits of the human race. It is perhaps in order to relinquish this curiosity about each other that thousands of years ago our forefathers invented an extremely sophisticated medium of communication – *language*. Linguistic ability not only makes us different from the rest of the animal kingdom but is also central to the sense of identity that unites nations, cultures and ethnic groups. The same curiosity about ourselves gives birth to numerous important as well as interesting questions about this ability itself. Some of the most intriguing ones among these are “How did language evolve?”, “Why do languages change over time?”, “What are the universal characteristics of the thousands of mutually unintelligible languages that co-exist at a given time?” and “How does a child almost effortlessly acquire all the intricacies of language in the very early stages of development?” Various disciplines have joined in the search for the answers to the above questions in a collaborative and systematic approach. As a consequence of this collective effort, one argument that has gained enormous popularity in recent times is that language is a *complex adaptive system*, which has evolved through the process of self-organization in order to serve the purpose of human communication needs [147]. In fact, the main premise of *synergetic linguistics* [84,85,86] is that language is a self-organizing and self-regulating system and its (a) existence,

(b) properties, and (c) change can be successfully explained within this framework. The symmetries observed across languages are therefore, primarily an outcome of the dynamic interdependence of the structure and the functions of a language [86]. The aforementioned efforts have repeatedly pointed to the fact that the emergent complexity of a linguistic phenomenon can be understood by treating language as a physical system.

Like any physical system, a linguistic system (i.e., language) can be viewed from three different perspectives [11]. At one extreme, it is a collection of *utterances* that are produced and perceived by the speakers of a linguistic community during the process of communication with the other speakers of that community. This is analogous to the *microscopic* view of a thermodynamic system, where every utterance and its corresponding context together render the identity of the language, that is, its grammar. At the other extreme, a language can be described by a lexicon and a set of grammar rules. This is equivalent to the *macroscopic* view of a thermodynamic system. Sandwiched between these two extremes, one can also conceive of a *mesoscopic* view of language, where the different linguistic entities such as phonemes, syllables, words or phrases form the basic units of the system and the grammar is an emergent property resulting from the complex interactions among these units. Figure 1.1 presents a hypothetical illustration of these three levels of a linguistic system.

In the recent years, *complex networks* have proved to be an extremely suitable framework for modeling the structure and dynamics of various large-scale systems primarily at the level of mesoscopy. Examples of well-studied naturally-occurring networks include biological, ecological and social networks such as metabolic networks [77, 131], gene regulatory networks [6], protein interaction networks [21, 76], signalling networks [106], epidemic networks [121, 122], food webs [48, 164], scientific collaboration networks [113], movie-actor networks [130] and acquaintance/friendship networks [2, 112]. Similarly, there have also been a lot of studies on man-made networks, which mainly include communication networks and transportation infrastructures such as the Internet [52], WWW [3, 9], p2p networks [87, 4, 109, 110, 111], power grid [10], airlines [68] and railway networks [137]. An extensive survey of the different theoretical and empirical developments of the field have been presented in [8, 41, 42, 114].

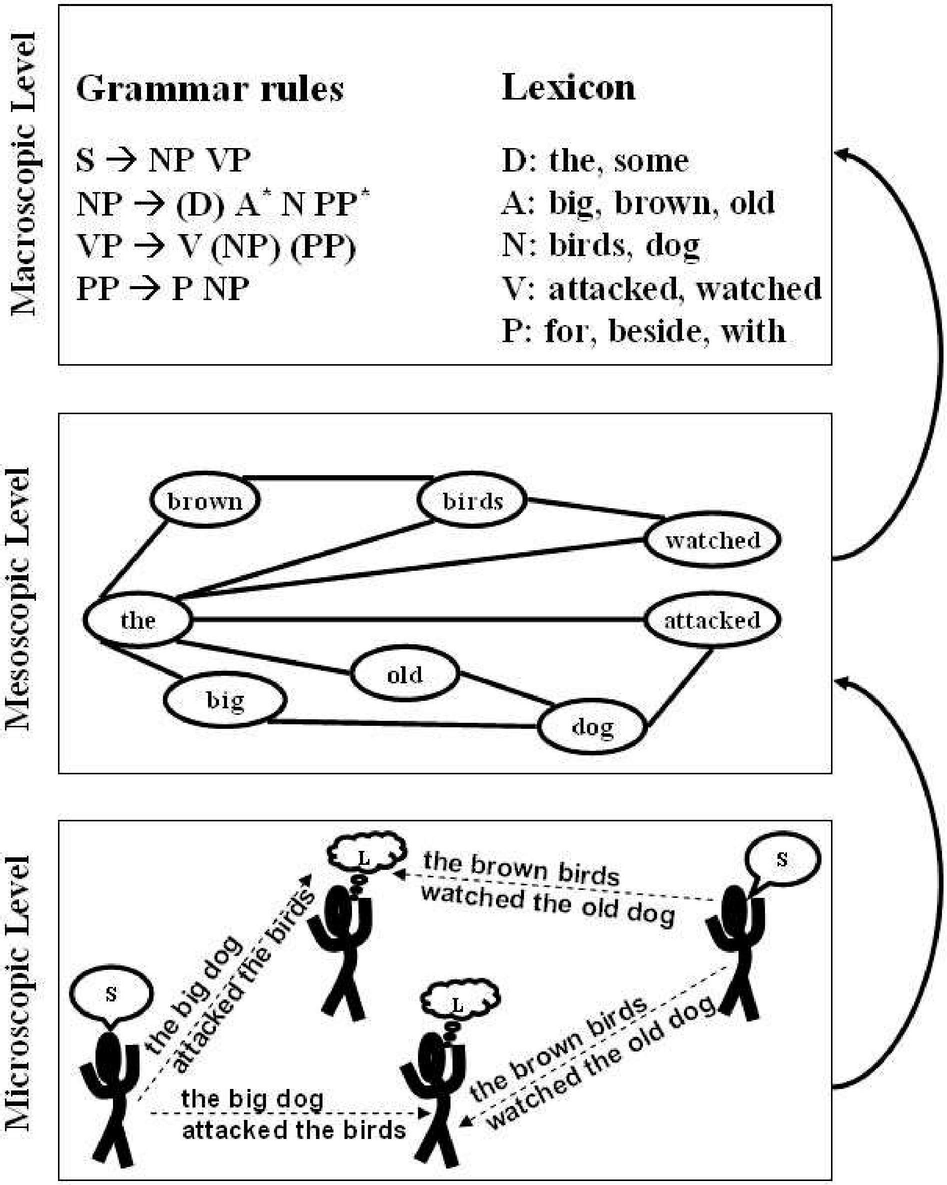


Figure 1.1: A hypothetical illustration of the three levels of a linguistic system. In the microscopic level, two example sentences, "the brown birds watched the old dog" and "the big dog attacked the birds" are being uttered by the speakers (S) and heard by the listeners (L). The broken arrows in this level indicate direction of communication. In the mesoscopic level, the words in the above two example sentences are represented by nodes and a pair of nodes are linked if they are adjacent to each other in either of these two sentences. The network, so formed, is representative of the complex patterns of interaction among the words in a language. At the macroscopic level, the properties of the two sentences can be represented using a set of grammar rules and a small lexicon

Since human language is one of the most appropriate examples of a complex system, principles of network theory have proved to be extremely suitable for modeling as well as explaining the emergence of various intricate linguistic phenomena. In fact, it is due to this reason that within a very short period of time the study of linguistic networks, to understand the structure and the evolutionary dynamics of language, has gained a lot of momentum (see [5, 25, 26, 65, 78, 140, 143, 151] for references). The primary motivation for the computational methodology adopted in this work is the burgeoning success of the aforementioned studies (some of which are discussed below) that not only investigate but also substantiate numerous linguistic properties within the framework of complex networks. More precisely, we show how this computational framework helps us in addressing one of the most important problems in phonology that involves modeling and explaining the structure, dynamics and emergence of human speech sound inventories across the languages of the world.

In this chapter, we review some of the most popular studies on linguistic networks, which have been the principal source of inspiration for the work presented here (section 1.1). This is followed by a brief history of the problem addressed in this thesis (section 1.2). A detailed survey of the same will be presented in Chapter 2. Section 1.3 outlines the main objectives of this thesis and section 1.4 summarizes the salient contributions of the work. The organization of the thesis is presented in section 1.5.

1.1 Linguistic Networks

The study of linguistic networks at the level of mesoscopy can be broadly classified into three different categories based on the purpose of construction of these networks. These categories are

(i) *Lexical networks* that are constructed to explore the organization of the “mental lexicon” (i.e., the repository of word forms, which are assumed to reside in the human brain).

(ii) *Word co-occurrence networks* that are constructed to study the evolution of the syntactic structure of a language.

(iii) *Phonological networks* which are built to determine the universal properties of the sound structure of linguistic systems.

In the rest of this section, we will briefly describe some of the most important studies (referring, wherever applicable, to the other relevant ones) in each of the aforementioned categories. Note that the definitions of the standard statistical properties of complex networks used in this chapter as well as in the rest of this thesis are provided in Appendix C.

1.1.1 Lexical Networks

In 1940, Seashore and Eckerson [136] reported that the average vocabulary size of an educated adult is 150,000. It is quite surprising to note that native speakers can navigate this huge lexicon and almost instantaneously recognize (usually in less than 200 milliseconds) a word of their language. Consequently, there are two important questions associated with the mental lexicon (ML): (a) how are the words stored in the long term memory, i.e., how ML is organized, and (b) how are these words retrieved from ML. The above questions are highly inter-related – to predict the organization one can investigate how words are retrieved from ML and vice versa.

The inherent complexity of the problem has motivated a lot of researchers in the past to investigate the organization of ML in the framework of complex systems and more specifically, complex networks (see [37, 65, 78, 140, 153, 157] for references). In all of these studies, ML is modeled as a network of inter-connected nodes, where each node corresponds to a word form and the inter-connections can be based on any one of the following:

(i) *Phonological Similarity*: A large scale phonological similarity based ML can be represented as a complex network in which the nodes correspond to word forms and two nodes (read words) are connected by an edge if they differ only by the addition, deletion or substitution of one or more phonemes [65,78,153,157]. For instance, the words cat, bat and rat may be connected since they differ by the substitution of a single phoneme. [78] reports one of the most popular studies, where the author constructs a Phonological Neighborhood Network (PNN) based on a slight variation of the above definition in order to unfurl the organizing principles of ML. The author shows that PNN is characterized by a high clustering coefficient (0.235) but at the same time exhibits long average path length (6.06) and diameter (20). The above results indicate that, like a small-world network, the lexicon has many densely inter-connected neighborhoods. However, connections between two nodes from two different neighborhoods are harder to find unlike in small-world networks. An intuitive explanation for such a structure of ML is as follows [104].

Low mean path lengths are necessary in networks that need to be traversed quickly, the purpose of the traversal being search in most cases. However, in the case of ML, the search is not expected to inhibit those nodes that are neighbors of the immediate neighbors of the stimulus but are non-neighbors of the stimulus itself and are therefore, not similar to the stimulus. Hence, it can be conjectured that, in order to search in PNN, traversal of links between distant nodes is usually not required. In contrast, the search involves an activation of the structured neighborhood that share a single sub-lexical chunk that could be acoustically related during the process of word recognition.

(ii) *Semantic Similarity*: One of the classic examples of semantic similarity based networks is the Wordnet [53] lexicon. In this network, concepts (known as synsets) are the nodes and the different semantic relationships between them are represented through the edges. [140] analyzes the structure of the noun network extracted from the English Wordnet database (version 1.6). The semantic relationships between the nouns can be primarily of four types (i) hypernymy/hyponymy (e.g., animal/dog), (ii) antonymy (e.g., day/night), (iii) meronymy/holonymy (e.g., eye/body), and (iv) polysemy (e.g., the concepts “the main stem of a tree”, “the body excluding the head and neck and limbs”, “a long flexible snout as of an elephant” and “luggage consisting of a large strong case used when traveling or for storage” are connected to each other due to the polysemous word "trunk" which can mean all of these). Some of the important findings from this work are – (a) semantic relationships in this network are scale-invariant, (b) the hypernymy tree forms the skeleton of the network, (c) inclusion of polysemy re-organizes the network into a small-world, (d) subgroups of fully connected meanings become regions of higher traffic (i.e., nodes with maximum number of paths passing through them), and (e) in presence of polysemous edges, the distance between two nodes across the network is not in correspondence with the depth at which they are found in the hypernymy tree.

(iii) *Orthographic Similarity*: Like phonological similarity networks, one can also construct networks based on orthographic similarity, where the nodes are the words and the weight of the edge between two nodes is defined by the edit distance between the words corresponding to those two nodes. Such networks have been studied to investigate the difficulties involved in the detection and correction of spelling errors that are made by humans while typing [37]. In the aforementioned work, the authors construct orthographic similarity based networks (SpellNet) for three different languages (Bengali, Hindi and English) and analyze them to show that (a) for a particular language, the probability of real word errors is proportional to the average weighted degree of the corresponding SpellNet, (b) for a particular language, the hardness of non-word error correction is correlated to the average clustering coefficient of the corresponding SpellNet, and (c) the basic topological properties of SpellNet are invariant in nature for all the languages.

Other relevant studies pertaining to the structure and dynamics of various lexical networks may be found in [98, 65, 153, 157, 148].

1.1.2 Word Co-occurrence Networks

In this section, we present a brief review of some important studies on word co-occurrence networks where the nodes represent words and two nodes are connected by an edge if the words corresponding to them co-occur in a language in certain context(s). Most of these studies attempt to explore the evolution of the syntactic structure of a linguistic system. In this category, we shall mainly focus on *word collocation networks* and their application in unsupervised induction of the grammar of a language.

(i) *Word Collocation Network*: One of the most fundamental and well-studied examples of co-occurrence networks are the *word collocation networks*, where two words are connected if they are neighbors, that is they collocate, in a sentence [25]. Two types of networks – the *unrestricted* and the *restricted* ones – have been constructed for English in [25] from The British National Corpus. In the unrestricted network all the collocation edges are retained while in the restricted one only those edges are retained for which the probability of the occurrence of the edge is higher than in the case where two words collocate independently. The authors report that (a) both the networks exhibit small-world properties. The mean path lengths are small (around 2 to 3) and the clustering coefficients are high (0.69 in case of the unrestricted network and 0.44 for the restricted one), (b) for both the networks, the degree distributions follow a two regime power-law. The degree distribution of the 5000 most connected words exhibit a power-law with exponent -3.07, which is very close to that predicted by the Barabási-Albert model [13]. These findings led the authors to posit that the word usage in human languages is preferential in nature, where the frequency of a word determines the comprehensibility and production capability. From (a) and (b) together they conclude that evolution of language has resulted in an optimal structure of the word interactions that facilitate easier and faster production, perception and navigation of the words.

In a separate study, Dorogovtsev and Mendes [46] propose a preferential attachment based growth model to explain the emergence of the two regime power-law degree distributions obtained for the aforementioned networks. In this model, at every time step t a new word (i.e., a node) enters the language (i.e., the network) and connects itself preferentially to one of the pre-existing nodes. Simultaneously, ct (where c is a positive constant) new edges are grown between pairs of old nodes, which are also selected preferentially. Through a series of experiments and mathematical analysis the authors show that the power-law exponents predicted by the model are close to those exponents that have been reported in [25].

There have also been studies on the properties of collocation networks of languages other than English (see [31, 79] for references). These studies show that the basic structural properties (e.g., scale-free, small-world, assortative) are similar across all languages. Therefore, together they qualify as linguistic universals and call for well-founded psycho-linguistic accounts for their emergence and existence.

(ii) *Unsupervised Grammar Induction*: One of the most interesting applications of collocation networks, reported in [144], is the unsupervised induction of the grammar of a language. Understanding the process of language acquisition is one of the greatest challenges of modern science. Even at the very early stages of development, infants can pick up all the intricacies of the language they are exposed to quite accurately and effortlessly. This is one of the strongest observations in support of the instinctive capacities of human beings towards language [125], which Chomsky calls the Universal Grammar [36]. In [144], the authors propose a very simple algorithm for learning hierarchical structures from the collocation graph built from a raw text corpus. A brief description of this algorithm, which they call ADIOS is presented below.

A directed collocation network is constructed from the corpus, where the words are the nodes and there is a directed edge from node u to node v if node (read word) v follows node (read word) u in a sentence. Therefore, each sentence is represented by a directed path in the network. The algorithm iteratively searches for *motifs* that are shared by different sentences. A linguistic motif is defined as a sequence of words, which tends to occur quite frequently in the language and also serves certain special functions. For instance, "X is same as Y" is a very commonly occurring motif in English, where X and Y can be substituted by a large number of words and the whole sequence can be embedded in various parts of a sentence. The authors define the probability of a particular structure being a motif in terms of network flows. Once the motifs are extracted, the algorithm proceeds to identify interchangeable motifs and merge them into a single node. Consequently, in each step the network

> becomes smaller and a hierarchical structure emerges. This structure in turn can be presented as a set of phrase structure grammar rules.

For other studies on word co-occurrence networks the reader is referred to [26, 27, 28, 30].

1.1.3 Phonological Networks

In the preceding sections, we have described how complex networks can be used to study the different types of interactions (e.g., phonological, syntactic, semantic) among the words of a language. Networks can also be constructed to study the properties of different sub-lexical units. One such study is presented in [143] where the authors construct a network of Portuguese syllables from two different sources: a Portuguese dictionary (DIC) and the complete work of a very popular Brazilian writer – Machado de Assis (MA). Each node in this network is a syllable and links are established between two syllables each time they are shared by a word. The authors show that the networks have (a) a low average path length (DIC: 2.44, MA: 2.61) and (b) a high clustering coefficient (DIC: 0.65, MA: 0.50). Further, both the networks exhibit a power-law behavior. Since in Portuguese the syllables are close to the basic phonetic units unlike in English, the authors argue that the properties of the English syllabic network should be different from that of Portuguese. The authors also conjecture that since Italian has a strong parallelism between its structure and syllable hyphenization it is possible that the Italian syllabic network has properties close to that of the Portuguese network pointing to certain cross-linguistic similarities.

1.2 The Problem of Consonant Inventories

The most basic units of human languages are the speech sounds. The repertoire of sounds that make up the sound inventory of a language are not chosen arbitrarily, even though the speakers are capable of perceiving and producing a plethora of them. In contrast, the inventories show exceptionally regular patterns across the languages of the world, which is arguably an outcome of the self-organization that goes on in shaping their structure [119]. Earlier researchers have proposed various functional principles to explain this self-organizing behavior of the sound inventories. The most important among these are as follows.

(i) *Maximal perceptual contrast* [96], which implies that the phonemes as well as the other linguistic units (e.g., syllables, words) of a language should be maximally distinct from each other, because this facilitates proper perception of the individual linguistic units in a noisy environment.

(ii) *Ease of articulation* [44,96], which states that the structure of a language should facilitate expression and dissemination of information at the expense of minimal energy spent on the part of the speaker. Some of the general implications of this principle are: frequent words are shorter in length; the sound systems of all languages are formed of certain universal (and highly frequent) sounds that do not use complicated articulatory gestures, etc.

(iii) *Ease of learnability* [44], which states that a language should be easily learnable in order to propagate through the generations. Consequences of this principle include facts that linguistic structures are mostly regular and irregularities, if any, are observed for only extremely frequent linguistic units (e.g., some very frequent verbs in English are irregular).

These principles are applied to language as a whole, thereby, viewing it from the macroscopic level. In fact, the organization of the vowel inventories across languages has been quite satisfactorily explained in terms of the single principle of maximal perceptual contrast through linguistic arguments [158], numerical simulations [92,94,134] as well as genetic algorithms [80]. With the advent of highly powerful computers, it has also been possible to model the micro-level dynamics involving a group of (robotic) speakers and their interactions and this in turn has proved to be highly successful in explaining how the vowel inventories originated and self-organized themselves over the linguistic generations [44].

Right from the beginning of the 20th century, there have been a large number of linguistically motivated attempts [20,39,154,155] in order to explain the emergence of the regularities that are observed across the consonant inventories. However, unlike the case of vowel inventories, majority of these efforts are limited to the investigation of certain specific properties primarily because of the inherent complexity of the problem. The complexity arises from the fact that (a) consonant inventories are usually much larger in size and are characterized by much more articulatory/acoustic features than the vowel inventories, and (b) no single force is sufficient to explain the organization of these inventories; rather a complex interplay of forces collectively shape their structure. Thus, a versatile modeling methodology, which is hitherto absent in the literature, is required so that the problem can be viewed and solved from an alternative perspective.

In the next chapter, we shall present a more detailed history of the problem of sound inventories and in particular, consonant inventories, which is the central theme of this thesis.

1.3 Objectives

The primary objective of this thesis is to develop a computational framework for simulating the structure and dynamics of the consonant inventories of the world's languages. More specifically, we model the self-organization of these inventories through a complex network approach. The choice of this approach is motivated by (a) its enormous success in explaining various dynamical properties of language (examples of which have been already discussed in the earlier sections), and (b) its easy applicability in modeling this particular problem.

Some of the typical questions that we would like to answer in the course of this thesis are as follows.

(i) *Representation of the Inventories*: The first question that one needs to answer is how can the structure of the consonant inventories be accurately represented within the framework of complex networks. This is indeed a very important problem, because all the results obtained as well as the predictions made can be heavily influenced by the underlying scheme of representation.

(ii) *Analysis of the Inventory Structure*: Once a suitable representation scheme is chosen, the next crucial question is how to conduct the analysis in order to extract meaningful results. In particular, one needs to answer (a) which statistical properties of the network(s) should be studied in order to discover the different cross-linguistic patterns that manifest across the consonant inventories, (b) what are the basic principles that could be responsible for the formation of these patterns, and (c) how can these principles be systematically quantified in order to figure out the extent to which they drive the origins of these patterns.

(iii) *Synthesis of the Inventory Structure*: A third and an equally important problem is to explain the emergence of the different statistical properties (obtained from the analysis) by means of generative mechanisms that are usually based on various models of network growth. The typical questions that one needs to answer here are (a) what can be a suitable synthesis model for explaining the statistical properties of the network, (b) how can such models be analytically solved to have a better understanding of the dynamics, (c) what are the linguistic correlates of each of these models with reference to the consonant inventories, and (d) what is the physical significance of the parameters involved (if any) in each of these models.

Although the thrust of this work is on consonant inventories, we also aim to investigate certain well-known properties of the vowel inventories within the same computational framework. The objective of this is twofold – (a) to show that the formalism proposed here is generic and is useful in studying the evolution and emergence of human speech sound inventories, and (b) to report interesting new observations about the vowel inventories apart from validating the results presented by the earlier researchers.

1.4 Contributions

In this work, we show how the structure of the consonant inventories can be represented, analyzed as well as synthesized within the framework of complex networks. For this purpose, we construct two networks, one of which is based on the occurrence of consonants across languages while the other is based on co-occurrence of the consonants across languages. A brief report (which we shall elaborate in the forthcoming chapters) on the studies of these two networks and the results obtained thereby, are presented below.

Occurrence Network of Consonants

We represent the inventories as a bipartite network in which one of the partitions consists of nodes corresponding to the languages while the other partition consists of nodes corresponding to the consonants. There is an edge between the nodes of these two partitions if a particular consonant occurs in a particular language. An exhaustive study of this network reveals various interesting results as follows.

(i) The size of the consonant inventories (indicated by the distribution of the degrees of the language nodes) follow a β-distribution.

(ii) The distribution of occurrence of the consonants over languages (i.e., the degree distribution of the consonant nodes in the network) follow a well-behaved probability distribution.

(iii) A synthesis model based on preferential attachment (i.e., a language node attaches itself to a consonant node depending on the current degree (k) of the consonant node) coupled with a tunable randomness component can explain the emergence of the degree distribution of the consonant nodes.

(iv) The emergent degree distribution obtained from the synthesis model can be analytically shown to approach a β-distribution in the asymptotic limits.

Co-occurrence Network of Consonants

After studying the properties of occurrence of consonants, the next apparent step is to investigate their co-occurrence properties. For this purpose, we construct a network in which the nodes are the consonants and an edge between two nodes (read consonants) signifies their co-occurrence likelihood across languages. Some of the important findings from this study are summarized below.

(i) The co-occurrence distribution of the consonants across languages (i.e., the degree distribution of the consonant nodes in the co-occurrence network) is again found to follow a well-behaved probability distribution.

(ii) The clustering coefficient of the co-occurrence network is very high, a property commonly observed in social networks [113,130] that is indicative of the presence of a large number of densely connected neighborhoods (formed by groups of consonants).

(iii) Community structure analysis of this network reveals strong patterns of co-occurrence of consonants that are prevalent across the languages of the world.

(iv) Languages exhibit an economic behavior by using a small number of articulatory/acoustic features and maximizing the combinatorial possibilities of these features in order to generate a large number of consonants. This behavior, often termed as *feature economy* [20, 39, 45, 105], leads to the formation of the consonant communities. An information theoretic quantification of this principle further shows the extent to which it is responsible for the community formation.

(v) The emergent degree distribution of the co-occurrence network can be shown to be sensitive to the distribution of the consonant inventory sizes even though the degree distribution of the occurrence network does not depend on the same.

(vi) The clustering coefficient of the co-occurrence network can be explained through a synthesis model that is based on both preferential attachment and *triad* (i.e., fully-connected triplet) formation. This process of triad formation actually imposes a large number of triangles onto the generated network thereby creating many densely connected neighborhoods and increasing the clustering coefficient.

Apart from exploring various significant properties of the consonant inventories, we also employ our computational methodology to study the structure of the vowel inventories. Some of our observations are

(i) The topological properties of the occurrence and co-occurrence networks constructed from the vowel inventories are found to be largely similar to that of the consonant inventories. In particular, preferential attachment plays the key role in the emergence of their structure.

(ii) Community analysis of the co-occurrence network of vowels indicate that the small size vowel inventories tend to be organized based on the principle of maximal perceptual contrast – an observation that is in agreement with those reported by the earlier researchers [44, 92, 94, 134].

(iii) On the other hand, the larger vowel inventories reflect a considerable extent of feature economy – an observation that has been made by a school of linguists earlier [20, 39], but quantitatively substantiated here.

(iv) Co-occurrences based on implications (one vowel implying the presence of another) are prevalent across the vowel inventories and their presence is again a consequence of feature economy. This property has also been noted by linguists earlier; however, it has been quantitatively established here.

It is worthwhile to mention here that this thesis also contributes significantly to the development of the modeling techniques that are used in general, in the field of complex networks. For example, in most of the bipartite networks studied in the past both the partitions are assumed to grow over time (see [24, 113, 130] for references). Nevertheless, the occurrence network of consonants introduced here is a special class of bipartite network in which one of the partitions remains almost fixed over time (i.e., the partition of consonants) while the other can grow unboundedly (i.e., the partition of languages). This subclass of bipartite networks can be extremely useful in representing, analyzing as well as synthesizing *discrete combinatorial systems* (DCS) [125], where the basic building blocks are a finite set of elementary units (e.g., consonants) and the system is a collection of potentially infinite number of discrete combinations of these units (languages). In fact, two of the greatest wonders of evolution on earth, life and language, are examples of DCS. In case of living systems, for instance, the elementary units are the nucleotides or codons while their discrete combinations give rise to the different genes. In case of language, the elementary units are the letters or words and the discrete combinations are the sentences formed from them. Therefore, the network growth models that we propose in this work essentially attempt to mimic the evolution of a DCS. Note that analytical treatment of these growth models are not straightforward because, the average degree of the fixed partition diverges with time and hence, stationary state assumptions, that are commonly made in the literature to solve these types of problems (see [114] for reference), are no longer applicable here. This thesis, therefore, opens up many new challenging theoretical problems in the area complex networks and more specifically, bipartite networks.

Finally, to summarize the contributions of this thesis in a single sentence, *we have shown that the self-organization and the emergence of the structure of human speech sound inventories can be successfully studied within the framework of complex networks. Thus, we believe that in future, this computational framework can serve as an extremely powerful tool in modeling the structure and dynamics of several linguistic phenomena, which are as complex as the one presented here and for which no satisfactory explanation exists.*

1.5 Organization of the Thesis

The thesis is organized into seven chapters.

Chapter 2 presents the history of the problem of sound inventories. It describes the human articulatory apparatus and the representation scheme for a phonological system. This is followed by a concise review of the different linguistic and computational studies pertaining to the organization of the sound inventories, which together form the basic motivation for this work.

Chapter 3 centers around the study of the occurrence network of consonants. It outlines the construction procedure for the network and describes the data source used for this construction. This is followed by an analysis of some of the interesting topological properties of the network. A synthesis model is then proposed and analytically solved to explain these properties. Finally, we employ this model to investigate the dynamics within and across five major language families of the world.

Chapter 4 investigates in detail the properties of the co-occurrence network of the consonants. It begins with a study of some of the important topological properties of this network. Suitable refinements of the synthesis model presented in Chapter 3 are also proposed to explain the emergence of these topological properties.

Chapter 5 presents the community structure analysis of the co-occurrence network of consonants. It discusses the algorithm for community detection and identifies the role of feature economy in the formation of the communities. Furthermore, it describes an information theoretic approach to quantify feature economy so as to determine the extent to which this factor governs the community formation in consonant inventories.

Chapter 6 presents a detailed study of the topological properties of the occurrence and co-occurrence network of vowels. Furthermore, it outlines the community analysis of the co-occurrence network and reports various interesting as well as important observations about the vowel inventories apart from validating the results that are already documented in the literature.

Chapter 7 concludes the thesis by summarizing the contributions and pointing to a few topics of future research that have been opened up from this work.

There are also a few appendices that present an annotated list of publications from this work, a complete list of publications by the candidate, a glossary of definitions and an alphabet of phonetic notations.

Chapter 2

Background

Almost all human languages make use of speech sounds as a primary medium for conveying meaning. The physical speech signal is usually represented as a sequence of abstract minimal units called *phonemes*, which render a distinction in meaning between the words of a language. Such distinctions are usually reflected by word pairs (also minimal pairs) which mean different things, but differ only in one sound. For instance, in English, /m/[1] and /n/ are phonemes because, word pairs such as “rum” and “run” have different meanings. The repertoire of phonemes that the speakers of a language use to bring about these meaning distinctions is termed as the *phoneme inventory* (or alternatively, the *sound inventory*) of that language. Interestingly, the phoneme inventories of the world’s languages exhibit a large number of universal tendencies. In fact, these universal tendencies are among the best-researched universal properties of language [20, 39, 44, 80, 92, 94, 96, 134]. The linguistic questions that are addressed in this thesis are mostly taken from those that have been unveiled by this extensive research. Furthermore, the outcomes from this research have also been used in verifying whether the computational models proposed here are able to produce results that are compatible with what is known about natural languages.

[1] Throughout the thesis we shall represent phonemes using the IPA (International Phonetic Alphabet) symbols. The complete list of the IPA symbols is provided in Appendix D (adapted from http://www.langsci.ucl.ac.uk/ipa/fullchart.html).

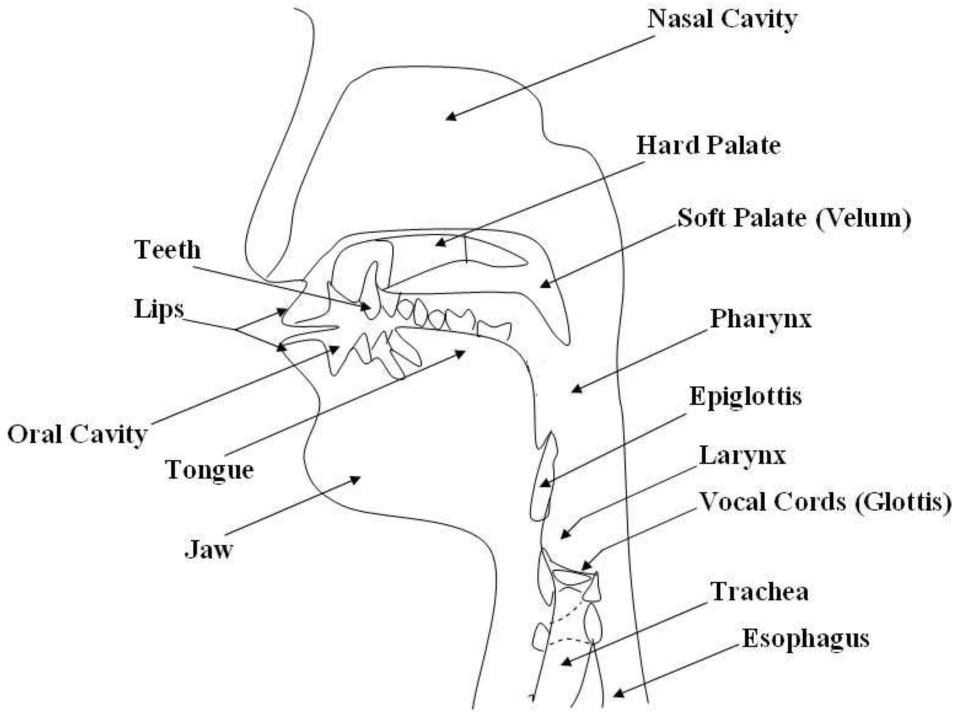


Figure 2.1: Human vocal system

In this chapter, we shall attempt to briefly outline this long history of research on the universal properties of the phoneme inventories of human languages. We begin, in section 2.1, with a brief description of the human vocal system and the mechanism of speech production. In the same section, we also discuss how speech sounds can be appropriately represented as phonemes in terms of a set of articulatory/acoustic features extracted from them. In section 2.2, we point out several regularities that are observed across the phoneme inventories of the world's languages. In the following section, we describe some of the linguistic as well as computational attempts that have been made in the past in order to explain the emergence of these regularities. In section 2.4, we summarize the state-of-the-art in this area of research and also identify the gaps, which in turn form the primary motivation for the work presented in this thesis.

2.1 Human Vocal System and Speech Production

Human speech is produced by the vocal organs shown in Figure 2.1[2]. The *vocal tract*, which plays the role of a resonance tube during speech production mainly consists of three cavities namely, the *pharynx*, the *nasal cavity*, and the *oral cavity* (see Figure 2.1). The shape of the vocal tract is altered by the *soft palate* (*velum*), the *tongue*, the *lips* and the *jaw*. These organs together are collectively known as the *articulators*. The phenomenon of shaping the structure of the vocal tract for producing different speech signals is known as *articulation*. During speech production, the air flow from the lungs is forced through the *glottis* (between the *vocal cords*) and the *larynx* to the vocal tract. From the oral and the nasal cavities the air flow exits through the nose and the mouth, respectively. The vocal cords can act in several different ways, the most important function being the modulation of the air flow by rapidly opening and closing, thereby, generating a buzzing sound from which the vowels and the voiced consonants are produced. Voiceless sounds, on the other hand, are produced if this rapid vibration is absent (see [88] for a more detailed description of the speech production system).

A phoneme can be described by a collection of articulatory features [154] that can be broadly categorized into three different types namely, the *manner of articulation*, the *place of articulation* and *phonation*. Manner of articulation is concerned with the flow of air, that is, the path it takes and the degree to which it is impeded by the constrictions of the vocal tract. Some of the most important manners of articulation are (a) plosives – these phonemes result from blocking the vocal tract by closing the lips and the nasal cavity, (b) fricatives – these phonemes are generated by constricting the vocal tract at some point and forcing the air stream to flow at a velocity that is appropriate for producing turbulence, (c) affricates – these phonemes are produced by a combination of the plosive and the fricative phonemes, and (d) nasals – these phonemes are generated when the vocal tract is closed but the velum opens a route for the air stream to the nasal cavity.

[2] This figure has been drawn by the author himself. The guidelines for the figure has been taken from http://www.telecom.tuc.gr/~ntsourak/tutorial_acoustic.htm.

Place of articulation specifies the active speech organ and also the place where it acts. Vowels are usually described in terms of tongue position and lip rounding. The significant places of articulation for consonants are the lips (*bilabial*), the lips and the teeth (*labio-dental*), the teeth (*dental*), the upper gums (*alveolar*), the hard palate (*palatal*), the soft palate (*velar*), and the glottis (*glottal*).

Phonation specifies whether rapid vibrations are produced in the vocal cords during the articulation of a particular phoneme. If such vibrations are produced during the articulation of a phoneme then it is said to be vocied. In contrast, if a phoneme is produced without the vibration of the vocal cords then it is called voiceless.

Apart from these three major classes, there are also secondary articulatory features that are used to describe some specific phonemes found in certain languages. A few representative examples of phonemes as a collection of features are shown in Table 2.1. Note that sometimes more than one features are required in a particular category to perfectly describe a phoneme (see [102]). For instance, there are phonemes that can be best expressed (a) as a combination of two places of articulation (e.g., dental-alveolar, palato-alveolar, labial-velar), or (b) using more than one secondary feature (e.g., a phoneme can be labialized and velarized at the same time).

It is worthwhile to mention here that there is a lively debate about the cognitive reality of the articulatory features with some researchers [99] claiming that they do not exist at all. Nevertheless, recent experiments in neuroscience [128] show that during speech production and perception specific motor circuits are recruited in the brain that reflect distinctive features of the speech sounds encountered. Therefore, in this light, one might interpret articulatory features as motor programs.

In the following section, we shall discuss about some of the regularities that are observed across the phoneme inventories of the world's languages.

Table 2.1: Examples of phonemes (consonants and vowels) represented as a collection of features

Voewls	Position of Tongue	Tongue Height	Lip Roundedness
/i/	front	high	unrounded
/a/	central	low	unrounded
/u/	back	high	rounded

Consonants	Manner	Place	Phonation	Secondary Features
/t/	plosive	alveolar	voiceless	–
/d/	plosive	alveolar	voiced	–
/t^{w}/	plosive	alveolar	voiceless	labialized
/t^{j}/	plosive	alveolar	voiceless	palatalized

2.2 Regularities of the Phoneme Inventories

The phoneme inventories of the world's languages exhibit remarkable regularities. Although the human vocal tract is capable of producing an amazing variety of sounds, any single language only uses a small subset of them. The phonemes that a particular language use are not chosen randomly from the possible sounds that the human vocal tract can generate. In contrast, some phonemes recur more frequently across languages than others. For instance, if a language has only three vowels then these are usually /i/, /a/ and /u/ [44, 92, 94]. Similar observations can also be made for the case of consonants. Certain consonants like /m/ and /k/ are present in almost all languages while others, such as /ʡ/ and /ʢ/ are extremely rare [39, 44]. Some other equally important observations are (a) all languages have at least two of the voiceless plosives /p/,/t/ and /k/ [20, 39], (b) voiceless nasals only occur in languages that have voiced nasals [39], and (c) in the series of voiced plosives /b/, /d/, and /g/, /g/ is most likely to be missing [19, 20].

Regularities also arise from the fact that the phoneme inventories tend to be symmetric. For example, if a language has a front rounded vowel of a specific tongue height then it tends to have a corresponding back unrounded vowel of the same height. In case of consonants, if a language makes a distinction between voiced and voiceless plosives then it tends to do so at each place of articulation. In general, if a language makes use of certain place of articulation and manner of articulation features then usually all the combinatorial possibilities of these features are used rather than a subset of them (see [20, 39] for further references).

The above observations collectively imply that certain inventories of phonemes are usually favored, while others are systematically avoided. The regularities across the phoneme inventories are interesting because, their emergence calls for an explanation. Perhaps, the most important questions are – "why are certain sound patterns recurrent?", "what are the possible causes of symmetry across the phoneme inventories?", and "how can these properties of recurrence and symmetry be systematically investigated and assessed?". In the next section, we shall briefly review some of the popular studies conducted in the past in order to explain the evolution and the emergence of the regularities across the phoneme inventories.

2.3 Explanations for Cross-Linguistic Regularities

Several attempts have been made in the past to build a theory that can explain the structure of the phoneme inventories, which are primarily based on the physical and psychological properties of human speech production and perception. In this section, we shall present a concise report of both linguistic as well as computational studies that have been conducted by the earlier researchers in order to reason the emergence of the observed patterns across these inventories. While some of these studies employ purely linguistic insights to explain the inventory structure, others attempt to develop computational models grounded in the linguistic theories in order to explain the emergence of this structure.

Most of the studies based on linguistic insights are largely inspired by the concepts of *generative phonology* proposed by Noam Chomsky and Morris Halle in 1968 [35]. According to this view, phonological representations are sequences of segments made up of distinctive features (i.e., features that distinguish phonemes from one another). There have been several attempts following the advent of this theory to show that distinctive features are the key elements that determine the structure of the phoneme inventories [20, 32, 33, 39, 149, 150]. Cross-linguistic studies have also been carried out to show that these features are not innate; in contrast, they are acquired during language learning [108].

More recently, researchers have also started building computer models based on various well-studied linguistic principles in order to predict the structure of the inventories as a whole [92, 134, 44, 80]. In such studies various optimization techniques are employed which in turn result in the emergence of the inventory structure.

2.3.1 Linguistic Explanations

In this section, we shall highlight some of the popular linguistic attempts that have been made by the past researchers to develop a theory for the emergence of the sound patterns. We discuss four such theories that have proved to be quite useful in describing various structural patterns found across the phoneme inventories.

Theory of Feature Economy

One of the central findings of the earliest work on phonology is that language inventories tend to be structured in terms of correlations based on the features that characterize the phonemes present in them [70, 105, 155]. In order to explain this tendency *feature economy* was proposed as the organizing principle of the phoneme inventories (see [39] and the references therein). According to this principle, languages make use of a small number of distinctive features and maximize their combinatorial possibilities to generate a large number of phonemes [39]. Stated differently, a given phoneme will have a higher than expected chance of occurrence in those inventories, where a majority of its features have distinctively appeared in the other phonemes. The idea is illustrated in Table 2.2 for a set of four consonants.

There have been many attempts to investigate as well as establish the statistical significance of this principle mainly through linguistic insights [20, 39]. A preliminary mathematical formulation of this principle have also been provided in [39]. In this

Table 2.2: Example of feature economy. The table shows four plosives. If a language has in its consonant inventory any three of the four consonants listed in this table, then there is a higher than average chance that it will also have the fourth consonant of the table in its inventory

plosive	voiced	voiceless
dental	/d̪/	/t̪/
bilabial	/b/	/p/

study, the author defines the term *economy index* to measure the extent of feature economy across the phoneme inventories. Given an inventory of S phonemes that are characterized by F features, the economy index E is given by the expression

$$E = \frac{S}{F} \tag{2.1}$$

Feature economy can be thought of as the tendency to maximize E either by maximizing the number of phonemes S or by minimizing the number of features F. Note that this definition does not capture the extent to which the features contribute to discriminate the different phonemes in an inventory. The less discriminative the features are on an average the more closer should be the phonemes in the feature space and higher should be the feature economy.

The author further examines the co-occurrence of various pairs of plosives sharing the manner of articulation features but differing in the place of articulation and shows that languages having one member of the pair tend to have the other with a very high probability.

These studies and the observations made from them led the researchers to generalize feature economy as one of the most important principles governing the regularities across the phoneme inventories.

Optimality Theory

The central idea of optimality theory [107,127] is that the observed forms of language arise from the interaction between conflicting constraints. There have been some attempts to apply this theory to explain the structure of the sound inventories [20,55]. The inventory structure, in this theoretical framework, can be expressed in terms of a set of *constraints*, each of which can be *violated* if it is crucially *dominated* by a stronger constraint. The interaction between the constraints are based on a principle of *strict ranking*, i.e., a high-ranked constraint will always outweigh any number of lower-ranked constraints. For instance, a speaker will turn /r/ into /ʀ/ or /ɹ/ if the constraint expressing the articulatory effort of /r/ dominates the constraint that aims for perceptual clarity. On the other hand, if the clarity constraint is ranked higher then the speaker will faithfully pronounce /r/. In fact, in [20] the author shows that if the articulatory and perceptual principles are expressed directly as constraints in the production and perception models of the speakers of a language then the desired properties of their interactions that are necessary to explain the structure of the inventories will follow from the principles of optimality theory.

Quantal Theory

The study of the acoustic theory of speech production reveals that in general, if one changes the configurations of the vocal tract then the acoustic output also changes accordingly. However, the relationship between the articulatory parameters and the acoustic output produced is not linear. For certain positions of the articulators, a small displacement in the position causes only a negligible change in the acoustic output, while for certain other positions an equivalent displacement causes a significant change in the acoustic output. The idea is illustrated in Figure 2.2. The figure shows that there is a large acoustic difference between the regions I and III; however, within regions I and III the acoustic parameter is almost insensitive to the change in the articulatory parameter.

