Complex network techniques for discovering structures in functional data

Goh, Woon Peng

2017

Goh, W. P. (2017). Complex network techniques for discovering structures in functional data. Doctoral thesis, Nanyang Technological University, Singapore.

<http://hdl.handle.net/10356/71754>

<https://doi.org/10.32657/10356/71754>

Complex Network Techniques for Discovering Structures in Functional Data

by

GOH Woon Peng

A thesis submitted to the Nanyang Technological University
in partial fulfillment of the requirement for the degree of
Doctor of Philosophy

in the
Interdisciplinary Graduate School
Complexity Institute

2017

“Research is the process of going up alleys to see if they’re blind.”

Barstow Bates, October 1967

Acknowledgements

In this culmination of my 4-year research program, I am especially grateful to my doctoral advisor Assoc. Prof. Cheong Siew Ann for his support, wisdom, and positivity that made this thesis and the path to graduation possible. I would also like to thank my project collaborators Dr. Dennis Kwek, Prof. David Hogan, Dr. Jennifer Tan, and Dr. Elizabeth Koh who kindly shared their data and domain expertise that realized two of the projects described in this thesis. I am thankful also to the members of my Thesis Advisory Committee, Prof. Luke Kang-Kwong and Assoc. Prof. Kevin Xiao, for their guidance and suggestions over the years.

I am extremely fortunate to have my friends and colleagues Darrell, Boon Kin, Teck Liang, Wenyan, James, Chee Yong, and Misha from the Statistical and Multi-scale Simulation Laboratory with whom I enjoyed endless hours of productive discussions and also hours of not-so-productive chatter. Their companionship made the days when research led me to blind alleys much more bearable. I wish them luck in their future endeavours and, for some, their imminent graduation.

I must also not forget the support of my family, Ma, Pa, Eng Eng, Lan Lan, Stephen, Boon, and other extended members; and my friends. Their encouragement enabled me to pursue this research.

I thank also the Interdisciplinary Graduate School for their generous scholarship that funded this research.

Contents

Acknowledgements	ii
List of Figures	v
List of Tables	vi
Abstract	vii
1 Introduction	1
2 Network Analysis of Real-world Data	6
2.1 Network Basics	7
2.2 Extracting Networks from Data	8
2.2.1 Explicit Networks	10
2.2.2 Inferred Networks	10
2.3 Network Filtering	13
2.3.1 Threshold and Significance Filtering	13
2.3.2 Topological Filtering	14
2.3.3 Others	15
2.4 Quantitative Analysis	16
2.4.1 Random Graphs	16
2.4.2 Macroscopic Quantities	16
2.4.3 Mesoscopic Properties	21
2.4.4 Microscopic Quantities	22
3 Transition Networks of Classroom Teaching Practices	24
3.1 Introduction	24
3.2 Data	25
3.3 Methods	26
3.3.1 Transitions into First Activations	26
3.3.2 Measuring Significance	29
3.3.3 Network Visualization and Filtering	30
3.3.4 Dynamical Motifs	32
3.4 Results	32
3.4.1 Prevalent Transitions in Networks of Teaching Practices	32
3.4.2 Network Properties	35
3.4.3 Motifs of Teaching Practices	36

3.5	Discussion	37
3.6	Chapter Conclusions	39
	Appendix	42
4	Networks of ‘Collective Creativity’ Skills During Problem-Solving	45
4.1	Introduction	45
4.2	Collective Creativity in Problem Solving	46
4.3	Data	47
4.4	Methods	48
4.4.1	Network Inference from Dialogue Data	49
4.4.2	Minimum Preserved-Reachability Graph	51
4.4.3	Betweenness and Functional Betweenness	52
4.5	Results and Discussion	54
4.5.1	Transition Networks of CC Dialogue Types	54
4.5.2	Communities and Betweenness of CC Dialogue Types	56
4.5.3	Comparison with Time-Integrated Analysis	59
4.5.4	Subgroups in Successful and Unsuccessful Teams	61
4.6	Chapter Discussions Conclusions	64
	Appendix	66
5	Functional Shortcuts in Language Co-occurrence Networks	72
5.1	Introduction	72
5.2	Syntactic Network of Languages	74
5.3	Methods	76
5.3.1	Motif Extraction Algorithm	76
5.3.2	Equal-Cost Null Model for Network Shortcuts	78
5.3.3	Null Rate of Pattern Statistics	78
5.4	Results	78
5.4.1	Motif Detection in Language Networks	78
5.4.2	Network Distances	82
5.4.3	Taxonomy of MEX Patterns	82
5.5	Discussion	84
5.6	Chapter Conclusions	86
	Appendix	87
6	Conclusions	91
	Bibliography	94

List of Figures

2.1	Examples of network types.	8
2.2	Bridges of Königsberg problem.	9
2.3	Structural and functional brain networks.	9
2.4	Pearson’s correlations of time series and MST of the DJIA.	12
2.5	Construction of the MST and examples of planar graphs.	15
2.6	Random rewiring procedure for producing small-world networks.	17
2.7	Effect of degree distribution on network topology.	19
2.8	The 13 possible 3-node motifs.	22
3.1	Method of counting transitions.	27
3.2	Comparison of Markov probabilities and colour map of correlations.	28
3.3	Unfiltered network of significant transitions.	31
3.4	Time-resolved transition chart.	33
3.5	Transition networks of teaching practices.	34
3.6	Strength and degree distributions.	36
3.7	Common lesson motifs.	41
4.1	Conversion of line-based coding data to turn-based.	49
4.2	Simulated $\log \widehat{RR}$ distribution.	51
4.3	Minimum preserved-reachability graph filtering.	52
4.4	Illustration of functional betweenness.	54
4.5	MPRGs of successful and unsuccessful groups.	55
4.6	Unfiltered neighbourhoods.	56
4.7	Scatter plot of the differences in spontaneity s and generativity g	59
4.8	Normalized production rates of dialogue types.	61
4.9	Hierarchical clustering dendrograms.	62
4.10	Histogram of dialogue durations.	62
4.11	Overlapped filtered transition graphs of S-1 and S-2 subgroups.	63
4.12	Overlapped filtered transition graphs of S-1 and U-1 subgroups.	63
5.1	Toy co-occurrence network and pseudograph.	75
5.2	Results of MEX embedding.	80
5.3	Cumulative distributions of degree and selectivity.	81

List of Tables

3.1	Monte Carlo sampling example.	30
3.2	Network response to variations in significance cut-off.	32
3.3	Centrality rankings.	35
4.1	Dimensions, components, and categories of CC dialogue types.	47
4.2	Breakdown of CPS outcomes.	48
4.3	Network response to changes in $p_{threshold}$	51
4.4	Betweenness scores of successful and unsuccessful teams.	58
4.5	Spontaneity and generativity scores.	60
5.1	Table of network measures.	79
5.2	Motif templates of the USEC.	83

Abstract

Complex systems are systems with many interacting components that give rise to their rich dynamical behaviours. Complex networks provide a powerful and versatile perspective on complex systems, allowing us to define aspects such as hierarchy, flows, communities, and causality in terms of network topologies. In this thesis, we adopt such techniques of network analysis to interpret, represent, and reveal the internal organizations of three real-world complex data sets. Specifically, two examples used networks to represent and understand the dynamical processes of, respectively, teaching and problem-solving, and the third took a network approach to elucidate the syntactic organization of human language. In these examples, we use both existing and original techniques to infer networks from data, to filter dense networks into their most essential components, and to describe topological features of the network that correspond to the dynamics of their underlying systems. We show how network visualizations allow for an intuitive interpretation of complex dynamical data and the straightforward identification of generative drivers of dynamical sequences. Beyond visual interpretations, quantitative descriptions of the networks such as node betweenness and motif densities also reveal the internal logic of the systems.

Chapter 1

Introduction

The most interesting scientific, engineering, and social challenges that we face today are complex. Systems as diverse as the collective behavior of insect colonies [1], the syntax and semantic structure of human language [2], and the impact of policies in response to climate change [3] have all been bestowed upon this label of ‘complexity’. The overriding challenge in understanding these problems is that they contain a large number of interacting components. Exacerbating this difficulty is that one cannot dismantle these systems into their components, analyze the components separately, and then from the individual findings expect to gain deep insight into the systems as wholes. What we have just described is the reductionist approach to science that has been dominant since the 1600s. René Descartes in 1637 wrote in his *Discourse on the Method* that he aims “to divide all the difficulties under examination into as many parts as possible, and as many as were required to solve them in the best way” and “to conduct [his] thoughts in a given order, beginning with the simplest and most easily understood objects, and gradually ascending, as it were step by step, to the knowledge of the most complex” [4]. Reductionism has greatly shaped the scientific method since then and has enjoyed enormous successes in the last centuries to explain, among other things, the motions of the heavenly bodies and the atomic structures of matter.

However, it is increasingly clear to scientists that dealing with complex problems demand holistic system-level approaches. A canonical example where this approach is mandatory is in explaining the behaviour of insect colonies: A single ant is an extremely unsophisticated creature [5] with rudimentary survival instincts to seek food, fight enemies, reproduce, etc. However, an ant colony, through the interaction of simple individual actions between tens of thousands to millions of ants [6], is somehow capable of building fantastically complicated nests up to a million times their body size [7] with

features like advanced climate-control mechanisms [8]. All these are accomplished without any centralized command and control and *emerge* from the collective interactions at the system-level.

Today, we study such *complex systems* by adopting a holistic paradigm and recognize that in such systems, their dynamics are dictated by their web of interactions. Complexity has been discussed as early as 1948 when Warren Weaver distinguished between what he considered the three types of scientific problems: problems of simplicity, problems of *disorganized* complexity, and problems of *organized* complexity [9]. In his view, problems of ‘simplicity’ deals with systems with only a few variables (usually two). In the physical sciences, ‘simple’ problems like motion and gravity are the ones we first learnt to quantify with the tools of Newtonian mechanics and classical field theories. *Disorganized* complexity refers to systems with large numbers of parts and where the interaction between parts are more or less random (i.e. disorganized). He considered such problems to be mostly solved with 20th-century statistical mechanics that has provided us with remarkably accurate predictions of the ensemble behaviour of such systems. Problems of *organized* complexity, as the name suggests, contain also large numbers of interacting parts but are not arranged in complete randomness as to allow for a trivial statistical description. These systems were, and still are, not very well understood.

Despite the decades of efforts in investigating them, there is still no universally accepted definition of a complex system [10, 11]. It is generally agreed upon that a complex system is a large interacting network of individually simple components without central control [10, 11]. Such interactions often give rise to *emergent* [10, 11] and hard-to-predict behaviour like the collective intelligence of ant colonies we discussed earlier. Many complex systems, like the stock market [12] and, again, insect colonies [10], also possess elements of adaptation through learning or evolution. Some call such systems complex *adaptive* systems [10, 11, 13]. (The climate, though, is an example of a complex but non-adaptive system [10].) Another fairly common property is that complex systems can be history-dependent (or contingent) [11] in that prior states may have disproportionate magnitudes of influence on present states due to mechanisms like hysteresis or positive feedback. Other common but perhaps not universal properties of complex systems are their openness to influence by events outside the system and that they seem to switch between ordered and disordered states at will (both [11]).

In this thesis we are most interested in the network view of complex systems — it being the most widely accepted and versatile for analysis. Whereas traditional reductionism sought to simplify systems into isolated components with a ‘divide and conquer’ strategy, the network approach tends to reduce the components and interactions in a complex system into a collection of nodes and links. This simplified picture retains the overall

organization of the system to take what Murray Gell-Mann calls a “crude look at the whole” [14]. It is a powerful and versatile tool that allows us to define many aspects of a complex system such as hierarchy, flows, communities, and causality in terms of network topologies. In smaller systems, just plotting out the networks and examining them by eye can already give us a powerful overview of the structures that exist within the systems. However, beyond such crude ‘eyeball analysis’, a myriad of ways has been developed to quantify and characterize network topologies and describe with precision how the system is organized.