Stevens' *quantal* theory of speech [149, 150] is based on the above observation.

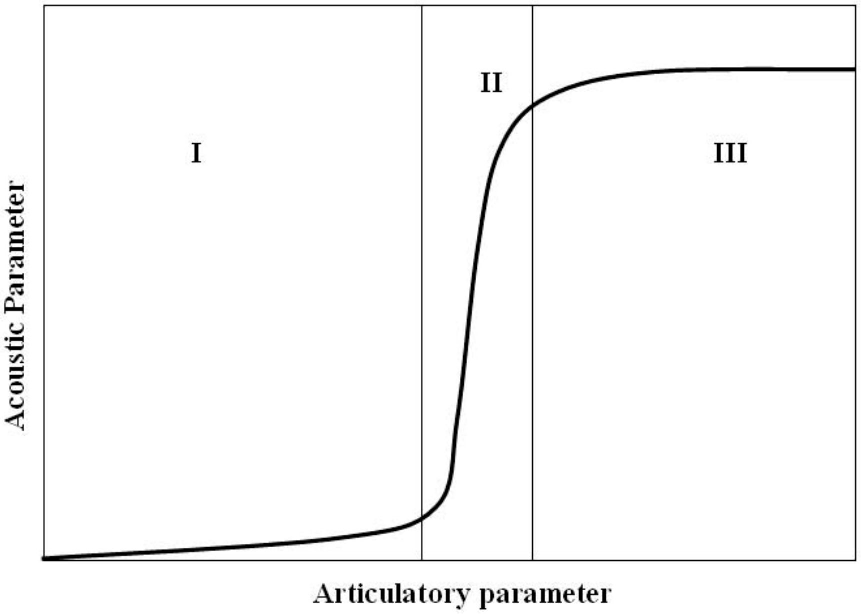


Figure 2.2: The change in the acoustic parameter with respect to the articulatory parameter

According to this theory, linguistic contrasts essentially involve differences in 'quantal' regions (i.e., differences in regions I and III in Figure 2.2). In other words, all quantal regions define contrastive phonemes or at least, quantal distinctions are preferred. There are various arguments in support of the claim that quantal regions influence the structure of the phoneme inventories. Some of these are (a) articulations need not be precise for producing a particular acoustic output, (b) continuous movement through the quantal regions will always yield an acoustic steady state, and (c) minor articulatory errors will not affect perception.

It is interesting to note that the quantal theory is essentially a theory of distinctive features, where each plateau-like region can be considered as a correlate of a particular distinctive feature. It therefore explains, from independent physical, physiological and psychological arguments why certain distinctive features are expected to be present in natural languages.

Theory of Distinctive Region

The theory of *distinctive region* developed by René Carré [32, 33] for explaining the structure of the phoneme inventories is based on the assumption that human speech communication is a near-optimal solution to the physical problem of producing communication over an acoustic channel using a deformable acoustic tube. The theory further assumes that for an optimal communication system maximal acoustic differences can be produced with minimal articulatory gestures.

In this model, the articulatory movements are defined in terms of the linear and orthogonal deformations of a uniform acoustic tube. Using mathematical techniques the author then calculates those deformations that cause maximal acoustic distinctions. The model is able to find distinctions that result in an acoustic space exactly corresponding to the vowel space of human languages. The uniform tube gets divided into four distinct regions corresponding to the regions of the vocal tract that are responsible for vowel production. The author further shows that the model can be extended to predict the places of articulation for consonant production by observing the maximal changes taking place in the *formant frequencies*[3]. In this case, the uniform tube gets separated into eight regions, each corresponding to a place of articulation for the consonants.

It is worthwhile to mention here that all the above theories attempt to show that certain feature distinctions are beneficial and therefore, preferred over others across the languages of the world. However, none of them actually predict phoneme inventories as a whole, that is, which vowels and consonants should appear in an inventory of a certain size. In other words, it remains to be answered that given a set of phonemes, what are the chances of this set being a natural language inventory.

[3] Formants are the distinguishing and meaningful frequency components of human speech. A set of formants, orthogonal to each other, defines the acoustic space. For instance, it has been found that most often the first two formants f_1 and f_2 are sufficient to disambiguate the vowels. Therefore, the vowel space can be thought of as a two dimensional Euclidean space defined by the first two formants.

2.3.2 Computational Explanations

In this section, we shall outline a few representative studies that have been conducted in the past to predict sound inventories as a whole. These studies attempt to develop computational models which are again based on various well-known linguistic principles. The computational techniques can be broadly classified into three different categories namely, (a) functional optimization – phoneme inventories are derived through the optimization of certain functions, (b) multi-agent simulation – phoneme inventories emerge due to interactions between linguistic agents over hundreds of generations, and (c) evolutionary optimization – phoneme inventories evolve through certain simulated evolution techniques such as genetic algorithms.

Functional Optimization

One of the first attempts to predict vowel inventories as a whole without looking at the qualities of the individual sounds and their features was undertaken by Liljencrants and Lindblom [92]. They searched for optimal vowel inventories of different sizes by maximizing the perceptual contrast between the vowels in a fixed two-dimensional perceptual space. For this purpose, they defined an energy function E of the form

$$E = \sum_{i=1}^{n-1} \sum_{j=0}^{i-1} \frac{1}{r_{ij}^2} \qquad (2.2)$$

where n is the total number of vowels in the system and r_{ij} is the Euclidean distance between the vowels i and j in the vowel space defined by the first and the second formants. The function E essentially adds the inverse square of all the distances between all pairs of vowels in the system. The authors tried to minimize this function so that the vowels get spread as evenly as possible in the vowel space. The simulation begins with a predefined number of vowels that are scattered randomly near the center of the available Euclidean space. In every step the positions of the vowels are perturbed (restricted within the vowel space that is limited by what can be produced by the vocal tract) and if as a result, the value of E gets reduced then the new state of the system is saved. The procedure is repeated until there is no further reduction in the value of E. The authors showed that the resultant vowel inventories are very close to the most common vowel inventories of similar size that are found in real languages. For instance, the seven–vowel inventory that they obtain closely resembles the vowel inventory of Italian.

Later, Schwartz *et al.* [134] proposed the *dispersion-focalization* theory for an improved calculation of the perceptual distances between the different vowels. Consequently, the predicted vowel inventories showed a better match with realistic vowel inventories than in the previous case.

The same method can also be applied in order to explain the emergence of the consonant inventories. However, as Ohala [117] points out, models exclusively based on the principle of maximal perceptual contrast make "patently false" predictions about the structure of the consonant inventories. For instance, such a model would predict that a 7–consonant inventory should include something like the set /ɖ k' ts ɬ m r |/. Nevertheless, languages with very few consonants (e.g., the Polynesian languages) do not have such an exotic consonant inventory; rather, languages that do possess these consonants, such as Zulu, also have a great variety of other consonants of each type, i.e., ejectives, clicks, affricates, etc. Therefore, apart from perceptual distance, one needs to also add other factors like the criterion of minimal articulatory complexity into the simulation model. This immediately introduces many more parameters into the model and makes it extremely hard to build and realize computer simulations for explaining the structure of the consonant inventories.

Functional optimization techniques have also been used in modeling the emergence of combinations of speech sounds. In [93] the authors presented a system that created a set of syllables from a given (large) set of possible consonants and vowels based on the criteria of both minimal articulatory complexity as well as maximal perceptual contrast. For the right choice of parameters, repertoires of "phonemically coded" (meaning that only a small subset of the available vowels and consonants would be used in a combinatorial way) realistic syllables would emerge. However, this system was limited to one type of syllable (consonant-vowel) and there were a large number of parameters that made the computer simulations much harder and more controversial to build.

Multi-agent Simulation

The earliest model that used simulation of a population of agents in order to explain the structure of the vowel inventories was the AGORA-model proposed by Hervé Glotin [61, 62]. This model is based on a community of talking "robots" termed as *carls* (Cooperative Agents for Research in Linguistics). Each *carl* has a repertoire of vowels that are represented in the model through articulatory features as well as acoustically in terms of a set of formant frequencies. The *carls* are equipped with an articulatory model to produce acoustic signals during communication. In each step of the simulation, two *carls* are chosen randomly from the population and they both randomly select and articulate a vowel from their repertoire. They then search for the vowel, which is closest to the one that they hear. They shift this vowel so that its acoustic signal is closer to the one heard by them and at the same time, shift all other vowels present in their repertoire away from this signal. A fitness parameter is calculated based on the amount of shifting that a *carl* does. The less the shifting, the more it is in agreement with the vowel inventories of the others and therefore, the fitter it will be. After a number of interactions among the *carls*, the less fit ones are removed from the population and the fittest ones are used to reproduce offsprings in a way similar to genetic algorithms (see [63] for reference). After several simulation steps the population is usually found to converge to a common repertoire of vowels (typically four to five in number) that is very close to what is observed in real languages. However, the convergence of this model is not guaranteed. Furthermore, in this model the agents are assumed to be conscious optimizers (which is usually not true for the speakers of a language), who are able to push the vowels in their vowel inventories away from each other. In essence, this makes it equivalent to Liljencrants and Lindblom's [92] model described earlier. As the agents perform local optimization through the computation of fitness functions, the interactions among them are not necessary for shaping the structure of the emergent vowel inventories. In fact, an agent talking to itself should be able to reproduce the same results.

The most popular study in this category is perhaps the one presented by Bart de Boer in [44]. The model proposed in this study is based on the first principles of communication, where each agent has the simple abilities of identifying and imitating the vowels produced by the other agents. To this end, the individual agents are able to articulate, perceive and maintain a repertoire of vowels. Vowels are represented as prototypes using the basic articulatory features of position, height and rounding. During articulation, this representation is converted into a corresponding acoustic representation based on formant frequencies with some additional noise to ensure that no two signals produced are exactly same. During perception, the vowel sound that is heard is mapped to the prototype that is closest to the signal. The interactions between the agents is limited to an idealized episode of imitation based on language games introduced by Steels in [146]. Each episode of the game consists of four distinct steps of production, imitation, evaluation and modification as follows.

(i) The first agent (i.e., the initiator or the speaker) utters a vowel randomly chosen from its repertoire.

(ii) The second agent (i.e., the imitator or the listener) hears this vowel, searches for it in its own repertoire, and produces the prototype that is closest to the signal it heard.

(iii) The initiator indicates to the imitator through an extra-linguistic signal whether the imitation corresponds to the original vowel produced by the initiator.

(iv) Based on the consequences of an episode, both the agents update their repertoire.

With each of the vowels in an agent's repertoire, there is an associated score, which is slightly increased if a vowel is used in a successful imitation game. On the other hand, if the game is unsuccessful, then this score is reduced and the imitator either adds the new vowel into its repertoire or shifts the unsuccessful vowel in such a way that it is close to the acoustic signal of the vowel that it had heard from the initiator. Apart from these, each agent regularly does certain housekeeping tasks, which include removal of vowels that are used minimum number of times, merging of vowels that are close to each other, and occasionally adding of a new random vowel to its repertoire.

Through the simulations of this model the author shows that the emergent vowel inventories largely correspond with those that are found across real languages. For instance, the five–vowel inventory, which emerged from the simulations, occurs in 88% of the world's languages that have five vowels.

The author further extends this framework to explain the emergence of more complex utterances like the consonant-vowel syllables. In this case, the possible syllable onsets (i.e., the beginning of a syllable incorporating a single or a group of consonants) comprise seven consonants while the possible nuclei (i.e., the central part of a syllable which is most commonly a vowel) of the syllable comprise nineteen vowels. The rules of the imitation game played in the population of agents using consonant-vowel syllables are similar to the ones presented above for the vowels except for one important difference. In case of the vowels, new vowel prototypes that were added or updated depending on the outcome of an imitation game, were improved solely on the basis of the distance between a particular prototype and the acoustic signal perceived by an agent. In contrast, the syllable prototypes are improved not only based on the acoustic distance but also on the articulatory effort involved in producing complex syllables. However, as the author himself observes, the results obtained from the experiments with syllables are "somewhat confusing". The outcomes of the simulations are largely dependent on the parameter settings that determine the way in which the distance between the syllables are calculated as well as the way in which the quality of the syllables are improved with respect to an acoustic signal. Although for the right choice of parameters it is possible to produce phonemically encoded repertoires of syllables, most of these are unrealisitc and extremely infrequent across the languages of the world.

Evolutionary Optimization

The most popular modeling paradigm in this category are genetic algorithms (henceforth GA) (see [63] for reference). GA is a technique based on the way evolution works in nature. Instead of keeping track of a single optimal solution, the algorithm keeps track of a population of solutions. Each solution is represented at two levels – (a) the level at which the fitness of the solutions are evaluated (i.e., the *phenotype*), and (b) the level at which the solutions are recombined and mutated for reproduction (i.e., the *genotype*). Solutions with high fitness are allowed to reproduce off-springs, while the bad solutions (i.e., solutions with low fitness values) are removed from the population.

In [80] the authors use a GA model to search for the optimal configuration of vowel inventories. The model consists of a population of chromosomes, each representing a possible vowel inventory. The individual vowels are encoded using the basic articulatory features of position (p), height (h), and rounding (r). While the first two features are assumed to be continuous within the range [0,1], the last feature is considered to be binary-valued. Therefore, each vowel can be represented as an assemblage of these three features. For instance, if for a particular vowel $p = 0.1$, $h = 0.3$, and $r = 1$ then we can represent it as three-tuple (0.1,0.3,1). Although this encoding method allows for an infinite number of vowels, the authors assume that there is only a limited number of inventories of vowel prototypes from which the system can select candidate configurations. In this GA model, one-point crossover (recombination) and one-point mutation are used. For instance, Figure 2.3 shows how the simulation of three–vowel inventories take place. Parent$_1$ in the figure is a hypothetical vowel inventory consisting of three vowels represented by (0.0,0.5,1), (0.1,0.3,1), and (0.2,0.7,0), respectively. Similarly, Parent$_2$ is a hypothetical inventory of three vowels encoded as (0.5,0.0,0), (0.3,0.5,0), and (0.3,0.2,0), respectively. Crossover takes place between two randomly chosen vowels and in the process the two chromosomes exchange their vowels as shown in the figure. Next, a mutation takes place at random in one of the vowels of the off-springs that are produced (e.g., the third vowel of the Off-spring$_2$ in the figure has been mutated). The fitness of each chromosome after every reproductive phase is computed using the energy function introduced in [92] as well as the dispersion-focalization principle reported in [134]. The off-springs with high fitness values are allowed to breed, while the ones with low fitness are removed from the population. The idea behind this policy is that chromosomes coding for high quality solutions shall multiply in the population, whereas chromosomes coding for bad solutions shall disappear.

Comparison of the real and the emergent inventories show that only the most frequently observed three– and four–vowel inventories are predicted by the system.

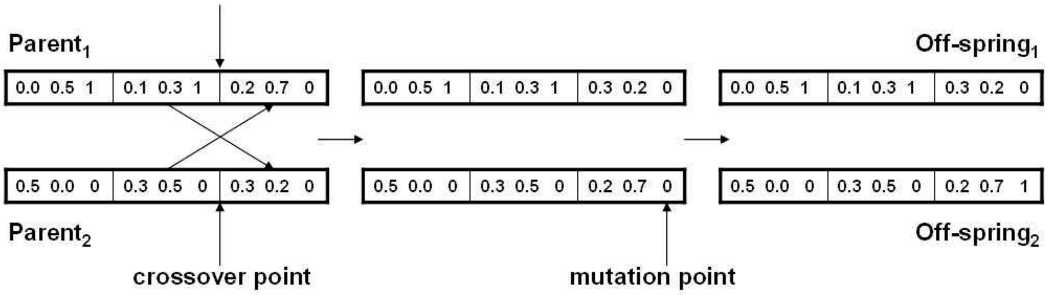


Figure 2.3: Crossover and mutation in the simulation of (hypothetical) three–vowel inventories

Other predictions, for inventories of larger size, do not match well with real inventories due to the limitations of the simple GA model used by the authors in producing more than a single optimal vowel inventory.

In [132] the authors adopt a genetic algorithm framework to explain the emergence of syllable structures that are found across human languages. In this case, the population is composed of a set of candidate strings (the "words" of a language) which are selected on the basis of a number of functional criteria and then mutated, crossed and multiplied to reproduce the next generation of strings. After a number of generations, strings that closely resemble the structure of human syllables are found to emerge. The most important problem with this model is that the constraints that govern the structure of the human syllables are built in as selection criteria. Consequently, the model does not actually explain the structure of the syllables; in contrast, it only indicates that if the selection criteria are present, simulated evolution would be sufficient to produce syllables that obey the constraints.

2.4 Summary

In this chapter, we have briefly outlined the state-of-the-art of research in one of the central problems of phonology that involves predicting the structure of the human speech sound inventories. More specifically, we have described (a) the speech production mechanism, (b) how speech sounds can be represented as abstract units called phonemes, (c) the regularities observed across the phoneme inventories of the world's languages, and (d) linguistic as well as computational approaches to explain these regularities.

This review makes it clear that there have been several attempts made by the earlier researchers to explain the structure of the sound inventories of human languages. In fact, some of these studies have proved to be quite successful in accurately modeling the emergence of various universal properties of these inventories. However, another issue that is also apparent from the review is that it becomes increasingly difficult as one attempts to model the emergence of the inventories of more and more complex utterances such as consonants and syllables.

Therefore, the primary objective of this work is to develop a versatile modeling methodology that can, in general, serve as a computational framework for explaining the emergent structure of the sound inventories of human languages. More specifically, in the forthcoming chapters, we shall show how this methodology, grounded in the theories of complex networks, can be successfully employed to study the self-organization of the consonant inventories which is considered to be one of the more difficult problems in phonology. The fact that we are able to easily employ the same framework to successfully analyze the structure of the vowel inventories also, illustrates the generality of this modeling paradigm.

Since our primary focus throughout this thesis are the consonant inventories, it is worthwhile to mention here that we do not pretend to provide final and authoritative answers about the evolution and emergence of these inventories. However, we present an initial attempt to model and investigate this old but difficult problem in a highly sophisticated computational framework which in turn can be also used to successfully study other types of inventories.

Chapter 3

Analysis and Synthesis of the Occurrence Network of Consonants

In this chapter, we present the basic computational framework to represent, analyze and synthesize the consonant inventories of the world's languages. This framework belongs to a special class of complex networks where there are two different sets (or partitions) of nodes: the *bipartite* networks. An edge, in a bipartite network, connects nodes from one partition to the other, but never the nodes within the same partition. We represent the consonant inventories as a bipartite network namely, the **P**honeme-**La**nguage **Net**work or **PlaNet** where the nodes in the two partitions are labeled by the consonants and the languages respectively. Edges run between the nodes of these two partitions depending on whether a particular consonant occurs in the inventory of a particular language.

The construction of PlaNet as a bipartite network is motivated by similar modeling of various complex phenomena observed in society as well as nature. In most of these networks, however, both the partitions grow with time unlike PlaNet where the partition corresponding to the consonants remains relatively fixed over time while the partition corresponding to the languages grows with time. Typical examples include different types of collaboration networks such as (a) the movie-actor network [7, 10, 123, 130, 161] where movies and actors constitute the two respective partitions and an edge between them signifies that a particular actor acted in a particular movie, (b) the article-author network [14, 89, 113] where the two partitions respectively correspond to articles and authors and edges denote which person has authored which articles and (c) the board-director network [24, 152] where the two partitions correspond to the boards and the directors respectively and a director is linked by an edge with a society if he/she sits on its board. In fact, the concept of collaboration has also been extended to model such diverse phenomena as the city-people network [49] where an edge between a person and a city indicates that he/she has visited that city, the word-sentence network [25,66], the bank-company network [145] or the donor-acceptor network [142] (see section 3.6 for a detailed review on bipartite networks).

Several models have been proposed in literature to synthesize the structure of these bipartite networks, i.e., when both the partitions grow unboundedly with time [7, 66, 123, 130, 161]. The results of such growth models indicate that when an incoming movie node (in case of movie-actor networks) *preferentially* attaches itself to an actor node, the emergent degree distribution of the actor nodes follows a power-law (see [130] for details). This result is reminiscent of unipartite networks where *preferential attachment* leads to the emergence of power-law degree distributions (see [13] for details).

Although there have been some studies on non-growing bipartite networks [118, 50], those like PlaNet where one of the partitions remain fixed over time (i.e., the partition of consonants) while the other grows (i.e., the partition of languages) have received much less attention. Therefore, the primary objective of this chapter is to systematically analyze as well as synthesize the structure of PlaNet and thereby, explain the occurrence distribution of the consonants across languages.

The rest of the chapter is organized as follows. In section 3.1, we present the formal definition of PlaNet, describe the data source and outline its construction procedure. We analyze some interesting topological properties of PlaNet in the following section. In section 3.3, we present a synthesis model that can, quite accurately, reproduce the structure of PlaNet. The next section presents an analytical solution for the proposed synthesis model after certain simplifications. In section 3.5, we further construct five different networks that respectively represent the consonant inventories belonging to the five major language families namely, the Indo-European (IE-PlaNet), the Afro-Asiatic (AA-PlaNet), the Niger-Congo (NC-PlaNet), the Austronesian (AN-PlaNet) and the Sino-Tibetan (ST-PlaNet). We analyze as well as synthesize these networks in order to examine the dynamics within and across the language families. In the next section, we present a detailed review on various studies related to complex bipartite networks and, wherever possible, draw a comparison between them and the one proposed by us here. In section 3.7, we summarize some of the important contributions of this chapter, outline a few linguistic interpretations of the model and identify certain limitations with reference to the network construction and synthesis, most of which are addressed in the next chapter.

3.1 Definition and Construction of PlaNet

In this section, we present a formal definition of PlaNet. This shall serve as the working definition throughout the rest of the thesis. We also present a detailed description of the data source used for the construction of PlaNet and discuss the methodology adopted for this construction.

3.1.1 Definition of PlaNet

PlaNet is a bipartite graph $G = \langle\, V_L, V_C, E_{pl}\, \rangle$ consisting of two sets of nodes namely, V_L (labeled by the languages) and V_C (labeled by the consonants); E_{pl} is the set of edges running between V_L and V_C. There is an edge $e \in E_{pl}$ from a node $v_l \in V_L$ to a node $v_c \in V_C$ iff the consonant c is present in the inventory of language l. Figure 3.1 presents a hypothetical example illustrating the nodes and edges of PlaNet.

3.1.2 The Data Source

The source of data for this work is the UCLA Phonological Segment Inventory Database (UPSID) [102]. The choice of this database is motivated by a large number

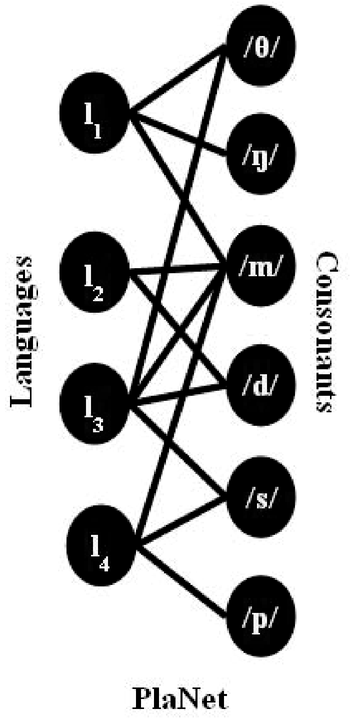


Figure 3.1: A hypothetical example illustrating the nodes and edges of PlaNet

of typological studies [44,69,88,96] that have been carried out in the past on UPSID. We have selected UPSID mainly due to two reasons – (a) it is the largest database of this type that is currently available and, (b) it has been constructed by selecting languages from moderately distant language families, which ensures a considerable degree of genetic balance.

The languages that are included in UPSID have been chosen in a way to approximate a properly constructed quota rule based on the genetic groupings of the world's extant languages. The quota rule is that only one language may be included from each small language family (e.g., one from the West Germanic and one from the North Germanic) but that each such family should be represented. Eleven major genetic groupings of languages along with several smaller groups have been considered while constructing the database. All these together add up to make a total of 317 languages in UPSID. Note that the availability as well as the quality of the phonological descriptions have been the key factors in determining the language(s) to be included from within a group; however, neither the number of speakers nor the phonological

Table 3.1: Some of the important features listed in UPSID

Manner of Articulation	Place of Articulation	Phonation
tap	velar	voiced
flap	uvular	voiceless
trill	dental	
click	palatal	
nasal	glottal	
plosive	bilabial	
r-sound	alveolar	
fricative	retroflex	
affricate	pharyngeal	
implosive	labial-velar	
approximant	labio-dental	
ejective stop	labial-palatal	
affricated click	dental-palatal	
ejective affricate	dental-alveolar	
ejective fricative	palato-alveolar	
lateral approximant		

peculiarity of a language has been considered.

Each consonant in UPSID is characterized by a set of articulatory features (i.e., place of articulation, manner of articulation and phonation) that distinguishes it from the other consonants. Certain languages in UPSID also consist of consonants that make use of secondary articulatory features apart from the basic ones. There are around 52 features listed in UPSID; the important ones are noted in Table 3.1. Note that in UPSID the features are assumed to be binary-valued (1 meaning the feature is present and 0 meaning it is absent) and therefore, each consonant can be represented by a binary vector.

Over 99% of the UPSID languages have bilabial (e.g., /p/), dental-alveolar (e.g., /t/) and velar (e.g., /k/) plosives. Furthermore, voiceless plosives outnumber the voiced ones (92% vs. 67%). According to [101], languages are most likely to have

8 to 10 plosives; nevertheless, the scatter is quite wide and only around 29% of the languages fall within the mentioned limits. 93% of the languages have at least one fricative (e.g., /f/). However, as [101] points out, the most likely number of fricatives is between 2 to 4 (around 48% of the languages fall within this range). 97% of the languages have at least one nasal (e.g., /m/); the most likely range reported in [101] is 2 to 4 and around 48% of the languages in UPSID are in this range. In 96% of the languages there is at least one liquid (e.g., /l/) but, languages most likely have 2 liquids (around 41%) [101]. Approximants (e.g., /j/) occur in fewer than 95% of the languages; however, languages are most likely to have 2 approximants (around 69%) [101]. About 61% of the languages in UPSID have the consonant /h/, which is not included in any of the categories already mentioned above. Some of the most frequent consonants in UPSID are, /p/, /b/, /t/, /d/, /tʃ/, /k/, /g/, /ʔ/, /f/, /s/, /ʃ/, /m/, /n/, /ɲ/, /ŋ/, /w/, /l/, /r/, /j/, /h/, and together they are often termed as the 'modal' inventory [101].

It is important to mention here that there are certain criticisms of this database especially related to representation of the phonemes [156]. The phoneme inventories in UPSID are represented using a feature-based classificatory system developed by the phoneticians primarily through the inspection of various observable facts about language (see [88] for a vivid description of this design methodology). Although there are questions regarding the authenticity of this representation mostly related to the existence of (abstract) features [156], in absence of any other alternative resource for the validation of our computational models we had to resort to UPSID. Note that it is hard to find what exactly could be a true representation of the phonemes and there is no consensus on such an issue even among the experts in the field. Nevertheless, the representation of UPSID can be at least said to be quite "faithful" if not the "true". This is because, numerous studies on this database show that various patterns reflected by it actually correlate perfectly with what is observed in nature. The results presented in this thesis further brings forth certain universal qualities of the feature-based classificatory system for describing the consonant and the vowel inventories, which do not manifest in case of the randomly constructed inventories. Most importantly, the structural regularities reported in the thesis were not presumed by the phoneticians while designing the classificatory system. Therefore, these nontrivial findings possibly point to universal properties of real languages that are getting reflected only because the classificatory system turns out to be a very appropriate way of representing the inventories. We understand that the statistics that we present might change if the experiments are carried out on a different data set. Therefore, we do not claim that the inferences drawn here are sacrosanct; rather they are only indicative. In this context, the trends in the results outlined here are more important than the exact values. We believe that for any choice of the data set the trends should remain similar and this being an interesting future research question related to the evolution of sound inventories, our results definitely have a crucial role in propelling it forward.

3.1.3 Construction Methodology

We have used UPSID in order to construct PlaNet. Consequently, the total number of language nodes in PlaNet (i.e., $|V_L|$) is 317. The total number of distinct consonants found across the 317 languages of UPSID, after appropriately filtering the *anomalous* and the *ambiguous* ones [102], is 541. In UPSID, a phoneme has been classified as anomalous if its existence is doubtful and ambiguous if there is insufficient information about the phoneme. For example, the presence of both the palatalized dental plosive and the palatalized alveolar plosive are represented in UPSID as palatalized dental-alveolar plosive (an ambiguous phoneme). According to popular techniques [124], we have completely ignored the anomalous phonemes from the data set, and included all the ambiguous forms of a phoneme as separate phonemes because, there are no descriptive sources explaining how such ambiguities might be resolved. Therefore, the total number of consonant nodes in PlaNet (i.e., $|V_C|$) is 541.

The number of edges in PlaNet (i.e., $|E_{pl}|$) is 7022. Thus, the connection density of PlaNet is $\frac{|E_{pl}|}{|V_L||V_C|} = \frac{7022}{317 \times 541} = 0.06$, which can also be thought of as the probability that a randomly chosen consonant occurs in a particular language. However, as we shall see below, the occurrence of the consonants does not depend upon a single probability value; rather, it is governed by a well-behaved probability distribution.

3.2 Topological Properties of PlaNet

In this section, we shall study the topological properties of PlaNet mainly in terms of the degree distributions of its two sets of nodes.

3.2.1 Degree of a Node

The degree of a node v, denoted by k_v, is the number of edges incident on v. Therefore, the degree of a language node v_l in PlaNet refers to the size of the consonant inventory of the language l. Similarly, the degree of a consonant node v_c in PlaNet refers to the frequency of occurrence of the consonant c across the languages of UPSID.

3.2.2 Degree Distribution of PlaNet

The degree distribution is the fraction of nodes, denoted by p_k, that have a degree equal to k [114]. In other words, it is the probability that a node chosen uniformly at random from the network (with N nodes) has a degree equal to k. The cumulative degree distribution P_k is the fraction of nodes having degree greater than or equal to k. Therefore,

$$P_k = \sum_{k'=k}^{\infty} p_{k'} \tag{3.1}$$

Note that the cumulative distribution is more robust to noise present in the observed data points, but at the same time it contains all the information encoded by p_k [114].

Degree Distribution of the Language Nodes

Figure 3.2 shows the degree distribution of the nodes in V_L where the x-axis denotes the degree of each language node expressed as a fraction of the maximum degree and

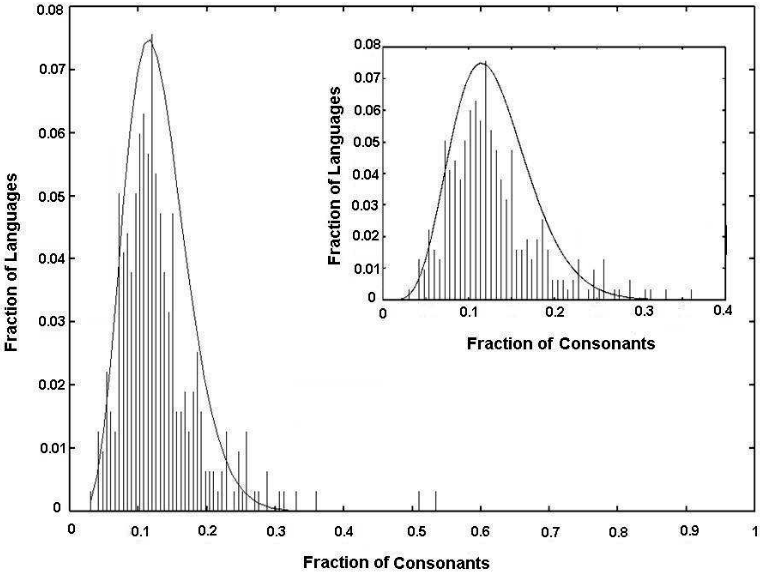


Figure 3.2: Degree distribution of the language nodes in PlaNet. The figure in the inset is a magnified version of a portion of the original figure

the y-axis denotes the fraction of nodes having a given degree.

Figure 3.2 indicates that the number of consonants appearing in different languages follow a β-distribution[1] (see [22] for reference) which is right skewed with the values of α and β equal to 7.06 and 47.64 (obtained using maximum likelihood estimation method) respectively. This asymmetry in the distribution points to the fact that languages usually tend to have smaller consonant inventory size, the best value being somewhere between 10 and 30. The distribution peaks roughly at 21 (which is its mode) while the mean of the distribution is also approximately 21 indicating that on an average the languages in UPSID have a consonant inventory of size 21 (approx.) [103].

[1] A random variable is said to have a β-distribution with parameters $\alpha > 0$ and $\beta > 0$ if and only if its probability mass function is given by, $f(x)= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$ for $0 < x < 1$ and $f(x)= 0$ otherwise. $\Gamma(\cdot)$ is the Euler's gamma function.

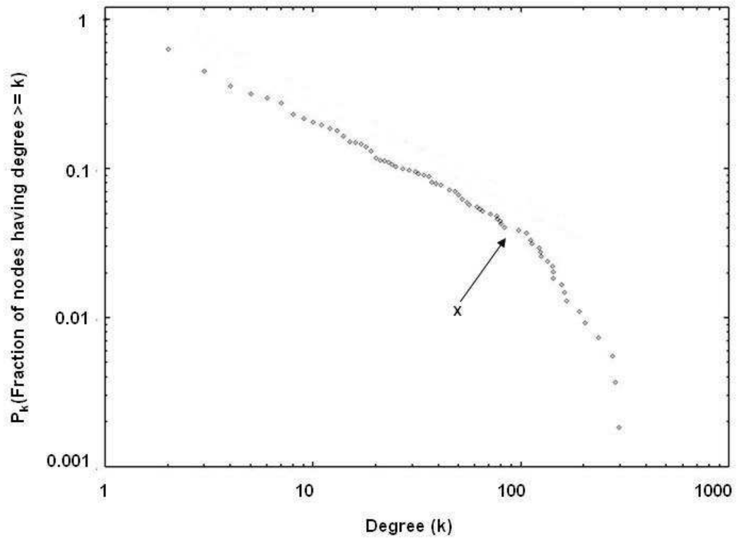


Figure 3.3: Degree distribution of the consonant nodes in PlaNet in doubly-logarithmic scale. The letter **x** denotes the cut-off point

Degree Distribution of the Consonant Nodes

Figure 3.3 illustrates the degree distribution plot for the consonant nodes in V_C in doubly-logarithmic scale. In this figure the x-axis represents the degree k and the y-axis represents the distribution P_k.

Figure 3.3 indicates that P_k follows a well-behaved distribution along with an exponential cut-off towards the tail. The cut-off point is shown by the letter **x** in the figure. We find that there are 22 consonant nodes which have their degree above the cut-off range (i.e., these are the extremely frequent consonants). In the forthcoming sections, we shall further explore the nature of this distribution.

An immediate question that follows is that what could be a possible reason for the emergence of this degree distribution. In most of the networked systems like the society, the Internet, the World Wide Web, and many others, *preferential attachment* (i.e., when "the rich gets richer") [13, 141] is known to play a crucial role in generating such distributions. With reference to PlaNet this preferential attachment can be interpreted as the heterogeneity in the choice of consonants by the speakers over linguistic generations. Consonants belonging to languages that are more prevalent among the speakers in one generation have higher chances of being transmitted to the speakers of languages of the subsequent generations than other consonants (see [18] for similar observations). Therefore, it may be argued that preferential attachment is a manifestation of the heterogeneity in the choice of consonants by the language speakers. In the next section, we attempt to develop a growth model based on preferential attachment coupled with a tunable randomness component that can mimic the degree distribution of the consonant nodes in PlaNet to a considerable extent.

3.3 The Synthesis Model

In this section, we present a synthesis model for PlaNet based on preferential attachment coupled with a tunable randomness component, where the distribution of the consonant inventory size, i.e., the degrees of the language nodes, is assumed to be known *a priori*. Note that this shall be a working assumption for all the synthesis models presented in the rest of the thesis.

Let us denote the degree of a language node $L_i \in V_L$ by d_i. The consonant nodes in V_C are assumed to be unlabeled, i.e., they are not marked by the distinctive features that characterize them. We first sort the nodes L_1 through L_{317} in the ascending order of their degrees. At each time step a node L_j, chosen in order, preferentially attaches itself with d_j *distinct* nodes (call each such node C_i) of the set V_C. The probability $Pr(C_i)$ with which the node L_j attaches itself to the node C_i is given by,

$$Pr(C_i) = \frac{\gamma k_i + 1}{\sum_{i' \in V_C'} (\gamma k_{i'} + 1)} \tag{3.2}$$

where, k_i is the current degree of the node C_i, $V_C^{'}$ is the set of nodes in V_C that are not already connected to L_j and γ is the tunable parameter that controls the amount of randomness in the system. The lower the value of γ the higher is the randomness. Note that $1/\gamma$ is a positive constant usually referred to as the *initial attractiveness* [47]. Algorithm 3.1 summarizes the mechanism to generate the syn-

Algorithm 3.1: Synthesis model based on preferential attachment

```
Input: Nodes $L_1$ through $L_{317}$ sorted in an increasing order of their degrees
for $t$ = 1 to 317 do
    Choose (in order) a node $L_j$ with degree $d_j$;
    for $c$ = 1 to $d_j$ do
        Connect $L_j$ to a node $C_i \in V_C$ to which it is not already connected
        following the distribution, $Pr(C_i) = \frac{\gamma k_i + 1}{\sum_{i' \in V_C'} (\gamma k_{i'} + 1)}$ where $V_C'$ is the set of
        nodes in $V_C$ (inclusive of $C_i$) to which $L_j$ is not yet connected, $k_i$ is the
        current degree of node $C_i$ and $\gamma$ is the tunable parameter;
    end
end
```

thesized version of PlaNet (henceforth PlaNet$_{syn}$) and Figure 3.4 illustrates a partial step of the synthesis process. In the figure, when language l_4 has to connect itself with one of the nodes in the set V_C it does so with the one having the highest degree (=3) rather than with others in order to achieve preferential attachment which is the working principle of our algorithm.

Apart from the ascending order, we have also simulated the model with descending and random order of the inventory size. The degree distribution obtained by considering the ascending order of the inventory size, matches much more accurately than in the other two scenarios. One possible reason for this might be as follows. With each consonant is associated two different frequencies: (a) the frequency of occurrence of a consonant over languages or the *type* frequency, and (b) the frequency of usage of the consonant in a particular language or the *token* frequency. Researchers have shown in the past that these two frequencies are positively correlated [23]. Nevertheless, our synthesis model based on preferential attachment takes into account only the type frequency of a consonant and not its token frequency. If language is considered to be an evolving system then both of these frequencies, in one generation, should play an important role in shaping the inventory structure of the next generation.

In the later stages of our synthesis process when the attachments are strongly preferential, the type frequencies span over a large range and automatically compen-

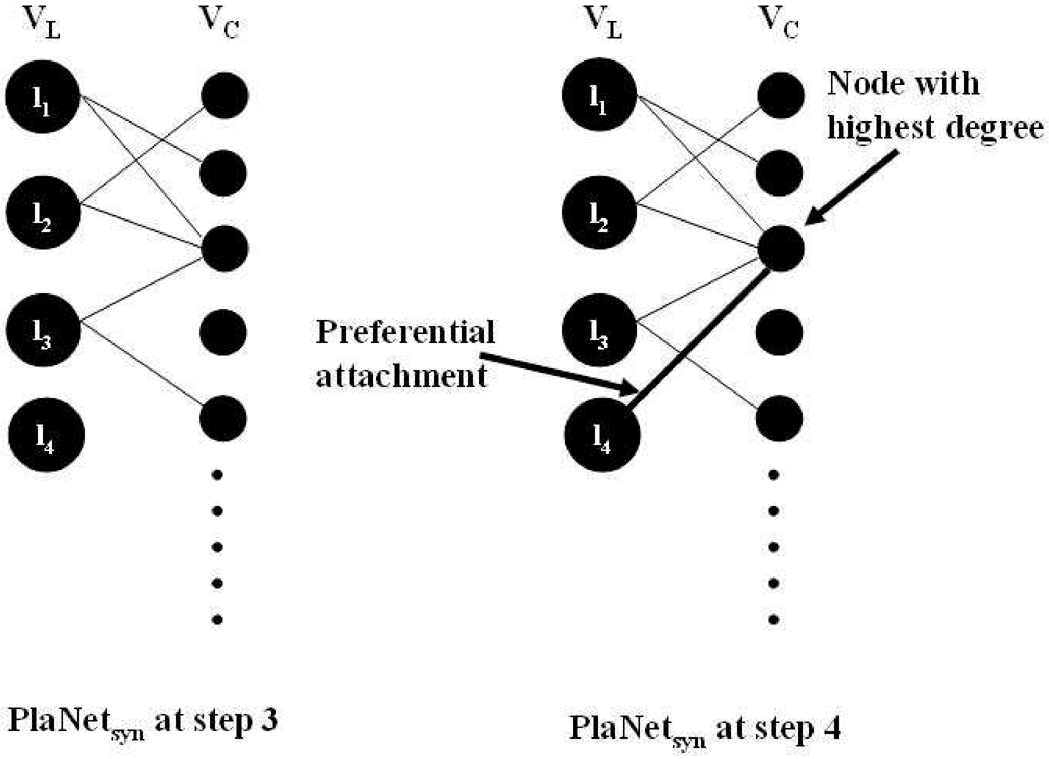


Figure 3.4: A partial step of the synthesis process

sate for the absence of the token frequencies (since they are positively correlated). However, in the initial stages of this process the attachments that take place are random in nature and therefore, the type frequencies of all the nodes are roughly equal. At this point it is the token frequency (absent in our model) that should discriminate between the nodes. This error due to the loss of information of the token frequency in the initial steps of the synthesis process can be minimized by allowing only a small number of attachments (so that there is less spreading of the error). This is perhaps the reason why sorting the language nodes in the ascending order of their degree helps in obtaining better results.

Simulation Results

We simulate the above model to obtain PlaNet$_{syn}$ for 100 different runs and average the results over all of them. We find that the degree distributions that emerge, fit the

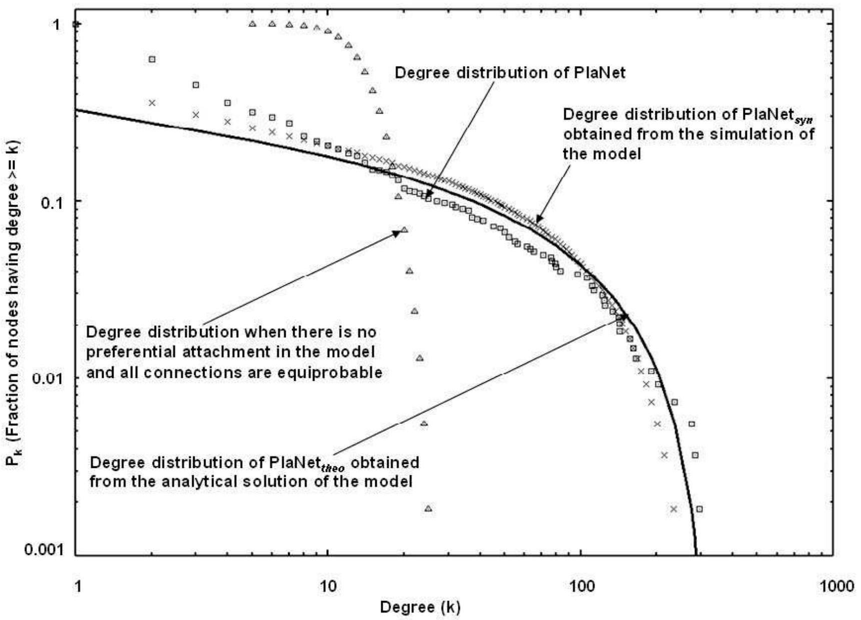


Figure 3.5: Comparison of the degree distribution (in doubly-logarithmic scale) of the consonant nodes in PlaNet with that of (a) PlaNet$_{syn}$ obtained from the simulation of the model ($\gamma = 14$) and (b) PlaNet$_{theo}$ obtained from the analytical solution of the model ($\gamma = 14$). The results are also compared with the case where there is no preferential attachment and all the connections are equiprobable

empirical data well for $\gamma \in [12.5,16.7]$, the best being at $\gamma = 14$ (shown in Figure 3.5). In fact, the mean error[2] between the real and the synthesized distributions for $\gamma = 14$ is as small as 0.03. In contrast, if there is no preferential attachment and all the connections to the consonant nodes are equiprobable (see Figure 3.5), then this error rises to 0.35.

[2] Mean error is defined as the average difference between the ordinate pairs (say y and $y^{'}$) where the abscissas are equal. In other words, if there are Y such ordinate pairs then the mean error can be expressed as $\frac{\sum|y-y^{'}|}{Y}$

3.4 The Analytical Solution for the Model

In this section, we attempt to analytically solve the model[3] presented above in order to derive a closed form expression for the degree distribution of the consonant nodes. We shall refer to this analytically derived version of PlaNet as PlaNet$_{theo}$. Let $p_{k,t}$ denote the probability that a randomly chosen node from the partition V_C has degree k after t time steps. It is difficult to solve this model because, unlike the popular preferential attachment based synthesis models for unipartite [13] and bipartite [130] networks, in this case, one cannot make the stationary state assumption $p_{k,t+1} = p_{k,t}$ in the limit $t \to \infty$. This is due to the fact that the average degree of the nodes in V_C diverges with time and consequently, the system does not have a stationary state.

Nevertheless, for certain simplifications of the model we can derive an approximate closed form expression for the degree distribution of the V_C partition of PlaNet$_{theo}$. More specifically, we assume that the degree of the nodes in the V_L partition is equivalent to their average degree and is therefore, a constant (μ). In other words, μ represents the average size of a consonant inventory or the average number of consonants present in human languages. We further assume that in a time step a language node can attach itself to a consonant node more than once. Although by definition, a consonant can occur in the inventory of a language only once, as we shall see, the result derived with the above assumption matches fairly well with the empirical data.

Under the assumptions mentioned above the denominator of the equation 3.2 can be re-written as $\sum_{i=1}^{N}(\gamma k_i + 1)$ where $N = |V_C|$. Further, since the sum of the degrees in the V_L partition after t steps ($= \mu t$) should be equivalent to that in the V_C partition therefore we have

$$\sum_{i=1}^{N} k_i = \mu t \tag{3.3}$$

[3] This is a joint work that has been carried out with two colleagues, Fernando Peruani and Monojit Choudhury, and one of my supervisors.

Notice that the average degree of the nodes in V_C after t steps is $\mu t/N$ which, as we have pointed out earlier, diverges with t because, N is fixed over time.

At time $t = 0$, all the nodes in V_C have a degree zero and therefore our initial condition is $p_{k,t=0} = \delta_{k,0}$, where $\delta_{k,0}$ is the Kronecker delta function [74]. We shall now solve the model for $\mu = 1$ and then generalize for the case $\mu > 1$.

3.4.1 Solution for $\mu = 1$

Since $\mu = 1$, at each time step a node in the V_L partition essentially brings a single incoming edge as it enters the system. The evolution of $p_{k,t}$ can be expressed as

$$p_{k,t+1} = (1 - \widetilde{P}(k,t))p_{k,t} + \widetilde{P}(k-1,t)p_{k-1,t} \tag{3.4}$$

where $\widetilde{P}(k,t)$ refers to the probability that the incoming edge lands on a consonant node of degree k at time t. $\widetilde{P}(k,t)$ can be easily derived for $\mu = 1$ using together the equations 3.2 and 3.3 and takes the form

$$\widetilde{P}(k,t) = \begin{cases} \frac{\gamma k+1}{\gamma t+N} & \text{for} \quad 0 \leq k \leq t \\ 0 & \text{otherwise} \end{cases} \tag{3.5}$$

for $t > 0$ while for $t = 0$, $\widetilde{P}(k,t) = \frac{1}{N}\delta_{k,0}$.

Equation 3.4 can be explained as follows. The probability of finding a consonant node with degree k at time $t+1$ decreases due to those nodes, which have a degree k at time t and receive an edge at time $t+1$ therefore acquiring degree $k+1$, i.e., $\widetilde{P}(k,t)p_{k,t}$. Similarly, this probability increases due to those nodes that at time t have degree $k-1$ and receive an edge at time $t+1$ to have a degree k, i.e., $\widetilde{P}(k-1,t)p_{k-1,t}$. Hence, the net increase in the value of $p_{k,t+1}$ can be expressed by the equation 3.4.

In order to have an exact analytical solution of the equation 3.4 we express it as a product of matrices

$$\mathbf{p}_{t+1} = \mathbf{M}_t \mathbf{p}_t = [\prod_{\tau=0}^{t} \mathbf{M}_\tau] \mathbf{p}_0 \tag{3.6}$$

where $\mathbf{p}_t$ denotes the degree distribution at time t and is defined as $\mathbf{p}_t = [p_{0,t}\, p_{1,t}\, p_{2,t}\ \ldots]^T$ (T stands for the standard transpose notation for a matrix), $\mathbf{p}_0$ is the initial condition expressed as $\mathbf{p}_0 = [1\ 0\ 0\ \ldots]^T$ and $\mathbf{M}_\tau$ is the evolution matrix at time τ which is defined as

$$\mathbf{M}_\tau = \begin{pmatrix} 1-\widetilde{P}(0,\tau) & 0 & 0 & 0 & \ldots \\ \widetilde{P}(0,\tau) & 1-\widetilde{P}(1,\tau) & 0 & 0 & \ldots \\ 0 & \widetilde{P}(1,\tau) & 1-\widetilde{P}(2,\tau) & 0 & \ldots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix} \tag{3.7}$$

Let us further define a matrix $\mathbf{H}_t$ as follows.

$$\mathbf{H}_0 = \mathbf{M}_0 \tag{3.8}$$

$$\mathbf{H}_t = \mathbf{M}_t \mathbf{H}_{t-1} = [\prod_{\tau=0}^{t} \mathbf{M}_\tau] \tag{3.9}$$

Thus we have,

$$\mathbf{p}_{t+1} = \mathbf{H}_t \mathbf{p}_0 \tag{3.10}$$

Since our initial condition (i.e., $\mathbf{p}_0$) is a matrix of zeros at all positions except the first row therefore, all the relevant information about the degree distribution of the consonant nodes is encoded by the first column of the matrix $\mathbf{H}_t$. The $(k+1)^{\text{th}}$ element of this column essentially corresponds to $p_{k,t}$. Let the entry corresponding to the i^{th} row and the j^{th} column of $\mathbf{H}_t$ and $\mathbf{M}_t$ be denoted by $h^t_{i,j}$ and $m^t_{i,j}$ respectively.

On successive expansion of $\mathbf{H}_t$ using the recursive definition provided in equation 3.9, we get (see Figure 3.6 for an example)

$$h^t_{i,j} = m^t_{i,i-1} h^{t-1}_{i-1,j} + m^t_{i,i} h^{t-1}_{i,j} \tag{3.11}$$

or,

$$h^t_{i,j} = (m^t_{i,i-1} m^{t-1}_{i-1,i-2}) h^{t-2}_{i-2,j} + (m^t_{i,i-1} m^{t-1}_{i-1,i-1} + m^t_{i,i} m^{t-1}_{i,i-1}) h^{t-2}_{i-1,j} + m^t_{i,i} m^{t-1}_{i,i} h^{t-2}_{i,j} \tag{3.12}$$

Since the first column of the matrix $\mathbf{H}_t$ encodes the degree distribution, it suffices to calculate the values of $h^t_{i,1}$ in order to estimate $p_{k,t}$. In fact, $p_{k,t}$ (i.e., the $(k+1)^{\text{th}}$ entry of $\mathbf{H}_t$) is equal to $h^t_{k+1,1}$. In the following, we shall attempt to expand certain values of $h^t_{k+1,1}$ in order to detect the presence of a pattern (if any) in these values. In particular, let us investigate two cases of $h^1_{2,1}$ and $h^2_{2,1}$ from Figure 3.6. We have

$$h^1_{2,1} = m^1_{2,1} h^0_{1,1} + m^1_{2,2} h^0_{2,1} = \left(1 - \frac{1}{N}\right)\left(\frac{1}{\gamma + N}\right) + \left(\frac{N-1}{\gamma + N}\right)\left(\frac{1}{N}\right) \tag{3.13}$$

or,

$$h^1_{2,1} = 2\frac{(N-1)}{(\gamma + N)N} \tag{3.14}$$

Similarly,

$$h^1_{2,1} = m^2_{2,1} m^1_{1,1} h^0_{1,1} + m^2_{2,2} m^1_{2,1} h^0_{1,1} + m^2_{2,2} m^1_{2,2} h^0_{2,1} \tag{3.15}$$

or,

$$h^1_{2,1} = 3\frac{(\gamma + N - 1)(N-1)}{(2\gamma + N)(\gamma + N)N} \tag{3.16}$$

A closer inspection of equations 3.14 and 3.16 reveals that the pattern of evolution

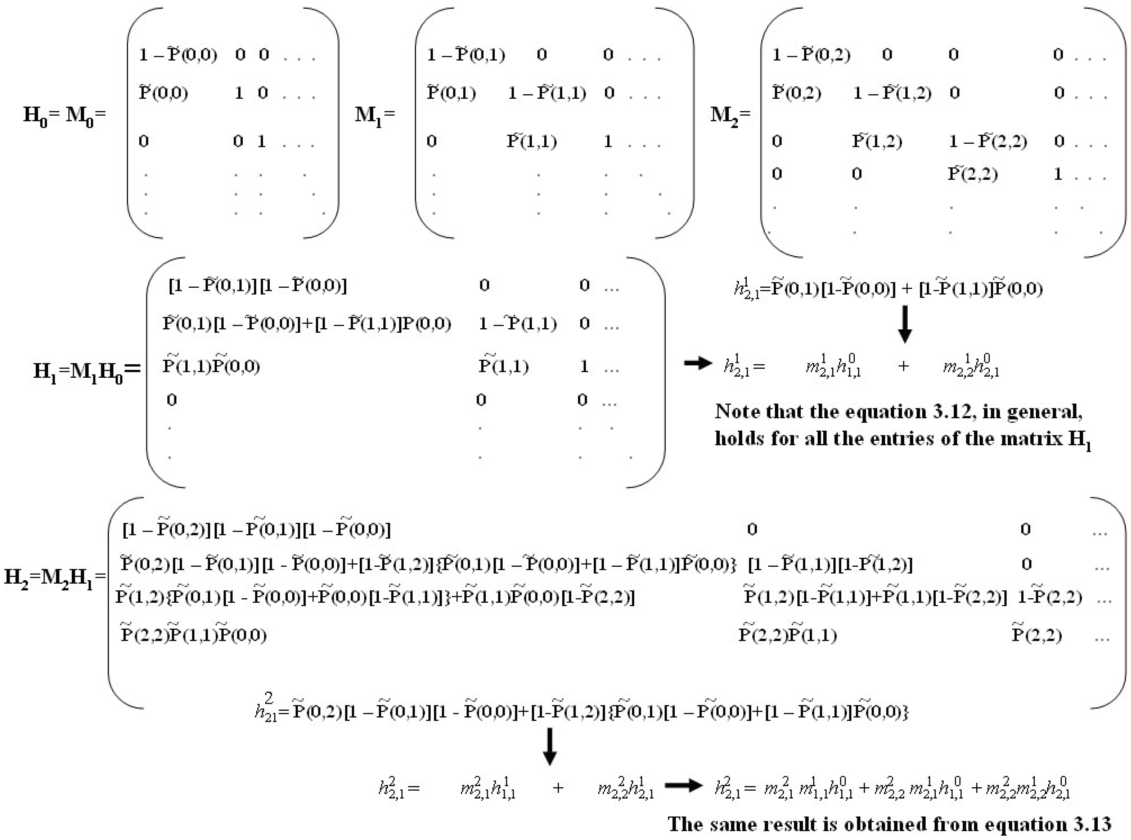

$H_0 = M_0 =$
$M_1 =$
$M_2 =$
$H_1 = M_1 H_0 =$
$H_2 = M_2 H_1 =$
$h^1_{2,1} = m^1_{2,1} h^0_{1,1} + m^1_{2,2} h^0_{2,1}$
Note that the equation 3.12, in general, holds for all the entries of the matrix H_1
$h^2_{2,1} = m^2_{2,1} h^1_{1,1} + m^2_{2,2} h^1_{2,1}$
$h^2_{2,1} = m^2_{2,1} m^1_{1,1} h^0_{1,1} + m^2_{2,2} m^1_{2,1} h^0_{1,1} + m^2_{2,2} m^1_{2,2} h^0_{2,1}$
The same result is obtained from equation 3.13

Figure 3.6: A few steps showing the calculations of equation 3.11 and 3.12

of this row, in general, can be expressed as

$$p_{k,t} = \left(\begin{array}{c} t \\ k \end{array} \right) \frac{\prod_{x=0}^{k-1} (\gamma x + 1) \prod_{y=0}^{t-1-k} (N - 1 + \gamma y)}{\prod_{w=0}^{t-1} (\gamma w + N)} \tag{3.17}$$

for $0 \leq k \leq t$ and $p_{k,t} = 0$ otherwise. Further, we define the special case $\prod_{z=0}^{-1}(\dots) = 1$. Note that if we now put $t = 2$, $k = 1$ and $t = 3$, $k = 1$ in 3.17 we recover equations 3.14 and 3.16 respectively.

Equation 3.17 is the exact solution of the equation 3.4 for the initial condition $p_{k,t=0} = \delta_{k,0}$. Therefore, this is the analytical expression for the degree distribution of the consonant nodes in PlaNet$_{theo}$ for $\mu = 1$.

In the limit $\gamma \rightarrow 0$ (i.e. when the attachments are completely random) equation 3.17 takes the form

$$p_{k,t} = \begin{pmatrix} t \\ k \end{pmatrix} \left(\frac{1}{N}\right)^{k} \left(1 - \frac{1}{N}\right)^{t-k} \tag{3.18}$$

for $0 \leq k \leq t$ and $p_{k,t} = 0$ otherwise.

On the other hand, when $\gamma \rightarrow \infty$ (i.e., when the attachments are completely preferential) the degree distribution of the consonant nodes reduces to

$$p_{k,t} = \left(1 - \frac{1}{N}\right) \delta_{k,0} + \frac{1}{N}\delta_{k,t} \tag{3.19}$$

3.4.2 Solution for $\mu > 1$

In the previous section, we have derived an analytical solution for the degree distribution of the consonant nodes in PlaNet$_{theo}$ specifically for $\mu = 1$. However, note that the value of μ is greater than 1 (approximately 21) for the real network (i.e., PlaNet). Therefore, one needs to analytically solve for the degree distribution for values of μ greater than 1 in order to match the results with the empirical data. Here we attempt to generalize the derivations of the earlier section for $\mu > 1$.

We assume that $\mu \ll N$ (which is true for PlaNet) and expect equation 3.4 to be a good approximation for the case of $\mu > 1$ after replacing $\widetilde{P}(k,t)$ by $\widehat{P}(k,t)$ where $\widehat{P}(k,t)$ is defined as

$$\widehat{P}(k,t) = \begin{cases} \frac{(\gamma k+1)\mu}{\mu\gamma t+N} & \text{for} \quad 0 \leq k \leq \mu t \\ 0 & \text{otherwise} \end{cases} \tag{3.20}$$

The term μ appears in the denominator of the equation 3.20 for $0 \leq k \leq \mu t$ because, in this case the total degree of the consonant nodes in PlaNet$_{theo}$ at any point in time is μt rather than t as in equation 3.5. The numerator contains a μ since at each time step there are μ edges that are being incorporated into the network rather than a single edge.

The solution of equation 3.4 with the attachment kernel defined in equation 3.20 can be expressed as

$$p_{k,t} = \left(\begin{array}{c} t \\ k \end{array} \right) \frac{\prod_{x=0}^{k-1} (\gamma x + 1) \prod_{y=0}^{t-1-k} (\frac{N}{\mu} - 1 + \gamma y)}{\prod_{w=0}^{t-1} (\gamma w + \frac{N}{\mu})} \tag{3.21}$$

for $0 \leq k \leq \mu t$ and $p_{k,t} = 0$ otherwise.

Given that $\mu \ll N$ we can neglect the term containing μ/N in the equation 3.21 and express the rest using factorials as

$$p_{k,t} = \frac{t!\eta!(t-k+\eta-\gamma^{-1})!(k-1+\gamma^{-1})!\gamma^{-1}}{(t-k)!k!(t+\eta)!(\eta-\gamma^{-1})!(\gamma^{-1})!} \tag{3.22}$$

where $\eta = N/\mu\gamma$. Approximating the factorials using Stirling's formula (see [1] for a reference), we get

$$p_{k,t} = \widetilde{A}(t,\gamma,\eta) \frac{(k-1+\gamma^{-1})^{k-1+\gamma^{-1}+0.5}(t-k+\eta-\gamma^{-1})^{t-k+\eta-\gamma^{-1}+0.5}}{k^{k+0.5}(t-k)^{t-k+0.5}} \tag{3.23}$$

where

$$\widetilde{A}(t,\gamma,\eta) = \frac{t^{t+0.5}\eta^{\eta+0.5}\gamma^{\gamma^{-1}-0.5}e}{\sqrt{2\pi}(t+\eta)^{t+\eta+0.5}(\eta-\gamma^{-1})^{\eta-\gamma^{-1}+0.5}} \tag{3.24}$$

is a term independent of k.

Since we are interested in the asymptotic behavior of the network such that t is very large, we may assume that $t \gg k \gg \eta > \gamma^{-1}$. Under this assumption, we can re-write the equation 3.23 in terms of the fraction k/t and this immediately reveals that the expression is approximately a β-distribution in k/t. More specifically, we have

$$p_{k,t} \approx \widehat{A}(t,\eta,\gamma)\mathrm{B}(k/t;\gamma^{-1},\eta-\gamma^{-1}) = \widehat{A}(t,\eta,\gamma)(k/t)^{\gamma^{-1}-1}(1-k/t)^{\eta-\gamma^{-1}-1} \quad (3.25)$$

where $\mathrm{B}(z;\alpha,\beta)$ refers to a β-distribution over variable z. We can generate different distributions by varying the value of γ in equation 3.25. We can further compute $P_{k,t}$ (i.e. the cumulative degree distribution) using equations 3.1 and 3.25 together.

Recall that in section 3.3 we have found through simulations that the best fit for the degree distribution emerges at $\gamma = 14$. Replacing μ by 21, t by 317, N by 541 and γ by 14 we obtain the degree distribution for the consonant nodes $P_{k,t}$ of PlaNet$_{theo}$. The bold line in Figure 3.5 illustrates the plot for this distribution in doubly-logarithmic scale. The figure indicates that the theoretical curve (i.e., the degree distribution of PlaNet$_{theo}$) matches quite well with the empirical data (i.e., the degree distribution of PlaNet). In fact, the mean error between the two curves in this case is as small as 0.03. It is worthwhile to mention here that since the degree distribution obtained from the simulation as well as the theoretical analysis of the model matches the real data for a very high value of γ there is a considerable amount of preferential attachment that goes on in shaping the emergent structure of PlaNet.

3.5 Dynamics of the Language Families

In this section, we investigate the dynamics within and across the consonant inventories of some of the major language families of the world. More specifically, for our investigation, we choose five different families namely the Indo-European, the Afro-Asiatic, the Niger-Congo, the Austronesian and the Sino-Tibetan. We manually sort the languages of these five groups from the data available in UPSID. Note that we have included a language in any group if and only if we could find a direct evidence of its presence in the corresponding family. We next present a brief description of each of these groups[4] and list the languages from UPSID that are found within them.

Indo-European: This family includes most of the major languages of Europe and south, central and south-west Asia. Currently, it has around 3 billion native speakers, which is largest among all the recognized families of languages in the world. The total number of languages appearing in this family is 449. The earliest evidences of the Indo-European languages have been found to date 4000 years back.

Languages: Albanian, Bengali, Breton, Bulgarian, Farsi, French, German, Greek, Hindi/Urdu, Irish, Kashmiri, Kurdish, Lithuanian, Norwegian, Pashto, Romanian, Russian, Sinhalese, Spanish[5].

Afro-Asiatic: Afro-Asiatic languages have about 200 million native speakers spread over north, east, west, central and south-west Africa. This family is divided into five subgroups with a total of 375 languages. The proto-language of this family began to diverge into separate branches approximately 6000 years ago.

Languages: Amharic, Angas, Arabic, Awiya, Dera, Dizi, Hamer, Hausa, Iraqw, Kanakuru, Kefa, Kullo, Margi, Ngizim, Shilha, Socotri, Somali.

Niger-Congo: Majority of the languages that belong to this family are found in the sub-Saharan parts of Africa. The number of native speakers is around 300 million and the total number of languages is 1514. This family descends from a proto-language, which dates back 5000 years.

Languages: Akan, Amo, Bambara, Bariba, Beembe, Birom, Bisa, Cham, Dagbani, Dan, Diola, Doayo, Efik, Ga, Gbeya, Igbo, Ik, Kadugli, Koma, Kpelle, Lelemi, Moro, Senadi, Tampulma, Tarok, Teke, Temne, Wolof, Zande, Zulu.

[4] Most of the information has been collected from the Ethnologue: http://www.ethnologue.com/ and the World Atlas of Language Structures: http://wals.info/.

[5] Interestingly, while preparing this set of Indo-European languages from UPSID, we did not find English.

Austronesian: The languages of the Austronesian family are widely dispersed throughout the islands of south-east Asia and the Pacific. There are 1268 languages in this family, which are spoken by a population of 6 million native speakers. Around 4000 years back it separated out from its ancestral branch.

Languages: Adzera, Batak, Chamorro, Hawaiian, Iai, Javanese, Kaliai, Malagasy, Roro, Rukai, Tsou, Tagalog.

Sino-Tibetan: Most of the languages in this family are distributed over the entire east Asia. With a population of around 2 billion native speakers it ranks second after Indo-European. The total number of languages in this family is 403. Some of the first evidences of this family can be traced 6000 years back.

Languages – Ao, Burmese, Dafla, Hakka, Jingpho, Karen, Lahu, Mandarin, Taishan.

We use the consonant inventories of the language families listed above to construct five bipartite networks – IE-PlaNet (for Indo-European family), AA-PlaNet (for Afro-Asiatic family), NC-PlaNet (for Niger-Congo family), AN-PlaNet (for Austronesian family) and ST-PlaNet (for Sino-Tibetan family). The number of nodes and edges in each of these networks are noted in Table 3.2.

Table 3.2: Number of nodes and edges in the four bipartite networks corresponding to the four language families

Networks	$\|V_L\|$	$\|V_C\|$	$\|E_{pl}\|$
IE-PlaNet	19	148	534
AA-PlaNet	17	123	453
NC-PlaNet	30	135	692
AN-PlaNet	12	82	221
ST-PlaNet	9	71	201

We next attempt to fit the degree distribution of the five empirical networks with the analytical expression derived for $P_{k,t}$ in the previous section. For all the experiments, we set $N = 541$, $t =$ number of languages in the family under investigation and $\mu =$ average degree of the language nodes in the PlaNet representing the family under

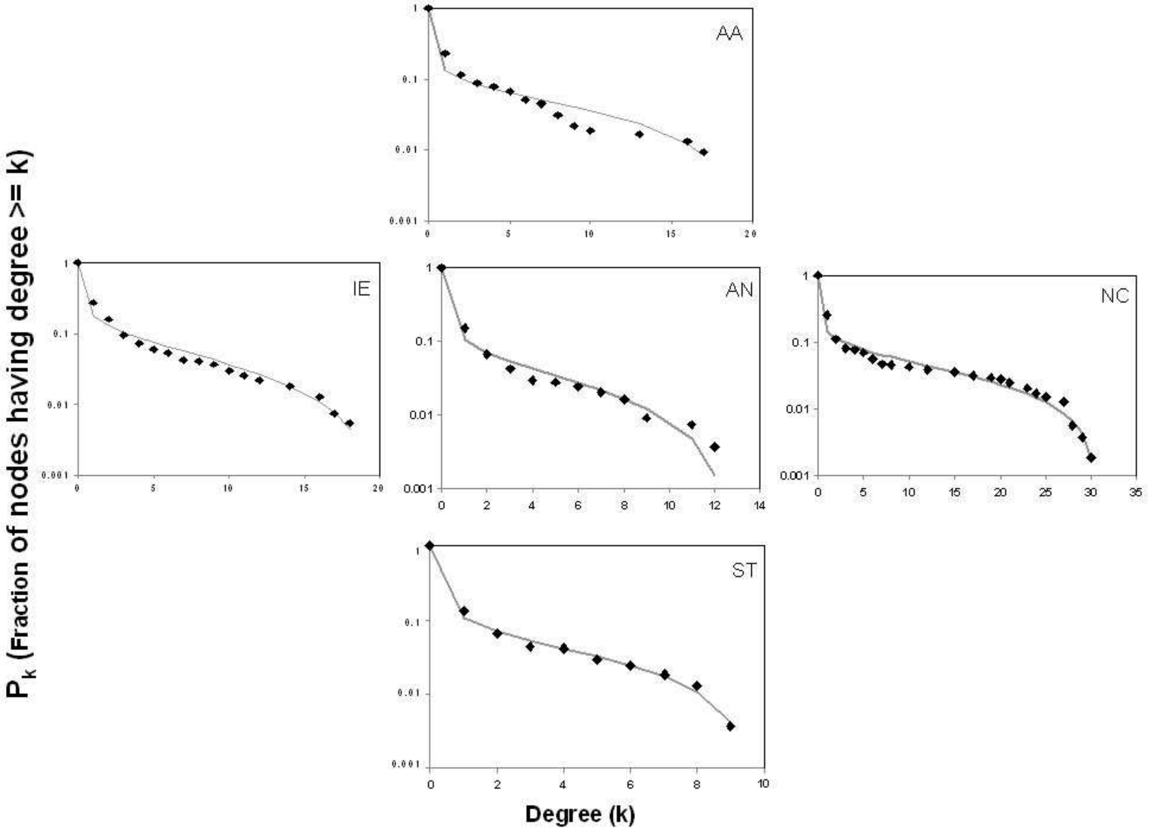


Figure 3.7: The degree distribution of the different real networks (black dots) along with the best fits obtained from the analytical expression for $P_{k,t}$ (grey lines). For all the plots the y-axis is in log-scale

investigation. Therefore, given the value of k we can compute $p_{k,t}$ and consequently, $P_{k,t}$, if γ is known. We vary the value of γ such that the mean error between the degree distribution of the real network and the equation is least. The best fits obtained for each of the five networks are shown in Figure 3.7. The values of γ corresponding to these fits are noted in Table 3.3.

The results indicate that the value of γ for PlaNet is lower than that of all the individual networks corresponding to the language families. Therefore, it may be argued that the preferential component within a language family is stronger than across families. Note that this is true only for real linguistic families and not for any arbitrary group of languages. In fact, if one randomly selects a set of inventories to represent a family then for a large number of such sets the average value of γ is 14.7

Table 3.3: The values of γ obtained for the best fits for each family together with the age of the families

Families	γ	Age (in years)
Austronesian	33.3	4000
Niger-Congo	28.6	5000
Sino-Tibetan	28.6	6000
Afro-Asiatic	26.0	6000
Indo-European	18.0	4000 (or 8000)

which is close to that of PlaNet.

We further observe a very interesting positive correlation between the approximate age of the language family and the values of γ obtained in each case (see Table 3.3). The only anomaly is the Indo-European branch, which possibly indicates that this might be much older than it is believed to be. In fact, a recent study [12] has shown that the age of this family dates back to 8000 years. If this last argument is assumed to be true then the values of γ have a one-to-one correspondence with the approximate period of existence of the language families. As a matter of fact, this correlation can be intuitively justified – higher is the period of existence of a family higher are the chances of its diversification into smaller subgroups, which in turn increases the randomness of the system and therefore, the values of γ are found to be less for the older families.

3.6 Review on Bipartite Networks

A large number of real-world systems can be naturally modeled as a bipartite network. One of the most important examples are the social collaboration networks that are generally defined in terms of a set of people (known as *actors* in the social science literature) and a set of *collaboration acts*. Consequently, the bipartite network in this case is composed of two sets of vertices one corresponding to the actors and the other to the acts of collaboration. In the following, we shall review some of the well-studied collaboration networks namely the movie-actor network, the network of scientific collaborations and the network of company board directors sitting on the same board.

Movie-Actor Network: The two partitions of a movie-actor network are composed of the movies and the actors respectively and an edge signifies that a particular actor acted in a particular movie cast. In [123, 130], the authors constructed the movie-actor network from the Internet Movie Database (IMDB) and studied the degree distributions of both the movie and the actor partition. The distribution of the movie cast size (i.e., the number of actors in a movie) was found to exhibit an exponential decay. On the other hand, the number of movies in which an actor has played adjusted better to a power-law fit with an exponent close to 2.

Scientific Collaboration Network: In a scientific collaboration network, the two partitions correspond to scientists and scientific articles while an edge denotes that a particular scientist has (co)-authored a particular article. In [113, 123], the authors analyzed the collaboration networks of various scientific communities and showed that the distribution of the number of authors in a given article is exponential in nature. The distribution of the number of articles written by an author was found to roughly exhibit a power-law behavior.

Board-Director Network: In a board-director network, the two respective partitions are the boards of different companies and the directors while an edge signifies that a particular director sits on the board of a particular company. The properties of board-director networks have been extensively studied in [16, 130]. The results show that the distribution of the number of boards on which a single director serves can be adjusted by an exponentially decaying function.

The concept of collaboration has also been extended to model various other phenomena such as the city-people network [49], the bank-company network [145] and the word-meaning network [29]. A common observation is that these networks exhibit a scale-free topology and in particular, the degree distribution of the actor nodes follow a power-law behavior.

In order to explain the emergence of the degree distribution of the actor nodes the authors in [130] proposed a network growth model based on preferential attachment. The model can be described using three simple rules – (i) at each time step t a new movie with n actors (in the context of movie-actor network) is added; (ii) of the n actors playing in a new movie, m actors are assumed to be new, without any previous experience; (iii) the rest $n - m$ actors are chosen from the pool of "old" actors with a probability proportional to the number q of movies that they previously starred (i.e., the degree of the actor nodes at time t). Assuming n and m to be constants equal to their average values $\widehat{n}$ and $\widehat{m}$ respectively the authors analytically solved the model and showed that the degree distribution is proportional to $q^{-\lambda}$ where $\lambda = 2 + \frac{\widehat{m}}{\widehat{n}-\widehat{m}}$. Several variants of this model have also been proposed and analytically solved to establish that the degree distribution of the actor nodes indeed follow a power-law behavior.

It is important to note that in all the above networks both the partitions grow unboundedly with time unlike the case of PlaNet where one of the partitions corresponding to the consonants remains relatively fixed over time while the other partition can grow undoundedly. Therefore, the asymptotics of the two models are different and in particular one cannot make the steady state assumptions in the latter case because, the average degree of the consonant nodes (i.e., $\mu t/N$) diverges with time. This fundamental difference in the two models also manifests as a difference in the emergent degree distribution; while it is a power-law in the former case, it is a β-distribution in the latter case.

There are also certain non-growing models of bipartite networks primarily based on rewiring. For instance, one of the most popular rewiring based models has been described by Evans and Plato in [50] (henceforth the EP Model). In this study, one of the partitions, which the authors refer to as the set of *artifacts*, is fixed. The nodes in the other partition are referred to as *individuals*, all of which have degree one. In the EP model, there are fixed number of edges; at every time step, an artifact node is selected following a distribution Π_R and an edge that is connected to the chosen artifact is picked up at random. This edge is then rewired to another artifact node which is chosen according to a distribution Π_A. During the rewiring process the other end of the edge is always attached to the same *individual* node. The authors derive the exact analytical expressions for the degree distribution of the artifact nodes at all times and for all values of the parameters for the following definitions of the removal and attachment probabilities:

$$\Pi_R = \frac{k}{E}; \, \Pi_A = p_r \frac{1}{N} + p_p \frac{k}{E}$$

where E, N and k stands for the number of edges, the number of artifacts and the degree of an artifact node respectively. Furthermore, p_r and p_p, which add up to one, are positive constants (model parameters) that control the balance between random and preferential attachment.

One of the most important differences between the EP model and our model is that the total number of edges in the latter case diverges with time. Consequently, if we rewrite the attachment probability for our model (equation 3.5) in a form similar to that of Π_A, we obtain the following expressions for the parameters p_r and p_p.

$$p_r = \frac{1}{1+\gamma t/N}; \, p_p = \frac{\gamma t/N}{1+\gamma t/N}$$

Clearly, as $t \to \infty$, $p_r \to 0$ and $p_p \to 1$, whereas in the EP model these parameters are fixed. Thus, apart from the two extreme cases of $p_r = 0$ and $p_r = 1$, the two models are fundamentally different.

3.7 Summary

In this chapter, we have introduced a computational framework in order to investigate various interesting properties of the consonant inventories of human languages. We have dedicated the preceding sections for the following.

(i) Propose a bipartite network representation of the consonant inventories, namely PlaNet.

(ii) Describe the data source and the construction procedure for the network.

(iii) Analyze the topological properties of PlaNet. We find that the degree distribution of the consonant nodes in PlaNet is well-behaved with an exponential cut-off towards the tail.

(iv) Propose a synthesis model based on preferential attachment coupled with a tunable parameter γ controlling the randomness of the system in order to explain the emergence of the degree distribution of the consonant nodes in PlaNet.

(v) Analytically solve the synthesis model in order to find a closed form expression for the degree distribution of the consonant nodes. In particular, we observe that the degree distribution obtained from this theoretical analysis asymptotically tends to a β-distribution with time. Further, a very high value of γ necessary to fit the distribution with empirical data points to the fact that preferential attachment plays a significant role in shaping the structure of the consonant inventories.

(vi) Investigate the dynamics within and across the consonant inventories of five major language families of the world namely, Indo-European, Afro-Asiatic, Niger-Congo, Austronesian and Sino-Tibetan. We find that the preferential component is stronger within a linguistic family than it is across the families.

Linguistic Significance of the Model

A possible reason behind the success of our model in explaining the distribution of the occurrence of consonants is the fact that language is a constantly changing system and preferential attachment plays a significant role in this change. The sociolinguist Jennifer Coates remarked that this linguistic change occurs in the context of linguistic heterogeneity. She explained that "… linguistic change can be said to have taken place when a new linguistic form, used by some sub-group within a speech community, is adopted by other members of that community and accepted as the norm." [40]. In this process of language change, those consonants that belong to languages that are more prevalent among the speakers of a generation have higher chances of being transmitted to the speakers of the subsequent generations [18]. An explanation based on this observation that assumes an initial disparity in the distribution of the consonants across languages can be intuitively formulated as follows – let there be a community of N speakers communicating among themselves by means of only two consonants say /k/ and /g/. Let the number of /k/ speakers be m and that of /g/ speakers be n. If we assume that each speaker has l descendants and that language inventories are transmitted with high fidelity, then after i generations, the number of /k/ speakers should be ml^i and that of /g/ speakers should be nl^i. Now if $m > n$ and $l > 1$ then for sufficiently large values of i we have $ml^i \gg nl^i$. Stated differently, the /k/ speakers by far outnumber the /g/ speakers after a few generations even though the initial difference between them is quite small. This phenomenon is similar to that of preferential attachment where language communities get attached to, i.e., select consonants that are already highly preferred. The parameter γ in this case may be thought of as modeling the randomness of the system that creeps in due to accidental errors which might occur during language transmission.

Furthermore, the fact that the choice of consonants within the languages of a family is far more preferential than it is across the families is possibly an outcome of shared ancestry. In other words, the inventories of genetically related languages are similar (i.e., they share a lot of consonants) because they have evolved from the same parent language through a series of linguistic changes, and the chances that they use a large number of consonants used by the parent language is naturally high.

Although the bipartite network formulation presented in this chapter faithfully captures the properties of occurrence of consonants across languages, it does not suitably reflect such other advanced properties as the distribution of co-occurrence or patterns of co-occurrence of consonants across languages. Consequently, the synthesis model proposed here also does not employ sophisticated techniques to replicate these properties. These limitations bring us to the central objective of the next chapter where we shall propose another novel complex network representation of the inventories, derived from PlaNet, that can capture the co-occurrence likelihood of the consonants over languages. In particular, we shall analyze this network to extract many other interesting properties of the consonant inventories. We shall also attempt to develop more involved network growth models in order to explain the emergence of these properties.

Chapter 4

Analysis and Synthesis of the Co-occurrence Network of Consonants

In the previous chapter, we proposed a bipartite network representation of the consonant inventories and investigated the topological properties of this network. We also presented a preferential attachment based growth model that can explain, quite successfully, the distribution of the occurrence of the consonants across the world's languages.

An immediate question that comes up is how do the consonants co-occur with each other across different languages. In this chapter, therefore, we take a step further and attempt to analyze in detail the co-occurrence properties of the consonants. Such co-occurrence properties, in general, can be suitably captured from any bipartite collaboration network by constructing a network of shared collaboration acts, the so called *one-mode projection* onto the actor nodes alone. The links in this network represent the "intensity" of collaboration between a pair of actors. In order to determine the co-occurrence properties of the consonants, we project PlaNet on the consonant nodes and thereby, derive a new network called the **Pho**neme-Phoneme **Net**work or **PhoNet** which is a network of consonants where a pair of nodes are linked as many times as they are found to occur together across the inventories of different languages. Therefore, PhoNet is a weighted *unipartite* network of consonants where an edge between two nodes signifies their co-occurrence likelihood over the consonant inventories.

In fact, there are a number of studies related to the one-mode projections of real-world bipartite networks. For instance, it has been shown that for a movie-actor collaboration network, the degree distribution of the actor nodes in the bipartite as well as the one-mode network follow a power-law [123, 130]. Similarly, in case of a scientific collaboration network, it has been observed that the degree distribution of the author nodes shows a fat-tailed behavior in both the bipartite network as well as the one-mode projection [130]. In case of board-director networks it has been found that the degree distribution of the director nodes in the bipartite and the one-mode network can be roughly fitted using exponential functions [16, 130]. Furthermore, it has been also shown that all these real-world networks are characterized by a high clustering coefficient [123,130]. Various models such as the one reviewed in section 3.6 of Chapter 3 and others like [67, 123] have also been proposed and analytically solved to explain the power-law degree distribution and the high clustering coefficient of the one-mode projection.

The study of the one-mode network PhoNet can give us important insights into the organization of the consonant inventories and therefore we dedicate this chapter to analyze and synthesize the different topological properties of this network. In particular, we investigate two of the most crucial properties of the network namely the degree distribution and the clustering coefficient. We observe that the degree distribution follows a well-behaved probability distribution (although not a power-law) and the clustering coefficient of the network is quite high.

The theoretical predictions from the growth model presented in the previous chapter, however, does not seem to match well with those observed for PhoNet. Neither the degree distribution nor the clustering coefficient predicted by the model are found to be good approximations of the real data. Therefore, we attempt to identify the gap in our previous analysis and suitably refine it so as to accurately explain the empirical properties of the network.

More specifically, in the previous analysis, we assumed that the degree of the nodes in the language partition of PlaNet (i.e., the size of the consonant inventories) is equal to the average degree and therefore a constant. Although this assumption does not affect the degree distribution of the bipartite network, it affects the degree distribution of the one-mode projection. It is important to mention here that this assumption has also been made in most of the earlier studies related to bipartite networks [123, 130]. Nevertheless, real-world systems present us with instances where there is a necessity to relax this assumption for a more accurate analysis; for instance, not all movies have the same cast size and not all languages have the same consonant inventory size. In essence, one has to think of this size as a random variable that is being sampled from a particular distribution and indeed it is possible to analytically show that this distribution actually affects the emergent degree distribution of the one-mode network.

The clustering coefficient, on the other hand, obtained from the analysis of our model is found to be much lower than that of PhoNet. The reason for this deviation is that the number of triangles present in the real network is significantly higher than that generated by the model. We therefore, attempt to refine the model by incorporating *triad* (i.e., fully connected triplet) formation into it coupled with preferential attachment. The motivation behind the triad formation process is the fact that such a model for growing unipartite networks have been found to lead to increased clustering [72]. We analytically show that the clustering coefficient can be indeed tuned based on the probability of triad formation (a parameter in our revised model). Consequently, for a certain range of this probability, the clustering coefficient predicted by the revised model closely matches with the real data.

The rest of the chapter is organized as follows. In section 4.1, we present the formal definition of PhoNet and outline its construction procedure. The topological properties of the network are analyzed in section 4.2. In the next section, we analytically compute the degree distribution and the clustering coefficient for the model presented in the previous chapter and show that they do not match with the real data. In section 4.4, we extend the theoretical framework and show that the degree distribution of the one-mode network is sensitive to the distribution of the node degrees of the growing partition of a bipartite network like PlaNet. We refine the network growth model in section 4.5 by incorporating triad formation and show that there is a significant increase in the clustering coefficient due to this refinement. In section 4.6, we summarize our observations, propose certain linguistic interpretations of our results and identify certain extensions of this study, which are mostly dealt with in the next chapter.

4.1 Definition and Construction of PhoNet

PhoNet is the one-mode projection of PlaNet onto the consonant nodes, i.e., a network of consonants in which two nodes are linked by an edge with weight as many times as they co-occur across languages. Hence, it can be represented by a graph $G = \langle\ V_C,\ E_{ph}\ \rangle$, where V_C is the set of consonant nodes and E_{ph} is the set of edges connecting these nodes in G. There is an edge $e \in E_{ph}$ if the two nodes (read consonants) that are connected by e co-occur in at least one language and the number of languages they co-occur in defines the weight of the edge e. Figure 4.1 illustrates the nodes and the edges of PhoNet through a hypothetical example. In the figure, the numerical values against the edges of PhoNet denote their corresponding weights. The numerical value against a particular node of PhoNet denotes its frequency of occurrence across languages (or alternatively its degree in PlaNet).

We obtain PhoNet by taking the one-mode projection of PlaNet (constructed from UPSID in the previous chapter) onto the V_C nodes. Consequently, the number of nodes in PhoNet (i.e., $|V_C|$) is 541. In order to give the reader an idea of the complex structure resulting from this construction we present a partial illustration of PhoNet in Figure 4.2. All edges in this figure have an edge-weight greater than or equal to 50. The number on each node corresponds to a particular consonant. For instance, node number 508 corresponds to /g/ whereas node number 540 represents /k/.

The actual number of edges (ignoring the weights on them) in PhoNet, that is $|E_{ph}|$, is 30412. The connection density of PhoNet (assuming that its edges are unweighted) is $\frac{|E_{ph}|}{\binom{|V_C|}{2}} = \frac{2\times30412}{541\times540} = 0.2$, and this is the probability with which two

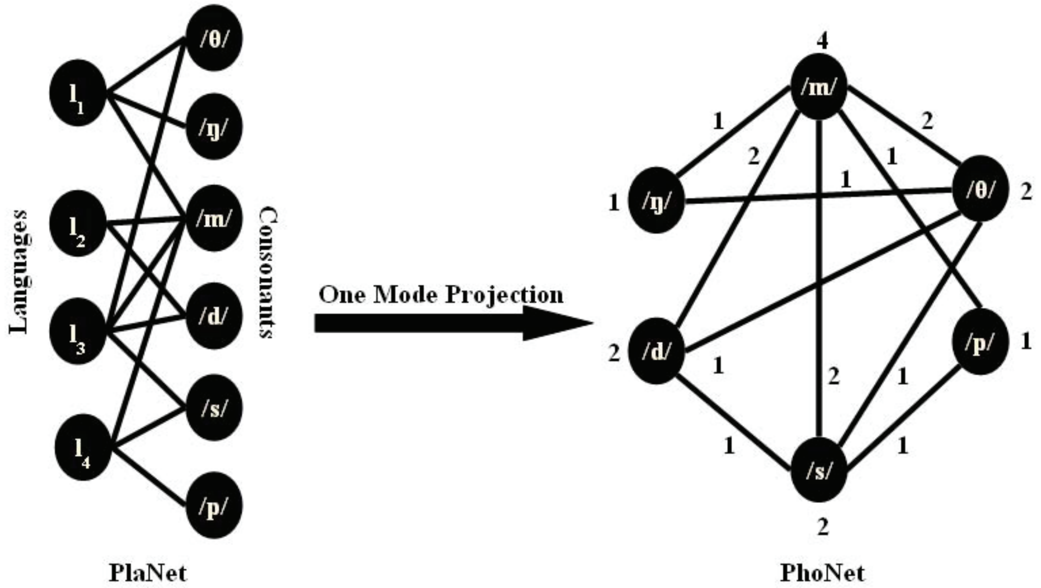


Figure 4.1: A hypothetical example illustrating the nodes and edges of PhoNet – the one-mode projection of PlaNet

randomly chosen consonants co-occur in at least one of the languages. However, as we shall see, this co-occurrence is not simply governed by a single probability; instead it follows a well-behaved probability distribution.