With this realization, we saw in the past two-and-a-half decades a boom in research attempting to understand networks of real-world systems quantitatively. The earliest works, particularly those done by Watts and Strogatz [15], Barabási and Albert [16], and Newman [17, 18], featured networks that are prominent in our daily experience. They included the power grid network, collaboration networks (in acting and in academia), the World Wide Web (WWW), etc. These early papers found that the topologies in empirical networks are markedly different from randomized networks but are at the same time similar to each other. In particular, many are found to have small-world [15] and scale-free [16] properties. We see in these peculiar topologies a validation of Weaver’s organized complexity wherein ‘organization’ takes the form of these non-trivial network topologies. More than mere mathematical features, these network properties have very real implications in the dynamics of the systems they represent. For example, the small-worldness of the Internet suggests efficiency in that data never needs to pass through many nodes to reach the entire network. However, the scale-free network property, also seen in the Internet, means that while it is robust against random failures it is also vulnerable against targeted attacks [19].

Then, there are also not-so-obvious networks extracted from systems in which the network structures are not explicitly defined. One such example is Mantegna’s pioneering work in extracting networks of influence in the financial markets [20]. Financial markets are highly complex systems where the movement of asset prices (stocks in the example cited here) are determined by a vast number of factors which we can think of as different layers in a multi-layered network of interactions. We can postulate that among these layers of influence, the way price movements are coupled with each other plays a significant role. However, unlike the power grid, collaboration, and WWW networks, this influence network cannot be explicitly observed in the financial market (we espouse more on this concept in Chapter 2). Therefore, Mantegna turned to examining the data output by financial market systems, namely the time series of stock prices, and used the correlations in their fluctuations to reveal the hierarchical structures in the ‘hidden’ networks of influence. From his networks, he established that stock price movements are clustered according to industry sectors. While this logical discovery may not be

unexpected, it nevertheless serves as evidence that meaningful interaction networks can be viably extracted from data.

We focus, in this thesis, on the application of network techniques to analyze empirical data of complex systems. In many complex systems, like the aforementioned financial markets, their networks of interactions cannot be directly observed. The alternative then is to probe the systems by measuring a large number of variables (i.e. the system's outputs) and then examine how these variable interact to *infer* interaction networks. These relationships may be directed if we are interested in transitions, causation, or activations in the system dynamics or undirected when we examine similarities or co-occurrences. Such networks of variable dependencies represent crude reconstructions of (or part of, depending on what and how many variables are measured) the underlying complex systems. We believe that there remains much potential in this inferential approach to extract the relationships in such 'hidden-networked' systems and to make these systems accessible to the tools of network analysis. Thus, in this thesis we aim to address the problem of gleaning useful insights using this approach on several complex systems. Specifically, we direct our attention to two examples of using networks to represent and understand the dynamical processes of, respectively, teaching and problem-solving; and an example of taking a network approach to elucidate the syntactic organization of human language.

The thesis is organized into 6 chapters with this being the first.

In Chapter 2, we present a review of the literature and essential tools of network analysis. We discuss about the methods used to infer networks from data, the filtering techniques that simplify networks, and the various quantitative ways of characterizing network topologies.

Chapter 3 details how we used directed networks to represent and understand the dynamical processes of teaching and learning. The hypothesis that underlies this project is that there is a web of dependencies in how different teaching and learning practices lead to one another that can be empirically discovered from data. To do this, we developed methods to infer significant transitions between practices relying on the principle of 'first-activations' and techniques to discard spurious transitions in the network by using a first-order null Markov model. To these networks, we applied topological filtering algorithms to reveal the main transition pathways of the systems and also extracted motifs of higher-order (3 or more variables) transitions.

Chapter 4 continues with the theme of Chapter 3 in creating a suite of tools to understand dynamical processes from a network perspective. It too deals with data from the realm of

education. In this chapter we inferred networks from the temporal production of expert-defined ‘collective creativity’ dialogue categories by students when they solve problems collaboratively. We believe that successful problem-solving is a process of purposefully using the relevant strategies, knowledge, and skills to understand and conceptualize the problem. Thus, the dialogue that accompanies this process would not proceed in a helter-skelter (i.e. random) manner but instead have a logic in its sequences which we depict as directed networks. Since the data here is differentiated between the successful and unsuccessful problem-solving groups, we were able to perform direct network-to-network comparisons of the successful problem-solving system and its unsuccessful counterpart. Here, we expanded on the techniques developed in Chapter 3. In particular, we modified the network inference methodologies as necessitated by the subtle differences between the problems and introduced the minimum preserved reachability graph — an novel graph filtering method that preserves reachability in filtered directed graphs. We also developed a measure of *functional* betweenness for network nodes in this chapter which helps identify important (i.e. central) variables in a dynamical system. It is possible to retroactively apply the methods developed in this chapter to the study in Chapter 3.

In Chapter 5, we consider the problem of how we can examine the syntactic organization in natural language by using unsupervised network analysis methods. We demonstrated that word co-occurrence networks are typified by the presence of coherent motifs which can be loosely interpreted as coherent patterns of language use. Such motifs appear much more frequently in empirical networks than in equivalent randomized networks. We then showed that motifs function as efficient and effective shortcuts in language networks, potentially explaining why we are able to generate and decipher language expressions so rapidly. Finally, we explored the possible link between these network motifs and constructions in Construction Grammar and speculated on the mechanisms behind the emergence of constructions in the early stages of language acquisition. The analysis performed in this chapter is quite different from those in Chapters 3 and 4 (for instance, the co-occurrence network of words is explicitly defined in the data and does not require inferential methods for its construction). However, the methods of identifying and characterizing patterns in a dynamical process (i.e. the production of utterances) described in this chapter expands our toolbox of network methodologies to describe even more complex and hierarchical dynamical systems.

Chapter 6 concludes our contributions.

Chapter 2

Network Analysis of Real-world Data

In recent years, it has become popular to use networks to visualize and analyze the structures within large data sets obtained from complex systems. Networks are especially useful for representing the extent and magnitude of pair-wise interactions among the many components in such systems. Network-based interpretation of data has found success in this non-exhaustive list of examples: (i) in finance, where the correlation networks of dynamical quantities like stock returns and exchange rates reveal the topology of interaction among entities such as firms [20, 21] and currencies [22]; (ii) in sociology, where the analysis of social networks reveal that people interact within preferred communities [23, 24] and that conflict can cause the network to split along community-defined borders [25]; (iii) in neuroscience, in which we identified important hubs [26] and signal-regulating motifs [27] in brain networks; (iv) in infrastructure, where the realization that the scale-free network architecture of the Internet [28] and possibly that of the air transportation network [29] could mean that they are robust against failure but vulnerable against targeted attacks [19]; and (v) in visualizing dynamical processes in, for instance, gene regulation [30], psychometry [31], and even history [32].

In this chapter, we review and discuss the important developments and methodologies for network analysis on real-world systems. In view of the explosion in network research in the past two decades, it will be over-ambitious to attempt to list every published technique. Instead, we concentrate on the most established ones to provide the reader with a broad overview of the field. We begin by briefly defining what is a network along with some common terminologies used to describe them. Next, we list and review the techniques that have been employed to convert empirical data into networks. We consider that networks can either be explicitly extracted from data or have to be indirectly

inferred from it. We follow with a discussion on how networks are filtered to aid analysis. Finally, we review the quantitative measures used to characterize network topologies. We classify these measures into those describing the macroscopic, mesoscopic, and microscopic topologies of networks. Some other significant aspects of network science, like the growth models of complex networks and the modelling of dynamics mediated by such networks, are outside the scope of this thesis. For more thorough treatments of these topics and other methods of network analysis, the reader is encouraged to refer to a number of outstanding reviews and books including, but not limited to: Amaral *et al.* [33], Newman [34, 35], Barabási [36], Watts [37], Boccatelli *et al.* [38], and Estrada [39].

2.1 Network Basics

In the mathematical sense, a graph G , also called a network, is a set of vertices and edges

$$G = (V, E) \quad (2.1)$$

which is used to model pairwise relations between objects. Vertices are also referred to as nodes and edges can be termed links, connections, bonds, or ties. In the reminder of this thesis we use these terms interchangeably. A few simple networks are drawn in Fig. 2.1 for reference. The simplest graph is a set of nodes joined by *unweighted* and *undirected* edges. An unweighted edge $e = (x, y)$ joining nodes x and y signifies some relationship between the nodes without specifying any magnitude or other properties of this relationship. It being an undirected edge means there is no directionality in the relationship between x and y . More complicated graphs can involve either weighted edges, directed edges, or both. For example, the lending network of banks is best represented with a weighted (to show the loan amounts) and directed (to identify the lenders/borrowers) network. Graphs can be *cyclic* when there is a path originating from a node which eventually joins back to itself either in the directed or undirected sense. This is possible in the lending network. Citation networks are however *acyclic* since one can only cite works done in the past. Graphs can also be fully connected which means that every node is connected to each other (either in the directed or undirected sense). Excluding self linkages, such networks contain $N(N-1)/2$ (N being the number of nodes) edges in the case of undirected graphs and have $N(N-1)$ edges otherwise. Other exotic network types have also been proposed — notably the multigraph [40] where parallel edges between nodes are allowed and the hypergraph [40] in which edges can join more than 2 nodes.

FIGURE 2.1: Examples of network types. **(A)** shows an undirected and unweighted network with 2 circuits — a 3-node circuit and a 4-node circuit. **(B)** shows an undirected and weighted network with thicker edges representing high edge weights. **(C)** shows an unweighted but directed graph. It contains one 3-node cycle. The 4-node enclosed structure is not a cycle because the edges do not point in the same direction. Note also the reciprocated edges (i.e. pointing in both directions) in the top right of the graph. **(D)** shows a weighted and directed graph.

2.2 Extracting Networks from Data

The first instance of solving an empirical problem with a network approach came in 1736 when Euler addressed the problem of crossing the seven bridges of Königsberg in a single non-looping path. His negative resolution of it [41] laid the foundation of modern graph theory. In formulating the problem, Euler converted cartographic data into a simplified picture of a network by replacing each land mass with a node while bridges that joined the land masses are represented as edges (see Fig. 2.2).

Subsequently, it has become popular in several fields to condense what is often lengthy and hard-to-visualize data into a more digestible network. This is especially driven by the growing availability of huge data sets and the computational power to process them. In many instances, the conversion from data to network is a relatively simple process often guided by a physical intuition of how the data maps into a network representation

FIGURE 2.2: Bridges of Königsberg problem. The cartographic details of the map is discarded and simplified into a multigraph with edge(s) between connected landmasses. Images by users Bogdan Giuscă and Booyabazooka obtained from https://en.wikipedia.org/wiki/Seven_Bridges_of_Königsberg.

i.e. the connection in these networks are *explicitly* specified in the data. In other cases, the network structure is not immediately obvious from the data and have to be *inferred* with indirect methods. Bullmore and Sporns similarly illustrates the distinction between explicit and inferred networks in their treatment of brain networks [42, 43]. They consider that the *structural* brain network is explicitly constructed by charting the anatomical connections in the synapses and that the *functional* network is inferred indirectly with correlation or causality models (see Fig. 2.3).

FIGURE 2.3: Structural and functional brain networks. Reproduced from [42] with permission from Nature Publishing Group.