4.2 Analysis of the Topological Properties of PhoNet

In this section, we shall analyze some of the topological properties of PhoNet. In particular, we shall investigate two important properties of the network – the degree distribution and the clustering coefficient.

4.2.1 Weighted Degree

For a weighted graph like PhoNet, the degree q of a node v is defined as the sum of the weights of the edges that are incident on v. This is also sometimes referred to as the *weighted degree* of the node v. The unweighted (or plain) degree k of the node v,

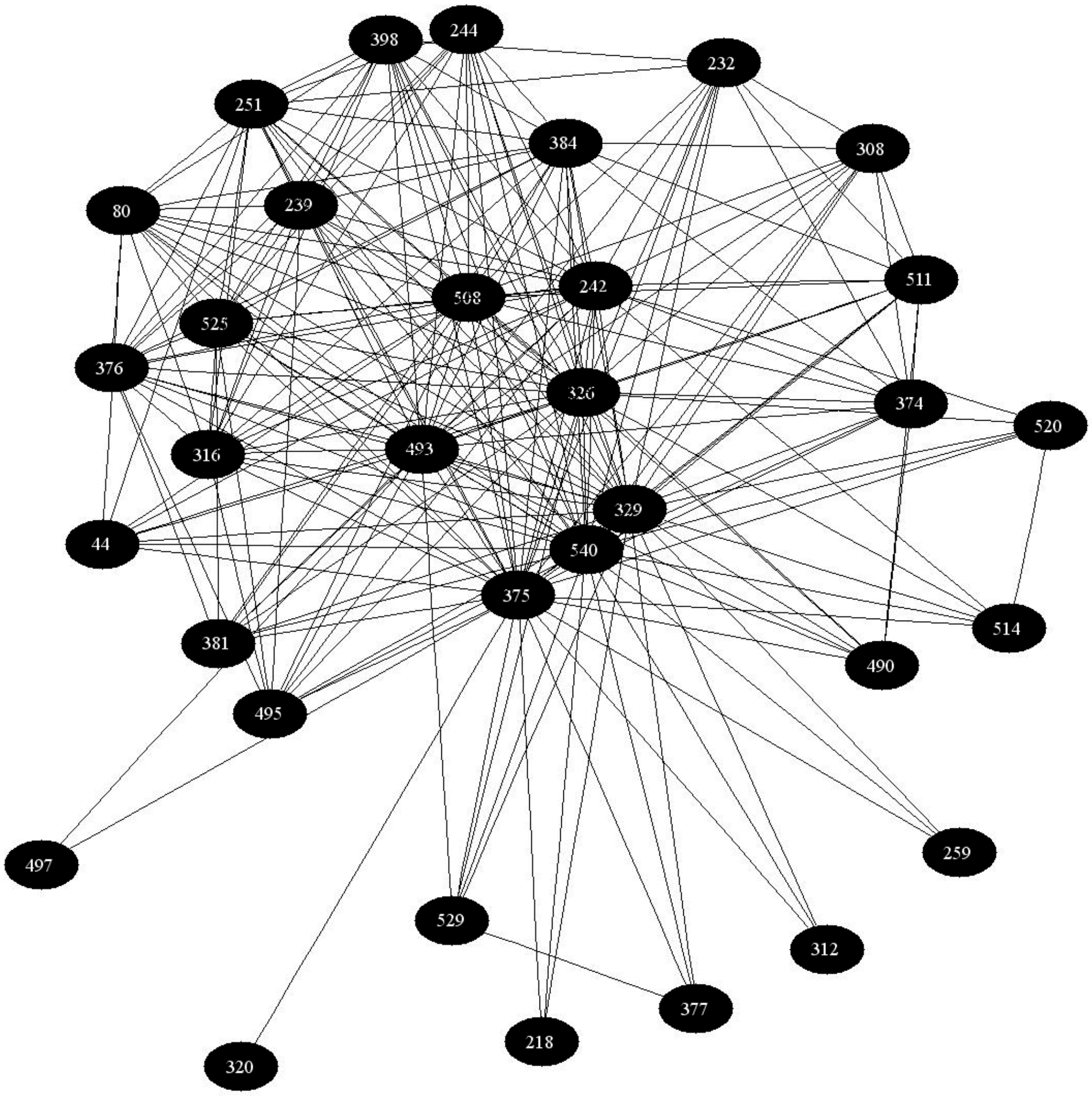


Figure 4.2: A partial illustration of PhoNet constructed from UPSID

on the other hand, is the number of edges (ignoring the weights on them) that are incident on v. In the rest of the chapter, we shall mainly refer to the weighted degree unless otherwise mentioned.

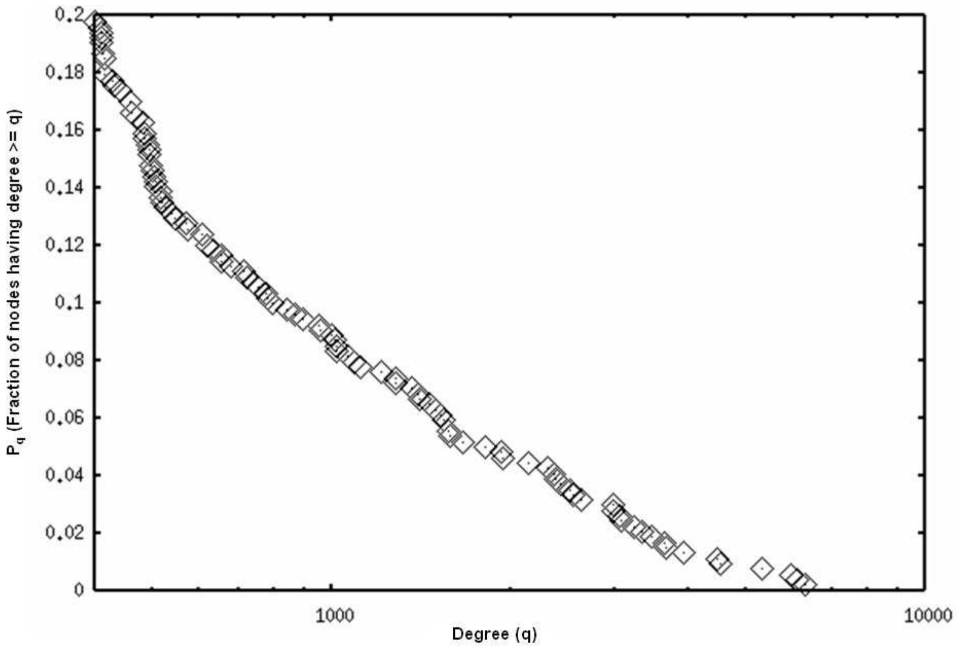


Figure 4.3: Degree distribution of the nodes in PhoNet. The x-axis is in logarithmic scale

4.2.2 Degree Distribution of PhoNet

Figure 4.3 shows the degree distribution plot for the nodes of PhoNet (x-axis is in logarithmic scale). As we have observed in case of PlaNet, this distribution is not exactly a power-law. We shall discuss in further detail about the properties of this distribution in the forthcoming sections of this chapter.

4.2.3 Clustering Coefficient of PhoNet

The clustering coefficient for a node i is the proportion of links between the nodes that are the neighbors of i divided by the number of links that could possibly exist between them. For instance, in a friendship network it represents the probability that two friends of the person i are also friends themselves. Therefore, the larger the number of triangles formed by the neighbors of i, the higher is the clustering coefficient (see [114] for further reference). For a weighted graph such as PhoNet, this definition has been suitably modified in [15]. According to this definition, the clustering coefficient for a node i is,

$$c_i = \frac{1}{\left(\sum_{\forall j} w_{ij}\right)(k_i - 1)} \sum_{\forall j,l} \frac{(w_{ij} + w_{il})}{2} a_{ij} a_{il} a_{jl} \tag{4.1}$$

where j and l are neighbors of i; k_i represents the plain degree of the node i; w_{ij}, w_{jl} and w_{il} denote the weights of the edges connecting nodes i and j, j and l, and i and l respectively; a_{ij}, a_{il}, a_{jl} are boolean variables, which are true iff there is an edge between the nodes i and j, i and l, and j and l respectively. The formula for c_i in this equation essentially counts for each triplet formed in the neighborhood of the vertex i, the weight of the two participating edges of the vertex i. The normalization factor $\left(\sum_{\forall j} w_{ij}\right)(k_i - 1)$ accounts for the weight of each edge times the maximum possible number of triplets in which it may participate, and it ensures that $0 \leq c_i \leq 1$.

The clustering coefficient of the network (c_{av}) is equal to the average clustering coefficient of the nodes. The value of c_{av} for PhoNet is 0.89, which is significantly higher than that of a random graph with the same number of nodes and edges (0.08). Note that similar characteristics are also observed in many other social networks [113, 130]. This in turn, indicates that the probability of co-occurrence of two consonants that have a common neighbor in PhoNet is much greater than expected by random chance.

4.3 Predictions from the Previous Model

In this section, we analytically derive the expression for the degree distribution as well as the clustering coefficient of the theoretical model presented in section 3.4 and show that the predictions do not match well with the real data.

4.3.1 Expression for the Degree Distribution

It is easy to analytically calculate the degree of the V_C nodes in the one-mode network (henceforth PhoNet$_{theo}$) if we assume the degree of each V_L node is equal to a constant μ (i.e., the average inventory size). Consider a node $u \in V_C$ that has degree k in the bipartite network. Therefore, u is connected to k nodes in V_L and each of these k nodes are in turn connected to $\mu - 1$ other nodes in V_C. Defining the degree of a node as the number of edges attached to it, in the one-mode projection, u has a degree of $q = k(\mu - 1)$. Consequently, the degree distribution $p_u(q)$ of the one-mode network is given by

$$p_u(q) = \begin{cases} p_k & \text{if } k = q/(\mu - 1) \\ 0 & \text{otherwise} \end{cases} \tag{4.2}$$

Note that this mapping simply implies that $p_u(q = 0) = p_0$, $p_u(q = \mu - 1) = p_1$, $p_u(q = 2(\mu - 1)) = p_2$, …, $p_u(q = j(\mu - 1)) = p_j$. Figure 4.4 compares the degree distribution predicted by the equation 4.2 with the empirical degree distribution of PhoNet. The figure clearly indicates that the degree distribution of PhoNet$_{theo}$ largely differs from that of PhoNet.

4.3.2 Expression for the Clustering Coefficient

Here we attempt to analytically calculate the clustering coefficient of PhoNet$_{theo}$. Each time a node $u \in V_C$ is selected for attachment, a neighborhood of μ-1 other nodes is created in the one-mode projection. Therefore, the number of triangles attached to u for a particular iteration is $[(\mu\text{-}1)(\mu\text{-}2)]/2$. Further, each iteration is independent of the earlier ones and if there are k iterations (since the degree of u in the bipartite network is k) then the total number of triangles attached to u should be $[k(\mu\text{-}1)(\mu\text{-}2)]/2$. On the other hand, if the degree of the node u in the one-mode is q then the total number of possible triangles should be $[q(q\text{-}1)]/2$. Thus, the clustering

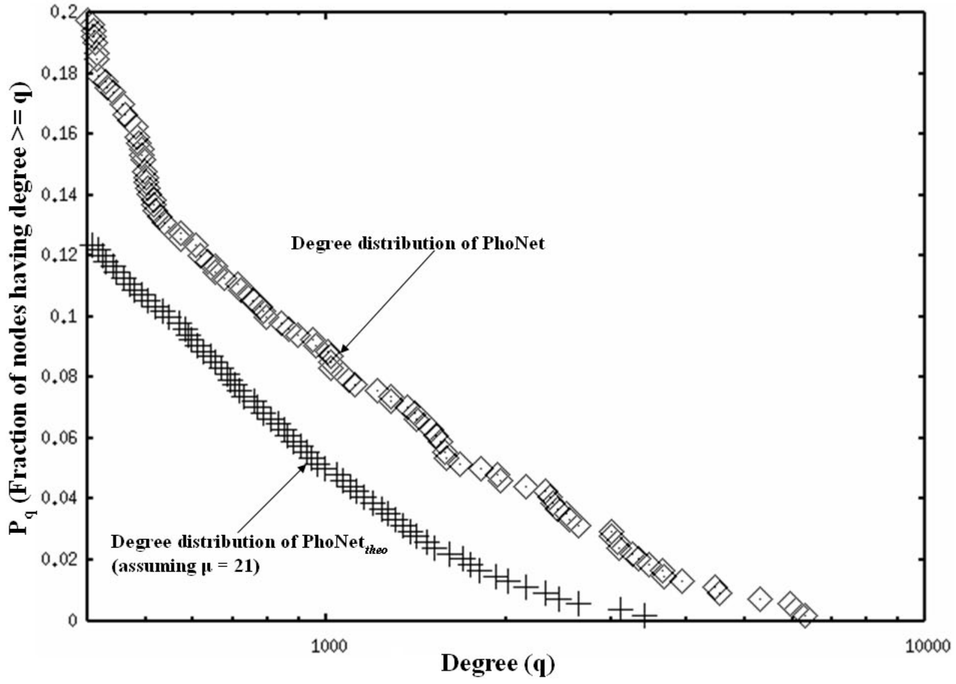


Figure 4.4: Comparison of the degree distribution of PhoNet and PhoNet$_{theo}$. The x-axis is in the logarithmic scale

coefficient $[C_u]^{PA}$ (PA stands for preferential attachment) is

$$[C_u]^{PA} = \frac{[k(\mu-1)(\mu-2)]/2}{[q(q-1)]/2} \tag{4.3}$$

Since in this case $q = k(\mu\text{-}1)$ so

$$[C_u]^{PA} = \frac{[k(\mu-1)(\mu-2)]/2}{[k(\mu-1)(k(\mu-1)-1)]/2} \tag{4.4}$$

or,

$$[C_u]^{PA} = \frac{(\mu-2)}{k(\mu-1)-1} \tag{4.5}$$

It is interesting to note that for our model, the above formula can be also shown to be equivalent to the equation 4.1. It is known that $\sum_{\forall j} w_{ij} = q = k(\mu\text{-}1)$. k_u, which denotes the number of distinct neighbors, is also equal to $k(\mu\text{-}1)$ assuming that the neighborhoods created for u each time are independent. Therefore, the weight of an edge (u, j) for any neighbor j of u is equal to 1. So, $(w_{uj}+w_{ul})/2 = (1+1)/2 = 1$. Since there are ($\mu$-1) neighbors of u and we are picking a pair of neighbors at a time (j and l), the number of times that the summation $(w_{uj}+w_{ul})/2$ takes place is [(μ-1)(μ-2)]/2. Nevertheless, each neighbor is considered twice in the formula (j and l) and (l and j), and thus the total number of times this summation takes place is 2[(μ-1)(μ-2)/2] = (μ-1)(μ-2). Further, this happens each of the k times the node u is selected. In other words, $\sum_{\forall j,l} \frac{(w_{uj}+w_{ul})}{2} a_{uj} a_{ul} a_{jl} = k(\mu\text{-}1)(\mu\text{-}2)$. Putting the parts together we have the clustering coefficient $[C_u]^{BAR}$ (BAR refers to Barrat *et al.* [15]) as follows

$$[C_u]^{BAR} = \frac{k(\mu-1)(\mu-2)}{k(\mu-1) \times (k(\mu-1)-1)} \tag{4.6}$$

or,

$$[C_u]^{BAR} = [C_u]^{PA} = \frac{(\mu-2)}{k(\mu-1)-1} \tag{4.7}$$

The clustering coefficient of the whole network is the average over all the nodes u (i.e., $\forall u$) present in it. Using equation 4.7 we find that the clustering coefficient of PhoNet$_{theo}$ is 0.37 which is close to that obtained from the simulation of the model (0.35). However, we have already observed in section 4.2 that the clustering coefficient of PhoNet is as high as 0.89 and therefore differs largely from the above result. The primary reason for this deviation is that real networks like PhoNet indicate the presence of a large number of triangles and this fact is not taken into account by our model.

Therefore, it turns out that neither the degree distribution nor the clustering coefficient predicted by the model matches with that of PhoNet. In the rest of this chapter, we shall attempt to suitably extend the analytical framework so as to accurately explain these topological properties of PhoNet.

4.4 Extension of the Analytical Framework to Match the Degree Distribution

In this section, we show that the node degrees in V_L actually affect the emergent degree distribution of the one-mode projection even though the degree distribution of the bipartite network is not affected. In particular, we relax the assumption that the node degrees in V_L are equal to a constant μ; in contrast, we consider them as random variables that are being sampled from a distribution. The degree q of u in the one-mode projection is dependent on this sampling distribution while the degree k in the bipartite network is not as long as the mean degree of the nodes in V_L remains μ. Note that under such an assumption, the equation 3.3 holds so that the denominator of the equation 3.2 is again equal to $\mu\gamma t + N$ as in the earlier case and thus, k remains unchanged.

4.4.1 Method for Computing $p_u(q)$

Let us assume that the node degrees in V_L are being sampled from a particular distribution f_d. Let us call the probability that the node u having degree k in the bipartite network ends up as a node having degree q in the one-mode projection $F_k(q)$. If we now assume that the degrees of the k nodes in V_L to which u is connected to are d_1, d_2, $\ldots$, d_k then we can write

$$q = \sum_{i=1\ldots k} (d_i - 1) \tag{4.8}$$

The probability that the node u is connected to a node in V_L of degree: d_1 is $d_1 f_{d_1}$, d_2 is $d_2 f_{d_2}$, $\ldots$, d_k is $d_k f_{d_k}$. One might apply the *generating function* (GF) formalism introduced in [116] to calculate the degree distribution of the V_C nodes in the one-mode projection as follows. Let $f(x)$ denote the GF for the distribution of the node degrees in V_L. In other words, $f(x) = \sum_d f_d x^d$. Similarly, let $p(x)$ denote the GF for the degree distribution of the V_C nodes in the bipartite network, i.e.,

$p(x) = \sum_k p_k x^k$. Further, let $g(x)$ denote the GF for the degree distribution $p_u(q)$ of the one-mode projection. Therefore, $g(x) = \sum_q p_u(q)x^q$. The authors in [116] (see equation 70) have shown that $g(x)$ can be correctly expressed as

$$g(x) = p(f'(x)/\mu) \tag{4.9}$$

If f_d and p_k are distributions for which a closed form is known for $f(x)$ and $p(x)$ (e.g., if both f_d and p_k are Poisson-distributed) then it is easy to derive a closed form solution for $g(x)$. However, in our case, p_k is β-distributed as shown in equation 3.25 and there is no known closed form expression for $p(x)$. Therefore, it is difficult to carry out the theoretical analysis any further using the GF formalism. Another way to approach the problem would be to calculate a generic expression for $p_u(q)$ from the first principles. We shall therefore attempt to obtain such an expression, propose a suitable approximation for it and then check for its dependence on the choice of f_d. As we shall see, in many cases, it is even possible to obtain a closed form solution for the expression of $p_u(q)$. The appropriately normalized probability that the node u in the bipartite network is connected to nodes of degree d_1, d_2, …, d_k in V_L is $\left(\frac{d_1 f_{d_1}}{\mu}\right)\left(\frac{d_2 f_{d_2}}{\mu}\right)\ldots\left(\frac{d_k f_{d_k}}{\mu}\right)$ (each such connection is independent of the others). Under the constraints $d_1 + d_2 + \cdots + d_k = q$, we have

$$F_k(q) = \sum_{d_1+d_2+\cdots+d_k=q} \frac{d_1 d_2 \ldots d_k}{\mu^k} f_{d_1} f_{d_2} \ldots f_{d_k} \tag{4.10}$$

We can now add up these probabilities for all values of k weighted by the probability of finding a node of degree k in the bipartite network. Thus we have,

$$p_u(q) = \sum_k p_k F_k(q) \tag{4.11}$$

or,

$$p_u(q) = \sum_k p_k \sum_{d_1+d_2+\cdots+d_k=q} \frac{d_1 d_2 \ldots d_k}{\mu^k} f_{d_1} f_{d_2} \ldots f_{d_k} \tag{4.12}$$

For the rest of the analysis, we shall assume that $d_1 d_2 \ldots d_k$ is approximately equal to μ^k. In other words, we assume that the arithmetic mean of the distribution is close to the geometric mean, which holds if the variance of the distribution is low. We shall shortly discuss in further details the bounds of this approximation. However, prior to that, let us investigate, how this approximation helps in advancing our analysis. Under the assumption $\frac{d_1 d_2 \ldots d_k}{\mu^k} = 1$, $F_k(q)$ can be thought of as the distribution of the sum of k random variables each sampled from f_d. In other words, $F_k(q)$ tells us how the sum of the k random variables is distributed if each of these individual random variables are drawn from the distribution f_d. This distribution of the sum can be obtained by the *iterative convolution* (see [64] for details) of f_d for k times. If the closed form expression for the convolution exists for a distribution, then we can obtain an analytical expression for $p_u(q)$.

4.4.2 Effect of the Sampling Distribution f_d

In the following, we shall attempt to analytically find an expression for $p_u(q)$ assuming different forms of the distribution f_d. As we shall see, $F_k(q)$ is different for each of this form, thereby, making the degree distribution of the nodes in the one-mode projection sensitive to the choice of f_d. Furthermore, we shall show that the degree distribution of PhoNet can be indeed matched better if one assumes the actual distribution of the consonant inventory sizes rather than fixing this size to a constant μ.

It is important to mention here that since in the expression for q (equation 4.8) we need to subtract one from each of the d_i terms therefore the distribution $F_k(q)$ has to be shifted accordingly. We shall denote the approximate and shifted version of $F_k(q)$ by $\mathring{F}_k(q)$.

(i) *Normal distribution*: If f_d is a normal distribution of the form $N(\mu,\, \sigma^2)$ then the sum of k random variables sampled from f_d is again distributed as a normal distribution of the form $N(k\mu,\, k\sigma^2)$. Therefore, $\mathring{F}_k(q)$ is given by

$$\mathring{F}_k(q) = N(k\mu - k, k\sigma^2) = N(k(\mu - 1), k\sigma^2) \tag{4.13}$$

If we substitute the density function for N we have

$$\mathring{F}_k(q) = \frac{1}{\sigma\sqrt{2\pi k}} \exp\left(-\frac{(q-k(\mu-1))^2}{2k\sigma^2}\right) \tag{4.14}$$

Hence, $p_u(q)$ is given by

$$p_u(q) = \frac{1}{\sigma\sqrt{2\pi}} \sum_k p_k k^{-0.5} \exp\left(-\frac{(q-k(\mu-1))^2}{2k\sigma^2}\right) \tag{4.15}$$

(ii) *Delta function*: Let f_d be a delta function of the form

$$\delta(d,\mu) = \begin{cases} 1 & \text{if } d = \mu \\ 0 & \text{otherwise} \end{cases} \tag{4.16}$$

Note that this boils down to the case where the degree of each V_L node is a constant μ and therefore $\frac{d_1 d_2 \ldots d_k}{\mu^k}$ is exactly equal to 1. If this delta function is convoluted k times then the sum should be distributed as

$$\mathring{F}_k(q) = \delta(q, k\mu - k) = \begin{cases} 1 & \text{if } q = k\mu - k \\ 0 & \text{otherwise} \end{cases} \tag{4.17}$$

Therefore, $p_u(q)$ exists only when $q = k(\mu-1)$ or $k = q/(\mu-1)$ and we have

$$p_u(q) = \begin{cases} p_k & \text{if } k = q/(\mu-1) \\ 0 & \text{otherwise} \end{cases} \tag{4.18}$$

Note that this is the only case that we had dealt with earlier (see equation 4.2).

(iii) *Exponential distribution*: If f_d is an exponential distribution of the form $E(\lambda)$ where $\lambda = 1/\mu$ then the sum of the k random variables sampled from f_d is known to be distributed as a gamma distribution of the form $\Gamma(q;\, k,\, \mu)$. Therefore, we have

$$\mathring{F}_k(q) = \Gamma(q; k, \mu - 1) \tag{4.19}$$

Substituting the density function we have

$$\mathring{F}_k(q) = \frac{\lambda^{'} \exp{(-\lambda^{'} q)} (\lambda^{'} q)^{k-1}}{(k-1)!} \tag{4.20}$$

where $\lambda^{'} = 1/(\mu - 1)$. Hence, $p_u(q)$ is given by

$$p_u(q) = \lambda^{'} \sum_k p_k \frac{\exp{(-\lambda^{'} q)} (\lambda^{'} q)^{k-1}}{(k-1)!} \tag{4.21}$$

(iv) *Power-law distribution*: There is no known exact solution for the sum of k random variables each of which is sampled from f_d that is power-law distributed with exponent λ_i. However, as noted in [162,163], asymptotically the tail of the distribution obtained from the convolution is dominated by the smallest exponent (i.e., $minimum(\lambda_1, \lambda_2, \ldots, \lambda_k)$). Note that due to this approximation the resultant degree distribution should indicate a better match with the stochastic simulations towards the tail. We have

$$\mathring{F}_k(q) \approx kq^{-minimum(\lambda_1, \lambda_2, \ldots, \lambda_k)} \tag{4.22}$$

However, since we are sampling from the same distribution each time so $\lambda_1 = \lambda_2 = \cdots = \lambda_k = \lambda$ and

$$\mathring{F}_k(q) \approx kq^{-\lambda} \tag{4.23}$$

Consequently, $p_u(q)$ can be expressed as

$$p_u(q) = \sum_k p_k k q^{-\lambda} \tag{4.24}$$

Figure 4.5(a) shows the degree distribution of the V_C nodes in the bipartite network assuming that f_d is a (i) normal, (ii) delta, (iii) exponential and (iv) power-law distribution each having the same mean ($\mu = 22$). Note that although we carried out our analysis using continuous functions the numerical simulations that we present in Figure 4.5 are performed using their discrete counterparts (i.e., we use probability mass functions rather than probability density functions for our simulations). In all cases, $N = 1000$, $t = 1000$ and $\gamma = 2$. For stochastic simulations, the results are averaged over 100 runs. All the results are appropriately normalized. Figure 4.5(a) shows the degree distributions of V_C nodes of the bipartite networks generated through simulations when f_d is a (i) normal distribution ($\mu = 22$, $\sigma = 13$), (ii) delta function ($\mu = 22$), (iii) exponential distribution ($\mu = \frac{1}{\lambda} = 22$) and (iv) power-law distribution ($\lambda = 1.16$, mean $\mu = 22$); Figure 4.5(b) shows the degree distributions of the one-mode projections of the bipartite networks in (a); Figure 4.5(c) compares the simulations (blue dots) and equation 4.15 (pink dots) for $\mu = 22, \sigma = 13$; green dots indicate the case where V_L nodes have a constant degree $\mu = 22$; brown dots show how the result deteriorates when $\sigma = 1300$; Figure 4.5(d) compares the simulations (blue dots) and equation 4.18 (pink dots) for $\mu = 22$; Figure 4.5(e) compares the simulations (blue dots) and equation 4.21 (pink dots) for $\frac{1}{\lambda} = 22$; yellow dots show the plot of equation 4.27; green dots indicate the case where V_L nodes have a constant degree $\mu = 22$ (given as a reference to show that even the approximate equation 4.27 is better than it); Figure 4.5(f) compares the simulations (blue dots) and equation 4.24 (pink dots) for $\gamma = 1.16$, $\mu = 22$; yellow dots show the plot of equation 4.28; green dots indicate the case where V_L nodes have a constant degree $\mu = 22$ (again given as a reference to show that it is worse than the equation 4.28).

Figures 4.5(a) and (b) together imply that the degree distribution of the one-mode projection varies depending on f_d although the degree distribution remains unaffected for all the bipartite networks generated as long as the means of the different f_d chosen are the same. Figures 4.5(c)–(f) show that the analytically obtained expressions are in good agreement with the simulations. Note that in case of power-law, while the heavy tail matches perfectly, the low degree zone deviates slightly which is a direct consequence of the approximation used in the convolution theory for power-law.

In many cases it is possible to derive a closed form expression for $p_u(q)$. One can think of $p_k\mathring{F}_k(q)$ as a function F in q and k, i.e., $p_k\mathring{F}_k(q) = F(q, k)$. If $F(q, k)$ can be exactly (or approximately) factored into a form like $\widehat{F}(q)\widetilde{F}(k)$ then $p_u(q)$ becomes

$$p_u(q) = \widehat{F}(q) \sum_k \widetilde{F}(k) \tag{4.25}$$

Changing the sum in equation 4.25 to its continuous form we have

$$p_u(q) = \widehat{F}(q) \int_0^\infty \widetilde{F}(k) dk = A\widehat{F}(q) \tag{4.26}$$

where A is a constant. In other words, the nature of the resulting distribution is dominated by the function $\widehat{F}(q)$. For instance, in case of exponentially distributed f_d, with some algebraic manipulations and certain approximations one can show that (see the yellow dots in Figure 4.5(e))

$$p_u(q) \approx A \exp\left(\frac{q}{\mu - 1}\right) \tag{4.27}$$

Similarly, in case of power-law one can show that (see the yellow dots in Figure 4.5(f))

$$p_u(q) \approx Aq^{-\lambda} \tag{4.28}$$

Therefore, it turns out that when this factorization is possible, the resulting degree distribution of the one-mode projection is largely dominated by that part of the convolution which is only dependent on q.

Matching the Degree Distribution of PhoNet

In the previous chapter, we have observed in section 3.2 that the size of the consonant inventories (i.e., the degrees of the V_L nodes in PlaNet) are β-distributed with parameters α and β equal to 7.06 and 47.64. It is hard to obtain a closed form solution for the

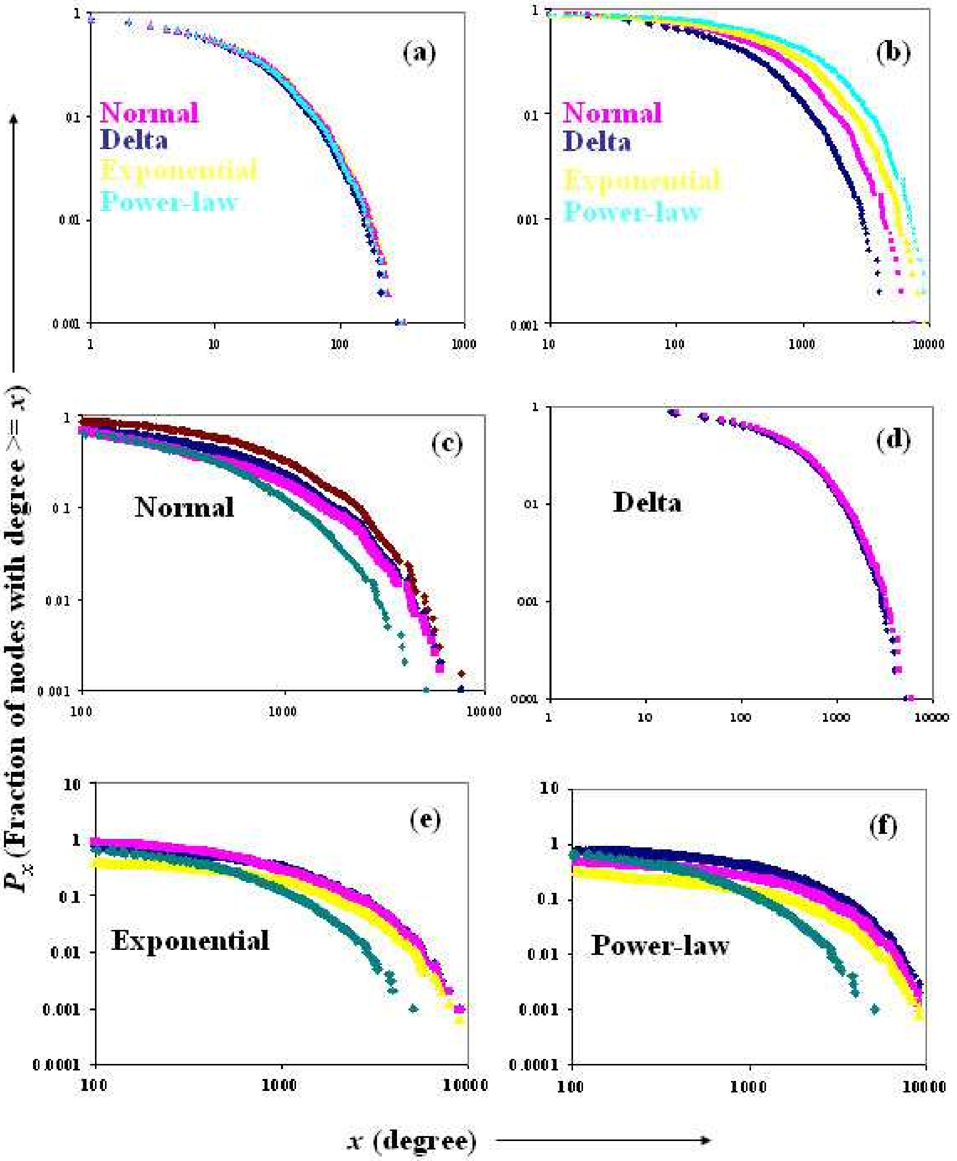


Figure 4.5: Degree distributions of bipartite networks and corresponding one-mode projections in doubly-logarithmic scale

convolution of different families of β-distributions and, thereby, derive an analytical expression for the degree distribution of the consonant nodes in the one-mode projection. However, one can apply numerical methods to convolve a β-distribution and use the result of this numerical simulation to predict the degree distribution of the one-mode projection. In fact, the degree distribution of the synthesized version of PhoNet (henceforth PhoNet$_{syn}$) which is actually the one-mode projection of PlaNet$_{syn}$ that

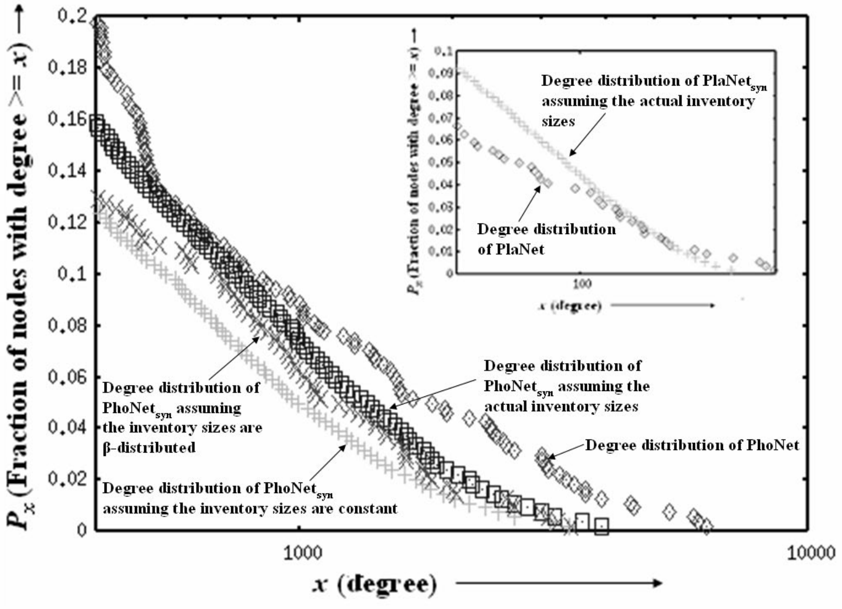


Figure 4.6: Comparison between the degree distribution of PhoNet and PhoNet$_{syn}$. The inset compares the degree distribution of the consonant nodes in PlaNet and PlaNet$_{syn}$ obtained assuming the actual consonant inventory sizes. The x-axis is in the logarithmic scale

is obtained assuming the inventory sizes to be β-distributed (see Figure 4.6) matches the real data better than if the sizes are assumed to be a constant. Furthermore, if the actual distribution of the inventory sizes is used for the stochastic simulations (rather than using the β-distribution) then the degree distribution of PhoNet$_{syn}$ (see in Figure 4.6) becomes almost same as the empirical one. The mismatch that still remains results out of the mismatch occurring at the bipartite level (see the inset of the Figure 4.6).

4.4.3 Approximation Bounds

In the previous section we carried out our analysis assuming that the geometric mean of degrees of the V_L nodes is approximately equal to the arithmetic mean. Here we shall investigate in details the bounds of this approximation. We shall employ the GF formalism to find the necessary condition (in the asymptotic limits) for our approximation to hold. More precisely, we shall attempt to estimate the difference in the means (or the first moments) of the exact and the approximate expressions for $p_u(q)$ and discuss when this difference is negligible which in turn serves as a necessary condition for the approximation to be valid. Let us denote the generating function for the approximate expression of $p_u(q)$ as $g_{app}(x)$. In this case, the GF encoding the probability that the node $u \in V_C$ is connected to a node in V_L of degree d is simply $f(x)/x$ and consequently, $\mathring{F}_k(q)$ is given by $(f(x)/x)^k$. Therefore,

$$g_{app}(x) = \sum_k p_k \left[\frac{f(x)}{x}\right]^k = p(f(x)/x) \tag{4.29}$$

Now we can calculate the first moments for the approximate and the exact $p_u(q)$ by evaluating the derivatives of $g_{app}(x)$ and $g(x)$ respectively at $x = 1$. We have

$$\begin{aligned} g'_{app}(1) &= \frac{d}{dx} p(f(x)/x)|_{x=1} \\ &= p'(f(x)/x)(f'(x)/x - f(x)/x^2)|_{x=1} \\ &= p'(1)(\mu/1 - 1/1) \\ &= (t/N)\mu(\mu - 1) \end{aligned} \tag{4.30}$$

Note that $p'(1)$ is the mean degree of the V_C nodes which is $\mu t/N$, $f'(1)$ is the mean degree of the V_L nodes and hence equal to μ, and $f(1) = p(1) = g(1) = 1$. Similarly,

$$\begin{aligned} g'(1) &= \frac{d}{dx} p(f'(x)/\mu)|_{x=1} \\ &= p'(f'(x)/\mu)f''(x)/\mu|_{x=1} \\ &= p'(1)f''(1)/\mu \\ &= (t/N)f''(1) \\ &= (t/N)\mu(\mu - 1) + t/N\sigma^2 \end{aligned} \tag{4.31}$$

Thus, the mean of the approximate $p_u(q)$ is smaller than the actual mean by $(t/N)\sigma^2$. Clearly, for $\sigma = 0$, the approximation gives us the exact solution, which is indeed the case for delta functions. Also, in the asymptotic limits, if $\sigma^2 \ll N$ (with a scaling of $1/t$), the approximation holds good. As the value of σ increases, the results deteriorate (see the brown dots in Figure 4.5(c)) because, the approximation does not hold any longer.

4.5 Refinement of the Model to Match the Clustering Coefficient

Real-world networks (mostly the socially rooted ones) exhibit the presence of a large number of triangles or mutually linked triples of nodes [113]. In other words, a commonly observed characteristic of many real networks is that if a node is linked to either node of a connected pair of vertices, then it is highly likely that it is also linked to the other node of the pair. The local clustering coefficient actually measures this connectedness of a node's neighbors with each other. The coefficient for the given node changes only when a new node gets linked to both to it and to one of its neighbors. Therefore, it turns out that if one can tune the number of triangles in a network growth model, it is possible to actually tune the local clustering of a node and consequently, the clustering coefficient of the whole network.

In order to generate networks with tunable clustering the concept of *triad* (i.e., fully connected triplet) formation was introduced in [72]. By regulating the triad formation step through a parameter of the model the authors showed that it is indeed possible to tune the amount of clustering in the growing network. They further establish that the emergent degree distribution of the network follows a power-law behavior. In the following, we refine our model in order to incorporate triad formation and, thereby, explain the clustering coefficient of PhoNet. For the specific case of our model, we also analytically show that this process leads to increased clustering.

4.5.1 The Triad Formation Model

A triad refers to a set of fully connected triplet (i.e., a triangle) of nodes. The triad model builds upon the concept of *neighborhood* formation. Two consonant nodes u_1 and u_2 become neighbors if a language node at any step of the synthesis process attaches itself to both u_1 and u_2. Let the probability of triad formation be denoted by p_t. At each time step a language node $L_j \in V_L$ makes the first connection preferentially to a consonant node $u_i \in V_C$ to which L_j is not already connected using the kernel defined in equation 3.2. For the rest of the (μ-1) connections L_j attaches itself preferentially to only the neighbors of u_i to which L_j is not yet connected with a probability p_t. Consequently, L_j connects itself preferentially to the non-neighbors of u_i to which L_j is not yet connected with a probability (1-p_t). The neighbor set of u_i gets updated accordingly. In essence, each step of network growth is dependent on the history of the connections made (i.e., the neighborhood formed) in the earlier steps. This phenomenon leads to the formation of a large number of triangles in the one-mode projection thereby increasing the clustering coefficient of the resultant network. A step of the synthesis process illustrating the concept of triad formation is outlined in Figure 4.7. In the figure, if the language node L_4 (which has degree 3) has the initial connection a_1 (due to preferential attachment) then according to the revised model the following connections would be a_2 and a_3 respectively in that order. This series of connections increases the co-occurrence by allowing the formation of a triad in the one-mode projection. The bold line in the figure indicates the edge that completes the triad.

4.5.2 Analysis of the Clustering Coefficient for the Triad Model

The triad model is different from the previous model because, it reduces the number of distinct neighbors of a node u formed in each step. Note that comparing the earlier (EAR) and the current (CUR) cases (see Figure 4.8) everything in the formula for $[C_u]^{BAR}$ (equation 4.1) effectively remains same except for the term k_u which represents the number of the distinct neighbors of u. Therefore, in order to estimate

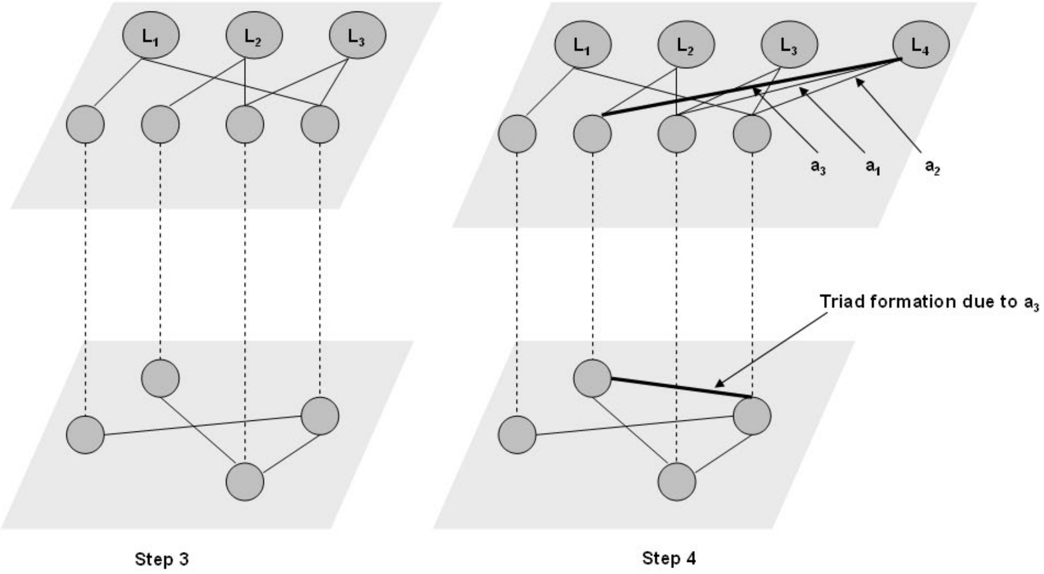


Figure 4.7: A partial step of the synthesis process illustrating the concept of triad formation

the new clustering coefficient we need to only estimate k_u. The expected number of distinct neighbors of u that are being formed now at each time step are (those that are entering from the non-neighbor to the neighbor set of u)

Step 1: (μ-1);

Step 2: (μ-1)+(μ-1)(1-p_t) = (μ-1)(2-p_t);

Step 3: (μ-1)(2-p_t)+(μ-1)(1-p_t) = (μ-1)(3-2p_t);

$\vdots$

Step k: (μ-1)[k-(k-1)p_t].