2.2.1 Explicit Networks

In some systems, especially in physically networked systems, converting data into networks is straightforward. They include the power grid network [44], the network of airline routes [29, 33], the rail network [45], and neural networks [46]. The conversion is easy because the nodes (generator/transformers/substations, airports, rail stations, and neural cells) and edges joining nodes (power lines, flight routes, rail tracks, and axons) correspond with our mental picture of networks.

Explicit networks are also extracted from data where the relationships between entities are unambiguously declared. In sociology, the prototypical example is a network of declared friendships among people [47]. Building such networks used to be slow, laborious, and often produces rather small graphs. However, with the growing use of online social networking platforms in the modern era, extracting explicit networks of social connections has become an exercise of scanning through lists of online ‘friends’ [48]. The directed network of scientific citations [49] is also explicit since the bibliography in each work leaves no ambiguity if a directed connection exists between papers. The network of Internet pages connected via hyperlinks [50] also constitutes such an explicit network.

Another method of defining explicit networks is through the direct observation of interaction between the components of systems. For example, a directed connection exists in the food web we if know that one species consumes another [51]. (Although, there may sometimes be uncertainties if a species only consumes another unintentionally or as a last resort.)

Lastly, affiliation networks are also explicit. Early examples of social affiliation networks linked two persons if both were present at certain social events [25, 52]. More recent affiliation networks dealt with scientific collaboration connecting individuals who appeared in the same publication [53] and with film collaboration connecting actors (or actresses) who worked on the same film [33]. The data used to extract such relationships contain no ambiguity — Miss Johnson and Mrs. Harris either both attended the party on November 20th or they did not. Thus we consider such networks explicit.

2.2.2 Inferred Networks

One turns to indirect methods of inferring networks from the data of complex systems when it is not possible to directly observe the interactions among their components. This is true in many systems such as in the equity markets, foreign exchange, brain, and genomic interactions where the components are more or less already identified but the

networks of interactions and influences remain hidden. Fortunately, the increasing availability of high-throughput data in these domains have enabled scientists to confidently mine for correlative and causative relationships in these systems that help to define the structures of their corresponding networks.

Undirected Inferred Networks

The construction of undirected financial networks constitutes some of the earliest works using this concept of inference. Particularly, the pioneering paper by Mantegna in 1999 [20] used Pearson's correlation [54]

$$\rho_{i,j} = \frac{\langle Y_i Y_j \rangle - \langle Y_i \rangle \langle Y_j \rangle}{\sqrt{(\langle Y_i^2 \rangle - \langle Y_i \rangle^2)(\langle Y_j^2 \rangle - \langle Y_j \rangle^2)}} \quad (2.2)$$

to measure the similarity in fluctuations in the log returns Y_i and Y_j of a pair of stocks i and j and used the similarities between all pairs to define the system's underlying network. Fig. 2.4A gives an intuitive picture of how the correlation works and Fig. 2.4B shows one of the networks constructed by Mantegna. From this picture, he was able to infer the structures of influence in the equity markets and conclude that price movements are similar within industry sectors as we mentioned already in Chapter 1.

Following Mantegna's lead, a number of other authors have went on to produce other types of financial networks using correlations. Notables ones include: Onnela *et al.* who showed that correlation networks shrink during market crisis [55] and Bonanno *et al.* who varied the time horizons in computing $\rho_{i,j}$ and demonstrated that the networks changes non-trivially (i.e. inter-sector correlations decay faster than intra-sector correlations) under such variations [56]. The latter work is a reminder that selecting appropriate time horizons in correlation analysis requires the careful considerations of the timescales of the relationships we want to investigate.

To capture nonlinear (but still monotonic) correlations between variables, one can use Spearman's correlation [57]. It is mostly used to define connections in gene co-expression networks [58, 59]. Less common are information theoretic quantities, such as mutual information (MI), that define the non-linear interdependence between variables. Network inference with MI stretches back to 1998 when Liang *et al.* proposed to reverse engineer functional genetic network architectures using mutual information [60].

Despite the myriad of ways one can use to calculate similarities to define network edges, it is important to point out that in many cases their differences are academic and the methodologies in fact return identical, or very similar, networks [61].

FIGURE 2.4: Pearson's correlations of time series and MST of the DJIA. In (A) we demonstrate how Pearson's correlation captures synchronous fluctuations in time series. The column of scatter plots are the parametric plots of the X and Y time series. The absolute correlations are high if the parametric scatter plots fall in a straight line. (B) is the minimum spanning tree of the Dow Jones Industrial Average. The distances between nodes is set at $d_{i,j} = \sqrt{2(1 - \rho_{i,j})}$ i.e. highly correlated stocks are placed close together. Panel (B) is reproduced from [20] with permission from Springer.

Directed Inferred Networks

In inferring directed networks, we are frequently interested in the causative effects of one variable on the other (while keeping in mind that correlation does not always imply causation). Most causality models make use of the simple fact that cause precedes effect. For example, it is easy to imagine how Pearson's correlation can be slightly modified to incorporate temporal directionality: We can compute $\rho(Y_{i,t}, Y_{j,t+h})$ which is the correlation of the time series of variable i with that of variable j translated forward

by h time units. If there is a significant correlation, it means that the fluctuations of i in the future reflects the movement of j in the past. Such analysis has been applied to financial data of stock returns [62] and foreign exchange rates [63]. Any strong edges in such networks could represent an opportunity to exploit market inefficiencies. Unfortunately, [62] found that the characteristic time in which there are significant lead-lag correlations was only in the order of minutes, indicating that the markets were indeed quite efficient.

Information theoretic measures also have their time-directed versions in Granger causality [64] and transfer entropy [65] (which is a generalization of the former to non-linear signals). The heaviest usage of such measures come in deducing the effective brain connectivity networks [66–69]. More elaborate models to determine causation in networks, like the Dynamic Causal Model [70], have also been suggested.

2.3 Network Filtering

The human eye has proved to be a powerful analytical tool since the beginnings of network studies [34]. This, of course, is provided that the networks under study are not too big (usually $N \leq 10^2$) and not too dense. However, this is not the case in many networks and especially in inferred networks which are typically fully connected and practically intractable by eye. Luckily, it is not usually necessary to retain all these edges in the network for a large number of them are lowly weighted edges representing vanishingly weak connections. In network filtering, the basic idea is to discard the less important information (usually edges) of the network and to retain its most crucial structures. Network filtering yields more benefits than simple ‘eyeball’ analysis of networks. For instance, clustering in densely connected graphs is a considerable computational challenge [71] and pre-filtering a dense network into a sparse one accelerates the algorithm. Here, we consider threshold and significance filtering methods, topological filtering methods, and also briefly discuss a few others.

2.3.1 Threshold and Significance Filtering

One of the most obvious solutions in reducing the number of edges in weighted networks is to introduce a weight threshold and eliminate the edges below this threshold [72, 73]. The advantage of this approach is its simplicity and applicability to both directed and undirected graphs. However, imposing a global threshold introduces a characteristic scale [74] in real networks that are often scale-free [16]. It can potentially discard edges important at local scales but nevertheless below the artificial threshold we set. Moreover,

this procedure may fragment the network into unconnected components [73] which could either be a desired or undesired outcome. Other variants of threshold filtering methods, with similar advantages and drawbacks, include retaining only a certain number of the most heavily weighted edges [21] or removing weak edges until the network achieves a selected edge density [75].

Beyond plain threshold filtering, one can also compute the significance of edge weights and filter out those that fail the significance test. In attempting to extract a syntactic network of words, Ferrer i Cancho and Solé [76] began with an explicitly defined co-occurrence network where words adjacent in the sentences of a corpus are joined. To determine if their adjacency is significant, they considered if $p_{ij} > p_i p_j$ which means that the observed probability of words i and j being adjacent is greater than the product of their independent probabilities. The failure to satisfy this condition means that the co-occurrence connection between those words is probably spurious and should be dropped from the syntactic network. This is an interesting example where a network of less well-defined relationship (i.e. syntax) is distilled from an explicitly defined (co-occurrence) network.

2.3.2 Topological Filtering

Topological filtering methods remove the less important connections of the network while respecting certain topological constraints. The *minimum spanning tree* (MST) and the *planar maximally filtered graphs* (PMFG) are the most widely used. Both guarantee the connectedness of undirected graphs after filtering. Both techniques can be applied to directed graphs by ignoring their directionality but in such cases, there is no guarantee that the reachability of the original network is preserved. In Chapter 4, we propose an alternate filtering algorithm that preserves reachability in directed graphs.

The MST of a network can be extracted by following Kruskal’s algorithm [77] (see Fig. 2.5A). The resulting network retains only $N - 1$ edges and excludes any loops. This could be problematic in systems where loops indicate important feedback-type mechanisms. That being said, MST filtering has been employed successfully in many applications. Notably, from the MST-filtered network of stocks (see Fig. 2.4) in the earlier mentioned paper by Mantegna [20], it is easy to perceive the industrial-sector-wise hierarchies by eye. The PMFG, proposed by Tumminello *et al.* [78], keeps more edges ($3N - 2$) and information than the MST. It extracts a subgraph from the original network which can, in principle, be drawn on a plane without any crossing of edges (see Fig. 2.5B-D). To extract the PMFG, we can modify the condition in step (iii) of Kruskal’s algorithm to allow the addition of an edge if it does not violate the graph’s planarity.

FIGURE 2.5: Construction of the MST and examples of planar graphs. The solid lines in (A) show a MST built with Kruskal's algorithm from a decreasing order of weights. The algorithm is as follows: (i) Begin the MST with a unconnected network of only the nodes of the original network without its edges. (ii) Rank the edges in order of increasing or decreasing weight, according to the network desired. (iii) Starting from the edge with the largest (or smallest) weight, add it to the MST provided that the edge added does not produce a circuit (a loop) in the network. (iv) repeat (iii) until the MST is connected to all nodes. The edges represented by dashed lines with weights 8, 4, and 1 failed the condition in (iii) and are not admitted into the MST as they would have induced loops. (B), (C), and (D) shows the properties of a planar graph. Even though the edges in (B) intersect, the graph is planar. This is because the lines can easily be redrawn such that they do not intersect like in (C). However, if the new blue edge is added like to the graph in (D), there is no way to redraw the edges or rearrange the nodes such that the links never intersect. Therefore the graph in (D) is not planar. Reproduced from [79], a work published by the thesis author and collaborators.

2.3.3 Others

The *Disparity* filter [80] and the *Global Statistical Significance* (GloSS) filter form an interesting class of filtering techniques that considers both local and global information in deciding which edges to retain and discard. The gist in these ideas is to identify, for each node, edges incident on it that carry a disproportionately large share of the node's strength.

There are also filtering methods that can be applied on unweighted graphs. The transitive reduction [81] of a directed graph finds the sub-graph with the fewest possible edges while maintaining its *reachability*. In citation networks, the transitive reduction highlights most immediate causal links [82]. Another option of simplifying unweighted graphs is the k -core decomposition of a network [83] which iteratively filters away nodes with lower-than- k degrees.

Finally, instead of reducing the number of edges in the networks, coarse-graining achieves simplification through reducing the number of nodes in them [84]. Nodes can be merged based on their sharing of some property defined outside the network [84] or for belonging to certain network communities [85] or sub-graph arrangements [84].

2.4 Quantitative Analysis

Large networks of tens of thousands and up to millions of nodes and edges are common nowadays. These networks defy the ‘eyeball’ analysis, and some are even impossible to actually draw [34]. Instead, quantifiable measures have been developed to distill crucial information about a network’s topology without actually seeing it. Tens, if not hundreds, of such measures have been put forward. In this review we focus on a reduced scope of the most known ones and those relevant to the thesis.

2.4.1 Random Graphs

We prefix the discussion on quantitative network measures with a short introduction to random networks. Complex networks are termed so because they carry non-trivial features not seen in random networks. Random graphs are first studied by Rapoport [86] and then by Erdős and Rényi [87]. Algorithms to generate random graphs are very simple. Most begin with n number of nodes and then decide for all possible pairs (undirected or directed, the algorithms work either way) whether or not to join them with probability p , resulting in $G_{n,p}$. In the undirected case, it returns a network with $m \approx \frac{1}{2}pn(n-1)$ edges.

2.4.2 Macroscopic Quantities

Macroscopic features of a network are those describing its global characteristics. Here, we discuss about the path lengths, the degree distributions, and some other macroscopic quantities of networks and complex networks.

Paths Lengths

In 1967, Stanley Milgram conducted an ingenious experiment [88, 89] that is one of the first known attempts to estimate the path lengths in social networks. To several randomly selected individuals in the United States, he requested that they forward a letter to a target individual in Boston. If they did not personally know the target, as was likely to be the case, then they should forward the letter to an acquaintance whom they

think will have a better chance of knowing the target. Of the letters that reached the destination, Milgram found that the average number of times that they were forwarded was remarkably few ($l \approx 6$). This led to the popular saying some decades later that the world is connected by just “six degrees of separation” [90]. More importantly for us, it is the first experimental demonstration of what is now known as the *small-world effect*.

To quantify average path lengths in a network, we compute its mean geodesic (i.e. shortest) distance between node pairs

$$l = \frac{1}{\frac{1}{2}n(n+1)} \sum_{i>j} d_{ij} \quad (2.3)$$

where d_{ij} is the shortest distance from node i to j . Consider geometrically regular graphs where all nodes have the same number of r edges [91] (see left graph in Fig. 2.6). In such graphs, edges always only connect nodes locally. As we vary the network size n , we can expect that the average path length l in regular graphs to change linearly with n . At the other extreme, we consider a large random graph with average degree $\langle k \rangle$. From any vertex v , roughly $\langle k \rangle^l$ other vertices are l steps away from v . Since the number of reachable vertices increases exponentially, we can inverse the relationship and deduce that l scales with $\log n$. Logarithmic or slower scaling is a prerequisite to consider a network small-world. The other necessary condition is high transitivity which we discuss later in the same subsection. This property of transitivity is not seen in random networks.

FIGURE 2.6: Random rewiring procedure for producing small-world networks. Left: A $r = 4$ regular graph with long characteristic path length and high transitivity. Middle: With the rewiring of a few edges, average path length l decreases drastically while transitivity remains high. This network is small-world. Right: With a majority of edges rewired the graph becomes a random graph with small l but has also small C . Reproduced from [15] with permission from Nature Publishing Group.

Logarithmic scaling in l has been detected in several real-world networks [92]. In fact, there is even suspicion that l in scale-free networks (see following section) scale even slower than $\log n$ [93, 94]. In any case, small-world effects have important ramifications in dynamical systems mediated by such networks. Small-world networks of the Internet [95, 96] and the brain [97] suggest that information is transmitted from source to target efficiently with few intermediaries. However, on a more worrying note, the same feature in air-transportation networks could mean that diseases have the potential to spread worldwide very quickly if we do not act in time [98].

Watts and Strogatz [15] devised a simple rewiring procedure to transform a regular network into a small-world one by introducing a small number of shortcuts or by randomly rewiring edges in the network (see Fig. 2.6). In doing so, they demonstrated that in a regular, or close-to-regular, network, every shortcut introduced has a highly nonlinear effect of shrinking l by a large amount.

Degree Distribution

The degree distributions of real-world networks have been the object of focus in many research papers in recent times. The degree k_i of a vertex i is simply the number of edges incident on it. For directed networks, one can compute the in-degree k_i^{in} and out-degree k_i^{out} separately. $p(k)$ gives the probability that a node in the network has degree k . In Section 2.4.1, we described the $G_{n,p}$ random graph which has a binomial degree distribution (see Fig. 2.7)

$$p(k) = \binom{n-1}{k} p^k (1-p)^{(n-1)-k}. \quad (2.4)$$

It approximates to the Poisson distribution if $n \gg 1$ and the normal distribution if $np \gg 1$. The random graph has a characteristic scale in the sense that the average degree is np . However, the degree distributions of many real networks deviate significantly from this model. In particular, many are seen to follow a power-law distribution with $p(k) \sim k^{-\gamma}$ [16, 28] while some others seem to decay exponentially $p(k) \sim e^{-k}$ [33, 45]. An example of a network with power-law degree distribution is given also in Fig. 2.7.

Networks with power-law degree distributions have come under special scrutiny [16, 34, 92]. They are called *scale-free* because $p(k)$ varies with a constant term under scaling of k i.e. $p(ak) = bp(k)$, $b = a^{-\gamma}$. In most cases the scaling exponent γ is in the range $2 < \gamma < 3$. The implications of power-law degree distributions in networks are much greater than mere mathematical curiosities. To begin, the ‘fat-tails’ in power-laws mean that nodes with very high degrees are much more common in such distributions. This is consistent with our experience that hubs are present in the networks we interact with

FIGURE 2.7: Effect of degree distribution on network topology. The top left plot depicts a bell curve (normal) distribution of network degrees. It is an approximation of the Poisson distribution of degrees in random networks when $np \gg 1$. A stylized random network is drawn next to the distribution. The plot at the bottom left depicts a power-law distribution of network degrees and alongside a (also stylized) scale-free network is shown. Scale-free networks have a high proportion of high-degree hubs (shaded in grey). Plots are reproduced from [36] with permission from Perseus Publishing. Networks by Carlos Castillo obtained from https://en.wikipedia.org/wiki/Scale-free_network.

day-to-day. For instance, we are familiar with hub airports that the air traffic network is organized around as well as search engine hubs when navigating web pages. Scale-free networks like the Internet are also characterized by high robustness in response to randomized failures [19] — the network remains connected even when a relatively high proportion of nodes are randomly removed/disabled in the network as most would have been just low-degree periphery nodes. Conversely, such networks are vulnerable when subjected to a targeted attack of hubs — the deletion/disabling of just a few hubs can potentially fragment a network or significantly increase its average path length l .

The best known mechanisms for generating scale-free networks are the *cumulative advantage* growth model by Price [99] and the *preferential attachment* growth model by Barabási and Albert [16]. Both operate on the assumption that new nodes in the network prefer to attach themselves to nodes of high degrees (thus the model names) as

the network grows.

Others

Watts and Strogatz, besides observing that many real-life networks have logarithmically-scaled path lengths, also discovered that they are much more *clustered* (or transitive, we prefer the latter term) than random graphs [15]. The combination of both properties makes these networks small-world. Transitivity $C = \frac{3N_\Delta}{N_3}$, first defined in sociology, is the density of triangles in the network and is obtained by taking the ratio of the frequency of triads (3-node circuits) N_Δ to triples (connected 3-node sub-graphs, for example, a 3-node chain) N_3 [100]. In social networks, it is not difficult to deduce how the triangular sub-graph is a prominent structure. Often, the friend of your friend is probably also your friend — this translates to a 3-vertex circuit in the social network. Besides measuring the global transitivity C , one can also measure the local transitivity C_i and examine the relationship between it and the node's degree k_i . It was found that $C(k)$ scales with $k^{-\beta}$ with the exponent lying between 0.5 and 1 in some networks [101], indicating the presence of *hierarchical organizations*.

Assortativity measures whether there is a tendency for nodes tend to connect to others of the same type. The similarity of nodes can be defined in a number of ways. For instance, degree assortativity is the tendency of nodes to connect to others of similar degrees. This is quantified by measuring the Pearson's correlation of the degrees of linked pairs [102]. Newman showed that social networks tend to be assortative and technological networks, like the Internet, tend to be disassortative [102]. Assortativity in social networks should come as no surprise since people prefer to associate with others of similar age, race, language, etc. [103].

Finally, *modularity* is a quantity that measures how strongly are networks divided into distinct modules [104]. Given that nodes assigned with classification c (which can be interpreted as a module ID), one can compute the modularity

$$Q = \frac{1}{m} \sum_{j>i} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta(c_i, c_j) \quad (2.5)$$

where the $\delta(c_i, c_j)$ is the Kronecker delta. It returns the difference in the fraction of edges connecting nodes with identical c (i.e. of the same module) and that of the expected fraction (i.e. if the nodes were randomly connected). Highly modular organizations discovered in many social networks [105] suggest that people do not just associate with others of similar traits but also form closely knitted communities in their preferred groupings.

2.4.3 Mesoscopic Properties

We consider the mesoscopic properties of networks to be those at the scale involving a number of vertices in the graph.

Network Communities

Many networks are organized around communities. For instance, it is intuitive that social networks are organized around groups of people with similar interests/traits [103]. Community detection reveals such important structures in network analysis. One of the earliest example is work done by Zachary (mentioned earlier) who algorithmically found two distinct communities in the social network of a karate club which he investigated [25]. Moreover, the predicted communities matched almost exactly to the two factions the club split into after a dispute.

In unweighted networks, one of the modern methods of community detection is through *modularity maximization*. Whereas in Eq. 2.5 we computed Q based on a given set of c , in modularity maximization the aim is to find a set of c such that Q reaches a maximum value. Girvan and Newman proposed a method to arrive at this optimal separation without having to test all combinatorial possibilities of c through the iterative removal of edges with high-betweenness (see Section 2.4.4) [104]. Many newer algorithms exist too [106–108] with some being able to detect overlapping communities [85, 109, 110]. Meaningful communities have been discovered in co-authorship networks [53, 109] and collaboration networks [24] among many others.

In fully connected weighted networks, the preferred community detection technique is *hierarchical clustering* [111]. In this technique, we build dendrograms by iteratively joining vertex pairs with the highest edge weights. The variants of hierarchical clustering methods differ mainly in linkage criteria i.e. how weights (or ‘distance’) are measured from a cluster to another cluster (or vertex). Popular ones include using the largest distance between the two entities (complete-linkage), the average distance (average-linkage), and the smallest distance (single-linkage). The communities are determined by selecting an appropriate threshold to ‘cut’ the dendrograms. Mantegna in [20] used single-linkage hierarchical clustering to discern distinct groups of stocks in the DJIA.

Network Motifs

Milo *et al.* [27] proposed a highly original method of investigating the mesoscopic properties in networks by looking at *network motifs*. Motifs are small (usually 3 or 4 nodes)

directed sub-graphs that appear with higher-than-expected frequencies in real-world networks. Their paper considered mainly the roles of the 13 possible 3-node sub-graphs (see Fig. 2.8). If a certain sub-graph appears significantly more often in a network than predicted from its equivalent random graph model, the sub-graph is considered a network motif. They showed that transcription networks and neuronal networks have high incidences of the so-called ‘feed-forward loop’ motif. A possible function of this motif in signalling networks is to guard against false signals. Furthermore, Sporn and Kötter showed that the over and under-representation of certain motif types in brain networks is universal across species [112], reinforcing the important functions of these sub-structures. Motif analysis has also been extended to weighted networks in [113].

FIGURE 2.8: The 13 possible 3-node motifs. We highlight also the ‘feed-forward loop’. Reproduced from [27] with permission from AAAS.

2.4.4 Microscopic Quantities

We consider the centrality measures that describe a vertex’s importance in the network the microscopic network quantities.