Thus, we can write

$$[[C_u]^{BAR}]_{EAR} = \frac{\alpha}{(k_u - 1)} = \frac{\alpha}{k(\mu - 1) - 1} \tag{4.32}$$

and,

$$[[C_u]^{BAR}]_{CUR} = \frac{\alpha}{(k_u - 1)} = \frac{\alpha}{(\mu - 1)[k - (k-1)p_t] - 1} \tag{4.33}$$

where α is the non-changing part. Hence,

$$\frac{[[C_u]^{BAR}]_{CUR}}{[[C_u]^{BAR}]_{EAR}} = \frac{k(\mu-1)-1}{(\mu-1)[k-(k-1)p_t]-1} \tag{4.34}$$

We have already shown that $[C_u]^{BAR}]_{EAR} = [C_u]^{PA}$ (equation 4.7) and so

$$[[C_u]^{BAR}]_{CUR} = \frac{[k(\mu-1)-1](\mu-2)}{\{(\mu-1)[k-(k-1)p_t]-1\}[k(\mu-1)-1]} \tag{4.35}$$

or,

$$[[C_u]^{BAR}]_{CUR} = \frac{(\mu-2)}{(\mu-1)[k-(k-1)p_t]-1} \tag{4.36}$$

Note that for $p_t = 1$, the same neighbor set gets chosen every time and so

$$[[C_u]^{BAR}]_{CUR} = \frac{(\mu-2)}{(\mu-1)-1} = 1 \tag{4.37}$$

On the other hand, when $p_t = 0$, and there is no neighborhood formation,

$$[[C_u]^{BAR}]_{CUR} = \frac{(\mu-2)}{k(\mu-1)-1} = [C_u]^{PA} \tag{4.38}$$

For $0 < p_t < 1$, we have

$$[[C_u]^{BAR}]_{CUR} > [C_u]^{PA} \tag{4.39}$$

Therefore, it turns out that the clustering coefficient of the network indeed increases if the triad formation step is incorporated into our preferential attachment based model.

It is important to mention here that the introduction of the parameter p_t into the model also affects the degree distribution of the V_C nodes in the bipartite network. However, all other model parameters remaining same, for a particular range

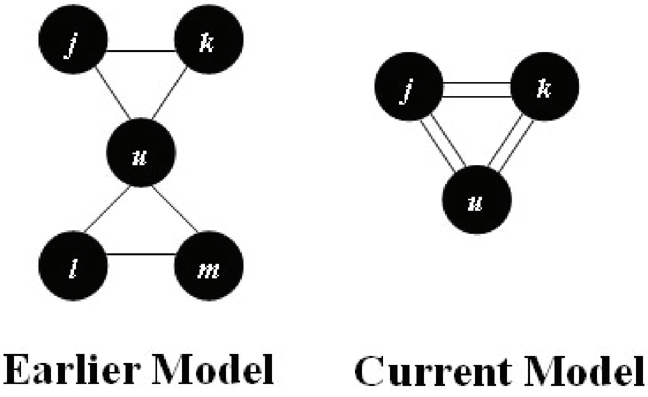


Figure 4.8: Comparison of the formation of new triangles in the simple preferential attachment based model (earlier model) and the model based on triad formation (current model)

of values of p_t the mean error between the degree distribution obtained in case of the simple preferential attachment based model and the triad model is quite low (see Figure 4.9(a)). Stochastic simulations show that the clustering coefficient is also quite high in this range of p_t values (see Figure 4.9(b)). Beyond this range, although the clustering coefficient increases, the mean error between the degree distributions rises. In other words, there exists a range of p_t values in which the degree distribution does not deteriorate much while the clustering coefficient increases significantly, that is, both the objectives are satisfied.

In fact, for the specific case of PhoNet, without affecting the degree distribution much (see Figure 4.10), the clustering coefficient that one can achieve with the triad model is 0.85 (within 3.5% of the empirical network) for $p_t \in [0.8,\ 0.9]$.

4.6 Summary

In this chapter, we analyzed and synthesized certain topological properties of the co-occurrence network of consonants which is the one-mode projection of the bipartite network that we defined in the previous chapter to represent the consonant inventories. Some of our important observations are

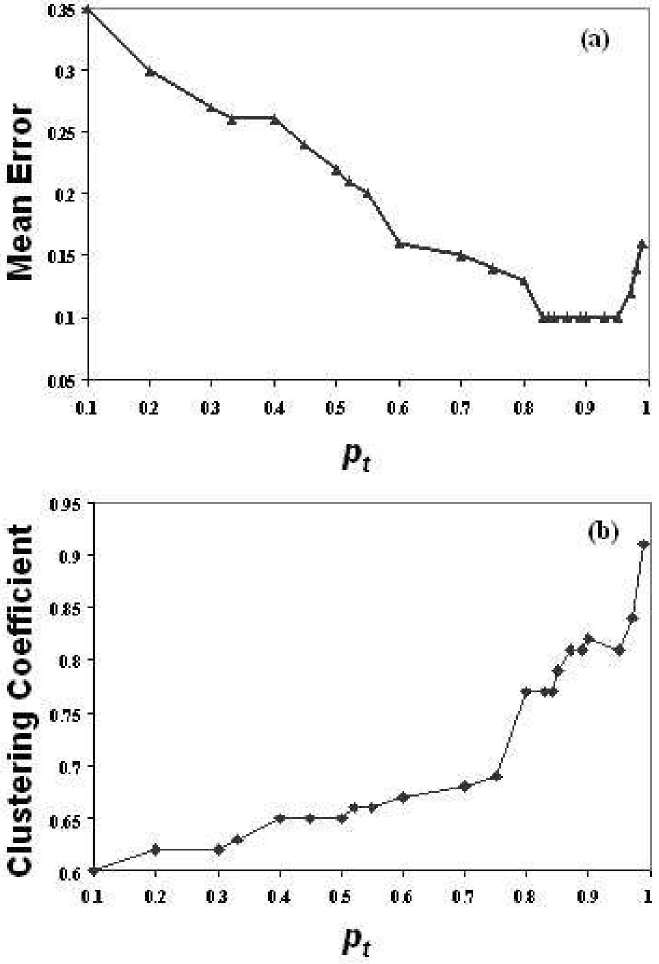


Figure 4.9: The effect of p_t on the degree distribution of the V_C nodes in the bipartite network and on the clustering coefficient of the corresponding one-mode projection. The model parameters are: $N = 500$, $t = 500$, $\mu = 22$ and $\gamma = 10$. (a) The mean error between the degree distribution of the V_C nodes in the bipartite network generated using the simple preferential attachment model and that using the triad model for various values of p_t; (b) the clustering coefficients of the one-mode projections of the bipartite networks generated in (a) versus p_t

(i) Although the emergent degree distribution of the consonant nodes in PlaNet is not sensitive to distribution of the size of the consonant inventories, the degree distribution of PhoNet is affected by this distribution.

(ii) The clustering coefficient of PhoNet is quite high like many other social networks indicating the presence of a large number of triangles in the network.

(iii) One can successfully explain the high clustering coefficient of PhoNet by having

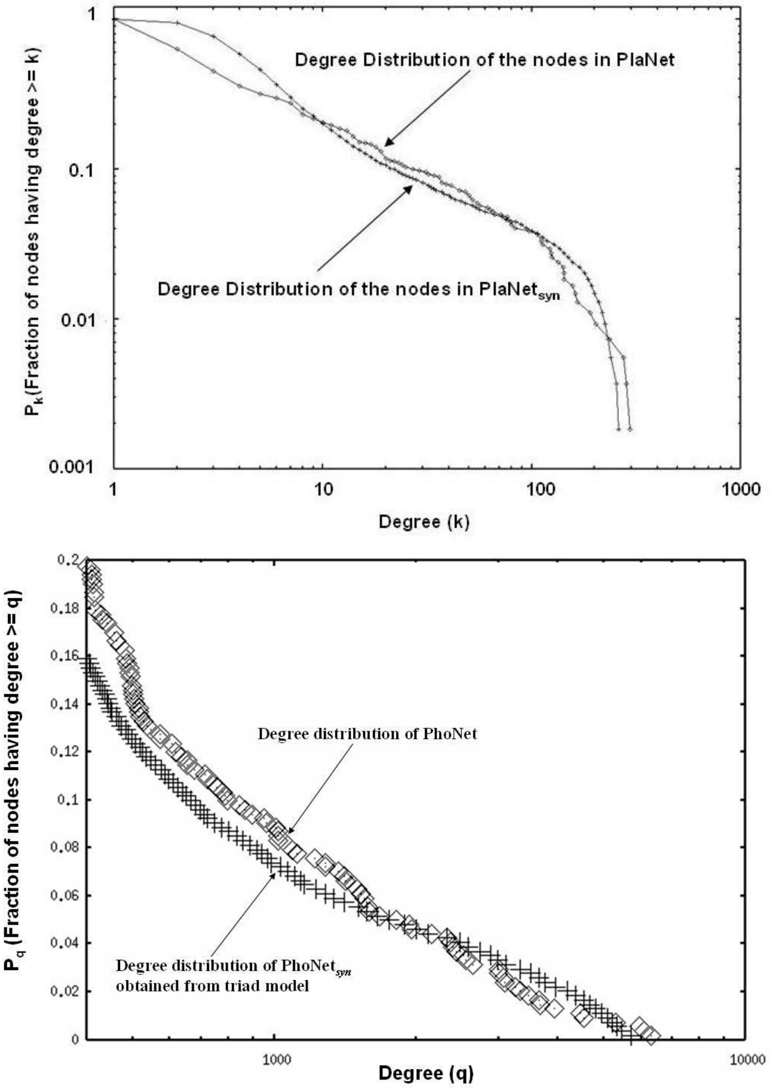


Figure 4.10: Degree distribution of PlaNet$_{syn}$ and PhoNet$_{syn}$ obtained from the triad model along with their corresponding real counterparts. For PlaNet$_{syn}$ the degree distribution is in doubly-logarithmic scale and for PhoNet$_{syn}$ the x-axis is in logarithmic scale

a sophisticated network growth model based on preferential attachment coupled with triad formation.

Linguistic Significance of the Refined Model

One can again find a possible association of the triad model with the phenomenon of language change. If a group of consonants largely co-occur in the languages of a generation of speakers then it is very likely that all of them get transmitted together in the subsequent generations [18]. The process of triad formation in our model is actually a reflection of this fact. Since the value of p_t that we obtain is quite high, it may be argued that such transmissions in groups are largely prevalent in nature.

It is interesting to note that whereas triad formation among consonants takes place in a top-down fashion as a consequence of language change over linguistic generations, the same happens in a social network in a bottom-up fashion where actors come to know one another through other actors and, thereby, slowly shape the structure of the whole network. Moreover, unlike in a social network where a pair of actors can regulate (break or acquire) their relationship bonds, if the co-occurrence bond between two consonants breaks due to pressures of language change it can be never acquired again[1]. Such a bond breaks only if one or both of the consonants are completely lost in the process of language change and is never formed in future since the consonants that are lost do not reappear again. In this context, Darwin in his book, *The Descent of Man* [43], writes "A language, like a species, when once extinct never reappears."

Although the direction of growth in a social network is different from the network discussed here, both of them target to achieve the same configuration. It is mainly due to this reason that the principle of preferential attachment along with that of triad formation is able to capture to a large extent the self-organizing behavior of the consonant inventories.

In the next chapter, we shall delve deeper into the co-occurrence properties of the consonants. In particular, we shall attempt to capture the "patterns of co-occurrence" of the consonants by finding groups/communities in which they tend to occur highly across the language inventories. We shall also figure out the precise reason for such pattern formation and quantitatively justify our argument.

[1] There is a very little chance of reformation of the bond if by coincidence the speakers learn a foreign language which has in its inventory one of the consonants lost due to language change.

Chapter 5

Patterns of Co-occurrence across the Consonant Inventories

An important observation that has been repeatedly made about the consonant inventories is that consonants tend to occur in pairs which show strong correlation in terms of their articulatory/acoustic features. In other words, consonants have a tendency to form groups or communities that effectively reflect their patterns of co-occurrence across the languages of the world. As we had already pointed out in Chapter 2, it was in order to explain these trends that feature economy was proposed as the basic organizing principle of the consonant inventories. In this chapter, we present a method to automatically discover the patterns of co-occurrence of the consonants across languages and also introduce an information theoretic measure for feature economy in order to establish that it is actually the driving force that leads to the emergence of these patterns. More specifically, for the purpose of capturing these patterns, we analyze PhoNet from the perspective of a social network where consonants are thought to exhibit community structures. Note that unlike in Chapters 3 and 4, here we assume that the consonant nodes are labeled, i.e., they are marked by the features that characterize them. Interestingly, the consonants forming the communities reflect strong correlations in terms of their features, which points to the fact that feature economy binds these communities. Another important observation is that if we treat each individual consonant inventory as a community in itself, then the amount of redundancy present across them (in terms of the features) is found to be (almost) constant irrespective of the inventory size.

This chapter is organized as follows. In section 5.1, we present a brief review on community analysis of complex networks. In the same section, we modify the community identification algorithm proposed by Radicchi *et al.* [129] for weighted networks like PhoNet and thereby, identify the consonant communities. In section 5.2, we test the goodness of the communities detected by our algorithm and observe that the constituent consonants forming these communities frequently occur is similar groups in real languages also. We present a mathematical formulation for the concept of feature economy in section 5.3 and suitably employ this formula to show that indeed the consonant communities obtained from PhoNet are significantly more economic than the case where the consonant inventories are assumed to have been generated randomly. In the next section, we extend our formula for feature economy to mathematically express the redundancy across the consonant inventories and show that it remains fixed irrespective of the consonant inventory size thereby, unfolding an universal structural property of these inventories. Finally, in section 5.5, we summarize the contributions of this chapter and point out some of the linguistic implications of our results.

5.1 Identification of Community Structures

There is a large volume of literature suggested by computer scientists, physicists as well as sociologists that describe various methods for identifying communities in a network [54, 60, 71, 81, 115, 126, 129]. This is mainly because, the ability to find communities within large networks in some automated fashion can provide a considerable insight into the organization of the network. Communities in a web graph, for instance, might correspond to sets of web sites dealing with related topics [54], while communities in a biochemical network might correspond to functional units of some kind [71]. In the following, we review a few community-finding algorithms proposed in the areas of computer science, sociology and more recently in complex networks (for an extensive survey on community structure analysis see [57]). We further describe the Radicchi *et al.* [129] algorithm and its extension for the community analysis of PhoNet.

5.1.1 Review of Community-Finding Algorithms

Clustering or community-finding deals with the detection of intrinsic groups present in a network. A loose definition of clustering therefore could be "the process of organizing vertices into groups whose members are similar in some way". Consequently, a cluster/community is a collection of vertices which are "similar" among them and "different" from vertices belonging to the other clusters. Most of the clustering algorithms assume that the number of edges among the vertices within a group (i.e., similar vertices) by far outnumber the edges between vertices from different groups. This problem of clustering turns out to be extremely important in various disciplines including computer science, sociology and network theory. Some of the popular algorithms that have been devised by the researchers in each of the above disciplines to tackle this problem are presented below.

Algorithms in Computer Science

In computer science, community-finding corresponds to the *graph partitioning* problem that involves dividing a graph into pieces such that the pieces are of about the same size and there are only a few connections in between the pieces. Various approaches have been proposed in the literature including those based on (i) different optimization techniques as in [81, 100], (ii) spectral bisection [139] and (iii) message passing [58].

Optimization: One of the classic examples of graph bisection using optimization techniques is the Kernighan-Lin algorithm [81]. This is a greedy optimization method that assigns a benefit function Q to divisions of the network and then attempts to optimize this benefit over possible divisions. The benefit function is equal to the number of edges that are present within the two groups minus the number of edges that run between them. As an initial condition, one has to specify the size of the two groups into which the network is to be divided and the start configuration for the groups (may be random). The algorithm proceeds in two stages. In the first stage, a vertex is chosen from each of the groups and the change in the benefit function is calculated if these two vertices are swapped. The pair that maximizes this change is chosen and swapped. This process is repeated however, with a limitation that a vertex that has been swapped once is locked and is never swapped again. Once all the vertices in a group have been locked this stage of the algorithm ends. In the second stage, the sequence of swaps made are revisited so as to find out the point where Q was maximum. The corresponding configuration is chosen to be the bisection of the graph.

Another popular method based on optimization is the *k-means* clustering [100]. Here the number of clusters is preassigned to a value, say k. Each vertex is embedded as a point (x_i) in a metric space and a distance measure is defined between the points in this space. The distance measure indicates the dissimilarity between the pairs of vertices. The goal is to partition the points into k sets ($\mathbf{S} = \{S_1, S_2, \ldots, S_k\}$) so as to minimize a cost function. The cost function takes the following form

$$\underset{\mathbf{S}}{argmin} \sum_{i=1}^{k} \sum_{x_j \in S_i} || \, x_j - c_1 \, ||^2$$

where c_1 is the *centroid* (i.e., the center) of the points in S_i. The k-means clustering problem can be solved using the Lloyd's algorithm [97]. The algorithm begins with an initial distribution of centroids that are as far as possible from each other. In the first iteration, each vertex is assigned to a centroid. The centers of mass of the k clusters formed are then computed and this new set of centroids allows for a new classification of the vertices in the second iteration, and so on. Within a few iterations the centroids become stable and the clusters do not change any more.

Spectral Bisection: Given an adjacency matrix $\mathbf{A}$ of a graph G, one can construct the *Laplacian* matrix $\mathbf{L}$ where $\mathbf{L} = \mathbf{D}$ - $\mathbf{A}$ and $\mathbf{D}$ is the diagonal matrix of node degrees.

In other words, each diagonal entry of the matrix $\mathbf{D}$ is $d_{ii} = \sum_j a_{ij}$ where a_{ij} are the entries of the matrix $\mathbf{A}$. All the other entries of matrix $\mathbf{D}$ are zero. Spectral clustering techniques make use of the spectrum of the matrix $\mathbf{L}$ to perform dimensionality reduction for clustering in fewer dimensions. One such technique is the Shi-Malik algorithm [139], commonly used for image segmentation. The algorithm partitions based on the eigenvector v corresponding to the second smallest eigenvalue of the matrix $\mathbf{L}$. This partitioning can be done in various ways, such as by taking the median m of the components of the vector v, and placing all nodes whose component in v is greater than m in one partition, and the rest in the other partition. The algorithm can be used for hierarchical clustering by repeatedly partitioning the subsets in this fashion.

Message Passing: A set of data points can be clustered via passing certain messages between the points. The key idea is to construct a similarity network with nodes representing the data points and the weights on the edges representing some similarity between the data points and then iteratively propagating messages from a node to its neighbors finally resulting into to the emergence of the clusters. A popular algorithm in this category is the one suggested by Frey and Dueck [58] that is based on finding data centers or "exemplars" through a method which they call "affinity propagation". In this algorithm, each node in the similarity network is assumed to be a potential exemplar and real-valued messages are exchanged between nodes until a high-quality set of exemplars and corresponding clusters gradually emerges. There are two kinds of messages that are exchanged between the nodes, namely "responsibilities" and "availabilities". Responsibilities $r(i, k)$ indicate the accumulated evidence for how well-suited the node k is to serve as the exemplar for the node i, taking into account other potential exemplars for the node i. Responsibilities are sent from node i to the candidate exemplar k. On the other hand, availability $a(i, k)$ indicates the accumulated evidence for how appropriate it would be for the node i to choose node k as its exemplar, taking into account the support from other nodes that node k should be an exemplar. Availabilities are sent from candidate exemplar k to the node i.

The algorithm consists of some simple update rules and messages are exchanged only between node pairs connected by an edge. There are three steps in which it proceeds – (i) update responsibilities given availabilities, (ii) update availabilities given the responsibilities, and (iii) monitor exemplar decisions by combining availabilities and responsibilities. Terminate if the local decisions stay (almost) unchanged over successive iterations.

Algorithms in Sociology

Sociologists concerned with the analysis of social networks have developed a large volume of literature that deals with the problem of community identification. The primary technique adopted is known as *hierarchical clustering* [135]. A hierarchical clustering on a network of n nodes produces a series of partitions of the network, P_n, P_{n-1}, ..., P_1. The first partition P_n consists of n single-node clusters, while the last partition P_1, consists of a single cluster containing all the n nodes. At each particular stage the method joins together the two clusters which are closest (or most similar) to each other. Note that at the first stage, this amounts to joining together the two nodes that are most similar to each other.

Differences between methods arise because of the different ways of defining the distance (or similarity) between clusters. The most popular similarity measure in the sociology literature is called *structural equivalence*. Two nodes in the network are said to be structurally equivalent if they have exactly the same set of neighbors (other than each other, if they are connected). In most cases, exact structural equivalence is rare and therefore one needs to define the degree of equivalence between node pairs which, may be done in several ways. One of them, known as *Euclidian distance* [160], is defined as

$$dist_{ij} = \sqrt{\sum_{k \neq j,i} (a_{ik} - a_{jk})^2}$$

where a_{ik} and a_{jk} are entries of the adjacency matrix. The more structurally similar the nodes i and j are, the lower is the value of $dist_{ij}$. Another commonly used similarity metric is the Pearson's correlation between rows (or columns) corresponding to the nodes i and j in the adjacency matrix [160].

There are also several techniques of joining the clusters such as the

(i) *Single linkage clustering*: One of the simplest form of hierarchical clustering method is single linkage, also known as the nearest-neighbor technique. The fundamental characteristic of this method is that distance between the clusters is defined as the distance between the closest pair of nodes, where only pairs consisting of one node from each cluster are considered.

(ii) *Complete linkage clustering*: The complete linkage, also known as the farthest-neighbor, clustering method is the opposite of single linkage. Distance between clusters is defined here as the distance between the most distant pair of nodes, one chosen from each cluster.

(iii) *Average linkage clustering*: In this case, the distance between two clusters is defined as the average of distances between all pairs of nodes, where each pair is made up of one node from each cluster.

Algorithms in Network Theory

Clustering or community analysis is also a very well-researched area in the field of complex networks. Various algorithms have been proposed in order to detect "natural" divisions of the vertices in the network. In the following, we shall outline a few of these techniques that have become quite popular in the recent times.

Spin Model based Algorithm: One of the very popular models in the field of statistical mechanics, usually applied to study various properties of ferromagnets, is the Potts model [166]. The model describes a system of spin interactions where a spin can be in n different states. If the interactions are ferromagnetic then at the ground state all the spins have the same value (i.e., they are aligned). However, if the interactions are anti-ferromagnetic then in the ground state of the system there are different spin values co-existing in homogeneous clusters. If the Potts spin variables are assigned to the vertices of a network where the interactions are between the neighboring spins then the communities can be recovered from the like-valued spin clusters in the system. In [133], the authors propose a method to detect communities where the network is mapped onto a n-Potts model with nearest neighbor interactions. The Hamiltonian (i.e., the total energy) of the model can be expressed as

$$H = -J \sum_{i,j} a_{ij} \delta(\sigma_i, \sigma_j) + \kappa \sum_{s=1}^{n} \frac{n_s(n_s-1)}{2}$$

where a_{ij} is an element of the adjacency matrix, δ is the Kronecker delta function [74], σ_i and σ_j stand for the spin values, n_s is the number of spins in the state s and J and κ are the coupling parameters. H is a sum of two competing terms: one is the classical ferromagnetic Potts model energy that favors spin alignment while the other peaks when the spins get homogeneously distributed. The ratio κ/J controls the relative importance of these two terms. H is minimized via the process of simulated annealing (see [83]) that starts from an initial configuration where spins are randomly assigned to the vertices and the number of states n is kept very high. The authors in [133] show that the procedure is quite fast and the results do not depend on n (as long as n is sufficiently high).

Random Walk based Algorithm: Random walks [74] on a graph can be very useful for detecting communities. The basic idea is that if the graph has strong community structures then a random walker spends a long time inside a community due to the high density of the internal edges and the number of paths that could be followed. In [167], the author uses random walks to define the distance between pairs of vertices. The distance $dist_{ij}$ between vertices i and j is defined as the average number of edges that a random walker has to cross to reach j starting from i. Vertices that are close to each other are likely to belong to the same community. The author defines a “global attractor” of a vertex i to be a vertex closest to i, that is, any vertex which is at a smallest distance from i. On the other hand, the “local attractor” of i are its nearest neighbors (i.e., vertices directly sharing an edge with i). Two types of communities are defined based on the local and the global attractors: i has to be put into the community of its attractor and also into the community of all other vertices for which i is an attractor. Communities have to be minimal subgraphs i.e., they cannot have smaller subgraphs that are also communities as per the previously mentioned definition. Other variants of random walk based clustering algorithms may be found in [73, 90, 168].

Clique Percolation based Algorithm: In certain real-world graphs a particular vertex may belong to more than one community. For instance, a polysemous word in a semantic network (see Chapter 1) can belong to each different community of words that corresponds to a particular sense of that polysemous word. In such cases, one needs to detect *overlapping* communities and the term "cover" is more appropriate than partition. The most popular technique to detect overlapping communities is the clique percolation method [120]. This method is based on the concept that the internal edges of a community together form a clique while inter-community edges do not form such cliques. If it were possible for a k-clique (i.e., a clique composed of k nodes) to move on a network then it would probably get trapped within its original community as it would not be able to cross the bottleneck formed by the inter-community edges. The authors define a few terms so as to realize this idea – (i) two k-cliques are *adjacent* if they share $k-1$ nodes, (ii) the union of adjacent cliques is a *k-clique chain*, (iii) two k-cliques are connected if they are a part of a k-clique chain and (iv) a *k-clique community* is the largest connected subgraph formed by the union of a k-clique and of all k-cliques that are connected to it. The identification of a k-clique community is carried out by making a k-clique "roll" over adjacent k-cliques, where rolling means rotating a k-clique about the $k-1$ vertices it shares with any of the adjacent k-cliques (see [51] for a method outlining the computation of k-cliques and their overlaps). Since, by construction, k-clique communities can share vertices therefore they can be overlapping.

The Girvan-Newman Algorithm: The Girvan-Newman algorithm [60] focuses on those edges in the network that are least "central", that is the edges which are most "between" the communities. The "edge betweenness" of an edge is defined as the number of shortest paths between pairs of nodes that run along it. If there are more than one shortest paths between a pair of nodes then each of these paths is assigned equal weight such that the total weight of all of the paths sums to unity. If a network contains communities that are only loosely connected by a very few inter-community edges, then all the shortest paths between different communities must go along one of these few edges. Thus, the edges connecting communities should have high edge betweenness (at least one of them). By removing these edges, the communities can be separated from one another thereby, revealing the underlying community structure of the network.

In short, the steps in which the algorithm proceeds are – (i) calculate the edge betweenness of all existing edges in the network, (ii) remove the edge with the highest betweenness, (iii) recalculate the betweenness of all edges affected by the removal, and (iv) repeat steps 2 and 3 until no edges remain.

The Radicchi *et al.* Algorithm: The algorithm of Radicchi *et al.* [129] counts, for each edge, the number of loops of length three it is a part of and declares the edges with very low counts as inter-community edges. The basic idea is that the edges that run between communities are unlikely to belong to many short loops, because, to complete a loop containing such an edge there needs to be another edge that runs between the same two communities, and such other edges are rare. Therefore, it should be possible to spot the between-community edges by looking for the ones that belong to an unusually small number of loops. Such edges can be iteratively removed to decompose the network into disjoint communities.

5.1.2 Community Analysis of PhoNet

We modify the Radicchi *et al.* [129] algorithm (henceforth termed as MRad) to make it suitable for weighted networks and subsequently use it to conduct the community structure analysis of PhoNet. There are mainly two reasons for choosing this algorithm – (a) it is fast, and (b) it can be easily modified to work for the case of weighted networks. The basis and the modification of the algorithm are as follows.

Basis: Edges that run between communities should not belong to many triangles, because, to complete the triangle containing such an edge there needs to be another edge that runs between the same two communities, and such other inter-community edges are, by definition, rare.

Modification for a Weighted Network: Nevertheless, for weighted networks, rather than considering simply the triangles (loops of length three) we need to consider the weights on the edges forming these triangles. The basic idea is that if the weights on the edges forming a triangle are comparable then the group of consonants represented by this triangle highly occur together rendering a pattern of co-occurrence, while if these weights are not comparable then there is no such pattern. In order to capture this property, we define the edge-strength S for each edge of PhoNet as follows. Let the weight of the edge (u, v), where u, $v \in V_C$, be denoted by w_{uv}. S can be expressed as a ratio of the form,

$$S = \frac{w_{uv}}{\sqrt{\sum_{i \in V_C - \{u,v\}} (w_{ui} - w_{vi})^2}} \tag{5.1}$$

if $\sqrt{\sum_{i \in V_C - \{u,v\}} (w_{ui} - w_{vi})^2} > 0$ else $S = \infty$. The expression for S indicates that the strength of connection between two nodes u and v depends on (i) the weight of the edge (u, v) and (ii) the degree to which the weights on the edges forming triangles with (u, v) are comparable. If the weights are not comparable then the denominator will be high, thus reducing the overall value of S. PhoNet can be then decomposed into communities by removing edges that have S less than a specified threshold (say η). The entire idea is summarized in Algorithm 5.1. Figure 5.1 illustrates the process of community formation.

It is important to mention here that while employing the above algorithm for community detection, we have neglected those nodes in PhoNet that correspond to consonants which occur in less than 5 languages[1] in UPSID. Since the frequency of occurrence of each such consonant is extremely low therefore, the communities they form can be assumed to be statistically insignificant. Furthermore, we have also

[1] This number has been decided through the manual inspection of the data.

Algorithm 5.1: The MRad algorithm

```
Input: PhoNet
for each edge $(u, v)$ do
    Compute
    $S = \frac{w_{uv}}{\sqrt{\sum_{i \in V_C - \{u,v\}} (w_{ui} - w_{vi})^2}}$
    if $\sqrt{\sum_{i \in V_C - \{u,v\}} (w_{ui} - w_{vi})^2} > 0$ else $S = \infty$;
end
Redefine the edge-weight for each edge $(u, v)$ by $S$;
Remove edges with edge-weights less than or equal to a threshold $\eta$;
Call this new version of PhoNet, PhoNet$_\eta$;
Find the connected components in PhoNet$_\eta$;
```

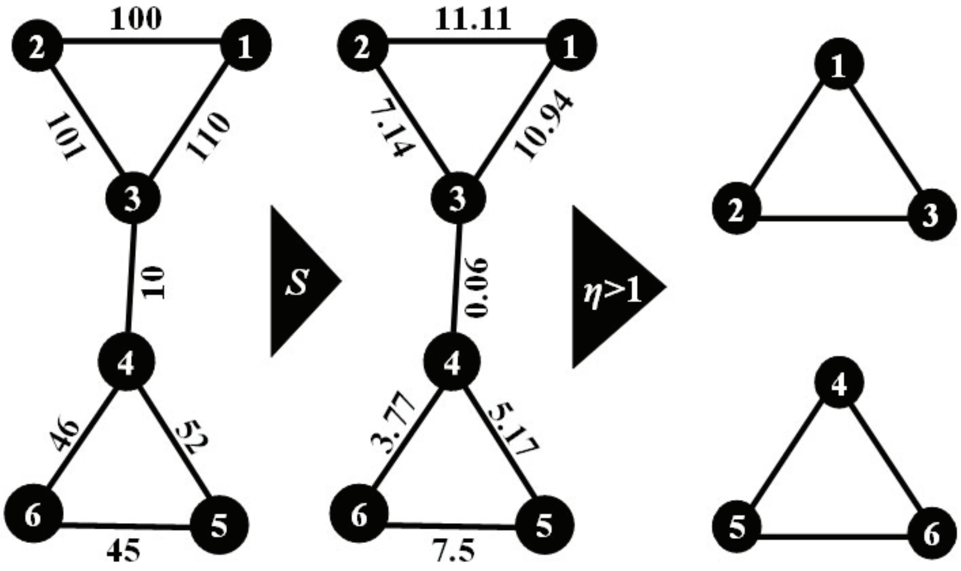


Figure 5.1: The process of community formation

removed those nodes that correspond to consonants which have a very high frequency of occurrence. Since such consonants co-occur with almost every other consonant (by virtue of their high frequency) the edge-strength S is likely to be high for the edges that connect pairs of nodes corresponding to these high frequency consonants. The value of S for these edges is much higher than η and as they do not get removed from the network therefore, they can pull in the nodes of two clearly disjoint communities

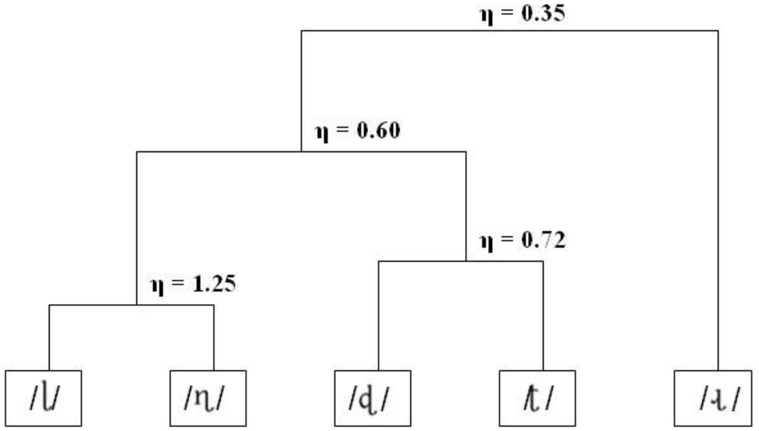


Figure 5.2: The dendrogram illustrates how the retroflex community of /ɖ/, /ʈ/, /ɳ/, /ɭ/ and /ɻ/ is formed with the change in the value of η

into a single community. For instance, we have observed that since the consonants /m/ and /k/ are very frequent, the edge connecting the nodes corresponding to them has a high edge-strength. The strong link between /m/ and /k/ forces the sets of bilabial and velar consonants which should ideally form two different communities to merge into a single community. Hence, we have removed nodes which correspond to consonants that occur in more than 130 languages[2] in UPSID (a total of 13 nodes). Note that even these 13 nodes which form a hub-like structure also indicate a high correlation among the features that characterize them thus attesting the presence of feature economy.

We can obtain different sets of communities by varying the threshold η. As the value of η decreases, new nodes keep joining the communities and the process becomes similar to hierarchical clustering [135]. Figure 5.2 shows a dendrogram, which illustrates the formation of the community of the consonants /ɖ/, /ʈ/, /ɳ/, /ɭ/ and /ɻ/ with the change in the value of η.

Some of the example communities obtained from our algorithm are noted in Table 5.1. In this table, the consonants in the first community are dentals, those in the second community are retroflexes, while the ones in the third are all laryngealized.

[2] This number has again been decided through the manual inspection of the data.

Table 5.1: Consonant communities

Community	Features in Common
/t/, /d/, /n/	dental
/ɖ/, /ʈ/, /ɳ/, /ɭ/, /ɻ/	retroflex
/w̰/, /j̰/, /m̰/	laryngealized

5.2 Evaluation of the Communities based on their Occurrence in Languages

In this section, we inspect whether the consonant communities detected from PhoNet by the MRad algorithm are actually found to occur significant number of times in such groups across the languages of UPSID.

For this purpose, we first arrange the consonants forming a community COM, of size N, in an ascending order of their frequency of occurrence in UPSID. We associate a rank R with each of the consonants in COM where the least frequent consonant (frequency calculated from UPSID) gets a rank $R = 1$, the second least gets a rank $R = 2$ and so on. Starting from rank $R = 1$, we count how many of the consonants in COM, occur in a language $L \in$ UPSID. Let the number of such consonants be M. We define the *occurrence ratio* O_L of the community COM for the language L to be

$$O_L = \frac{M}{N - (R_{top} - 1)} \tag{5.2}$$

where R_{top} is the rank of the highest ranking consonant that is found in L. The denominator of this ratio is $N - (R_{top} - 1)$ instead of N since it is not mandatory for a language to have a low frequency member of a community if it has the high frequency member; nevertheless, if the language already has the low frequency member of the community then it is highly expected to also have the high frequency member. For instance, let the community COM be formed of three consonants /k^w/, /k^h/ and /k/ (arranged in ascending order of their frequencies). When we inspect a language L, it is not necessary for it to have /k^w/ or /k^h/ if it has /k/ in its inventory; nevertheless

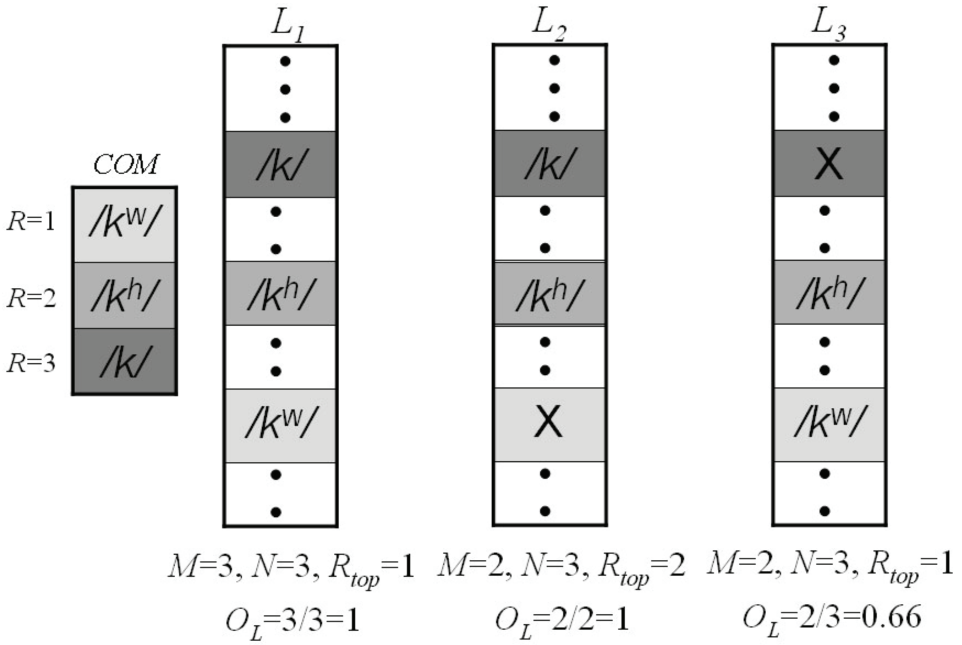


Figure 5.3: Calculation of occurrence ratio. A “X” indicates the absence of a particular consonant in a language

it is highly expected that if it already has /k^w/, it should also have /k/ and /k^h/ in its inventory (see [39] for further discussions about this linguistic phenomena). The idea is illustrated in Figure 5.3 through three example scenarios.

The average occurrence ratio O_{av} for the community COM can be obtained as follows,

$$O_{av} = \frac{\sum_{L \in UPSID} O_L}{L_{occur}} \tag{5.3}$$

where L_{occur} is the number of languages in UPSID that have at least one or more of the consonants occurring in COM. Figure 5.4 shows the average O_{av} of the communities obtained at a particular threshold η versus the threshold η. The curve clearly shows that the average O_{av} of the communities obtained from our algorithm for $\eta > 0.3$ is always more than 0.8. This in turn implies that, on an average, the communities obtained at thresholds above 0.3 occur in more than 80%[3] of the languages in UPSID.

[3] The expectation that a randomly chosen set of consonants representing a community of size

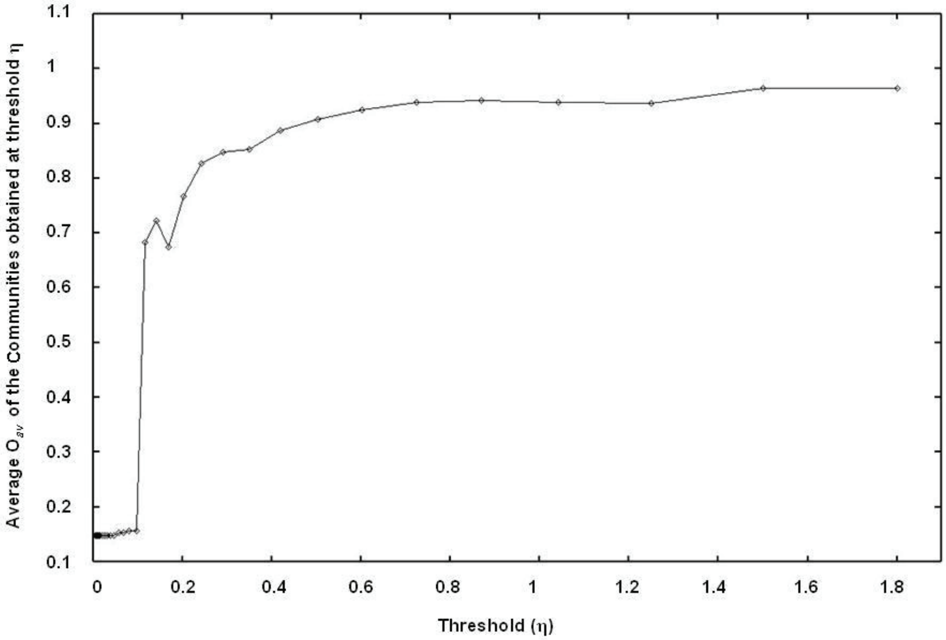


Figure 5.4: Average O_{av} of the communities obtained at a particular threshold η versus the threshold η

At thresholds below 0.3 the average O_{av} is found to fall gradually. This is because, very large size communities start emerging and the probability that all the constituent consonants of such a large community occur together across languages is very low. Thus, the value of M and consequently, O_L drops, thereby, reducing the overall value of the average O_{av}. In short, the communities obtained from our algorithm can be assumed to be true representatives of the patterns of co-occurrence of the consonants across languages.

between 2 to 5, occurs in a language, is 70% whereas the same is 89% for the communities obtained from PhoNet.

5.3 Feature Economy: The Binding Force of the Communities

In the earlier sections, we have mainly focused on the detection and evaluation of the communities emerging from PhoNet. In this section, we attempt to quantitatively show that feature economy (see Chapter 2 for definition) is the driving force that leads to the emergence of these communities.

The central idea behind this quantification is to determine the extent to which the features contribute to discriminate the consonants in a community. The less discriminative the features are on an average, the more close are the consonants and higher is the feature economy. In the following, we shall show how the discriminative capacity of the features in a community of consonants can be quantified through an information theoretic approach. We shall first state our assumptions and then outline the quantification process.