Degree, Strength, and Eigenvector Centralities

The conceptually simplest manner to determine the centrality of nodes in unweighted networks is to consider its *degree centrality* — the number of edges incident on the nodes. The global hubs of the networks can easily be picked out from vertices with the highest degrees (e.g. in air transportation networks [29]). A vertex of high degree centrality has many neighbours it interacts with and therefore it can influence, or be influenced by, several other components in its system. In directed networks, the in-degrees and out-degrees are often considered separately. Vertices with high in-degrees can be considered sinks of flow in the network and those with high out-degrees are sources of flow.

In weighted networks, the notion of degree centrality can be simply extended to consider the sum of weights incident on the nodes. This is known as *strength centrality*. In networks where the edge weights measure flow, strength centrality is an indicator of the

flow capacity of the node. For example, in [114], the strength centrality of the weighted airport network reflects its yearly passenger capacity.

Eigenvector centrality [115] is a form of higher-ordered degree centrality. Rather than merely counting how many neighbours a node has, it assigns weights to these neighbours according to their own centralities. A node with high eigenvector centrality is linked to many others whom are themselves highly linked. The leading eigenvector of a network's adjacency matrix gives the centrality values.

Betweenness Centrality

In many systems, networks are mediators of flows of information, energy, people, etc. *Betweenness centrality* makes the assumption that the shortest paths are always selected in network navigation. Nodes lying on a large number of such paths are considered to have high betweenness [116]. (In some literature 'centrality' refers by default to betweenness centrality. We prefer to make the distinction.) Betweenness centrality can pick out local hubs/gateways of access in networks missed by degree centrality [29].

Chapter 3

Transition Networks of Classroom Teaching Practices

3.1 Introduction

In 2010, the National Institute of Education, Singapore, (NIE) conducted a large-scale study of teaching practices in Singapore classrooms drawing on a representative sample of schools. The study had the goal of investigating how teachers teach in Singapore, why they teach the way they do, and the effects of their teaching on student learning in order to map and model the logic of pedagogical instruction. The data produced from this study is extremely high-dimensional (one dimension for each type of teaching practice investigated) with greatly varying prevalences — some practices are present up to 70% of the time in the time-segmented data while others appear only 0.001% of the time. Using commercial software like SPSS to analyse this large dataset for empirical sequences of pedagogical instruction proved to be computationally intensive as it explores the combinatorially large space of variables to find statistically significant transitions. In general, such complex temporal analysis of classroom pedagogy remains a major methodological challenge for educational research [117].

In this chapter, we demonstrate the viability of using network inference and filtering methods to extract and represent the real-life processes of teaching and learning that was captured in the data. Our aims are to (i) construct complex networks that typify activity transitions in the classroom, (ii) identify the prevalent pairwise transitions and important teaching practices, and (iii) extract sequences of activity transitions that frequently occur in the classroom. Our approach is both empirical and normative in that we focus, in the first instance, on establishing statistically weighted relationships between instructional practices and, in the second, because we assume that not all instructional

practices are equal and that from a normative perspective, there is a particular logic to instruction that teachers need to attend to if they are to optimise the intellectual quality of teaching and learning. As such, teaching and learning activities do not follow a random sequence but trace a purposeful logical path that teachers implement from their lesson plans, and simultaneously adapt in response to their monitoring of students' work. Therefore, the resulting network of transitions between teaching practices should contain detectable and meaningful topological structures that correspond to both effective and perhaps ineffective teaching progression. Our results show that teaching knowledge as truth and teacher-dominated talking serve as hubs for teaching practices in Singapore. These practices reflect a *transmissionist* model of teaching and learning. We also identify complex teacher-student-teacher-student interaction sequences of teaching practices that we considered as motifs.

The inference of significant transitions between variables (teaching practices) in this project is very much inspired by the works mentioned in Section 2.2.2 on the extraction of directed relationships from non-explicit data to construct networks. To the best of our knowledge, such network analysis has never been attempted on time-resolved pedagogical data. We also follow the lead of Mantegna [20], Tumminello *et al.* [78], and others in using topological filtering methods to filter the transition networks we constructed.

The work presented in this chapter was done in collaboration with members of the Office of Education Research, NIE, Dr. Dennis Kwek and Prof. David Hogan. They provided the time-resolved data of classroom teaching practices and the pedagogical interpretation of the results of our network analysis. A paper on this project was written and published in 2014 under [79]. Many passages in this chapter are reproduced verbatim from the paper.

3.2 Data

The data used in our analysis was obtained from the Core 2 Research Programme which is a comprehensive, large-scale, and multi-dimensional study on the state of instructional practices in Singapore classrooms conducted by NIE [118]. Specifically, the data used here is drawn from the 'observational and coding panel' of the programme which videotaped and collected data from English and Mathematics teachers in 15 Primary schools and 16 Secondary schools. Grades 5 (in Primary schools, also called Primary 5) and 9 (in Secondary schools, also called Secondary 3) were selected as these were considered to be years crucial to the development of skills and knowledge needed for the high stakes national examinations for Primary schools (the Singapore Primary School

Leaving Examination) and Secondary schools (the General Cambridge ‘O’ Level Examination). English and Mathematics were selected as key curriculum areas that have a significant influence on student social mobility; literacy and numeracy skills are often seen as key leverages to opening up educational and career pathways. In this work we focus on the analysis of the Mathematics lessons as a proof of concept.

Data was collected from April through November 2010 with 124 lessons in Primary 5 Mathematics and 171 lessons in Secondary 3 Mathematics. Teachers selected for observation were asked to nominate a unit of work – a full sequence of lessons around a particular topic, theme, or content area. The lessons are video and audio recorded to capture all whole class interactions and the majority of pair or group work assigned by teachers. Lessons are divided into three-minute ‘phases’, with an average one hour lesson having 20 phases. Trained subject specialists ‘code’ each phase in the lesson recordings for the presence (coded ‘1’) or absence (coded ‘0’) of some 500 observable teaching practices (we will sometimes refer to these teaching practices as ‘variables’) that are defined in the Singapore Pedagogy Coding Scheme 2 [119, 120]. Segmenting the lesson into phases allows for a temporal examination of instructional practices over the course of a lesson.

From these 500 variables, we selected 170 variables in the data for Primary 5 Mathematics and 167 variables for Secondary 3 which we believe to be most essential to the development of intellectual quality for our network analysis. The variables chosen extends current pedagogical research based on Hattie’s work on ‘Visible Learning’ which describes effective pedagogical practices [121], by focusing on the nature of students doing disciplinary knowledge work in the classroom: the epistemic focus of instructional tasks; the nature of knowledge practices (including the generation, representation, communication and justification of knowledge claims); the epistemic classroom talk that helps makes these knowledge claims explicit, transparent and visible to students; and the cognitive complexity of the knowledge work undertaken in instructional tasks, recognising the contested nature of knowledge claims [122]. The full lists of variables are too lengthy to be shown here and can be obtained, instead, from [119, 120].

3.3 Methods

3.3.1 Transitions into First Activations

The starting point of our analysis is to define what constitutes a transition from one variable to another and from this, develop a measure of transition frequency and significance. Previous work in network analysis tended to use measures like transfer entropy

[123, 124] or Pearson’s correlation [20] to denote the directed relationships between variable pairs (i.e. nodes). The major shortcoming in these measures when applied to this data set is that they do not consider the characteristic event durations of signals in the time series of the data. For instance, given the binary time signals of our data, a variable that is coded ‘1’ (i.e. active) continuously across several phases (i.e. time steps $t = t_0, t_0 + 1, t_0 + 2, \dots$) is assumed by these measures to represent distinct events at each of the steps. We argue that teaching practices that are active over consecutive phases are more likely to be singular events that happen to last longer than the time resolution of the coding (which we recall from Section 3.2 as being around 3 minutes).

FIGURE 3.1: Method of counting transitions. If we count all active state to active state transitions between variable pairs as shown in **(A)**, there will be 5 transitions from variable *A* to *B*. For clarity, *B* to *A* transitions are not shown. This method assumes that every phase in which a variable is active is independent of its previous phase(s). If we assume that variables are activated in blocks, i.e. first activated at a particular time and then self-sustained for a duration, then it becomes reasonable to count transition frequencies as shown in **(B)**. In **(B)**, transitions always result in a variable’s first activation. We count two *A* to *B* transitions and one *B* to *A* transition.

To formulate our definition of a transition from variable i to variable j , we assume: (i) the variable dynamics in our data set are well approximated by a lag-1 Markov process i.e. the states of the variables at phase $t + 1$ are conditioned on their states at phase t , and (ii) when a variable is active from phase t to $t + l$, but not present at $t - 1$, it is regarded as a single event first activated at phase t and lasting for duration l . For (i), Fig. 3.2A demonstrates that for a majority of variables, the ability to predict whether or not they are active at phase t is vastly improved by around one order of magnitude when we know the state of the variables at $t - 1$. We also see from Fig. 3.2B that the additional knowledge of their state at $t - 2$ did not improve significantly our prediction capability. Thus, under the assumption of inter-variable independence, we consider the lag-1 Markov process to be sufficient and also efficient in modelling the variable dynamics. In Fig. 3.2C, we showed that for a variable i which is active from t to $t + l$, it is much more likely that its active state at $t + k$ where $1 \leq k \leq l$ is due to itself than it

FIGURE 3.2: Comparison of Markov probabilities and colour map of correlations. In (A), the open circles are the observed probabilities P_i (the probability of observing a variable i being active at any phase), the upwards pointing triangles are sustaining probabilities $P(x_i(t) = 1|x_i(t-1) = 1)$ (the probability of observing an active i given that i is already active at a preceding phase), and the downwards pointing triangles are the first activation probabilities $P(x_i(t) = 1|x_i(t-1) = 0)$ (the probability of observing an active i given that i is inactive at a preceding phase). We ranked the variables in increasing order of observed probabilities. We see clearly that a vast majority of the sustaining probabilities are about an order of magnitude larger than their corresponding observed probabilities. That these variables have the tendency to self-sustain in the time-segmented data suggests that they are actually single events. In (B) we compare the lag-1 sustaining probabilities against the probabilities $P(x_i(t) = 1|x_i(t-1) = 1, x_i(t-2) = 0)$ and $P(x_i(t) = 1|x_i(t-1) = 1, x_i(t-2) = 1)$ (i.e. treating variables as a lag-2 Markov process). Unlike in (A), there is no universal increase in probabilities when the second preceding phase is accounted for. In (C) we show, on a colour map, the standard scores of observed conditional probabilities $P_{observed}(x_j(t) = 1|x_i(t-1) = 1)$ against their expected values if i and j were independent. If i and j were independent and characterized solely by their respective observed probabilities P_i and P_j , the expected conditional probabilities should be equal to the observed probability, i.e. $P_{expected}(x_j(t) = 1|x_i(t-1) = 1) = P_j$ and the distribution of these quantities should have a standard deviation of $\sigma = \sqrt{P_j(1-P_j)/N_i}$. The standard score is obtained from $Z = (P_{observed}(x_j(t) = 1|x_i(t-1) = 1) - P_j)/\sigma$. The colours on the map range from blue (lower than expected) to teal (equals expectation), and finally maroon (higher than expected). The dominant feature of this map is a diagonal of higher-than-expected conditional probabilities for self-transitions.

being conditioned by an external variable which leads to assumption (ii) that variables active over consecutive phases are regarded as single events. Therefore, we search for

inter-variable transitions only at the instances of first activations. A transition is thus defined as an event whereby a consequent variable is triggered (or first activated) by the activity of the precedent. Formally, we say a transition from i to j occurs at time t if j is first activated at time t with i active at $t - 1$ (i.e. $i(t - 1) = 1, j(t - 1) = 0, j(t) = 1$). The total number of transitions from variable i to j is therefore

$$N(j \leftarrow i) = \sum_{l=1}^L \sum_{t=1}^{T_l-1} x_i^l(t)(1 - x_j^l(t))x_j^l(t+1) \quad (3.1)$$

where $x_j^l(t)$ is the state of variable j at phase t of lesson l , L is the total number of lessons in the data set and T_l is the number of phases in lesson l . We illustrate examples of independence-assumed transitions and first-activation transitions in Fig. 3.1.

3.3.2 Measuring Significance

For two variables that are frequently active, the number of mutual first activations will likely be large even if they are causally unrelated. On the other hand, for two variables that are rarely active, even a small number of first activations may indicate strong causal relation between them if this number is larger than expected by chance. To test a transition frequency for statistical significance, we use a Monte Carlo sampling method described below.

Given a pair of variables i and j that we assume are uncorrelated, we can create synthetic signals of each variable with their lag-1 Markov models using the Monte Carlo sampling scheme explained in Table 3.1. These signals serve as null models that we can compare against the number of observed transitions from i to j . We generated $k = 10000$ synthetic signals $\hat{x}_{i,h}^l(t), h = 1, \dots, k$ for variable i and $\hat{x}_{j,h}^l(t), h = 1, \dots, k$ for variable j and computed the total number of transitions $\hat{N}_h, h = 1, \dots, k$ as described in Equation 3.1 for each pair of synthetic signals. The population of k synthetic transition frequencies form a distribution of $\hat{N}$.

Let p be the fraction of the population of $\hat{N}$ that is larger than or equal to the observed number of transitions $N_{observed}(j \leftarrow i)$. If $p \approx 0$ then $N_{observed}$ is far above the mean of this distribution — there is a far greater number of observed transitions from i to j than expected from the null model which assumed that variables i and j are independent. If $p \approx 0.5$, then $N_{observed}$ is close to the median of this distribution and can be well explained by the null model. Finally, if $p \approx 1$, then there are much fewer transitions than expected from the null model.

We take the significance to be $s = 1 - 2p$. By rescaling in this manner, $s \approx 0, 1, -1$ if $N_{observed}$ is, respectively, approximately the same as, greater than, or less than the

TABLE 3.1: Monte Carlo sampling example. Suppose $P(0|0) = 0.8$, $P(1|0) = 1 - P(0|0) = 0.2$, $P(0|1) = 0.3$, $P(1|1) = 1 - P(0|1) = 0.7$, and we start with $x(1) = 0$. Then to generate a sequence using these transition probabilities, we draw 5 random numbers $r = (0.52, 0.86, 0.67, 0.23, 0.91)$ uniformly from the interval $[0, 1)$ and allow $x(t) = 0$ when $r < P(0|0)$, $x(t-1) = 0$ and $x(t) = 1$ when $r < P(0|1)$, $x(t-1) = 1$. Since $x(1) = 0$, to obtain $x(2)$ we look at $P(0|0)$. Since $r(2) < P(0|0)$, $x(2) = 0$. However, $r(3) > P(0|0)$, giving $x(3) = 1$. Iterating further, since $x(3) = 1$, we look at $P(0|1)$ and $r(4) > P(0|1)$, $x(4) = 1$. Then $r(5) < P(0|1)$, $x(5) = 0$. Finally, $r(6) > P(0|0)$, and $x(6) = 1$. The sequence we get is therefore $x = (0, 0, 1, 1, 0, 1)$.

Phase	r	$P(0)$	x
1	—	—	0
2	0.52	$P(0 0) = 0.8$	0
3	0.86	$P(0 0) = 0.8$	1
4	0.67	$P(0 1) = 0.3$	1
5	0.23	$P(0 1) = 0.3$	0
6	0.91	$P(0 0) = 0.8$	1

value expected from the null model. We can then compute the significance-weighted transition frequencies by multiplying each transition frequency with its significance. The weighing procedure diminishes insignificant frequencies and makes lower-than-expected frequencies negative. The weighted scores are

$$N_{weighted}(i \leftarrow j) = sN(i \leftarrow j). \quad (3.2)$$

For our analysis, we admit only edges where the significance $s > 0.5$, i.e. $p < 0.25$ and thus there is 75% confidence that the transition is not spurious.

3.3.3 Network Visualization and Filtering

We visualize the weighted transition frequencies $N_{weighted}$ on complex directed networks (one for Grade 5 Mathematics and one for Grade 9 Mathematics) where each node in the network represents a variable (a teaching practice) and the directed edges between nodes denote significant transitions. If we show all the edges in these networks without any filtering, we get ‘fur balls’ like that shown in Fig. 3.3. The network is too densely interconnected and the important structures are obscured. Therefore it is essential to apply filtering schemes that remove the less important edges and highlight the more important ones. We use the minimum spanning tree (MST) and the planar maximally filtered graph (PMFG) filtering methods that were previously introduced in Section 2.3.2 because they guarantee connectedness in the resulting networks (at least in the undirected sense). Since we seek to identify transitional motifs (paths) in this work, it is important to choose methods that preserve connectedness.

FIGURE 3.3: Unfiltered network of significant transitions. Here, we show the network of all significant ($s > 0.5$) transitions between variables from the data of Grade 5 Mathematics. The network is drawn with the following features: (i) node diameters are proportionate to the observed probability P_i of their corresponding variables, (ii) arrowheads of the edges point in the direction from the precedent to the consequent, (iii) edge thicknesses are proportionate to the weighted frequencies with thick lines denoting high weighted transition frequencies, and (iv) the placement of the nodes are determined by a multilevel force algorithm [125] implemented through the graphing software Gephi. This algorithm assigns attractive and repulsive forces among nodes according to their connectedness and allows the graph to dynamically relax into a configuration of low energy. The multilevel treatment ensures that the graph relaxes into the global minimum energy configuration instead of being trapped in a local minimum. In this scheme, the distances between nodes are *roughly inversely* proportionate to the score between them i.e. a pair of node connected with a highly weighted edge will be placed closed together in the network. For this unfiltered network, the density of edges is so great that it is very difficult to discern any meaningful structure from it.

The stochastic computation of p using the Monte Carlo sampling method introduces errors in the weighted frequencies $N_{weighted}$. To test how sensitively such errors affect the filtered transition networks, we constructed 100 instances of transition networks for Grade 5 Mathematics classes each generated with an independent set of Monte Carlo trials. It was found that 99% of the edges in the PMFG networks and 98% of those in the MST networks are always preserved under the small fluctuations in each set of $N_{weighted}$. The transition network might also change under different levels of significance cut-offs — we chose a default of $s > 0.5$ but also tested the networks generated with significance cut-offs between $s > 0.3$ and $s > 0.7$. As seen in Table 3.2, even in the most extreme case of requiring $s > 0.7$, it still shares 82% of the edges with the $s > 0.5$ network. Moreover, if edge weights are taken into account, 96% of the total edge strength from

TABLE 3.2: Network response to variations in significance cut-off. The quantities u measure the proportion of edges shared by a network with the network generated by setting $s_{cutoff} = 0.5$ (i.e. the default value, hence the values at the $s_{cutoff} = 0.5$ are all 1). u_p measures the similarity between the PMFG-filtered networks and u_m does the same between the MST-filtered networks. The v quantities compute the proportion of edge strengths in the network that is also found in the $s_{cutoff} = 0.5$ network. The subscripts m and p indicate respectively the MST and PMFG filtered networks.

p_{cutoff}	0.35	0.30	0.25	0.20	0.15
s_{cutoff}	0.30	0.40	0.50	0.60	0.70
u_p	0.93	0.97	1.00	0.91	0.83
u_m	0.86	0.95	1.00	0.87	0.75
v_p	0.99	1.00	1.00	0.99	0.96
v_m	0.96	1.00	1.00	0.99	0.91

the $s > 0.5$ network is preserved. As such, the PMFG-filtered transition networks are very stable with respect to the semi-stochastic computation of $N_{weighted}$ and also stable with respect to varying s cut-offs. Although the tests were performed only with the Grade 5 transition networks, we believe that similar conclusions can be extended to the Grade 9 networks.

3.3.4 Dynamical Motifs

Beyond recognizing just the pairwise transitions between lesson variables, we also want to identify sequences of two or more significant transitions which we term dynamical motifs. These sequences enables us to track the progression of practices across longer durations than just two consecutive phases. To do so, we return to the time-segmented data of individual lessons and search for frequently occurring sequences of significant transitions. This procedure is illustrated Fig. 3.4. Dynamical motifs identified are sequences that were repeated over many lessons.

3.4 Results

3.4.1 Prevalent Transitions in Networks of Teaching Practices

In Figs. 3.5A and 3.5B, the filtered graphs of significant transition networks are shown for Grade 5 Mathematics and Grade 9 Mathematics respectively. We can observe that across both grade levels, teaching practices are organised around source (outgoing) hubs that correspond to *Knowledge as Truth* and *Instructional Activity (IA): Teacher-Dominated Talk*. These source hubs are characterized by large numbers of significant outgoing transitions. Although these hubs also tend to be the most frequent practices, their

FIGURE 3.4: The time-resolved chart of a limited set for variables for a given lesson is drawn here. Diagonal lines in the chart represent significant ($s > 0.5$) transitions defined in Section 3.3.1. There are several paths in the chart which involves two or more transitions (three or more variables) and we highlight three examples of such paths in blue, red, and yellow respectively. For each path on the time-resolved chart there is a corresponding path in the time-integrated transition network.

frequencies alone do not account for the large number of significant outgoing transitions. This suggests strongly that *Knowledge as Truth* and *Instructional Activity (IA): Teacher-Dominated Talk* are indeed drivers that initiate teaching sequences.

For Grade 5 (Fig. 3.5A), *IA: Teacher-Dominated Talk* is an especially generative node which leads to nodes such as *Knowledge as Truth* (via *Supervisory Monitoring*, where teachers monitor students' behaviour in class), *To Practice What was Learnt* (where students practice what teachers taught them), *Procedural Talk* (classroom talk about procedural knowledge), and *Doing Mathematics Activity* (students doing a mathematics task in a disciplinary manner). An attendant observation is that teacher responses during *Doing Mathematics Activity* tend to be short (one or two word responses such as “yes”, “good”, “well done”, etc.). But while *Doing Mathematics Activity* is a generative node in Grade 5, it all but disappears by Grade 9 (Fig. 3.5B). We also note the presence of sink (incoming) hubs such as *CD: To practice what was learnt*, *IA: Individual Seatwork*, and *IS: Monitoring — Individual* which attracts several significant incoming transitions. Unlike the source hubs, sink hubs are practices that occur much less frequently and are harder to detect. The network analysis methodology uniquely discerns these hubs with their special role.

At Grade 9, *Knowledge as Truth* and *IA: Teacher-Dominated Talk* are still present as large hubs, but additional hubs have appeared: *Procedural Talk*, *Procedural Knowledge Focus*, *Knowledge Communication Presentation* and *Knowledge Representation*. The latter two hubs refer to how students present and communicate mathematical ideas (typically when asked by the teacher to present a problem solution on the white board

FIGURE 3.5: Transition networks of teaching practices. The PMFG-filtered networks for **(A)** Grade 5 Mathematics and **(B)** Grade 9 Mathematics are drawn with features (i), (ii), and (iii) as described in Fig. 3.3. We use dark edges to highlight the MST backbone of the PMFG network. The node placement was determined by the same multilevel force algorithm [125] in Fig. 3.3 acted on the MST edges. We also indicate the community of the nodes in the graph by their colour. They are detected with a community-detection algorithm [108] that was built into the graphing software. We describe it in detail in the Appendix A3.1. The detection process, not guided by any pedagogy theory, nevertheless produced communities of variables that are in general agreement with theory.

in front of the classroom), and to how students represent mathematical ideas using different representational forms (typically abstract forms, rather than concrete or pictorial), respectively. Here, we find that while *IA: Individual Seatwork* and *IS: Monitoring Individual* remain as prominent sink hubs for Grade 9, *IS: Feedback to Individual Students* has replaced *CD: To practice what was learnt* in its role.