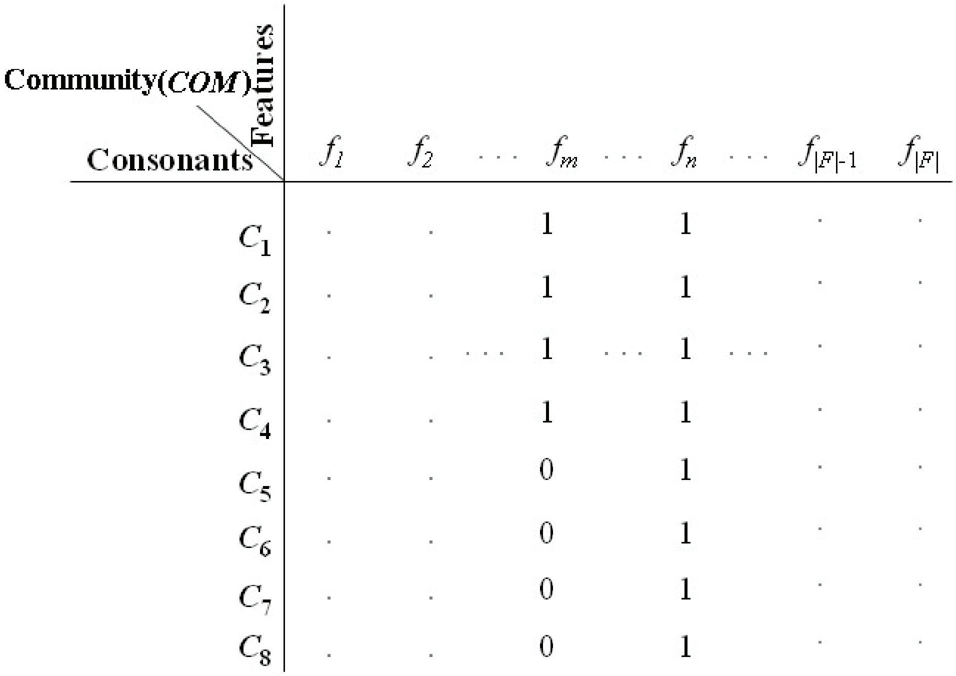
5.3.1 Assumptions

Let us assume that for a community COM of size N each consonant C_x $(1 \leq x \leq N)$ can be characterized by a set of features F. Further, let each feature $f_i \in F$ be independent (as in [39]) and binary-valued (as in [88, 101, 102, 103]). It is important to note that the assumption that the features are independent and binary is not always true. In the words of Ladefoged and Maddieson [88] "...the tongue has no clearly defined regions, and neither does the roof of the mouth; and there are, of course, similar problems in dividing the continua which underlie other phonetic parameters." However, they also state that "...it is very striking that languages often cut the continua in similar ways so that it is possible to equate a sound in one language with a similar sound in another." Along these lines, we assume that (a) the value of f_i is not dependent on the value of any other feature $f_{j \neq i} \in F$ and, (b) $f_i = 1$ if it is present in a consonant and 0 otherwise. Therefore, every consonant C_x in COM can be encoded as a binary vector such that $C_x = [f_1 \ f_2 \ \ldots \ f_{|F|-1} \ f_{|F|}]$, where $f_{1 \leq i \leq |F|} = \{0, 1\}$. In other words, for our analysis, each feature has a 1/0 value depending on whether a consonant uses it or not. Note that the assumption that the features are independent does not lead us to impossible configurations that is, a consonant, for instance, cannot have two primary places of articulation (e.g., both labial and velar). Since such configurations are never encountered in our data set, while one of the place-features would take up the value 1, the others shall take up the value 0. Stated differently, if there are n place-features and any one of them has the value 1, then it automatically implies that all the others have the value 0. It is not difficult to extend the idea to multi-valued features where a single discrete variable taking up different values would correspond to different places of articulation (we shall provide one representative result based on this multi-valued assumption in section 5.4 to show that the inferences that we draw still remain valid under this assumption). However, as the binary representation is easy to interpret and as there are no known standards for assigning values to a particular feature on a multi-valued scale, we shall continue our analysis with the binary representation.

5.3.2 The Quantification Process

Since each consonant is encoded as a binary vector therefore, the community COM is essentially a set of binary vectors as indicated by the example in Figure 5.5. If a feature takes the value 1 (or 0) for all the consonants in COM (such as f_n in Figure 5.5) then it does not contribute to differentiate the consonants and hence its *discriminative capacity* (henceforth DC) should be 0. On the other hand, if the feature takes up the value 1 for one half of the consonants in COM and the value 0 for the other half (such as f_m in Figure 5.5) then its contribution towards differentiating the consonants is maximum and hence, its DC should be 1. This observation leads us to the following information-theoretic definition of the DC of a feature in a particular community.

Let there be p_{f_i} consonants in COM in which the feature f_i is present (i.e., $f_i = 1$) and q_{f_i} consonants in which f_i is absent (i.e., $f_i = 0$). Given that the size of COM is N, the probability that $f_i = 1$ for a particular consonant chosen uniformly at



Community(COM) \ Features Consonants	f_1	f_2	⋯	f_m	⋯	f_n	⋯	$f_{\lvert F\rvert-1}$	$f_{\lvert F\rvert}$
C_1	.	.		1		1		.	.
C_2	.	.		1		1		.	.
C_3	.	.	⋯	1	⋯	1	⋯	.	.
C_4	.	.		1		1		.	.
C_5	.	.		0		1		.	.
C_6	.	.		0		1		.	.
C_7	.	.		0		1		.	.
C_8	.	.		0		1		.	.

Figure 5.5: A hypothetical community of eight consonants

random from COM is $\frac{p_{f_i}}{N}$. Consequently, the probability that $f_i = 0$ is $\frac{q_{f_i}}{N}$. Note that $\frac{p_{f_i}}{N} + \frac{q_{f_i}}{N} = 1$. Stated differently, f_i is an independent random variable, which can take values 1 and 0, and $\frac{p_{f_i}}{N}$ and $\frac{q_{f_i}}{N}$ define the probability distribution of f_i. Clearly, the discriminative capacity of the feature f_i in COM is dependent on the above probability distribution and the functional form that can appropriately quantify DC is the binary entropy H_{f_i} defined as [138]

$$H_{f_i} = -\frac{p_{f_i}}{N}\log_2\frac{p_{f_i}}{N} - \frac{q_{f_i}}{N}\log_2\frac{q_{f_i}}{N} \tag{5.4}$$

It is easy to show that H_{f_i} always produces the desired value of the DC of feature f_i in COM. For instance, if $f_i = 1$ for all the consonants in COM (like f_n in Figure 5.5) then $p_{f_i} = N$ and $q_{f_i} = 0$. Therefore, $DC = H_{f_i} = 0$, which should ideally be the case. The same argument also holds if $f_i = 0$ for all the consonants. In that case, $p_{f_i} = 0$ and $q_{f_i} = N$, which results in $DC = H_{f_i} = 0$. On the other hand, if $f_i = 1$ for one half of the consonants and 0 zero for the other half then $p_{f_i} = \frac{N}{2}$ as well as $q_{f_i} = \frac{N}{2}$ and consequently, $DC = 1$ as per the requirement. For

Table 5.2: The process of computing the value of F_E for the two hypothetical communities COM_1 ($F_E = 2.75$) and COM_2 ($F_E = 4.58$)

COM_1	voiced	dental	bilabial	velar	plosive	COM_2	voiced	dental	bilabial	nasal	retroflex	plosive
/b/	1	0	1	0	1	/b/	1	0	1	0	0	1
/d/	1	1	0	0	1	/n/	1	1	0	1	0	0
/g/	1	0	0	1	1	/ɖ/	1	1	0	0	1	0
p_f/N	1	0.33	0.33	0.33	1	p_f/N	1	0.67	0.33	0.33	0.33	0.33
q_f/N	0	0.67	0.67	0.67	0	q_f/N	0	0.33	0.67	0.67	0.67	0.67

any other combination of the values of p_{f_i} and q_{f_i}, the corresponding value of DC can be calculated using equation 5.4 and is bounded within the continuous range of [0, 1].

The total discriminative capacity of all the features in COM, which we shall call *feature entropy* (F_E) can be therefore, expressed as

$$F_E = \sum_{f_i \in F} H_{f_i} = \sum_{f_i \in F} \left(-\frac{p_{f_i}}{N} \log_2 \frac{p_{f_i}}{N} - \frac{q_{f_i}}{N} \log_2 \frac{q_{f_i}}{N}\right) \tag{5.5}$$

Since each feature f_i is assumed to be an independent random variable, F_E is essentially the joint entropy of the system. F_E can be thought of as an upper bound on the measure of the minimum number of distinctions that are important for a learner to pick up during the acquisition of the consonants in COM. Note that the lower the feature entropy the higher is the feature economy. The idea is illustrated in Table 5.2 through an example where F_E exhibited by the community COM_1 is lower than that of the community COM_2 because, in COM_1 the combinatorial possibilities of the features are better utilized by the consonants than in COM_2 (i.e., COM_1 is more economic than COM_2).

5.3.3 Experiments and Results

In order to establish the fact that feature economy is the key factor that drives the co-occurrence patterns of the consonants it is necessary to show that the communities obtained from PhoNet exhibit a significantly lower feature entropy than the case where the consonant inventories are assumed to have been generated randomly. For this purpose, we construct a random version of PhoNet (henceforth PhoNet$_{rand}$) and compare the communities obtained from it with those obtained from PhoNet in terms of feature entropy. We construct PhoNet$_{rand}$ as follows. Let the frequency of occurrence for each consonant C in UPSID be denoted by f_C. Let there be 317 bins each corresponding to a language in UPSID. f_C bins are then chosen uniformly at random and the consonant C is packed into these bins. Thus the consonant inventories of the 317 languages corresponding to the bins are generated[4]. Note that in such randomly constructed inventories the effect of feature economy should not be prevalent as there is no strict co-occurrence principle that plays a role in the process of inventory construction. Therefore, feature entropy in this case should be no better than what is expected by random chance. One can build PhoNet$_{rand}$ from these randomly generated consonant inventories in a procedure similar to that used for constructing PhoNet. The entire idea of constructing PhoNet$_{rand}$ is summarized in Algorithm 5.2.

Algorithm 5.2: Algorithm to construct PhoNet$_{rand}$

```
for each consonant $C$ do
    for $i$ = 1 to $f_C$ do
        Choose, uniformly at random, one of the 317 bins each of which
        corresponds to a language in UPSID;
        Pack the consonant $C$ into the bin so chosen if it has not been
        already packed into this bin earlier;
    end
end
Construct PhoNet$_{rand}$, similarly as PhoNet, from the new consonant
inventories (each bin corresponds to a new inventory) ;
```

We can apply the MRad algorithm to extract the communities from PhoNet$_{rand}$ similarly as in the case of PhoNet. Figure 5.6 illustrates, for all the communities obtained from PhoNet and PhoNet$_{rand}$, the average feature entropy exhibited by the communities of a particular size (y-axis) versus the community size (x-axis). The "average feature entropy exhibited by the communities of a particular size" can be

[4] This random model preserves the frequency of occurrence of the consonants across languages.

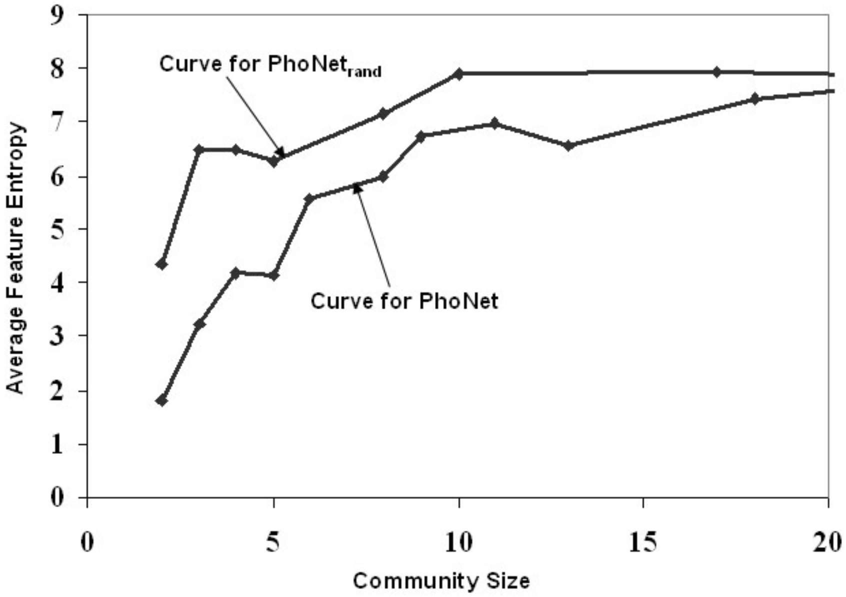


Figure 5.6: Average feature entropy of the communities of a particular size versus the community size for PhoNet as well as PhoNet$_{rand}$

calculated as follows. Let there be n communities of a particular size s obtained at all the different values of η. The average feature entropy of the communities of size s is $\frac{1}{n}\sum_{i=1}^{n} F_{E_i}$ where F_{E_i} signifies the feature entropy of the i^{th} community of size s. The curves in the figure make it quite clear that the average feature entropy exhibited by the communities of PhoNet are substantially lower than that of PhoNet$_{rand}$ (especially for a community size ≤ 20). As the community size increases, the difference in the average feature entropy of the communities of PhoNet and PhoNet$_{rand}$ gradually diminishes. This is mainly because of the formation of a giant community, which is similar for both PhoNet as well as PhoNet$_{rand}$. The above result indicates that the consonant communities in PhoNet are far more economic than what is expected by random chance. Note that if in contrast, the communities exhibit a feature entropy that is higher than that reflected by the randomly generated inventories then one can argue that on an average the features are more discriminative than expected by chance pointing to the prevalence of high perceptual contrast among the constituent nodes in the community (this shall become apparent from the analysis of the vowel communities presented in the next chapter).

In order to further strengthen the above argument one can inspect whether the

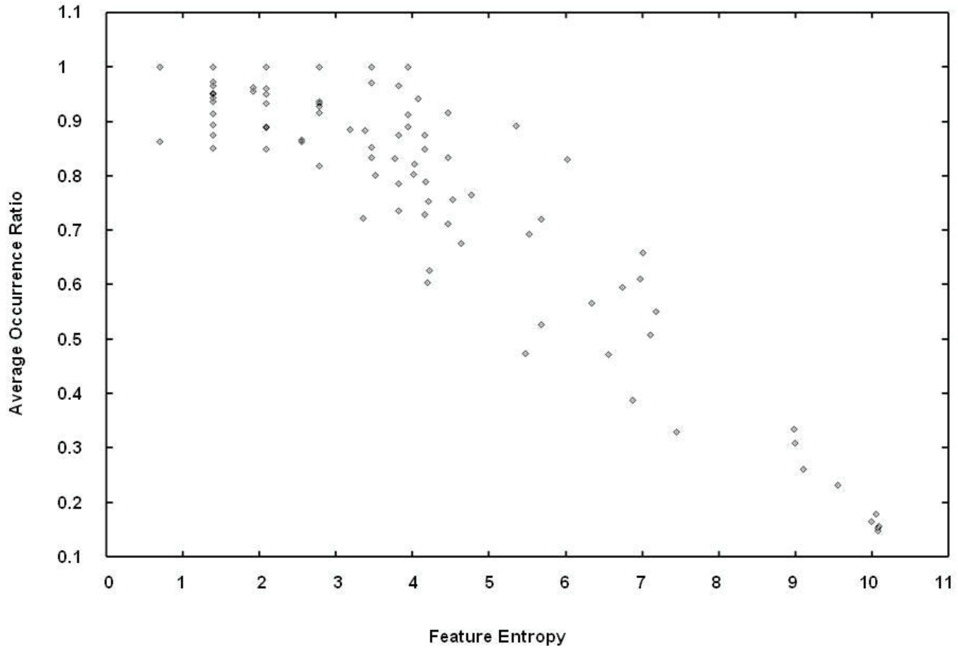


Figure 5.7: Average occurrence ratio (O_{av}) versus the feature entropy of the communities

consonants forming communities in PhoNet occur in real languages in such groups so as to minimize feature entropy. Figure 5.7 shows the scatter plot of the average occurrence ratio of the communities obtained from PhoNet (y-axis) versus the feature entropy of these communities (x-axis). Each point in this plot corresponds to a single community. The plot clearly indicates that the communities exhibiting lower feature entropy have a higher average occurrence ratio. For communities having feature entropy less than or equal to 3 the average occurrence ratio is never less than 0.7 which means that the consonants forming these communities occur together on an average in 70% or more of the world's languages. As feature entropy increases this ratio gradually decreases until it is almost close to 0 when feature entropy is around 10. Once again, this result fosters the fact that the driving force for the community formation is the principle of feature economy and languages indeed tend to choose consonants in order to maximize the combinatorial possibilities of the distinctive features, which are already available in the inventory.

5.4 The Redundancy across the Consonant Inventories

In this section, we shall treat each individual consonant inventory in UPSID as a community in itself and show that the redundancy (in terms of features) that exists in the representation of these inventories is fixed irrespective of the inventory size.

Note that in an artificially engineered system, one has the liberty to introduce redundancy by duplicating some of the critical components. However, this is not possible in case of phonological systems because, one cannot include arbitrary features in order to describe the system; in contrast, the description is thoroughly guided by the articulatory and perceptual constraints of the speakers. For an artificial system, it is quite easy to capture the notion of redundancy through different quantitative methods [138]. For a natural system (e.g., a phonological system), on the other hand, no such quantitative formulation exists and therefore there is a need to formally describe the redundancy across such systems. There have been a few attempts in this direction to measure the disorderedness (i.e., the entropy) of natural coding systems and, thereby, investigate their structural characteristics. In fact, the presence of redundancy has been attested at every level of a biological system such as in the codons [91], in the genes [165], and in the proteins [59]. For instance, [34] demonstrates how the concepts of information theory can be applied to genome analysis. This work specifically shows that the amount of disorderedness that exists in the genome sequence of *E. coli* is much less than that of a randomly constructed sequence of the same length. In [159] the authors measure the binary entropy of various tissue classes obtained from MRI scans in order to construct a probabilistic human anatomical atlas that can quite accurately describe anatomical variability.

Here we shall attempt to mathematically formulate the concept of redundancy using the definition of feature entropy introduced in the earlier section. This is followed by the experiments performed using this formulation and the results obtained from them.

5.4.1 Formulation of Redundancy

The feature entropy for a language inventory L can be easily calculated using the equation 5.5 with the exception that N now stands for the size of the inventory (instead of the size of a community). Note that for an optimal encoding (i.e., one that has no redundancy) of the consonants in L, the number of bits required should be $\lceil \log_2(N) \rceil$. For instance, in the example shown in Figure 5.5 the number of bits required to optimally encode the $N = 8$ consonants should be $\lceil \log_2(8) \rceil = 3 = |F|$. For such an encoding, every feature f_i takes the value 1 for one half of the consonants and 0 for the other half assuming that all the code words are exhaustively used. Therefore, we have $DC = 1$ for each feature, whereby, the value of F_E is $\lceil \log_2(N) \rceil$.

However, if the encoding is redundant, which is usually the case for many naturally occurring systems, then $F_E > \lceil \log_2(N) \rceil$. Normalizing F_E by $\log_2 N$ gives us the fraction of bits that are required in excess to encode all the consonants in L. We call this fraction *redundancy ratio* (RR) and formally define it as follows,

$$RR = \frac{F_E}{\log_2 N} \tag{5.6}$$

5.4.2 Experiments and Results

We measure the values of RR for different sets of consonant inventories chosen from the collection of 317 inventories available from UPSID. Set_{100}, Set_{200}, and Set_{250} respectively denote sets of 100, 200, and 250 consonant inventories chosen uniformly at random from the 317 real inventories. We have randomly constructed each set 10 different times and averaged the results over them. Set_{317} denotes the full collection of 317 inventories present in UPSID. The results are summarized in Figure 5.8 and Table 5.3.

One of the most important observations is that the scatter-plots for all the four different sets shown in Figure 5.8 indicate a Zipfian distribution [169] of the form: $y = Ax^{-\lambda}$, where x is the inventory size and y is the corresponding value of RR.

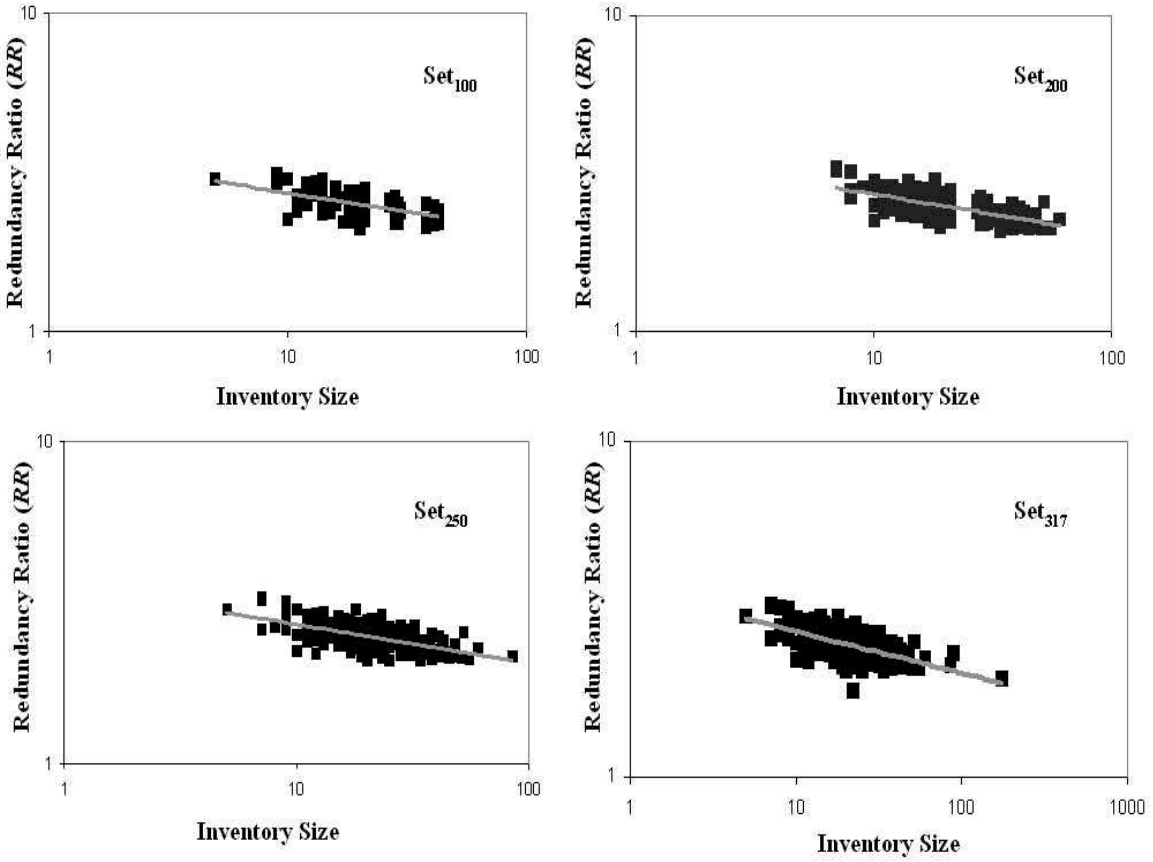


Figure 5.8: Inventory size versus RR in doubly-logarithmic scale for different sets of real inventories. The bold lines indicate the fits for the distributions

The approximate values of A and λ for all the plots are 3.60 and 0.12 respectively. In fact, the Pearson's correlation (r) is quite high between the actual values of RR and the values of y produced by replacing x with the size of the inventories in the equation $y = Ax^{-\lambda}$ for all the four sets ($r \approx 0.6$). Furthermore, it is interesting to note that the value of the Zipfian exponent λ is always close to zero indicating that RR is almost a constant irrespective of the inventory size. The low mean square error[5] ($MSE \approx 0.033$) around the line predicted by the power-law ($y = Ax^{-\lambda}$) further confirms this fact because, a high square error would actually mean that RR values are strongly dependent on the inventory size. The value of the standard deviation ($\sigma = 0.22$) also indicates that the variance of the distribution around the mean

[5] MSE between two distributions measures the average of the square of the "error". The error is the amount difference between a pair of ordinates (say y and y'), where the abscissas are equal. In other words, if there are N such ordinate pairs then MSE can be expressed as $\frac{\sum(y-y')^2}{N}$.

Table 5.3: Different statistical properties of the distributions of RR values (with respect to the size of the consonant inventories) presented in Figure 5.8. (A, λ): The fit parameters for a Zipfian distribution, where $y = Ax^{-\lambda}$ is the equation of the fit (denoted by the bold lines in Figure 5.8); μ: Mean value of RR in a distribution; σ: The standard deviation of a distribution; MSE: The mean square error around the line predicted by the power-law ($y = Ax^{-\lambda}$); r: The Pearson's correlation between the actual values of RR and the values of y produced by replacing x with the size of the inventories in the power-law equation

	Fit parameters (A, λ)	μ	σ	MSE	r
Set_{100}	(3.60, 0.12)	2.50	0.22	0.033	0.61
Set_{200}	(3.70, 0.12)	2.50	0.22	0.031	0.62
Set_{250}	(3.60, 0.12)	2.49	0.23	0.032	0.60
Set_{317}	(3.60, 0.12)	2.49	0.23	0.033	0.62

is quite low. The consistency of the results for any arbitrary subset of inventories possibly points to the robustness of UPSID. In conclusion, the feature-based encoding of the consonants generates a nearly constant redundancy across the inventories of the world's languages.

In fact, even if we slightly modify the phoneme representation assuming the place of articulation and the phonation features (see Table 3.1) to be multi-valued, the above inferences remain unaffected. Figure 5.9 shows the distribution of the RR values for this modified representation. The scatter-plot again indicates a Zipfian distribution where $A = 3.30$ and $\lambda = 0.09$. The value of λ is close to zero, thereby, attesting our earlier inference that the redundancy across the consonant inventories is constant with respect to the inventory size.

Note that if we compute the values of RR for the randomly generated inventories obtained from Algorithm 5.2 then the Zipfian exponent λ in this case is found to be 0.22. Therefore, there is an 83% (approx.) increase in the value of λ with respect to the real inventories. Thus, we can conclude that this universal property of fixed redundancy found across the consonant inventories of the world's languages is not a consequence by chance.

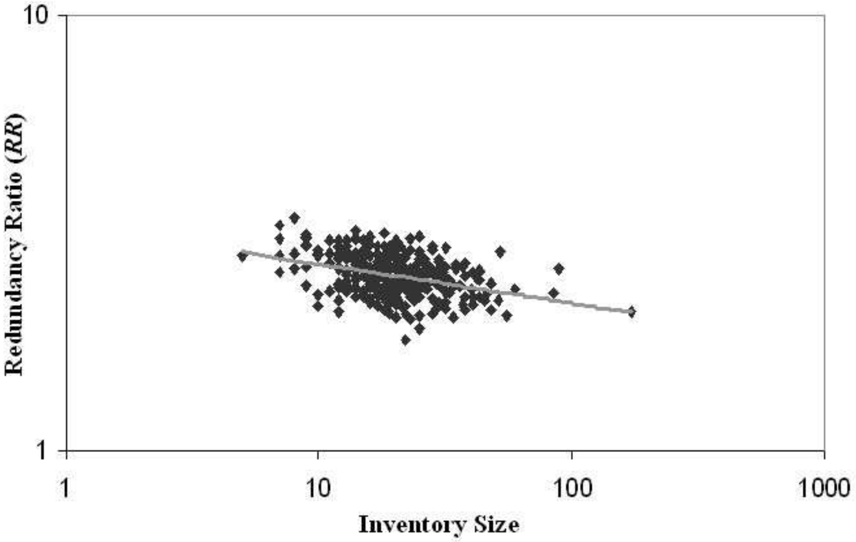


Figure 5.9: Inventory size versus *RR* in doubly-logarithmic scale for modified phoneme representation. The bold line indicates the fit for the distribution

5.5 Summary

In this chapter, we have explored the co-occurrence principles of the consonants across the inventories of the world's languages. In particular, we have shown the following.

(i) The patterns of co-occurrence of the consonants, reflected through communities in PhoNet, are observed in 80% or more of the world's languages.

(ii) The communities obtained are far more economic in terms of the constituent features than expected by random chance.

(iii) Those communities that maximize feature economy tend to occur more frequently (70% or higher number of times) across languages.

(iv) Redundancy ratio is *almost* an invariant property of the consonant inventories with respect to the inventory size.

Linguistic Implications

There are a number of linguistic implications of the work presented in this chapter. First of all, the automatic procedure for community detection provides an algorithmic definition of *natural classes* [35] of phonemes (Table 2.2 of Chapter 2 shows a natural class of plosives). This is significant because, there is no single satisfactory definition of such natural classes in literature [56]. The communities that we obtained from PhoNet are such natural classes and can be derived simply by regulating the threshold of the MRad algorithm.

Secondly, the discriminative capacity that we introduced for quantifying feature economy may be thought to imply the importance of learning to perceive the distinction between the presence and the absence of a feature f_i in an inventory L when it is transmitted from the speaker (e.g., the parent) to the learner (e.g., the child) during language acquisition. If L has equal number of example consonants with $f_i = 0$ and $f_i = 1$ (i.e., $DC = 1$) then it becomes more important to learn to perceive this distinction in order to successfully discriminate all the consonants present in the speaker's inventory than in the case where the distribution of examples is highly skewed ($DC = 0$). For instance, while an English speaker can discriminate between the interdental sound in the word “that” (i.e., the sound made by the part “th”) and the alveolar sound in the word “doctor” (i.e., the sound made by the part “d”) a Spanish speaker cannot. A Spanish speaker interprets both these sounds as dental. This is possibly because the alveolar feature is a non-discriminative one in Spanish (see [75] for an elaborate discussion on this topic).

The next important question is related to the origin and the general implication of feature economy observed across the consonant inventories. One possible way to answer this question would be to argue that it is the ease of learnability associated with an inventory that manifests as feature economy. Since the consonant inventories are usually large in size therefore, there are a lot of consonants to be learnt by a speaker in order to pick up the inventory. However, if the number of features that the speaker has to learn is small then even with the large size inventory the learnability effort is not very high. This is a probable reason for the consonant inventories to be economic. In this context, Lindblom [95] points out that learning a new form that overlaps with old patterns should involve storing less information in memory than acquiring one with nothing in common with old items, since part of the new motor score associated with the phoneme is already in storage.

The "fixed redundancy" indicates that although languages with larger inventory size tend to use more features, yet there is a bound on the rate at which the number of features increases. The power-law exponent, in this context, may be thought of as an indicator of how fast the redundancy ratio decays with respect to the inventory size. Since the value of this exponent is very low for consonant inventories, we argued that the ratio is independent of the inventory size.

In the next chapter, we shall show how the computational framework developed in Chapters 3, 4 and 5 can be successfully applied to investigate the structure and dynamics of the vowel inventories. Wherever necessary, we shall compare the results and thereby, point out the similarities as well as the differences with the consonant inventories. We shall also attempt to explain the reasons for these similarities and differences in the light of various linguistic theories.

Chapter 6

Network Methods applied to Vowel Inventories

In the last three chapters, we have shown how complex networks can be successfully used in modeling the structure and dynamics of the consonant inventories of the world's languages. In this chapter, we shall show that the mathematical framework developed is quite generic and can be easily employed to study the properties of the vowel inventories. We shall report some of the important results for the vowel inventories and compare the observations wherever required with those for the consonant inventories.

In order to represent the inventory structure, we define a bipartite network as in the case of the consonant inventories and call it the **V**owel-**La**nguage **Net**work or **VlaNet**. Once again, the nodes in the two partitions of VlaNet are labeled by the languages and the vowels while an edge signifies that a particular vowel occurs in the vowel inventory of a particular language. Subsequently, we analyze and synthesize the distribution of the frequency of occurrence of the vowels across the language inventories. As a following step, we construct the one-mode projection of VlaNet onto the vowel nodes and call it the **Vo**wel-Vowel **Net**work or **VoNet**. Clearly, VoNet is a network of vowels where a pair of nodes are connected as many times as the corresponding vowels are found to occur together across the inventories of different languages. We next investigate the topological properties of VoNet and attempt to explain the distribution of the co-occurrence of the vowels across the inventories.

Apart from the study of the topological properties, we also perform a detailed community structure analysis of VoNet. Interestingly, this investigation leads us to certain new observations about the organization of the vowel inventories apart from validating different results reported by the earlier researchers.

A general finding is that the topological properties of VlaNet as well as VoNet where the nodes are assumed to be unlabeled (i.e., they are not marked by the articulatory/acoustic features) are quite similar to that in the case of the consonants. However, community structure analysis of VoNet where the nodes are assumed to be labeled reveal certain interesting differences between these two basic types of human speech sound inventories. Differences also manifest in the study of the redundancy ratio of the vowel inventories.

The rest of the chapter is organized as follows. In section 6.1, we investigate the topological properties of the bipartite network VlaNet. The next section outlines the topological properties of the one-mode projection VoNet. In section 6.3, we perform community analysis of VoNet and make interesting inferences about the patterns of co-occurrence across the vowel inventories primarily through the application of the feature entropy metric defined in the previous chapter. We estimate the redundancy ratio of the vowel inventories and compare them with those of the consonant inventories in the next section. Finally, in section 6.5, we summarize the contributions of this chapter as well as outline certain linguistic implications of the results.

6.1 Topological Properties of VlaNet

VlaNet is a bipartite graph $G = \langle\ V_L,\ V_V,\ E_{vl}\ \rangle$ with two sets of nodes V_L and V_V representing the languages and the vowels respectively. An edge between a pair of nodes $v_l \in V_L$ and $v_v \in V_V$ implies that the vowel v occurs in the inventory of the language l. For the purpose of our analysis, we have constructed VlaNet using UPSID

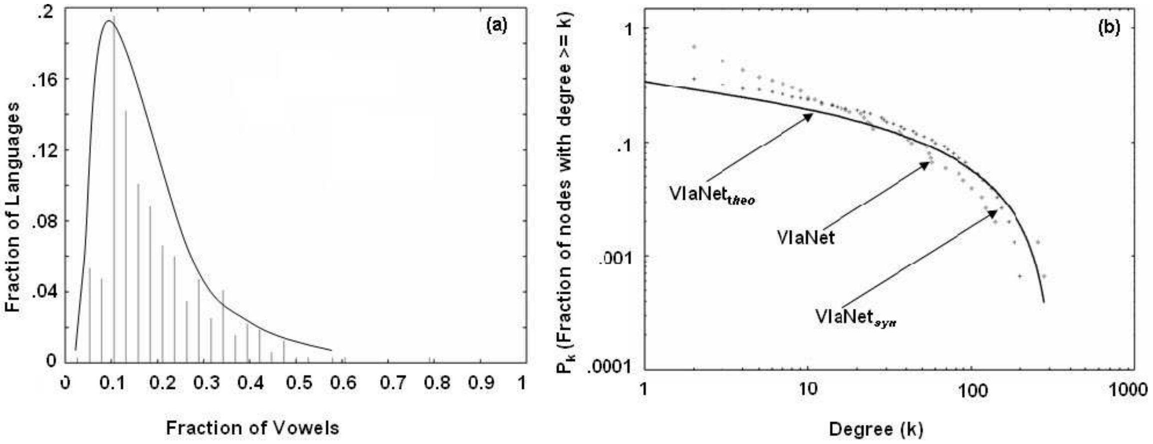


Figure 6.1: Degree distribution of (a) the language and (b) the vowel nodes in VlaNet, VlaNet$_{syn}$ and VlaNet$_{theo}$. The plots in (b) are in doubly-logarithmic scale

similarly as in the case of consonants. Consequently, the total number of nodes in V_L is 317 and that in V_V is 151. The total number of edges in the network so constructed is 2349.

Degree Distribution of the Language Nodes

Figure 6.1(a) shows the degree distribution of the language nodes where the x-axis denotes the degree of each language node expressed as a fraction of the maximum degree and the y-axis denotes the fraction of nodes having a given degree. The plot immediately shows that the size of the vowel inventories (i.e., the number of vowels in different languages) again follow roughly a β-distribution like the case of consonant inventories. However, the distribution in this case peaks at 4 (in contrast to 21 for consonants) indicating that a majority of the world's languages have 4 vowels in their inventories. Consequently, it may be stated that in most cases consonants present in a language by far outnumber the vowels.

Degree Distribution of the Vowel Nodes and its Synthesis

Figure 6.1(b) illustrates the degree distribution plot for the vowel nodes in V_V in doubly-logarithmic scale. We employ the model introduced in section 3.3 of Chapter 3 to synthesize the distribution of the occurrence of vowels across languages. We simulate the model to obtain VlaNet$_{syn}$ (i.e., the synthesized version of VlaNet) for 100 different runs and average the results over all of them. Good fits (in terms of mean error) emerge for $\gamma \in [14.1,18.5]$ with the best being at 14.4 (see Figure 6.1(b)). The mean error in this case is approximately 0.05. Figure 6.1(b) further shows the degree distribution of VlaNet$_{theo}$ (i.e., the theoretical version of VlaNet) obtained using the equations 3.1 and 3.25 for $\gamma = 14.4$. The low mean error ($\approx$ 0.06) between the degree distribution of the vowel nodes in VlaNet and VlaNet$_{theo}$ indicates that the distribution of occurrence of vowels across languages can be well-approximated by a β-distribution.

The high value of γ clearly indicates that preferential attachment, similarly as in the case of consonants, plays a very crucial role in shaping the emergent degree distribution of the vowel nodes in VlaNet.

6.2 Topological Properties of VoNet

VoNet is the one-mode projection ($G = \langle\ V_V,\ E_{vo}\ \rangle$) of VlaNet onto the vowel nodes in V_V. Consequently, the total number of nodes in VoNet is $|V_V| = 151$. There is an edge $e \in E_{vo}$ if the two nodes in V_V (read vowels) that are connected by e co-occur in at least one language inventory. The number of inventories that they co-occur in defines the weight of e. The total number of edges in VoNet (ignoring weights) is 2730.

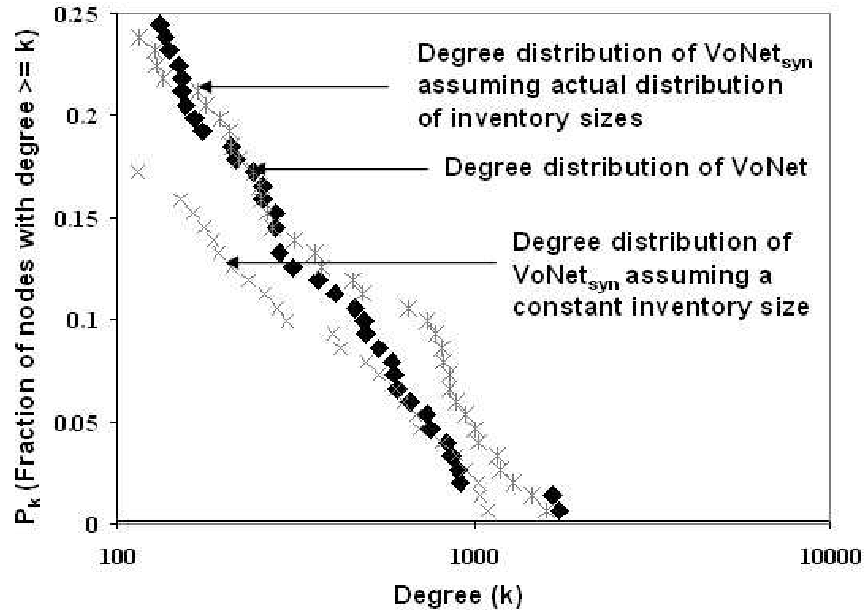


Figure 6.2: Comparison between the degree distribution of VoNet and VoNet_{syn}. The x-axis is in the logarithmic scale

Degree Distribution of VoNet and its Synthesis

Figure 6.2 shows the degree distribution of the nodes in VoNet. The distribution of the size of the vowel inventories is again found to affect the emergent degree distribution of VoNet_{syn} as observed for the case of consonants. Figure 6.2 compares the degree distribution of VoNet with VoNet_{syn} obtained by assuming (a) the inventory sizes to be fixed to a constant equal to the average size and (b) the actual distribution of the inventory sizes. The result indicates that the latter assumption (i.e., (b)) produces better match with the empirical data than the former one (i.e., (a)).

Clustering Coefficient of VoNet and its Synthesis

We compute the clustering coefficient of VoNet using the equation 4.1. The value of c_{av} is 0.86 which is significantly higher than a random graph with the same number of nodes and edges (0.09). This immediately points to the fact that a large number of triangles are also prevalent in VoNet similarly as in the case of the consonants. In

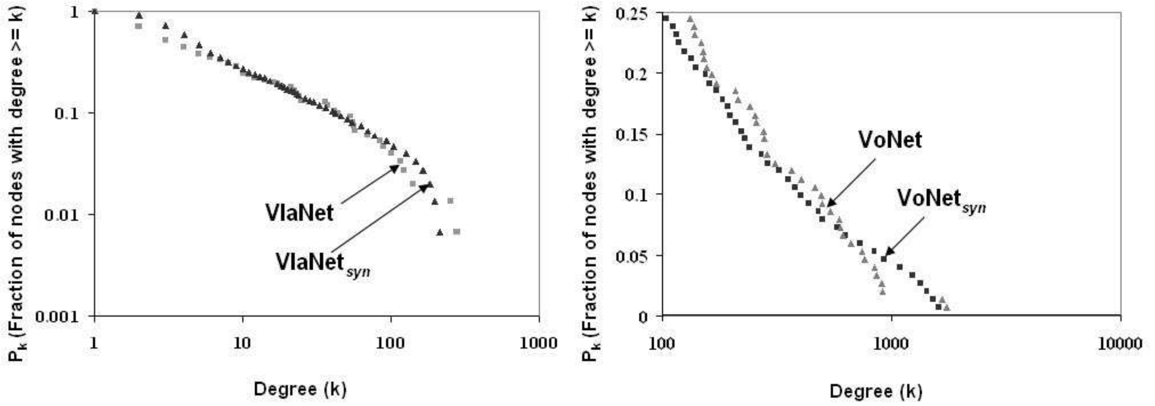


Figure 6.3: Degree distribution of VlaNet$_{syn}$ and VoNet$_{syn}$ obtained from the triad model along with their corresponding real counterparts. For VlaNet$_{syn}$ the degree distribution is in doubly-logarithmic scale and for VoNet$_{syn}$ the x-axis is in logarithmic scale

other words, on an average, there is a high probability that two vowel nodes having a common neighbor in VoNet themselves also co-occur frequently.

The triad model (see section 4.5 of Chapter 4), on the other hand, can be once again applied to explain the high clustering coefficient of VoNet as in the case of consonants. For values of p_t in the range [0.8, 0.9], without affecting the degree distribution much (see Figure 6.3), it is possible to achieve a clustering coefficient of 0.83 for VoNet$_{syn}$ which is within 3.5% of VoNet.

Therefore, it turns out that both the consonant as well as the vowel networks reflect largely similar topological characteristics with preferential attachment playing the most crucial role in their emergent structure. However, as we shall see in the next section, community structure analysis of VoNet reveals certain interesting differences between the two basic types of human speech sound inventories (i.e., consonants and vowels). In fact, differences are also apparent from the study of the redundancy ratio of the vowel inventories.

6.3 Community Analysis of VoNet

One of the central observations in the study of the vowel inventories has been that they are organized primarily based on the principle of maximal perceptual contrast [158]. In fact, a number of numerical studies based on this principle have been reported in literature [92, 94, 134]. Of late, there have been some attempts to explain the vowel inventories through multi-agent simulations [44] and genetic algorithms [80]; all of these experiments also use the principle of perceptual contrast for optimization purposes.

An exception to the above trend is a school of linguists [20, 39] who argue that perceptual contrast-based theories fail to account for certain fundamental aspects such as the patterns of co-occurrence of vowels based on similarity of features observed across the vowel inventories. Instead, they posit that the observed patterns, especially found in larger size inventories [20], can be explained only through the principle of feature economy.

We hypothesize the following organization of the vowel inventories based on the two orthogonal views stated above and systematically corroborate the hypothesis through the community analysis of VoNet.