3.4.2 Network Properties

We report in Fig. 3.6 the in/out strength and degree distributions of the unfiltered networks of significant transitions (i.e. the network in Fig. 3.3 for Primary 5 and the similarly defined network for Secondary 3). We observe that none of the in/out strength or degree distributions in either level seem to follow an exponential or power-law distribution as seen in many other real-world networks. We note that the strength distributions are more-or-less exponentially distributed for intermediate strengths but then decay faster than exponential towards the high-strength tail. The degree distributions decay somewhat linearly as seen in Figs. 3.6B and D. We hesitate to over-speculate the origins of these unusual strength and degree distributions except to point out that in our network, the nodes representing teaching practices in the system of teaching and learning are expert-defined. This is unlike, for example, the WWW network where there exist natural partitions in defining the components of the system which are the individual web pages. The observed distributions could be artifacts stemming from such artificially defined variables (i.e. system components).

We also give the top ranked teaching practices in terms of their degree and strength centralities in Table 3.3. We see that there is little reciprocity in the most highly ranked inwards and outwards centralities i.e. those with high outward centralities do not necessarily have high inward centralities.

TABLE 3.3: Centrality rankings of teaching practices. Descriptions of teaching practices by ID are given in the Appendix in Table A3.1.

Primary 5								
Rank	ID	s_{out}	ID	s_{in}	ID	k_{out}	ID	k_{in}
1	8	1.00	29	1.00	117	1.00	66	1.00
2	50	0.83	128	0.98	94	0.87	166	0.92
3	51	0.83	28	0.98	51	0.87	154	0.92
4	124	0.79	154	0.98	8	0.83	29	0.90
5	63	0.75	155	0.95	64	0.81	28	0.89
6	120	0.74	149	0.86	50	0.81	170	0.87

Secondary 3								
Rank	ID	s_{out}	ID	s_{in}	ID	k_{out}	ID	k_{in}
1	9	1.00	166	1.00	88	1.00	56	1.00
2	79	0.89	152	0.97	61	0.91	60	0.94
3	50	0.87	26	0.97	112	0.90	151	0.88
4	61	0.78	151	0.92	51	0.90	166	0.86
5	99	0.78	27	0.89	70	0.86	152	0.85
6	88	0.77	154	0.87	9	0.84	150	0.83

FIGURE 3.6: Strength and degree distributions of teaching practice networks. (A) and (C) shows the in/out strength distributions of the Primary 5 and Secondary 3 networks respectively. The horizontal axes are normalized by dividing with their maximum values. (B) and (D) shows the same for degree distributions. The blue plots are for the outward strengths/degrees and the orange plots are for the inward strengths/degrees as indicated in the legend.

3.4.3 Motifs of Teaching Practices

The most common length-3 and length-4 motifs for Grade 5 and Grade 9 Mathematics are shown in Fig. 3.7. A motif is said to be length- n if it consists of $n - 1$ successive significant transitions, but is not followed by an n th transition that is statistically significant. Some of the length-4 motifs contain a length-3 stem that branches out into more than one significant fourth transition.

As we can see, Grade 5 motifs typically show a common ‘cycle’ of teaching: The teacher starts talking about content (IA: *Teacher-Dominated Talk, Knowledge as Truth, Procedural Talk*); next, students are often given individual practice work (IA: *Individual Seat-work*); then, the teacher provides some feedback (*Comment/Remark*), returns to lengthy talk (IA: *Teacher-dominated Talk, Exposition*), or asks the whole class a closed question which typically has one correct answer (*Teacher Closed Question: WC*). Sometimes, the sequences carry on with more practice work (CD: *To Practice What Was Learnt*,

IA: Individual Seatwork) as well as the teacher’s monitoring of the student learning (*IS: Monitoring — Individual*, *IS: Formative Monitoring*, *IS: Supervisory Monitoring*) or providing individual feedback (*IS: Feedback to Individual Students*, *IA: Individual Talk to student*).

The Grade 9 motifs are largely similar to those in Grade 5, but with a greater focus on *Procedural Knowledge Focus* in the middle of the common motifs. Also, some Grade 9 lessons tend to end with the teacher providing a closing exposition (*Teacher Exposition*), lecture, and IRE (a form of talk where the teacher Initiates, the student Responds, and the teacher Evaluates) (*Lecture*, *Lecture + IRE*).

Interestingly, there are motif links such as *IA: Individual Seatwork* to *Teacher Exposition* for Grade 5, *Procedural Talk* to *IA: Individual Seatwork* for Grade 5, and *Procedural Knowledge Focus* to *IS: Monitoring — Individual* for Grade 9 which are present in the collection of common motifs in Fig. 3.7 while not appearing in their respective PMFG-filtered networks in Fig. 3.5. The PMFG filtering scheme picks edges in descending order of statistical significance. However, because the statistical significances are time-integrated, there is no guarantee that successive transitions in the PMFG appear one after the other during actual lessons. These time-integrated transitions may be part of two motifs, one ending at the node that the other starts at. This is why a time-resolved analysis must also be carried out to identify dynamical motifs consisting of successive transitions that are statistically significant.

Also, MST filtering preserves only a single possible undirected path between any pair of nodes while the PMFG is less restrictive and preserves several alternate paths that can be taken from one node to another. In this section, we showed that many important pedagogical motifs take place not on the MST but on the PMFG pathways. This suggests that MST filtering may be overly restrictive in representing processes in pedagogy.

3.5 Discussion

The transmissionist model of teaching views teaching as a straightforward and unproblematic process of transferring information (knowledge) from the mind of the teacher to the mind of the student [126–128]. In this model, knowledge is a collection of facts about the world and procedures to solve problems and learning is something akin to accumulating goods in the form of such knowledge. In the transition networks of both Grade 5 and Grade 9 Mathematics, *Knowledge as Truth* and *Instructional Activity (IA): Teacher-Dominated Talk* are seen as dominant hubs. *Knowledge as Truth* means that the knowledge that is presented in the lesson is viewed as non-contestable, non-negotiable

and objectively valid. As an Instructional Activity (IA), *Teacher-Dominated Talk* means simply that the teacher does most of the talking in the classroom. These practices operate most effectively in the transmissionist model of teaching when teachers need only to ensure that they cover the curriculum. However, in this setting students are not given the opportunity to engage in extended and productive discussions during lessons, and may be disengaged from the learning process.

However, as Ball [129] points out, mathematics instruction ought to involve a lot more than the transmission of facts and procedures. Above all, mathematics instruction should provide students with rich opportunities to engage in authentic, domain-specific mathematical practices. Thus, the transmissionist model is insufficient as a teaching and learning model for mathematics. In Grade 5, the strong presence of *Doing Mathematics Activity* as a generative hub is an important practice that educators hoped to see. *Doing Mathematics Activity* is inherently disciplinary in nature, and students need to acquire the mathematics-specific disciplinary skills, knowledge and disposition to understand mathematics deeply, debate, and discuss mathematical procedures and concepts, and appreciate the importance and relevance of mathematics. It captures the idea that mathematics learning does not consist purely of learning algorithmic procedures (captured in *Procedural Talk* and *Procedural Knowledge Focus*) and content (captured in *Factual Talk* and *Factual Knowledge Focus*), but requires students to perform discipline-specific knowledge practices as outlined by Ball. These activities are the most demanding of all activities, since they require the students to draw upon a range of mathematical knowledge and procedures/skills to complete some work that has no well-rehearsed way of completing the task.

The disappearance of *Doing Mathematics Activity* as a hub in Grade 9 suggests that at this level, the focus on what we consider to be a highly-valued disciplinary understanding of mathematics has been replaced by a focus on procedural knowledge (*Procedural Knowledge Focus*) as an effective and efficient means to present such knowledge, such as efficient problem solving methods. Yet, we believe that it is precisely because of this pedagogical approach focused on a highly proceduralised and efficient form of learning abstract mathematics that has made Singapore one of the top performing countries in international assessments that measure students ability in mathematics. Our observation confirms the PISA 2012 findings on Singapore Mathematics [130] which shows that there is a strong relationship between the teaching of Formal Mathematics (as opposed to Applied Mathematics or Word Problems) and student mathematical performance in the PISA tests.

Finally, teaching and learning must be understood as a process. Christie [131] talks of

the prototypical classroom lesson beginning with some form of teacher direction (Curriculum Initiation), followed by a middle portion where the teacher and students share direction (Curriculum Collaboration/Negotiation), and ending with students' independent activity (Curriculum Closure). In the motifs we identified for Grade 5 and Grade 9, the didactic of teacher-student-teacher as well as teacher-student-teacher-student sequences highlights the complex interaction cycles that occur in mathematics classrooms. Almost all lessons begin, as described above, with a teacher-led initiation (*Knowledge as Truth, Instructional Activity (IA): Teacher-Dominated Talk, Procedural Talk, Procedural Knowledge Focus*). Often, these motifs lead immediately to individual practice work (*IA: Individual Seatwork*), seemingly replacing the middle phase of Christie's pattern with student's independent activity. Some motifs then terminate with feedback from the teacher in the form of lengthy talk (*Teacher Comment/Remark, IA: Teacher-dominated talk, Teacher Exposition*) following student's practice work. These motifs suggest a teacher-dominated closure in the teaching pattern. However, in other motifs, we discovered a fourth movement which is a return to more practice work (*CD: To practice what was learnt*) coupled with teacher's monitoring of such practice work (*IS: Monitoring — Individual, IS: Formative Monitoring, IS: Supervisory Monitoring*).

In Grade 9, the greater focus on *Procedural Knowledge Focus* is in line with the high stakes environment that Grade 9 teachers and students are entrenched in, the Grade 10 General Cambridge 'O' Level Examination, and the necessity that students are familiar with the procedures and practice skills needed to perform well in the examination. Interestingly, when the focus is on *Procedural Knowledge Focus*, it is often followed by a more personalised approach where the teacher may provide individualised, specific, feedback (*IS: Feedback to Individual Students*) or learning support (*IS: Contextual & Flexible Procedural LS*). Such feedback and support are positive signs that rather than a summative exposition at the end (which does occur nevertheless), teachers and students have the opportunity to provide some feedback to one another. Expositions do end in about half of the motifs, likely due to the need to ensure that the important content and skills are reiterated and summarised, a necessary conclusion to the transmissionist model.

3.6 Chapter Conclusions

In this chapter, we presented the techniques we developed to infer directed networks from high-dimensional and multi-scale data based on transitions into first activations. In the discrete time-segmented data we worked on, events occur over a range of time scales. Some events are over in just a single time phase in the data while others last

for several phases. Our method takes into consideration such durations to define what constitutes a possible causal transition and thus a directed link in the network — one which a preceding variable triggers the first activation of a consequent. We think that this is a more accurate representation of transitions between events in time-segmented data than if we were to use a dynamic Bayesian approach that treats activations at every time phase independently. To test the transition frequencies for significance, we compared observed frequencies to synthetic ones using a Monte Carlo sampling method that creates synthetic signals with lag-1 Markov processes.