6.3.1 The Hypothetical Organization of the Vowel Inventories

According to our hypothesis, the two orthogonal views can be possibly linked together through the example illustrated by Figure 6.4. As shown in the figure, the bottom plane P constitutes of a set of three very frequently occurring vowels /i/, /a/ and /u/, which usually make up the smaller inventories and do not have any single feature in common. Thus, smaller inventories are quite likely to have vowels that exhibit a large extent of contrast in their constituent features. However, in bigger inventories, members from the higher planes (P' and P'') are also present and they in turn exhibit feature economy. For instance, in the plane P' consisting of the set of vowels /ĩ/, /ã/,

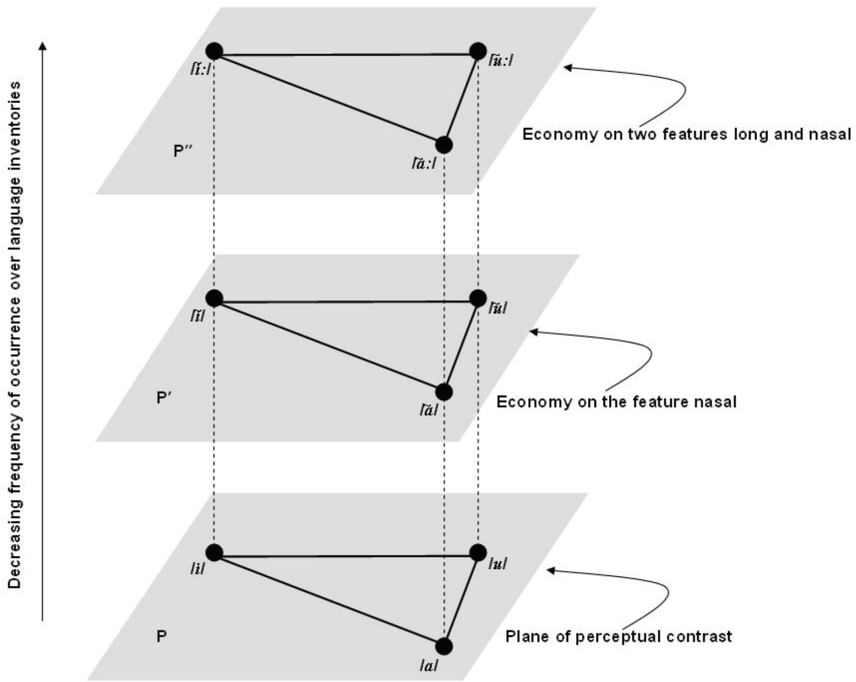

Decreasing frequency of occurrence over language inventories
/ĩ:/
/ũ:/
P″
/ã:/
Economy on two features long and nasal
/ĩ/
/ũ/
P′
/ã/
Economy on the feature nasal
/i/
/u/
P
/a/
Plane of perceptual contrast

Figure 6.4: The organization of the vowels (in decreasing frequency of occurrence) across the inventories indicated through different hypothetical planes

/ũ/, we find a nasal modification applied equally on all the three members of the set. This is actually indicative of an economic behavior of the larger inventories during the introduction of a new feature possibly for the reduction in the learnability effort of the speakers. The third plane P'' reinforces this idea by showing that the larger the size of the inventories the greater is the urge for this economy in the choice of new features. The figure also illustrates another interesting relationship that exists between the vowels across the planes (indicated by the broken lines). All these relations are representative of a common linguistic concept of markedness [39] in which the presence of a less frequently occurring vowel (say /ĩ/) implies the presence of another frequently occurring vowel (say /i/) in a language inventory (and not vice versa). In this co-occurring pair (/i/ and /ĩ/), the frequently occurring vowel (i.e., /i/) is usually referred to as the unmarked member, while the less frequent one (i.e., /ĩ/) is called the marked member. Note that these cross-planar relations are also indicative of feature economy because, all the features present in the frequent vowel are also shared by the less frequent one. In summary, while the basis of organization of the vowel inventories is perceptual contrast as indicated by the plane P in Figure 6.4, economic modifications of the perceptually distinct vowels takes place with the increase in the inventory size (as indicated by the planes P' and P'' in Figure 6.4).

In the following, we attempt to automatically capture the patterns of co-occurrence that are prevalent *in* and *across* the planes illustrated in Figure 6.4 through the community structure analysis of VoNet. We further employ the metric of feature entropy to estimate the extent of feature economy for a given set of vowels in a community. We observe that while the vowel communities within a plane exhibit lower feature economy, the communities across the planes display much higher feature economy. These findings, in turn, imply that the bigger vowel inventories are formed on the basis of feature economy, while the smaller ones are governed by the principle of maximal perceptual contrast. We also compare the extent of feature economy observed in real inventories to that in randomly generated inventories, which further corroborates the above findings.

6.3.2 Identification of the Communities and the Metric for Evaluation

We apply the MRad algorithm developed in Chapter 5 to extract the vowel communities from VoNet. Note that a community of vowels (as in the case of consonants) actually refers to a set of vowels, which occur together in the language inventories very frequently. For instance, if /i/, /a/ and /u/ form a vowel community and if /i/ and /a/ are present in any inventory then there is a very high chance that the third member /u/ is also present in the inventory.

After extracting the vowel communities from VoNet, we next investigate the driving force that leads to the emergence of these communities. According to our hypothesis, the principle of perceptual contrast as well as feature economy together shape the structure of the vowel inventories and in turn act as the driving forces for community formation. In order to establish this fact, one needs to define a quantitative measure to capture these two forces. We have already shown in the previous chapter

Table 6.1: F_E for the two different communities COM_1 ($F_E = 4$) and COM_2 ($F_E = 1$). The letters **h**, **f**, **b**, **r**, **u**, and **n** stand for the features high, front, back, rounded, unrounded, and nasalized respectively

COM_1	h	f	b	r	u
/i/	1	1	0	0	1
/u/	1	0	1	1	0
p_f/N	1	0.5	0.5	0.5	0.5
q_f/N	0	0.5	0.5	0.5	0.5

COM_2	h	f	u	n
/i/	1	1	1	0
/ĩ/	1	1	1	1
p_f/N	1	1	1	0.5
q_f/N	0	0	0	0.5

that the feature entropy metric faithfully captures the idea of feature economy. It is easy to show that this same metric can also be employed to capture the concept of perceptual contrast. Let a community COM consist of a set of perceptually distinct vowels, then larger number of bits should be required to represent the information in COM since in this case the set of features that constitute the vowels are more in number. Therefore, the higher the perceptual contrast, the higher is the feature entropy. The idea is illustrated through the example in Table 6.1. In the table, F_E exhibited by the community COM_1 is higher than that of the community COM_2, since the set of vowels in COM_1 are perceptually more distinct than those in COM_2. In general, if the feature entropy of a community is higher than what is expected by chance then it may be argued that the constituent vowels of the community are highly perceptually distinct from each other.

6.3.3 Experiments and Results

In this section, we describe the experiments performed and the results obtained from the community analysis of VoNet. In order to find the co-occurrence patterns in and across the planes of Figure 6.4 we define three versions of VoNet namely VoNet$_{hub}$, VoNet$_{rest}$ and VoNet$_{rest'}$. The construction procedure for each of these versions are presented below.

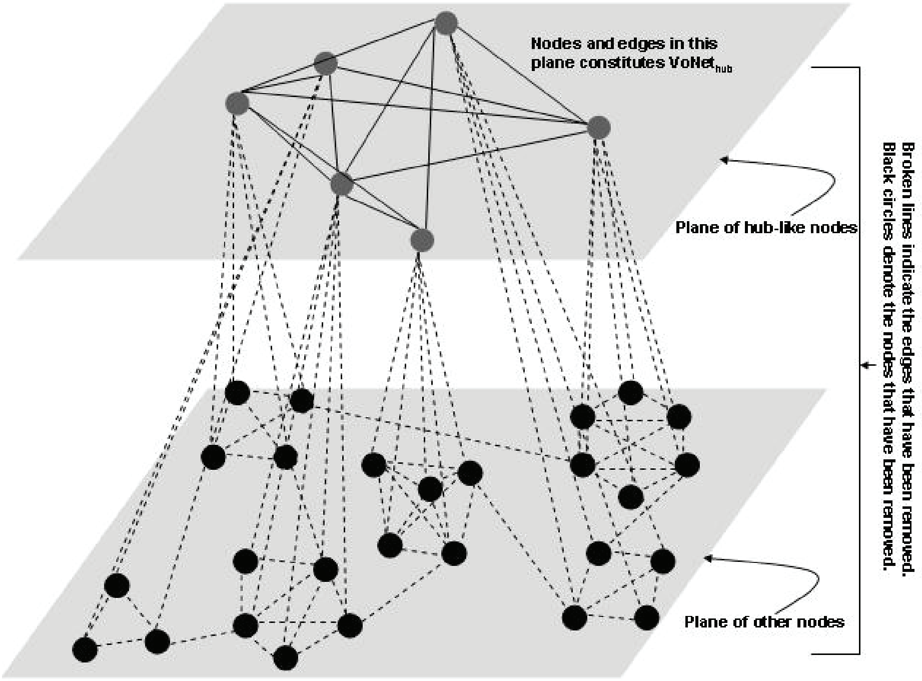

Nodes and edges in this plane constitutes VoNet$_{hub}$
Plane of hub-like nodes
Broken lines indicate the edges that have been removed.
Black circles denote the nodes that have been removed.
Plane of other nodes

Figure 6.5: The construction procedure of VoNet$_{hub}$ from VoNet

Construction of VoNet$_{hub}$: VoNet$_{hub}$ consists of the *hubs*, i.e. the nodes corresponding to those vowels having a very high occurrence frequency in UPSID[1]. We define a node as hub if the frequency of occurrence of the corresponding vowel in UPSID is greater than 120. Thus VoNet$_{hub}$ is a subgraph of VoNet composed of the hubs and the edges inter-connecting them. The rest of the nodes (having frequency less than 120) and the associated edges are removed from the network. We make a choice of this value (i.e., 120) for classifying the hubs from the non-hubs through an inspection of the distribution of the occurrence frequencies of the vowels in UPSID. Figure 6.5 illustrates how VoNet$_{hub}$ is constructed from VoNet. The number of nodes in VoNet$_{hub}$ is 9 corresponding to the vowels: /i/, /a/, /u/, /ɔ/, /ɛ/, /o/, /e/, /ŏ/ and /ĕ/.

[1] Hubs are nodes that have a very high degree. In VoNet, nodes having very high frequency of occurrence also have a very high degree. In particular, all the nodes included in VoNet$_{hub}$ has a degree $>$ 950 in VoNet, while the average degree of a node in VoNet is only 180 (approx.).

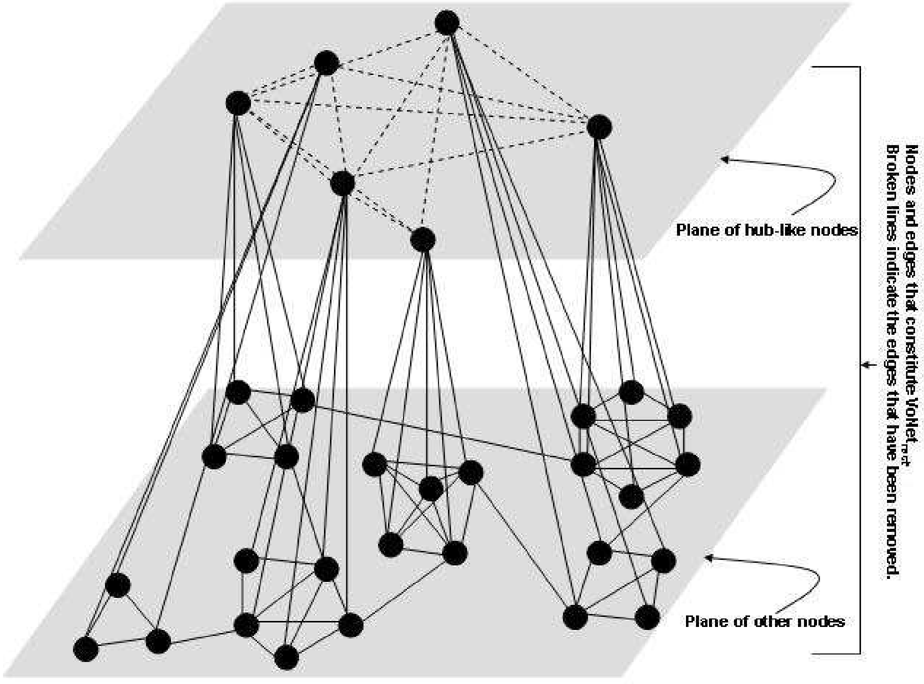

Plane of hub-like nodes
Nodes and edges that constitute VoNet$_{rest}$
Broken lines indicate the edges that have been removed.
Plane of other nodes

Figure 6.6: The construction procedure of VoNet$_{rest}$ from VoNet

Construction of VoNet$_{rest}$: VoNet$_{rest}$ consists of all the nodes as that of VoNet[2]. It also has all the edges of VoNet except for those edges that inter-connect the hubs. Figure 6.6 shows how VoNet$_{rest}$ can be constructed from VoNet.

Construction of VoNet$_{rest'}$: VoNet$_{rest'}$ again consists of all the nodes as that of VoNet. It consists of only the edges that connect a hub with a non-hub if the non-hub co-occurs more than ninety five percent of times with the hub. The basic idea behind such a construction is to capture the co-occurrence patterns based on markedness [39] (discussed earlier) that actually defines the cross-planar relationships in Figure 6.4. Figure 6.7 shows how VoNet$_{rest'}$ can be constructed from VoNet[3]. Note that since VoNet$_{rest'}$ has edges running between the frequent (i.e., unmarked) and the infrequent (i.e., marked) vowels, a com-

[2] We have neglected nodes corresponding to those vowels that occur in less than 3 languages in UPSID because, the communities they form do not reflect significant results. The number 3 has been decided arbitrarily based on the manual inspection of the data.

[3] The network does not get disconnected due to this construction since there is always a small fraction of edges that run between a hub and low frequency non-hubs.

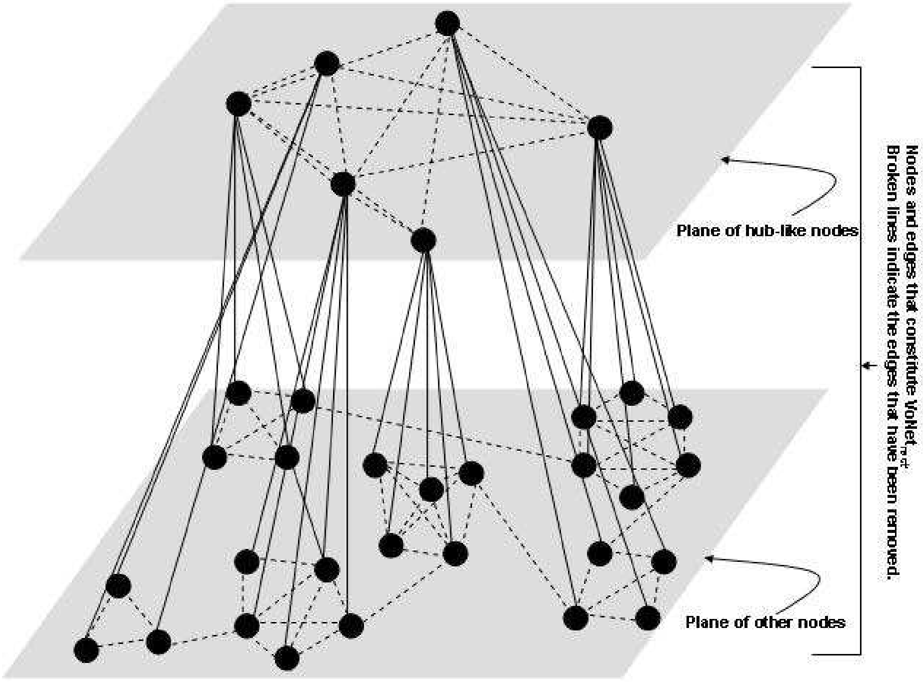

Plane of hub-like nodes
Plane of other nodes
Nodes and edges that constitute VoNet$_{rest'}$
Broken lines indicate the edges that have been removed.

Figure 6.7: The construction procedure of VoNet$_{rest'}$ from VoNet

Table 6.2: Some of the vowel communities obtained from VoNet$_{hub}$. The contrastive features separated by slashes (/) are shown within parentheses. Comma-separated entries (2nd column) represent the features that are in use from the three respective classes namely the height, the backness, and the roundedness

Community	Features in Contrast
/i/, /a/, /u/	(low/high), (front/central/back), (unrounded/rounded)
/e/, /o/	(higher-mid/mid), (front/back), (unrounded/rounded)

munity structure analysis of this network is expected to reveal the relationship between the marked and the unmarked pairs of vowels that co-occur frequently.

We separately apply the MRad algorithm on each of VoNet$_{hub}$, VoNet$_{rest}$ and VoNet$_{rest'}$ in order to obtain the respective vowel communities. Representative communities from Vonet$_{hub}$, VoNet$_{rest}$ and VoNet$_{rest'}$ are noted in Tables 6.2, 6.3 and 6.4 respectively.

Table 6.3: Some of the vowel communities obtained from VoNet$_{rest}$

Community	Features in Common
/ĩ/, /ã/, /ũ/	nasalized
/ĩː/, /ãː/, /ũː/	long, nasalized
/iː/, /uː/, /aː/, /oː/, /eː/	long

Table 6.4: Some of the vowel communities obtained from VoNet$_{rest'}$. Comma-separated entries (2nd column) represent the features that are in use from the three respective classes namely the height, the backness, and the roundedness

Community	Features in Common
/i/, /ĩ/	high, front, unrounded
/a/, /ã/	low, central, unrounded
/u/, /ũ/	high, back, rounded

Tables 6.2, 6.3 and 6.4 indicate that the communities in VoNet$_{hub}$ are formed based on the principle of perceptual contrast, whereas the formation of the communities in VoNet$_{rest}$ as well as VoNet$_{rest'}$ is largely governed by feature economy. We dedicate the rest of this section mainly to verify the above argument. For this reason, we present a detailed study of the co-occurrence principles of the communities obtained from VoNet$_{hub}$, VoNet$_{rest}$, and VoNet$_{rest'}$ primarily through the application of the feature entropy metric. In each case we compare the results with the random version of VoNet namely, VoNet$_{rand}$ constructed using the Algorithm 5.2 as in the case of consonants. Note that if the feature entropy for the communities obtained from the real language inventories is significantly higher than that of the random inventories then the prevalence of maximal perceptual contrast is attested; however, if the same is significantly lower then the prevalence of feature economy is attested.

Co-occurrence Principles of the Communities of VoNet$_{hub}$

The random counterpart of VoNet$_{hub}$ (henceforth VoNet$_{Rhub}$) is constructed from VoNet$_{rand}$ following the steps that were used to construct VoNet$_{hub}$ from VoNet. Figure 6.8 illustrates, for all the communities obtained from the clustering of VoNet$_{hub}$

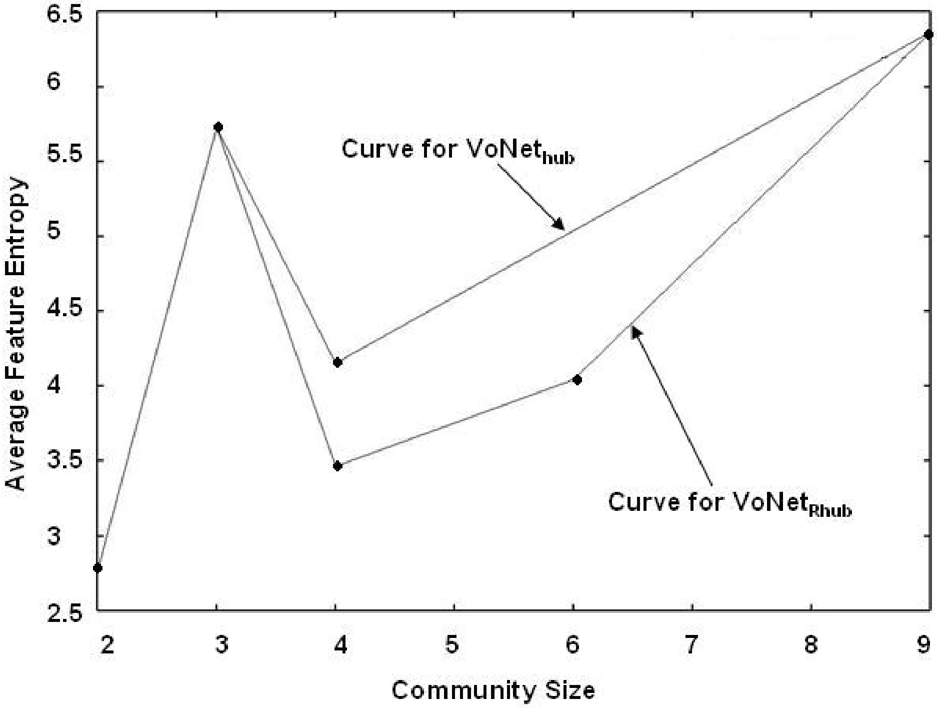


Figure 6.8: Average feature entropy of the communities of a particular size versus the community size for the case of VoNet$_{hub}$ as well as VoNet$_{Rhub}$

and VoNet$_{Rhub}$, the average feature entropy exhibited by the communities of a particular size (y-axis) versus the community size (x-axis).

A closer inspection of Figure 6.8 immediately reveals that the feature entropy exhibited by the communities of VoNet$_{hub}$ is higher as compared to that of VoNet$_{Rhub}$. The two curves intersect because, eventually for a low value of η, all the nodes in VoNet$_{hub}$ and VoNet$_{Rhub}$ form a single connected component, i.e., a single cluster. Since the set of hubs, which are defined solely in terms of occurrence frequency, is identical for VoNet and VoNet$_{rand}$, the feature entropy of the cluster formed of all the hubs together is also identical in both the cases.

Nevertheless, the number of data points in Figure 6.8 are fairly less and hence, it might not be alone sufficient to establish the fact that the communities in VoNet$_{hub}$ are formed based on the principle of perceptual contrast. Another possible way to investigate the problem would be to look into the co-occurrence principles of the smaller vowel inventories (of size $\leq$ 4) since they are mostly formed of the hubs. Table 6.5,

Table 6.5: Percentage frequency of occurrence of the members of the community /i/, /a/, and /u/, as compared to the percentage occurrence of other vowels, in smaller inventories. The last column indicates the average number of times that a vowel other than /i/, /a/, and /u/ occurs in the inventories of size 3 and 4

Inv. Size	% Occ. /i/	% Occ. /a/	% Occ. /u/	Avg. % Occ. other vowels
3	65	91	52	13
4	76	96	44	12

for instance, shows the percentage occurrences of the members of the community formed by /i/, /a/, and /u/, as compared to the average occurrence percentage of other vowels, in the inventories of sizes 3 and 4. The figures in the table points to the fact that the smaller inventories can be assumed to be good representatives of the communities obtained from VoNet$_{hub}$. We therefore compare the average feature entropy of these inventories as a whole with their random counterparts. Figure 6.9 illustrates the result of this comparison. The figure clearly shows that the average feature entropy of the vowel inventories of UPSID is substantially higher for inventory sizes 3 and 4 than that of those constructed randomly.

The results presented in Figures 6.8 and 6.9 together confirm that the communities in VoNet$_{hub}$ are formed based on the principle of maximal perceptual contrast.

Co-occurrence Principles of the Communities of VoNet$_{rest}$

Here we investigate whether the communities obtained from VoNet$_{rest}$ have a lower feature entropy than in case of the randomly generated vowel inventories. We construct the random version of VoNet$_{rest}$ (henceforth VoNet$_{Rrest}$) from VoNet$_{rand}$ and apply the MRad algorithm on it so as to obtain the communities. Figure 6.10 illustrates, for all the communities obtained from the clustering of VoNet$_{rest}$ and VoNet$_{Rrest}$, the average feature entropy exhibited by the communities of a particular size (y-axis) versus the community size (x-axis). The figure makes it quite clear that the average feature entropy exhibited by the communities of VoNet$_{rest}$ are substantially lower than that of VoNet$_{Rrest}$ (especially for a community size ≤ 7). As the community size increases, the difference in the average feature entropy of the commu-

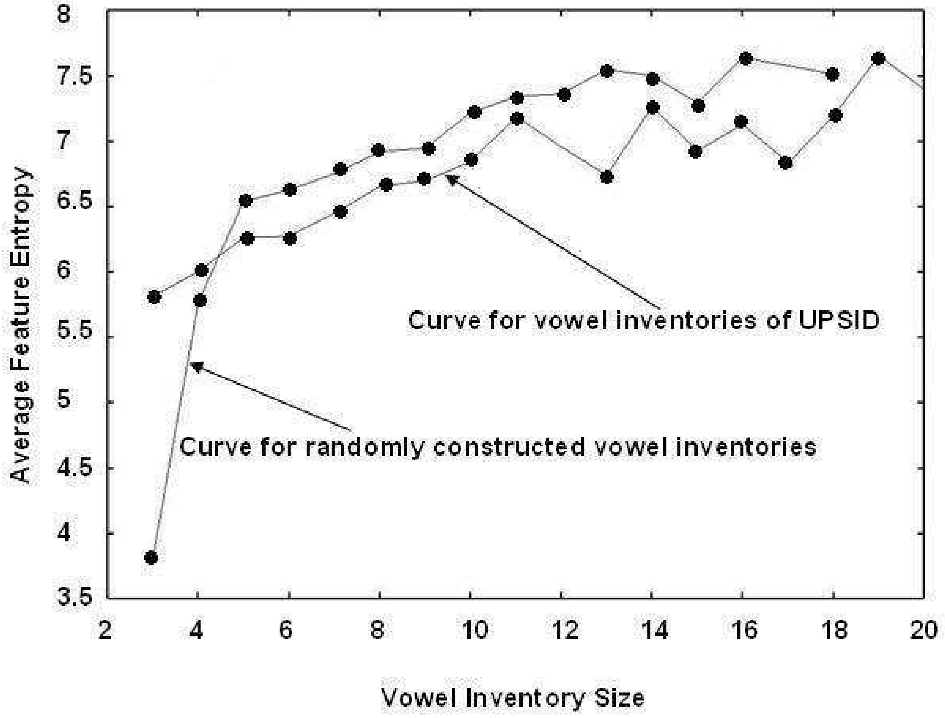


Figure 6.9: Average feature entropy of the real as well as the randomly generated vowel inventories of a particular size versus the inventory size

nities of VoNet$_{rest}$ and VoNet$_{Rrest}$ gradually diminishes. This is mainly because of the formation of a giant cluster, which is similar for both VoNet$_{rest}$ as well as VoNet$_{Rrest}$.

The above result indicates that the driving force behind the formation of the communities of VoNet$_{rest}$ is the principle of feature economy. It is important to mention here that the larger vowel inventories, which are usually composed of the communities of VoNet$_{rest}$, also exhibit feature economy to a large extent. This is reflected through Figure 6.9 where all the real inventories of size ≥ 5 have a substantially lower average feature entropy than that of the randomly generated ones.

Co-occurrence Principles of the Communities of VoNet$_{rest'}$

In this subsection, we compare the feature entropy of the communities obtained from VoNet$_{rest'}$ with that of its random counterpart VoNet$_{Rrest'}$ (constructed from VoNet$_{rand}$). Figure 6.11 shows the average feature entropy exhibited by the communi-

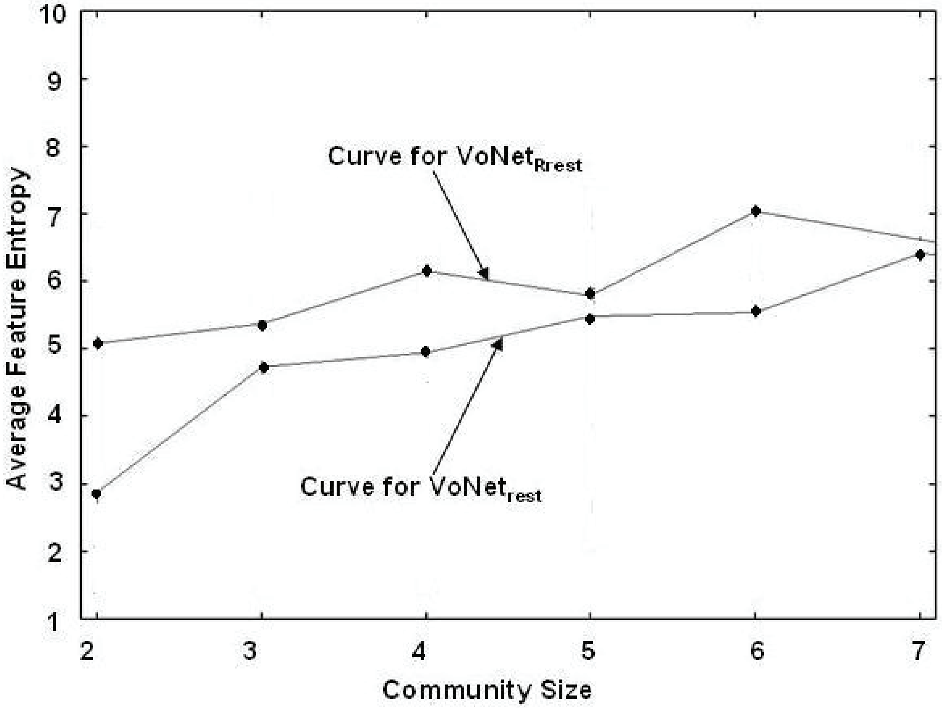


Figure 6.10: Average feature entropy of the communities of a particular size versus the community size for the case of VoNet$_{rest}$ as well as VoNet$_{Rrest}$

ties of a particular size (y-axis) versus the community size (x-axis) for both VoNet$_{rest'}$ and VoNet$_{Rrest'}$. The figure indicates that the average feature entropy exhibited by the communities of VoNet$_{rest'}$ are significantly lower than that of VoNet$_{Rrest'}$. This result immediately reveals that it is again feature economy that plays a key role in the emergence of the communities of VoNet$_{rest'}$.

6.4 Redundancy Ratio of the Vowel Inventories

In this section, we report the redundancy ratio of the vowel inventories and compare the results obtained with those of the consonant inventories. Figure 6.12 and Table 6.6 together summarize the results of the experiments with the vowel inventories. An important observation is that in this case the Zipfian exponent λ is more than three times what is obtained for the consonant inventories. The mean square error around the power-law line is also significantly high (0.110 compared to 0.033). Fur-

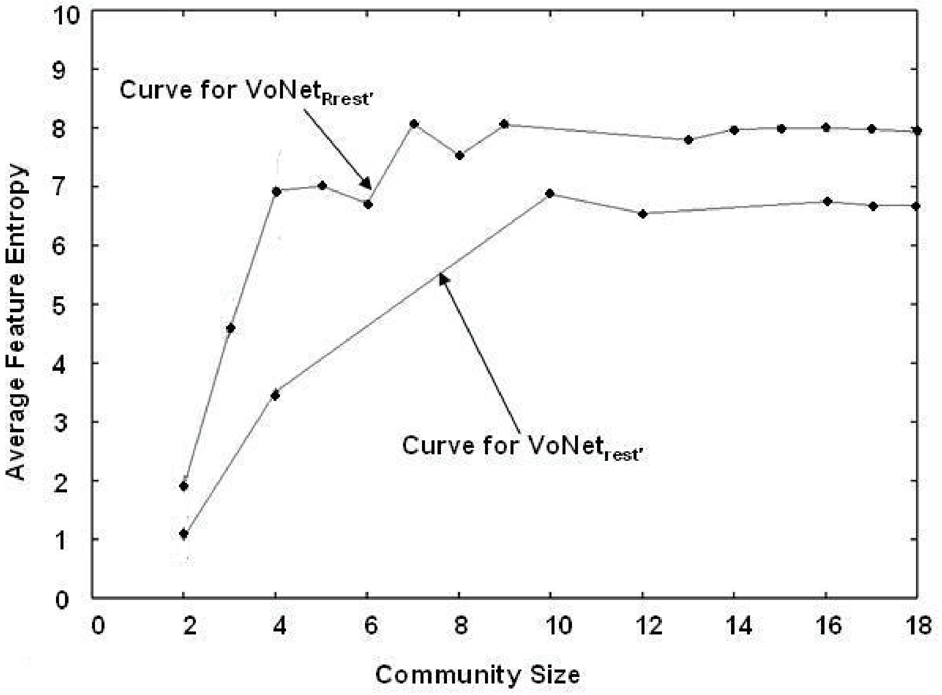


Figure 6.11: Average feature entropy of the communities of a particular size versus the community size for the case of VoNet$_{rest'}$ as well as VoNet$_{Rrest'}$

Table 6.6: Different statistical properties of the distributions of RR values for the vowel inventories. All the notations bear the same meaning as in Table 5.3

	(A, λ)	MSE	μ	σ
Vowel Inv. (full set)	(7.20, 0.40)	0.110	3.43	0.71
Vowel Inv. (Inv. size > 12)	(4.17, 0.17)	0.037	2.65	0.21

thermore, the standard deviation has increased from 0.22 to 0.71. This observation immediately points to the fact that the collection of vowel inventories, in general, lack the universal structural property of constant redundancy unlike the consonant inventories. However, if we inspect only the larger size vowel inventories then the results are quite similar to those obtained for the consonant inventories. For instance, if we consider only the vowel inventories with size > 12 then the Zipfian exponent λ drops to 0.17 while the mean square error and the standard deviation drop to 0.037 and 0.21 respectively. Therefore, for the vowel inventories, the range of inventory size that the data set covers, influences the nature of the distribution of the RR values.

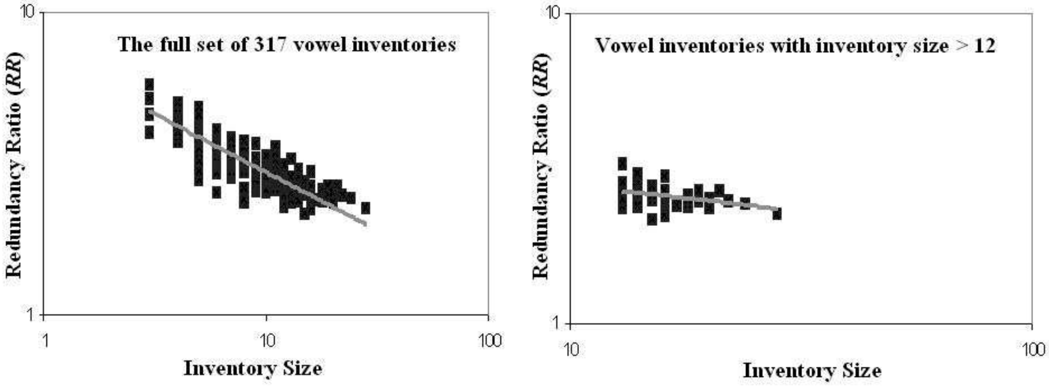


Figure 6.12: Inventory size versus RR in doubly-logarithmic scale for the full set of vowel inventories as well as for those that have inventory size > 12. The bold lines indicate the fits for the distributions

The above observation is possibly related to the fact that compositionality is a useful characteristic of larger sign systems while it is not so for the smaller sign systems (see [82] for reference). For smaller sign systems, non-compositionality provides better distinctiveness, but for larger systems, it is better to divide the system into a set of independent components that can be combined compositionally. Since consonant inventories are examples of large sign systems they seem to be organized into a small number of groups that could be arranged as matrices, e.g., Sanskrit, with plosives at five points of articulation by five manners of articulation. On the other hand, smaller vowel inventories e.g., /i/, /a/, /u/, are actually representatives of small sign systems and cannot be arranged as matrices like the consonant inventories. In general, for consonant inventories, we always observe the inherent three-dimensional compositionality of place of articulation, manner of articulation and phonation. For vowels, we observe non-compositionality in the smaller inventories while compositionality seems to prevail in the larger inventories.

6.5 Summary

In this chapter, we have shown how the computational framework developed in the previous chapters can be suitably applied to study the occurrence as well as the co-occurrence principles of the vowels across the inventories of the world's languages.

Some of our important findings from this work are,

(i) The topological properties of the vowel networks (where the nodes are assumed to be unlabeled) are to a large extent similar to the consonant networks. In principle, preferential attachment plays the most crucial role in the emergence of these properties.

(ii) The smaller vowel inventories (corresponding to the communities of VoNet$_{hub}$) tend to be organized based on the principle of maximal perceptual contrast.

(iii) On the other hand, the larger vowel inventories (mainly composed of the communities of VoNet$_{rest}$) reflect a considerable extent of feature economy;

(iv) Co-occurrences based on markedness (captured through the communities of VoNet$_{rest'}$) also reflect the presence of feature economy.

(v) Vowel inventories, in general, do not exhibit a constant redundancy ratio; however, if only the larger size inventories are considered then one can again observe that the redundancy ratio is almost fixed.

Linguistic Implications

The principles of feature economy and maximal perceptual contrast operate at a certain level of cognition, where speech sounds are assumed to be encoded in terms of relatively abstract elements (features) within a linguistic system. While feature economy tends to organize the linguistic data into a small number of groups, perceptual contrast tends to increase this number so as to minimize the level of confusion. If U be the set of linguistic units and C the categories that characterize these units then feature economy may be expressed as 'maximize U/C' (as suggested in [38]) while perceptual contrast as 'minimize U/C'. It is the interplay of these two optimization principles that shapes the structure of the vowel inventories. The speakers of smaller inventories can afford to choose perceptually distinct vowels because, there are a very few vowels to be learnt and hence the learnability effort is low. On the other hand, the larger inventories tend to be economic so that the effort of learnability does not increase considerably [95]. This is because, even though there are many vowels to be learnt yet due to the prevalence of feature economy the number of features that are actually to be learnt are less. In short, the difference in the average size of the consonant and the vowel inventories seems to be an important factor regulating the difference in their behavior as well as the overall organization.

It is important to mention here that all the results that we have presented throughout the thesis heavily depend on the assumptions in our models as well as on the data source. Consequently, the inferences (both statistical as well as purely linguistic) that we draw needs to be interpreted with caution. This, in fact, is a general drawback of any computational model. None of our inferences, therefore, should be assumed to be sacrosanct; in contrast, they are only indicative and meant to act as pointers for further research in computational phonology.

Chapter 7

Conclusion

One of the most fundamental problems in linguistics is the characterization and explanation of the universal patterns that are observed across human languages. In other words, the amount of variation across languages is constrained by these universal patterns. Such patterns embody a kind of ordered complexity similar to that found in the natural world of living systems and its other artifacts. Of late, sophisticated computational models have enabled researchers to explain the emergence of these patterns that manifest at different levels of linguistic structure (e.g., phonology, morphology, syntax and semantics).

In this thesis, we set out to identify as well as explain the emergence of the universal patterns found across the sound inventories of the world's languages. As we had already pointed out in Chapter 2, various computational models have been proposed by the past researchers to investigate these patterns and many interesting outcomes have been reported. Nevertheless, we also observed that it becomes increasingly difficult to model the structure of the inventories of mainly the complex utterances such as consonants and syllables. Therefore, in the last four chapters, we tried to develop a new computational model that can serve as a unified framework for studying the self-organization of the sound inventories which, in turn, explains the emergent sound patterns. In this context, we postulate that

> *complex networks can be suitably employed not only to represent the structure of the sound inventories but also to detect the universal patterns across them and explain the self-organizing process that leads to the emergence of these patterns. We believe that our thesis is reasonable because, such a computational framework actually allowed us to make a number of interesting inferences about the patterns across the sound inventories most of which were outlined in the previous chapters.*

In this chapter, we shall mainly attempt to summarize our contributions (section 7.1) and wrap up by pointing out some of the possible future directions of research that have been opened up by this thesis (section 7.2).

7.1 Summary of our Contributions

The objectives that we had laid out in the introduction of this thesis have been fulfilled. The first objective was to formulate a representation of the structure of the inventories. Towards this end, we proposed a bipartite network representation of the inventories. We further deduced the one-mode projection of this network which may be thought of as a more compact representation of the inventory structure.

The second objective was to devise suitable methods of analysis that would bring forth prevalent patterns across the sound inventories. In order to meet this objective, we investigated various topological properties of the bipartite network and its one-mode projection. We found from this analysis that (a) the average size of the vowel inventories is smaller than that of the consonant inventories, (b) the occurrence and co-occurrence of the consonants and the vowels across the language inventories follow well-behaved probability distributions, (c) the co-occurrence networks (i.e., the one-mode projections) of both consonants as well as vowels are characterized by high clustering coefficients which is an outcome of the presence of a large number of triangles in these networks, and (d) these triangles or the "tightly-knit communities" are formed on the basis of two functional principles – perceptual contrast and feature economy; while the consonant communities display mainly the presence of feature economy, vowel communities, depending on the frequency of occurrence of the constituent vowels, exhibit either perceptual contrast or feature economy.

The third and the last objective was to formulate growth models to synthesize the topological properties of the networks. Towards this end, we presented a preferential attachment based model for the bipartite network and showed that it can reproduce the distribution of occurrence of the consonants/vowels quite accurately. A theoretical analysis of the model revealed that actually this distribution asymptotically approaches a β-distribution. We further identified that the distribution of the size of the inventories affects the co-occurrence distribution of the consonant/vowels even though it does not affect the occurrence distribution. Subsequently, we incorporated this factor into our model so as to match the emergent co-occurrence distribution with the real data to a close approximation. Finally, we refined our preferential attachment based model to include the process of triad formation and, thereby, explained analytically as well as through simulations, the high clustering exhibited by the co-occurrence networks.

7.2 Future Directions

In this final section, we outline a few out of the many possible directions of future research that have been opened up by this thesis. Some of the specific problems that one might focus on could be the design of computationally tractable microscopic models from explaining the emergence of the consonant inventories or modeling of the structure and the dynamics of other types of inventories such as the syllable inventories. However, a more general aim could center around the development of a full-fledged computational framework (possibly) based on complex networks to study *self-organized phonology*. The primary objective would be to propose and validate theoretical models of speech sound self-organization covering different phonological phenomena including typology, markedness, learnability, phonotactics and various other processes.

Another line of research could progress in the direction of more sophisticated analytical treatment of the network growth models mainly by relaxing the different assumptions that we made for solving them. One might also attempt to theoretically derive other important topological properties and, in particular, the spectral properties (i.e., the eigenvalues and the eigenvectors) of the emergent networks.

Finally, network representations like those discussed in the thesis may be employed to tackle certain application-specific tasks in the areas of Natural Language Processing (NLP) and Information Retrieval (IR). One can construct various types of linguistic networks of words and use their different topological properties for the purpose of language modeling which is a very important problem in NLP. Furthermore, clustering of these networks can help in the identification of the word categories (either syntactic or semantic) in a completely unsupervised manner. Large amount of annotated data is a prime requirement for enhancing the performance of any NLP application. It is usually hard to avail such annotated data especially, for the resource poor languages. In such a scenario, unsupervised methods based on clustering of linguistic networks can be useful in the automatic creation of annotated data from raw corpus. This annotated data may be in turn used for bootstrapping supervised NLP algorithms. There have been already some initial work in this direction reported in [17]. Hyperlink structure analysis, analysis of blogs as well as analysis of query-logs are some of the areas in IR that can also benefit from the application of complex network based methods.

In summary, computational modeling seems to be a very fruitful way of doing research in evolutionary linguistics and as we have observed in this thesis the process of modeling and the associated analysis might turn out to be difficult, but it is certainly not impossible.

Bibliography

[1] M. Abramowitz and I. Stegun, editors. *Handbook of Mathematical Functions*. Dover Publications, New York, 1974.

[2] L. A. Adamic and E. Adar. Friends and neighbors on the web. *Social Networks*, 25(3):211–230, 2003.

[3] L. A. Adamic and B. A. Huberman. Power-law distribution of the world wide web. *Science*, 287:2115, 2000.

[4] L. A. Adamic, R. M. Lukose, and B. A. Huberman. Local search in unstructured networks. In S. Bornholdt and H. G. Schuster, editors, *Handbook of Graphs and Networks*. Wiley-VCH, 2003.

[5] M. E. Adilson, A. P. S. de Moura, Y. C. Lai, and P. Dasgupta. Topology of the conceptual network of language. *Phys. Rev. E*, 65:1–4, 2002.

[6] R. Albert. Boolean modeling of genetic regulatory networks. In E. Ben-Naim, H. Frauenfelder, and Z. Toroczkai, editors, *Complex Networks*, volume 650 of *Lecture Notes in Physics*, pages 459–481. Springer, 2004.

[7] R. Albert and A. L. Barabási. Topology of evolving networks: local events and universality. *Phys. Rev. Lett.*, 85:5234–5237, 2000.

[8] R. Albert and A. L. Barabási. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74:47, 2002.

[9] R. Albert, H. Jeong, and A.-L. Barabási. Diameter of the world-wide web. *Nature*, 401:130–131, 1999.

[10] L. A. N. Amaral, A. Scala, M. Barthélémy, and H. E. Stanley. Classes of small-world networks. *PNAS*, 97:11149–11152, 2000.

[11] P. Arhem, H. A. Braun, M. T. Huber, and H. Liljenstrom. *Micro-Meso-Macro: Addressing Complex Systems Couplings.* World Scientific, 2004.

[12] M. Balter. Early date for the birth of indo-european languages. *Science*, 302(5650):1490, 2003.

[13] A. L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286:509–512, 1999.

[14] A. L. Barabási., H. Jeong, R. Ravasz, Z. Néda, T. Vicsek, and A. Schubert. On the topology of the scientific collaboration networks. *Physica A*, 311:590–614, 2002.

[15] A. Barrat, M. Barthélemy, R. Pastor-Satorras, and A. Vespignani. The architecture of complex weighted networks. *PNAS*, 101:3747–3752, 2004.

[16] S. Battiston and M. Catanzaro. Statistical properties of corporate board and director networks. *Eur. Phys. J. B*, 38:345–352, 2004.

[17] C. Biemann. Unsupervised part-of-speech tagging employing efficient graph clustering. In *Proceedings of COLING/ACL 2006 Student Research Workshop*, pages 7–12, 2006.

[18] J. Blevins. *Evolutionary Phonology: The Emergence of Sound Patterns.* Cambridge University Press, 2004.

[19] J. Blevins. The importance of typology in explaining recurrent sound patterns. *Linguistic Typology*, 11:107–113, 2007.

[20] P. Boersma. *Functional Phonology.* The Hague: Holland Academic Graphics, 1998.

[21] D. Bu, Y. Zhao, L. Cai, H. Xue, X. Zhu, H. Lu, J. Zhang, S. Sun, L. Ling, N. Zhang, G. Li, and R. Chen. Topological structure analysis of the protein–protein interaction network in budding yeast. *Nucleic Acids Research*, 31(9):2443–2450, 2003.

[22] M. G. Bulmer. *Principles of Statistics*. Dover Publications, 1979.

[23] J. L. Bybee. Diachronic and typological properties of morphology and their implications for representation. In L. B. Feldman, editor, *Morphological Aspects of Language Processing*, pages 225–246. Lawrence Erlbaum Associates, Hillsdale, 1995.

[24] G. Caldarelli and M. Catanzaro. The corporate boards networks. *Physica A*, 338:98–106, 2004.

[25] R. Ferrer-i-Cancho and R. V. Solé. The small world of human language. *Proceedings of The Royal Society of London. Series B, Biological Sciences*, 268(1482):2261–2265, 2001.

[26] R. Ferrer-i-Cancho and R. V. Solé. Two regimes in the frequency of words and the origin of complex lexicons: Zipf's law revisited. *Journal of Quantitative Linguistics*, 8:165–173, 2001.

[27] R. Ferrer-i-Cancho and R. V. Solé. Patterns in syntactic dependency networks. *Phys. Rev. E*, 69:051915, 2004.

[28] R. Ferrer-i-Cancho. The structure of syntactic dependency networks: insights from recent advances in network theory. In V. Levickij and G. Altmman, editors, *Problems of Quantitative Linguistics*, pages 60–75. 2005.

[29] R. Ferrer-i-Cancho, O. Riordan, and B. Bollobás. The consequences of Zipf's law for syntax and symbolic reference. *Proceedings of the Royal Society of London Series B*, 272:561–565, 2005.

[30] R. Ferrer-i-Cancho. Why do syntactic links not cross? *Europhys. Lett.*, 76:1228–1235, 2006.

[31] R. Ferrer-i-Cancho, A. Mehler, O. Pustylnikov, and A. Díaz-Guilera. Correlations in the organization of large-scale syntactic dependency networks. In *TextGraphs-2: Graph-Based Algorithms for Natural Language Processing*, pages 65–72. Association for Computational Linguistics, 2007.

[32] R. Carré. 'Speaker' and 'Speech' characteristics: A deductive approach. *Phonetica*, 51:7–16, 1994.

[33] R. Carré. Prediction of vowel systems using a deductive approach. In *Proceedings of the ICSLP*, pages 434–437, 1996.

[34] T. Carter. An introduction to information theory and entropy. *http:// citeseer.ist.psu.edu/ carter00introduction.html*, 2000.

[35] N. Chomsky and M. Halle. *The Sound Pattern of English.* New York: Harper and Row, 1968.

[36] N. Chomsky. *The Minimalist Program.* MIT Press, Cambridge, MA, 1995.

[37] M. Choudhury, M. Thomas, A. Mukherjee, A. Basu, and N. Ganguly. How difficult is it to develop a perfect spell-checker? A cross-linguistic analysis through complex network approach. In *Proceedings of the Second Workshop on TextGraphs: Graph-Based Algorithms for Natural Language Processing*, pages 81–88. Association for Computational Linguistics, 2007.

[38] G. N. Clements. Feature economy in sound systems. *Phonology*, 20:287–333, 2003.

[39] G. N. Clements. The role of features in speech sound inventories. In E. Raimy and C. Cairns, editors, *Contemporary Views on Architecture and Representations in Phonological Theory.* Cambridge, MA: MIT Press, 2008.

[40] J. Coates. *Women, Men and Language.* Longman, London, 1993.

[41] L. da F. Costa, F. A. Rodrigues, G. Travieso, and P. R. V. Boas. Characterization of complex networks: A survey of measurements. *Advances in Physics*, 56(1):167–242, 2007.

[42] L. da F. Costa, O. N. Oliveira Jr., G. Travieso, F. A. Rodrigues, P. R. V. Boas, L. Antiqueira, M. P. Viana, and L. E. C. da Rocha. Analyzing and modeling real-world phenomena with complex networks: A survey of applications. *arXiv:0711.3199v2*, to appear.

[43] C. R. Darwin. *The Descent of Man, and Selection in Relation to Sex.* John Murray, London, 1871.

[44] B. de Boer. *Self-Organisation in Vowel Systems.* PhD Thesis, AI Lab, Vrije Universiteit Brussel, 1999.

[45] A. W. de Groot. Phonologie und phonetik als funktionswissenschaften. *Travaux du Cercle Linguistique de Prague*, 4:116–147, 1931.

[46] S. N. Dorogovtsev and J. F. F. Mendes. Language as an evolving word web. *Proceedings of the Royal Society of London B: Biological Sciences*, 268(1485):2603–2606, 2001.

[47] S. N. Dorogovtsev and J. F. F. Mendes. *Evolution of Networks: From Biological Nets to the Internet and WWW.* Oxford University Press, 2003.

[48] J. A. Dunne, R. J. Williams, N. D. Martinez, R. A. Wood, and D. H. Erwin. Compilation and network analyses of Cambrian food webs. *PLoS Biology*, 6(4):e102, 2008.

[49] S. Eubank, H. Guclu, V. S. A. Kumar, M. V. Marate, A. Srinivasan, Z. Toroczkai, and N. Wang. Modelling disease outbreaks in realistic urban social networks. *Nature*, 429:180–184, 2004.

[50] T. S. Evans and A. D. K. Plato. Exact solution for the time evolution of network rewiring models. *Phy. Rev. E*, 75:056101, 2007.

[51] M. G. Everett and S. P. Borgatti. Analyzing clique overlap. *Connections*, 21(1):49–61, 1998.

[52] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. *Computer Communications Review*, 29:251–262, 1999.

[53] C. Felbaum. *Wordnet, an Electronic Lexical Database for English.* MIT Press, Cambridge, MA, 1998.

[54] G. W. Flake, S. R. Lawrence, C. L. Giles, and F. M. Coetzee. Self-organization and identification of Web communities. *IEEE Computer*, 35:66–71, 2002.

[55] E. Flemming. *Auditory Representations in Phonology.* New York & London: Routledge, 2002.

[56] E. Flemming. Deriving natural classes in phonology. *Lingua*, 115(3):287–309, 2005.

[57] S. Fortunato. Community detection in graphs. *arXiv:0906.0612*, to appear.

[58] B. J. Frey and D. Dueck. Clustering by passing messages between data points. *Science*, 315(5814):972–976, 2007.

[59] L. L. Gatlin. Conservation of Shannon's redundancy for proteins. *Jour. Mol. Evol.*, 3:189–208, 1974.

[60] M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *PNAS*, 99:7821–7826, 2002.

[61] H. Glotin. *La Vie Artificielle d'une Société de Robots Parlants: Émergence et Changement du Code Phonétique.* DEA sciences cognitives-Institut National Polytechnique de Grenoble, 1995.

[62] H. Glotin and R. Laboissière. Emergence du code phonétique dans une societe de robots parlants. In *Actes de la Conférence de Rochebrune 1996 : du Collectif au social.* Ecole Nationale Supérieure des Télécommunications, Paris, 1996.

[63] D. E. Goldberg. *Genetic Algorithms in Search, Optimization, and Machine Learning.* Addison–Wesley, 1998.

[64] C. M. Grinstead and J. L. Snell. *Introduction to Probability*, chapter 7, pages 285–304. AMS Bookstore, 1997.

[65] T. Gruenenfelder and D. B. Pisoni. *Modeling the Mental Lexicon as a Complex Graph.* Ms. Indiana University, 2005.

[66] J. L. Guillaume and M. Latapy. Bipartite structure of all complex networks. *Inf. Process. Lett.*, 90(5):215–221, 2004.

[67] J. L. Guillaume and M. Latapy. Bipartite graphs as models of complex networks. *Physica A*, 371:795–813, 2006.

[68] R. Guimerà, S. Mossa, A. Turtschi, and L. A. N. Amaral. The worldwide air transportation network: Anomalous centrality, community structure, and cities' global roles. *PNAS*, 102:7794–7799, 2005.

[69] F. Hinskens and J. Weijer. Patterns of segmental modification in consonant inventories: A cross-linguistic study. *Linguistics*, 41(6):1041–1084, 2003.

[70] C. F. Hockett. *A Manual of Phonology.* University of Chicago Press, 1974.

[71] P. Holme, M. Huss, and H. Jeong. Subnetwork hierarchies of biochemical pathways. *Bioinformatics*, 19:532–538, 2003.

[72] P. Holme and B. J. Kim. Growing scale-free networks with tunable clustering. *Phys. Rev. E*, 65:026107, 2002.

[73] Y. Hu, M. Li, P. Zhang, Y. Fan, and Z. Di. Community detection by signaling on complex networks. *Phys. Rev. E*, 78:016115, 2008.

[74] B. D. Hughes. *Random Walks and Random Environments: Random walks*, volume 1. Oxford Science Publications, 1995.

[75] J. Ingram. Lecture notes on phonological development 1 (prelinguistic, first words). *http://www.emsah.uq.edu.au/linguistics/teaching/LING1005/*, 2005.

[76] H. Jeong, S. Mason, A. L. Barabási, and Z. N. Oltvai. Lethality and centrality in protein networks. *Nature*, 411:41–42, 2001.

[77] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A. L. Barabási. The large-scale organization of metabolic networks. *Nature*, 407:651–654, 2000.

[78] V. Kapatsinski. Sound similarity relations in the mental lexicon: Modeling the lexicon as a complex network. *Speech Research Lab Progress Report, Indiana University*, 2006.

[79] V. Kapustin and A. Jamsen. Vertex degree distribution for the graph of word co-occurrences in Russian. In *Proceedings of the Second Workshop on TextGraphs: Graph-Based Algorithms for Natural Language Processing*, pages 89–92. Association for Computational Linguistics, 2007.

[80] J. Ke, M. Ogura, and W. S.-Y. Wang. Optimization models of sound systems using genetic algorithms. *Computational Linguistics*, 29(1):1–18, 2003.

[81] B. W. Kernighan and S. Lin. An efficient heuristic procedure for partitioning graphs. *Bell System Technical Journal*, 49:291–307, 1970.

[82] S. Kirby. Syntax without natural selection: How compositionality emerges from vocabulary in a population of learners. In C. Knight, editor, *The Evolutionary Emergence of Language: Social Function and the Origins of Linguistic Form*, pages 303–323. Cambridge University Press, 2000.

[83] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220(4598):671–680, 1983.

[84] R. Köhler. *Zur Linguistischen Synergetik. Struktur und Dynamik der Lexik.* Bochum: Brockmeyer, 1986.

[85] R. Köhler. *Bibliography of Quantitative Linguistics.* Amsterdam: Benjamins, 1995.

[86] R. Köhler. Synergetic linguistics. In R. Köhler, G. Altmann, and R. G. Piotrovskií, editors, *Quantitative Linguistics. An International Handbook*, pages 760–775. Walter de Gruyter, Berlin, New York, 2005.

[87] S. Kulkarni, N. Ganguly, G. Canright, and A. Deutsch. A bio-inspired algorithm for location search in peer to peer network. In F. Dressler and I. Carreras, editors, *Advances in Biologically Inspired Information Systems: Models, Methods, and Tools*, Studies in Computational Intelligence, pages 267–282. Springer, 2007.

[88] P. Ladefoged and I. Maddieson. *The Sounds of the World's Languages.* Oxford: Blackwell, 1996.

[89] R. Lambiotte and M. Ausloos. N-body decomposition of bipartite networks. *Phys. Rev. E*, 72:066117, 2005.

[90] M. Latapy and P. Pons. Computing communities in large networks using random walks. In *Proceedings of the International Symposium on Computer*

and Information Sciences, Lecture Notes in Computer Science, pages 284–293. Springer-Verlag, 2005.

[91] A. M. Lesk. *Introduction to Bioinformatics.* Oxford University Press, New York, 2002.

[92] J. Liljencrants and B. Lindblom. Numerical simulation of vowel quality systems: the role of perceptual contrast. *Language*, 48:839–862, 1972.

[93] B. Lindblom, P. MacNeilage, and M. Studdert-Kennedy. Self-organizing processes and the explanation of language universals. In *Explanations for Language Universals*, pages 181–203. Walter de Gruyter & Co., 1984.

[94] B. Lindblom. Phonetic universals in vowel systems. In J. J. Ohala and J. J. Jaeger, editors, *Experimental Phonology*, pages 13–44. Orlando (FL): Academic Press, 1986.

[95] B. Lindblom. Systemic constraints and adaptive change in the formation of sound structure. In J. Hurford, editor, *Evolution of Human Language.* Edinburgh University Press, Edinburgh, 1996.

[96] B. Lindblom and I. Maddieson. Phonetic universals in consonant systems. In M. Hyman and C. N. Li, editors, *Language, Speech, and Mind*, pages 62–78, 1988.

[97] S. P. Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137, 1982.

[98] P. A. Luce and D. B. Pisoni. Recognizing spoken words: The neighborhood activation model. *Ear and Hearing*, 19:1–36, 1998.

[99] P. F. MacNeilage. *The Origin of Speech.* Oxford University Press, 2008.

[100] J. B. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297. University of California Press, Berkeley, 1967.

[101] I. Maddieson. *Working Papers in Phonetics*. Department of Linguistics, UCLA, N050, 1980.

[102] I. Maddieson. *Patterns of Sounds*. Cambridge University Press, 1984.

[103] I. Maddieson. In search of universals. In *Proceedings of the XIVth International Congress of Phonetic Sciences*, pages 2521–2528, 1999.

[104] W. Marslen-Wilson. Activation, competition, and frequency in lexical access. In G. Altmann, editor, *Cognitive Models of Speech Processing: Psycholinguistic and Computational Perspectives*, pages 148–173. Cambridge, MA: MIT, 1990.

[105] A. Martinet. *Ѐconomie des Changements Phonétiques*. Berne: A. Francke, 1955.

[106] S. Maslov and K. Sneppen. Specificity and stability in topology of protein networks. *Science*, 296:910–913, 2002.

[107] J. J. McCarthy. *A Thematic Guide to Optimality Theory*. Cambridge University Press, 2002.

[108] J. Mielke. *Emergence of Distinctive Features*. PhD thesis, The Ohio State University, 2004.

[109] B. Mitra, S. Ghose, and N. Ganguly. Effect of dynamicity on peer to peer networks. In *High Performance Computing (HiPC)*, Lecture Notes in Computer Science, pages 452–463. Springer, 2007.

[110] B. Mitra, F. Peruani, S. Ghose, and N. Ganguly. Analyzing the vulnerability of superpeer networks against attack. In *Proceedings of the 14^{th} ACM conference on Computer and Communications Security (CCS)*, pages 225–234, 2007.

[111] B. Mitra, F. Peruani, S. Ghose, and N. Ganguly. Measuring robustness of superpeer topologies. In *Proceedings of the 26^{th} annual ACM symposium on Principles of Distributed Computing (PODC)*, pages 372–373, 2007.

[112] J. Moody. Race, school integration, and friendship segregation in america. *American Journal of Sociology*, 107:679–716, 2001.

[113] M. E. J. Newman. Scientific collaboration networks. *Phys. Rev. E*, 64:016131, 2001.

[114] M. E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45(2):167–256, 2003.

[115] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Phys. Rev. E*, 69:026113, 2004.

[116] M. E. J. Newman, S. H. Strogatz, and D. J. Watts. Random graphs with arbitrary degree distributions and their applications. *Phy. Rev. E*, 64:026118, 2001.

[117] J. J. Ohala. Moderator's introduction to the Symposium on phonetic universals in phonological systems and their explanation. In E. Fischer-Jørgensen and N. Thorsen, editors, *Proceedings of the 9th ICPhS*, volume 3, pages 181–185. University of Copenhagen: Institute of Phonetics, 1980.

[118] J. Ohkubo, K. Tanaka, and T. Horiguchi. Generation of complex bipartite graphs by using a preferential rewiring process. *Phys. Rev. E*, 72:036120, 2005.

[119] P.-Y. Oudeyer. *Self-organization in the Evolution of Speech.* Oxford University Press, Oxford, 2006.

[120] G. Palla, I. Derényi, I. Farkas, and T. Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435:814–818, 2005.

[121] R. Pastor-Satorras and A. Vespignani. Epidemic spreading in scale-free networks. *Phys. Rev. Lett.*, 86:3200–3203, 2001.

[122] R. Pastor-Satorras and A. Vespignani. Epidemic spreading in scale-free networks. *Phys. Rev. E*, 65:036104, 2002.

[123] M. Peltomäki and M. Alava. Correlations in bipartite collaboration networks. *Journal of Statistical Mechanics: Theory and Experiment*, 1:P01010, 2006.

[124] V. Pericliev and R. E. Valdés-Pérez. Differentiating 451 languages in terms of their segment inventories. *Studia Linguistica*, 56(1):1–27, 2002.

[125] S. Pinker. *The Language Instinct: How the Mind Creates Language*. Harper-Collins, New York, 1994.

[126] A. Pothen, H. Simon, and K. P. Liou. Partitioning sparse matrices with eigenvectors of graphs. *SIAM J. Matrix Anal. Appl.*, 11:430–452, 1990.

[127] A. Prince and P. Smolensky. *Optimality Theory: Constraint Interaction in Generative Grammar*. Wiley-Blackwell, 2004.

[128] F. Pulvermüller, M. Huss, F. Kherif, F. M. d. P. Martin, O. Hauk, and Y. Shtyrov. Motor cortex maps articulatory features of speech sounds. *PNAS*, 103(20):7865–7870, 2006.

[129] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi. Defining and identifying communities in networks. *PNAS*, 101(9):2658–2663, 2003.

[130] J. J. Ramasco, S. N. Dorogovtsev, and R. Pastor-Satorras. Self-organization of collaboration networks. *Phys. Rev. E*, 70:036106, 2004.

[131] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A. L. Barabási. Hierarchical organization of modularity in metabolic networks. *Science*, 297:1551–1555, 2002.

[132] M. A. Redford, C. C. Chen, and R. Miikkulainen. Modeling the emergence of syllable systems. In *CogSci '98*, pages 882–886, 1998.

[133] J. Reichardt and S. Bornholdt. Detecting fuzzy community structures in complex networks with a Potts model. *Phys. Rev. Lett.*, 93(21):218701, 2004.

[134] J.-L. Schwartz, L.-J. Boë, N. Vallée, and C. Abry. The dispersion-focalization theory of vowel systems. *Journal of Phonetics*, 25:255–286, 1997.

[135] J. Scott. *Social Network Analysis: A Handbook*. Sage, London, 2000.

[136] R. H. Seashore and L. D. Eckerson. The measurement of individual differences in general English vocabularies. *Journal of Educational Psychology*, 31:14–38, 1940.

[137] P. Sen, S. Dasgupta, A. Chatterjee, P. A. Sreeram, G. Mukherjee, and S. S. Manna. Small-world properties of the Indian railway network. *Phys. Rev. E*, 67:036106, 2003.

[138] C. E. Shannon and W. Weaver. *The Mathematical Theory of Information.* University of Illinois Press, Urbana, 1949.

[139] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.

[140] M. Sigman and G. A. Cecchi. Global organization of the wordnet lexicon. *PNAS*, 99:1742–1747, 2002.

[141] H. A. Simon. On a class of skew distribution functions. *Biometrika*, 42:425–440, 1955.

[142] K. Sneppen, M. Rosvall, A. Trusina, and P. Minnhagen. A simple model for self-organization of bipartite networks. *Europhys. Lett.*, 67:349–354, 2004.

[143] M. M. Soares, G. Corso, and L. S. Lucena. The network of syllables in Portuguese. *Physica A*, 355(2-4):678–684, 2005.

[144] Z. Solan, D. Horn, E. Ruppin, and S. Edelman. Unsupervised learning of natural languages. *PNAS*, 102:11629–11634, 2005.

[145] W. Souma, Y. Fujiwara, and H. Aoyama. Complex networks and economics. *Physica A*, 324:396–401, 2003.

[146] L. Steels. A self-organizing spatial vocabulary. *Artificial Life*, 2(3):319–332, 1995.

[147] L. Steels. Language as a complex adaptive system. In M. Schoenauer, editor, *Proceedings of PPSN VI*, Lecture Notes in Computer Science, pages 17–26. Springer-Verlag, 2000.

[148] D. Steriade. Knowledge of similarity and narrow lexical override. *BLS*, 29:583–598, 2004.

[149] K. N. Stevens. The quantal nature of speech: Evidence from articulatory-acoustic data. In E. E. David Jr. and P. B. Denes, editors, *Human Communication: A Unified View*, pages 51–66. McGraw-Hill, New York, 1972.

[150] K. N. Stevens. On the quantal nature of speech. *Journal of Phonetics*, 17(1):3–45, 1989.

[151] M. Steyvers and J. B. Tenenbaum. The large-scale structure of semantic networks: statistical analyses and a model of semantic growth. *Cognitive Science*, 29(1):41–78, 2005.

[152] S. H. Strogatz. Exploring complex networks. *Nature*, 410:268–276, 2001.

[153] M. Tamariz. *Exploring the Adaptive Structure of the Mental Lexicon.* PhD thesis, Department of theoretical and applied linguistics, Univerisity of Edinburgh, 2005.

[154] N. Trubetzkoy. Die phonologischen systeme. *TCLP*, 4:96–116, 1931.

[155] N. Trubetzkoy. *Principles of Phonology.* University of California Press, Berkeley, 1969.

[156] B. Vaux and B. Samuels. Laryngeal markedness and aspiration. *Phonology*, 22(3):395–436, 2005.

[157] M. S. Vitevitch. *Phonological Neighbors in a Small World.* Ms. University of Kansas, 2004.

[158] W. S.-Y. Wang. The basis of speech. In C. E. Reed, editor, *The Learning of Language*, 1971.

[159] S. K. Warfield, J. Rexilius, P. S. Huppi, T. E. Inder, E. G. Miller, W. M. Wells, G. P. Zientara, F. A. Jolesz, and R. Kikinis. A binary entropy measure to assess nonrigid registration algorithm. In *Proc. of MICCAI*, pages 266–274, 2001.

[160] S. Wasserman and K. Faust. *Social Network Analysis.* Cambridge University Press, Cambridge, 1994.

[161] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393:440–442, 1998.

[162] C. Wilke, S. Altmeyer, and T. Martinetz. Large-scale evolution and extinction in a hierarchically structured environment. In *ALIFE*, pages 266–272, 1998.

[163] C. Wilke and T. Martinetz. Hierarchical noise in large systems of independent agents. *Phy. Rev. E*, 58:7101, 1998.

[164] R. J. Williams and N. D. Martinez. Simple rules yield complex food webs. *Nature*, 404:180–183, 2000.

[165] A. Woollard. Gene duplications and genetic redundancy in *c. elegans*. In *WormBook*, 2005.

[166] F. Y. Wu. The Potts model. *Rev. of Mod. Phys.*, 54:235–268, 1982.

[167] H. Zhou. Distance, dissimilarity index, and network community structure. *Phys. Rev. E*, 67:061901, 2003.

[168] H. Zhou and R. Lipowsky. Network brownian motion: A new method to measure vertex-vertex proximity and to identify communities and subcommunities. In *Proceedings of the Workshop on Computational Modeling of Transport on Networks*, Lecture Notes in Computer Science, pages 1062–1069. Springer-Verlag, 2004.

[169] G. K. Zipf. *Human Behaviour and the Principle of Least Effort*. Addison-Wesley, Cambridge, MA, 1949.

Appendix A

Publications from the Thesis

Publications from the work presented in the thesis: The publications are listed in chronological order with comments in parenthesis.

[1] M. Choudhury, A. Mukherjee, A. Basu, and N. Ganguly. Analysis and synthesis of the distribution of consonants over languages: A complex network approach. In *Proceedings of COLING–ACL*, 128–135, 2006. (A preliminary version of the synthesis model presented in Chapter 3)

[2] F. Peruani, M. Choudhury, A. Mukherjee, and N. Ganguly. Emergence of a non-scaling degree distribution in bipartite networks: A numerical and analytical study. *Euro. Phys. Lett.*, 79(2):28001, 2007. (The analytical treatment of the synthesis model presented in Chapter 3)

[3] A. Mukherjee, M. Choudhury, A. Basu, and N. Ganguly. Modeling the co-occurrence principles of the consonant inventories: A complex network approach. *Int. Jour. of Mod. Phy. C*, 18(2):281–295, 2007. (Community structure analysis of PhoNet presented in Chapter 5)

[4] A. Mukherjee, M. Choudhury, A. Basu, and N. Ganguly. Redundancy ratio: An invariant property of the consonant inventories of the world's languages. In *Proceedings of ACL*, 104–111, 2007. (Redundancy ratio analysis presented in Chapter 5)

[5] A. Mukherjee, M. Choudhury, A. Basu, and N. Ganguly. Emergence of community structures in vowel inventories: An analysis based on complex networks. In *Proceedings of ACL SIGMORPHON9*, 101–108, 2007. (A preliminary version of the community structure analysis of VoNet presented in Chapter 6)

[6] A. Mukherjee, M. Choudhury, S. Roy Chowdhury, A. Basu, and N. Ganguly. Rediscovering the co-occurrence principles of the vowel inventories: A complex network approach. *Advances in Complex Systems*, 11(3):371–392, 2008. (Detailed community structure analysis of VoNet presented in Chapter 6)

[7] A. Mukherjee, M. Choudhury, A. Basu, and N. Ganguly. Modeling the structure and dynamics of the consonant inventories: A complex network approach , In *Proceedings of Coling*, 601–608, 2008. (Synthesis models presented in Chapter 3 and Chapter 4)

[8] M. Choudhury, A. Mukherjee, A. Garg, V. Jalan, A. Basu, and N. Ganguly. Language diversity across the consonant inventories: A study in the framework of complex networks. In *Proceedings of EACL workshop on Cognitive Aspects of Computational Language Acquisition*, 51–58, 2009. (The dynamics within and across different language families presented in Chapter 3)

[9] A. Mukherjee, M. Choudhury, A. Basu, and N. Ganguly. Self-organization of sound inventories: Analysis and synthesis of the occurrence and co-occurrence network of consonants. *Journal of Quantitative Linguistics*, 16(2):157–184, 2009. (A detailed empirical analysis of some of the models presented in Chapter 3 and Chapter 4)

[10] A. Mukherjee, M. Choudhury, and N. Ganguly. Analyzing the degree distribution of the one-mode projection of alphabetic bipartite networks ($\alpha - BiN$s). *preprint: arXiv.org:0902.0702*. (Theoretical analysis of the one-mode projection presented in Chapter 4)

Other related publications: The publications are listed in chronological order.

[1] M. Choudhury, M. Thomas, A. Mukherjee, A. Basu, and N. Ganguly. How

difficult is it to develop a perfect spell-checker? A corss-linguistic analysis through complex network approach. In *Proceedings of HLT-NAACL workshop – TextGraphs-2*, 81–88, 2007.

[2] A. Mukherjee, M. Choudhury, and R. Kannan. Discovering global patterns in linguistic networks through spectral analysis: A case study of the consonant inventories. In *Proceedings of EACL*, 585–593, 2009.

[3] M. Choudhury and A. Mukherjee. The structure and dynamics of linguistic networks. In N. Ganguly, A. Deutsch, and A. Mukherjee, editors, *Dynamics on and of Complex Networks: Applications to Biology, Computer Science, Economics, and the Social Sciences*, Birkhauser, Boston, ISBN: 978-0-8176-4750-6, 2009.

[4] C. Biemann, M. Choudhury, and A. Mukherjee. Syntax is from Mars while semantics from Venus! Insights from spectral analysis of distributional similarity networks. In *Proceedings of ACL*, (accepted as short paper), 2009.

[5] M. Choudhury, N. Ganguly, A. Maiti, A. Mukherjee, L. Brusch, A. Deutsch, and F. Peruani. Modeling discrete combinatorial systems as alphabetic bipartite networks: Theory and applications. *preprint: arXiv.org:0811.0499*.

Appendix B

List of all Publications by the Candidate

Following is a list of all the publications by the candidate including those on and related to the work presented in the thesis. The publications are arranged in chronological order and in three sections – (i) books, (ii) journals and book chapters and (iii) conferences.

B.1 Books

[1] N. Ganguly, A. Deutsch, and A. Mukherjee, editors. *Dynamics on and of Complex Networks: Applications to Biology, Computer Science, Economics, and the Social Sciences*, Birkhauser, Springer, Boston, ISBN: 978-0-8176-4750-6, 2009.

B.2 Journals and Book Chapters

[1] D. Mukhopadhyay, A. Mukherjee, S. Ghosh, P. Chakraborty, and S. Biswas. A new approach for message hiding in steganography and audio hiding using

substitution technique. *Journal of The Institution of Engineers, Computer Engineering Division*, India, 86(2):41–44, 2005.

[2] F. Peruani, M. Choudhury, A. Mukherjee, and N. Ganguly. Emergence of a non-scaling degree distribution in bipartite networks: A numerical and analytical study. *Euro. Phys. Lett.*, 79(2):28001, 2007.

[3] A. Mukherjee, M. Choudhury, A. Basu, and N. Ganguly. Modeling the co-occurrence principles of the consonant inventories: A complex network approach. *Int. Jour. of Mod. Phy. C*, 18(2):281–295, 2007.

[4] M. Choudhury, R. Saraf, V. Jain, A. Mukherjee, S. Sarkar, and A. Basu. Investigation and modeling of the structure of texting language. Special issue of the *International Journal of Document Analysis and Recognition on Noisy Text Analytics*, Springer, 10(3-4):157–174, 2007.

[5] A. Mukherjee, M. Choudhury, S. Roy Chowdhury, A. Basu, and N. Ganguly. Rediscovering the co-occurrence principles of the vowel inventories: A complex network approach. *Advances in Complex Systems*, 11(3):371–392, 2008.

[6] A. Mukherjee, K. Chakraborty, and A. Basu. An adaptive virtual mouse for people with neuro-motor disorders. *Assistive Technology Journal of the Rehabilitation Engineering Society of North America*, 20(2):111–124, 2008.

[7] A. Mukherjee, M. Choudhury, A. Basu, and N. Ganguly. Self-organization of sound inventories: Analysis and synthesis of the occurrence and co-occurrence network of consonants. *Journal of Quantitative Linguistics*, 16(2):157–184, 2009.

[8] M. Choudhury and A. Mukherjee. The structure and dynamics of linguistic networks. In N. Ganguly, A. Deutsch, and A. Mukherjee, editors, *Dynamics on and of Complex Networks: Applications to Biology, Computer Science, Economics, and the Social Sciences*, Birkhauser, Boston, ISBN: 978-0-8176-4750-6, 2009.

[9] R. Pal, A. Mukherjee, P. Mitra, and J. Mukherjee. Modelling visual saliency using degree centrality. *IET Computer Vision* (*in press*).

[10] A. Mukherjee, M. Choudhury, and N. Ganguly. Analyzing the degree distribution of the one-mode projection of alphabetic bipartite networks ($\alpha - BiN$s). *preprint:* *arXiv.org:0902.0702*.

[11] M. Choudhury, N. Ganguly, A. Maiti, A. Mukherjee, L. Brusch, A. Deutsch, and F. Peruani. Modeling discrete combinatorial systems as alphabetic bipartite networks: Theory and applications. *preprint:* *arXiv.org:0811.0499*.

B.3 Conferences

[1] A. Mukherjee, S. Bhattacharya, K. Chakraborty, and A. Basu. Breaking the accessibility barrier: Development of special computer access mechanisms for the neuro-motor disabled in India. In *Proceedings of ICHMI*, 2004.

[2] A. Mukherjee, T. Bhattacharya, K. Chakraborty, and A. Basu. An intelligent tutoring system in promotion of Bangla literacy in the rural scenario. In *Proceedings of NCCPB*, 2005.

[3] A. Mukherjee, S. Bhattacharya, P. K. Halder, and A. Basu. A virtual predictive keyboard as a learning aid for people with neuro-motor disorders. In *Proceedings of ICALT*, 1032–1036, 2005.

[4] M. Choudhury, A. Mukherjee, A. Basu, and N. Ganguly. Analysis and synthesis of the distribution of consonants over languages: A complex network approach. In *Proceedings of COLING–ACL*, 128–135, 2006.

[5] A. Mukherjee and A. Basu. A Bangla predictive keyboard for people with neuro-motor disorders. In *Proceedings of ICCPB*, 2006.

[6] A. Mukherjee, A. Basu, and K. Chakraborty. Accessibility for all: Adaptive computer access mechanisms for the neuro-motor disabled in India, In *Proceedings of the ninety-third Indian Science Congress*, 2006.

[7] A. Mukherjee, M. Choudhury, A. Basu, and N. Ganguly. Redundancy ratio: An invariant property of the consonant inventories of the world's languages. In *Proceedings of ACL*, 104–111, 2007.

[8] A. Mukherjee, M. Choudhury, A. Basu, and N. Ganguly. Emergence of community structures in vowel inventories: An analysis based on complex networks. In *Proceedings of ACL SIGMORPHON9*, 101–108, 2007.

[9] M. Choudhury, M. Thomas, A. Mukherjee, A. Basu, and N. Ganguly. How difficult is it to develop a perfect spell-checker? A corss-linguistic analysis through complex network approach. In *Proceedings of HLT-NAACL workshop – TextGraphs-2*, 81–88, 2007.

[10] A. Gupta, A. Mukherjee, S. Chakraborty, and A. Basu. Mapping graphical user interfaces to scanning mechanisms: A fuzzy approach. In *Proceedings of AACC*, 2007.

[11] A. Mukherjee, M. Choudhury, A. Basu, and N. Ganguly. Modeling the structure and dynamics of the consonant inventories: A complex network approach , In *Proceedings of Coling*, 601–608, 2008.

[12] P. K. Bhowmick, A. Mukherjee, A. Banik, P. Mitra, and A. Basu. A comparative study of the properties of emotional and non-emotional words in the Wordnet: A complex network approach, In *Proceedings of ICON* (poster session), 2008.

[13] J. Nath, M. Choudhury, A. Mukherjee, C. Biemann, and N. Ganguly. Unsupervised parts-of-speech induction for Bengali, In *Proceedings of LREC*, 1220–1227, 2008.

[14] M. Choudhury, A. Mukherjee, A. Garg, V. Jalan, A. Basu, and N. Ganguly. Language diversity across the consonant inventories: A study in the framework of complex networks. In *Proceedings of EACL workshop on Cognitive Aspects of Computational Language Acquisition*, 51–58, 2009.

[15] A. Mukherjee, M. Choudhury, and R. Kannan. Discovering global patterns in linguistic networks through spectral analysis: A case study of the consonant inventories. In *Proceedings of EACL*, 585–593, 2009.

[16] C. Biemann, M. Choudhury, and A. Mukherjee. Syntax is from Mars while semantics from Venus! Insights from spectral analysis of distributional similarity networks. In *Proceedings of ACL*, (accepted as short paper), 2009.

Appendix C

Glossary of the Standard Statistical Properties of Complex Networks

Adjacency Matrix: Let G be a graph with n vertices. The $n \times n$ matrix $\mathbf{A}$ in which each entry $a_{ij} = 1$ if there is an edge between the vertices v_i and v_j in G and rest all entries are 0 is called the adjacency matrix of G.

Assortativity: Assortativity refers to the preference of the nodes in a network to be connected to other nodes that are similar or different in some way.

Assortativity coefficient: The assortativity coefficient is usually expressed in terms of the Pearson's correlation coefficient r between pairs of node degrees. Hence, positive values of r indicate a correlation between nodes of similar degree, while negative values indicate relationships between nodes of different degree.

Betweenness Centrality: Betweenness centrality of a node v is defined as the sum of the ratios of the number of shortest paths between vertices s and t through v to the total number of shortest paths between s and t. The betweenness centrality $g(v)$ of v is given by

$$g(v) = \Sigma_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

Centrality: The centrality of a node in a network is a measure of the structural importance of the node.

Clustering Coefficient: The clustering coefficient for a vertex v in a network is defined as the ratio between the total number of connections among the neighbors of v to the total number of possible connections among the neighbors. For a vertex i with a neighbor set N_i, the clustering coefficient is given by

$$C_i = \frac{|e_{jk}|}{k_i(k_i - 1)} : v_j, v_k \in N_i, e_{jk} \in E$$

Community: A community is a sub-graph, where in some reasonable sense the nodes in the sub-graph have more to do with each other than with the nodes which are outside the sub-graph.

Degree Centrality: Degree centrality is defined as the number of links incident upon a node.

Degree Distribution: The degree distribution of a network is defined as the probability distribution of the degree of a random node in the network.

Diameter: The diameter of a graph is defined as the maximum of all the shortest distances between any two nodes in the graph.

Eigenvector Centrality: Eigenvector centrality is a measure of the importance of a node in a network. It assigns relative scores to all nodes in the network based on the principle that connections to high-scoring nodes contribute more to the score of the node in question than equal connections to low-scoring nodes. Thus, the centrality of a node is proportional to the centrality of the nodes it is connected to and the definition turns out to be recursive.

Euclidean Distance: The Euclidean distance between two vectors a and b is defined as

$$ED(a,b) = \sum_i \sqrt{(a_i - b_i)^2}$$

Pearson's Correlation Coefficient: Pearson's correlation coefficient between two vectors x and y can be measured as

$$r = \frac{\Sigma xy - \frac{\Sigma x \Sigma y}{n}}{\sqrt{(\Sigma x^2 - \frac{(\Sigma x)^2}{n})(\Sigma y^2 - \frac{(\Sigma y)^2}{n})}}$$

Preferential Attachment: Preferential attachment refers to the fact that the more connected a node is, the more likely it is to receive new links. In other words, nodes with higher degree have stronger ability to grab links added to the network.

Random Graph: A random graph is a graph that is generated by some random process. The most common model used to generate such a graph is the Erdös-Rényi (E-R) model. In this model, each pair of n vertices is connected by an edge with some probability p. The probability of a vertex having degree k is given by

$$p_k = \binom{n}{k} p^k (1-p)^{n-k} \simeq \frac{z^k e^{-z}}{k!}$$

where $z = np$.

Scale-Free Network: The defining characteristic of the scale-free networks is that their degree distribution follows the Yule-Simon distribution - a power-law relationship defined by $p_k \sim k^{-\lambda}$.

Small-World Network: A small-world network is a network in which most nodes are not neighbors of one another, but most nodes can be reached from every other node by a small number of hops or steps. These networks show large clustering coefficient and a small average shortest path distance.

Zipf's Law: Zipf's law states that given some corpus of natural language utterances, the frequency of any word is inversely proportional to its rank in the frequency table.

Appendix D

International Phonetic Alphabet Chart

The International Phonetic Alphabet (IPA) is a system of phonetic notations based on the Latin alphabet that has been designed by the International Phonetic Association for a standardized representation of the sounds of spoken language. The IPA represents only those qualities of speech that are distinctive in spoken language such as phonemes, intonation and the separation of words and syllables. The latest version of the IPA was published in 2005 (see http://www.langsci.ucl.ac.uk/ipa/fullchart.html) and a snapshot of it is presented in the next page.

THE INTERNATIONAL PHONETIC ALPHABET (revised to 2005)

CONSONANTS (PULMONIC)

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d		ʈ ɖ	c ɟ	k ɡ	q ɢ	▨	ʔ ▨
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	ɴ	▨	▨
Trill	ʙ			r				▨	ʀ		▨
Tap or Flap		ⱱ		ɾ		ɽ		▨			▨
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateral fricative	▨	▨		ɬ ɮ						▨	▨
Approximant		ʋ		ɹ		ɻ	j	ɰ			▨
Lateral approximant	▨	▨		l		ɭ	ʎ	ʟ		▨	▨

Where symbols appear in pairs, the one to the right represents a voiced consonant. Shaded areas denote articulations judged impossible.

CONSONANTS (NON-PULMONIC)

Clicks		Voiced implosives		Ejectives	
ʘ	Bilabial	ɓ	Bilabial	ʼ	Examples:
ǀ	Dental	ɗ	Dental/alveolar	pʼ	Bilabial
ǃ	(Post)alveolar	ʄ	Palatal	tʼ	Dental/alveolar
ǂ	Palatoalveolar	ɠ	Velar	kʼ	Velar
ǁ	Alveolar lateral	ʛ	Uvular	sʼ	Alveolar fricative

OTHER SYMBOLS

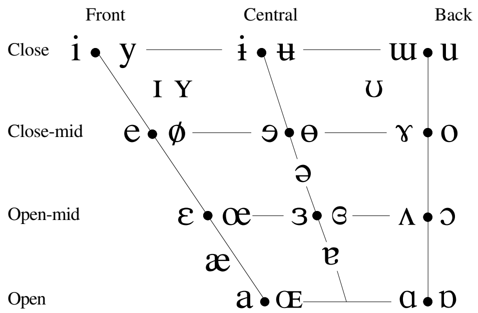
ʍ	Voiceless labial-velar fricative	ɕ ʑ	Alveolo-palatal fricatives
w	Voiced labial-velar approximant	ɺ	Voiced alveolar lateral flap
ɥ	Voiced labial-palatal approximant	ɧ	Simultaneous ʃ and x
ʜ	Voiceless epiglottal fricative		
ʢ	Voiced epiglottal fricative		
ʡ	Epiglottal plosive		

Affricates and double articulations can be represented by two symbols joined by a tie bar if necessary. k͡p t͜s

DIACRITICS Diacritics may be placed above a symbol with a descender, e.g. ŋ̊

◌̥	Voiceless	n̥ d̥	◌̤	Breathy voiced	b̤ a̤	◌̪	Dental	t̪ d̪
◌̬	Voiced	s̬ t̬	◌̰	Creaky voiced	b̰ a̰	◌̺	Apical	t̺ d̺
ʰ	Aspirated	tʰ dʰ	◌̼	Linguolabial	t̼ d̼	◌̻	Laminal	t̻ d̻
◌̹	More rounded	ɔ̹	ʷ	Labialized	tʷ dʷ	◌̃	Nasalized	ẽ
◌̜	Less rounded	ɔ̜	ʲ	Palatalized	tʲ dʲ	ⁿ	Nasal release	dⁿ
◌̟	Advanced	u̟	ˠ	Velarized	tˠ dˠ	ˡ	Lateral release	dˡ
◌̠	Retracted	e̠	ˤ	Pharyngealized	tˤ dˤ	◌̚	No audible release	d̚
◌̈	Centralized	ë	◌̴	Velarized or pharyngealized	ɫ			
◌̽	Mid-centralized	e̽	◌̝	Raised	e̝	(ɹ̝ = voiced alveolar fricative)		
◌̩	Syllabic	n̩	◌̞	Lowered	e̞	(β̞ = voiced bilabial approximant)		
◌̯	Non-syllabic	e̯	◌̘	Advanced Tongue Root	e̘			
˞	Rhoticity	ɚ a˞	◌̙	Retracted Tongue Root	e̙			

VOWELS



Where symbols appear in pairs, the one to the right represents a rounded vowel.

SUPRASEGMENTALS

ˈ	Primary stress	
ˌ	Secondary stress	ˌfoʊnəˈtɪʃən
ː	Long	eː
ˑ	Half-long	eˑ
◌̆	Extra-short	ĕ
|	Minor (foot) group	
‖	Major (intonation) group	
.	Syllable break	ɹi.ækt
‿	Linking (absence of a break)	

TONES AND WORD ACCENTS

LEVEL			CONTOUR		
e̋ or	˥	Extra high	ě or	˩˥	Rising
é	˦	High	ê	˥˩	Falling
ē	˧	Mid	e᷄	˧˥	High rising
è	˨	Low	e᷅	˩˨	Low rising
ȅ	˩	Extra low	e᷈	˧˦˧	Rising-falling
↓		Downstep	↗		Global rise
↑		Upstep	↘		Global fall

Index