We applied the techniques on data on teaching practices in Singapore obtained by the country's National Institute of Education. It represents a pioneering effort in performing time-resolved analysis of classroom pedagogy. First, we provided empirical evidence that events tended to self-sustain over time phases which suggests that such continuous activations are more likely to be singular events. From the networks of significant transitions extracted by our method, we were able to identify and verify the dominant transitions between instructional practices and also pick out dynamical motifs of more than two practices. We established that teaching practices of both Grade 5 and Grade 9 Mathematics lessons are organized around *Knowledge as Truth* and *Instructional Activity (IA): Teacher-Dominated Talk* hubs which exemplify the transmissionist model of teaching. In addition, a *Doing Mathematics Activity* hub is present in the Grade 5 transition network. This suggests that teaching practices at this level have incorporated exploratory elements into the pedagogy, a positive sign towards the goal of 21st century learning. In contrast, in Grade 9, disciplinary understanding of mathematics has been replaced by a focus on procedural knowledge, which nevertheless accounts for Singapore's strong performance in international benchmarks. The motifs we extracted from the network highlight cycles of complex teacher-student-teacher-student sequences with great similarity between Grade 5 and Grade 9.

FIGURE 3.7: Common lesson motifs. The most common motifs for both grades are listed here. For each grade, the first row lists the most common length-3 motifs that do not significantly extend to other variables and prevalence r gives the fraction of lessons in which the paths are present. The second and third rows present length-4 motifs. Motifs with broken arrows indicate the possible terminations of a common stem of three indicators. For these motifs we show r for each possible ending adjacent to the final node together with the total prevalence of all alternatives.

Acknowledgements

The data is drawn from a National Institute of Education project: OER 20/09DH, Core 2 Research Programme (Panel 3), funded by the Office of Education Research, National Institute of Education. The project's principal investigator is Prof. David Hogan, and the co-principal investigators are Dr. Phillip Towndrow and Dr. Dennis Kwek. Data was collected and coded by over 15 research assistants and associates, led by Prof. David Hogan, Dr. Phillip Towndrow, Dr. Dennis Kwek, and Dr. Ridzuan Rahim.

Appendix

A3.1 Louvain method for community detection

The Louvain method of community detection is a type of *modularity maximization* algorithm that discovers the optimal partition of communities in a network. In such maximization algorithms, a high-quality community of vertices is one with many edges within communities and few edges connecting across communities. We recall that the ‘goodness’ of this partition is given by the modularity Q ,

$$Q = \frac{1}{m} \sum_{j>i} (A_{ij} - \frac{k_i k_j}{2m}) \delta(c_i, c_j) \quad (2.5 \text{ revisited})$$

where c_i is the partition that node i belongs to and $\delta(c_i, c_j)$ is the Kronecker delta. It returns the difference between the fraction of edges connecting nodes with identical c and its expected value (i.e. if the nodes were randomly connected). For weighted networks, A_{ij} refers to the edge weight of (i, j) and $k_i = \sum_j A_{ij}$ is the strength of node i (or the node degree in the case of unweighted networks). The algorithm to maximize Q in the Louvain method is an agglomerative one that merges nodes/communities iteratively. In the first phase, the steps taken to group the nodes are:

1. First, assign each node to their individual community i.e. each community contains just one node.
2. For each node i , calculate the change in modularity Q if it were to merge with a neighbour j .
3. Merge i to the neighbour that maximizes the increase in Q , if Q cannot be increased, i stays in its own community.

4. Recompute Q to reflect the new merger (if any).
5. Repeat items 2 to 4 repeatedly for all nodes i until no more further improvement is made for Q . The nodes can be cycled through deterministically or randomly, the differences are minimal.

At the end of this first phase, Q reaches a local maximum which cannot be improved with *individual* mergers of nodes. The second phase provides for mergers at the coarse grained community level. To do so, a new network is defined whose nodes are the communities found during the earlier phase. The weights of the edges joining them are now the sum of weight between nodes that belong to the corresponding communities. The procedure described for the first phase can then be applied to this new coarse-grained network to find the next hierarchy of community structures. Naturally, the number of nodes in these networks at each pass reduces. The algorithm stops when Q achieves a global maximum.

A3.2 ID of Teaching Practices

Table A3.1: Teaching practices corresponding to the ID numbers of Table 3.3.

Primary 5	
ID	Teaching Practice
8	IA: Teacher-dominated talk
28	IA: Individual Seatwork
29	IA: Individual Talk to student
50	Teacher Exposition
51	Teacher Closed Question: WC
63	Student Short Response to Teacher
64	Student Medium Response to Teacher
66	Student Closed Question to Teacher: WC
94	Lesson Organisation/Activities/Resources Talk
117	Revision Activity
120	Doing Mathematics Activity
124	Knowledge Justification
128	CD: To practice what was learnt
149	CD: To compute/simplify
154	IS: Formative Monitoring
155	IS: Monitoring - Individual
166	IS: Planned & Fixed Procedural LS
170	IS: Contextual & Flexible Strategic LS

Secondary 3	
ID	Teaching Practice
9	IA: Teacher-dominated talk
26	IA: Individual Seatwork
27	IA: Individual Talk to student
50	Teacher Exposition
51	Teacher Closed Question: WC
56	Teacher Closed Question: Individual/Group
60	Teacher Extended Response: Individual/Group
61	Student Short Response to Teacher
70	Lecture + IRE
79	Procedural Talk
88	Lesson Organisation/Activities/Resources Talk
99	Knowledge as Truth
112	Comprehension/Knowledge Manipulation Activity
150	IS: Supervisory Monitoring
151	IS: Formative Monitoring
152	IS: Monitoring - Individual
154	IS: Feedback to Individual Students
166	IS: Contextual & Flexible Procedural LS

Chapter 4

Networks of ‘Collective Creativity’ Skills During Problem-Solving

4.1 Introduction

Successful problem-solving is a process that involves the deliberate application of relevant strategies, knowledge, and skills to understand, conceptualize, and then work out the problem. Particularly, in collaborative problem-solving (CPS), Tan *et al.* has proposed a suite of skills they termed *collective creativity* (CC) necessary for successful CPS. Although their findings have shown differences in the overall rate in which CC skills are expressed in successful and unsuccessful attempts at CPS, the challenge still remains in extracting and identifying the logical and significant sequences of skills used in empirical settings. It represents a significant knowledge gap in the real temporal processes that take place when people solve problems collaboratively. A better understanding of such dynamics can potentially help educators to tailor instructions and lessons to foster in students such crucial problem-solving skills for the 21st-century.

For this chapter, we turn again to network inference, filtering, and analysis techniques to visualize and quantify processes in real-world CPS. The functional data made available to us was a turn-resolved record of expressed CC skills by different student teams over the course of CPS. From this data, we inferred and tested for significance the most important transitions between CC skills and then use directed networks to represent these relationships. Our goal is to use these networks as an aid to understand the dynamics of how different collaborative problem-solving skills appear in sequence as students attend to their tasks. From a network-to-network comparison of the significant

transitions between successful and unsuccessful teams, we demonstrated that successful teams have network paths that lead to the acquiescence and acceptance of proposed solutions while unsuccessful teams frequently meandered around unproductive dialogue. This was also verified in structural and functional betweenness centrality scores of the different collaborative problem-solving skills represented as nodes in the networks. The measure of functional betweenness is another novel contribution of this chapter.

Our approach here is in principle similar to that in Chapter 3 where we performed network inference to extract functional networks that represent processes from empirical data and then used network filtering methods to render these networks easier to interpret. However, crucial refinements and improvements have been made to our methods here. In the first instance, the network inference methodology used here not only penalizes insignificant pairwise relationships (like in Chapter 3) but also rewards the highly significant ones. In the second instance, we adopt an original topological filtering method termed the *minimum preserved-reachability graph* which we believe is more suited for filtering directed networks than the MST used in Chapter 3. Moreover, the data in this chapter presents an additional dimension of interest because it distinguishes between the data obtained from successful and unsuccessful problem-solving teams. This allows us to perform network-to-network comparisons that directly showcase the differences between successful and unsuccessful problem-solving.

This is a project done with collaborators from the National Institute of Education, Singapore, Dr. Jennifer PL Tan and Dr. Elizabeth Koh. They provided the data and the theoretical justifications of operationalizing collaborative problem-solving skills into a set of 21 dialogue categories which can be found in [132] and also reproduced in Table 4.1. We intend to write up the results presented in this chapter for future publication.

4.2 Collective Creativity in Problem Solving

It is well established that individual problem-solving relies on individual creativity to generate solutions [133, 134]. However, when individuals work in a group to solve problems collaboratively, creativity within persons have to be supplemented with social and communicative competencies as well. This is a necessary combination that enables solution generation through mutual understanding and sharing of ideas [132]. These cognitive and social aspects of CPS require a suite of skills that is aggregated under the umbrella of *collective creativity* (CC) [132].

In CC, CPS skills fall under three major dimensions which are the meta-cognitive, the cognitive, and the socio-communicative. During CPS, these competencies manifest

themselves in inter-player interactions. Specifically, [132] proposed that they can be found in dialogue between players in CPS. To do this, [132] operationalized CC into 21 distinct dialogue categories which can be reliably recognized in transcribed dialogue from CPS situations by trained human experts. We reproduce the full list of CC categories in Table 4.1. The same work has shown evidence that, at the aggregate level, successful CPS teams display more positive CC skills than unsuccessful ones.

TABLE 4.1: Dimensions, components, and categories of CC dialogue types. We use ‘Solutn.’ as an abbreviation for ‘Solution’.

Dimensions	Components	Categories		
		ID	Abrv.	Description
Meta-cognition	Reflexive	0	RFD1	Monitoring
		1	RFD2	Planning
		2	RFD3	Regulation
Cognitive	Divergent Production	3	DPD1	Solutn. Generation-Epistemic
		4	DPD2	Solutn. Generation-Concrete
		5	DPD3	Solutn. Generation-Elaboration
	Anti-Divergent Production	6	DPDA1	Premature closure
		7	CPD1	Problem Defining
	Convergent Production	8	CPD2	Problem Analysis
		9	CPD3	Solutn. Evaluation-Acquiescence
		10	CPD4	Solutn. Evaluation-Checking
		11	CPD5	Solutn. Evaluation-Critique
		12	CPD6	Solutn. Evaluation-Justification
Socio-Communicative	Prosocial Interaction	13	PID-MGQ	Mutual Grounding-Questioning (Consultative) Questions
		14	PID-MGRF	Mutual Grounding-Responding (Responsiveness) First Responses
		15	PID2	Affective
		16	PID3	Cohesive-Task
		17	PID4	Cohesive-Playful
	Anti-social Interaction	18	PIDA2	Disaffective
		19	PIDA3	Uncohesive-Task
		20	PIDA4	Uncohesive-Playful

4.3 Data

The data used in this study was obtained from the ATC21S Singapore school trials which were conducted with the objective of evaluating and fostering 21st century skills through computer-based assessment tasks [135]. For the trials, a set of computer-based tasks for pairs were specifically developed to be ill-defined, and thus demanding creative problem-solving skills, and to require the collaboration and input of both players to be completed. Collaboration is enforced because neither of the players is provided with

enough information to solve the problems by himself or herself. We focus our analysis in this chapter on the ‘Small Pyramids’ and ‘Warehouse’ tasks. Brief descriptions of the tasks are given in the Appendix section of this chapter. In conducting the experiments, students were randomly paired with an anonymous peer with each person working on their individual computers. Under this physical separation, their only available mode of communication was via an online chat application. Non-dialogic interactions like hand gestures and body language can thus be ruled out as contributing factors to the test outcomes.

From the message logs, trained human experts were engaged to manually scan if any CC dialogue types were expressed. For each line, the dialogue types were coded as binary variables (‘0’ if absent, ‘1’ otherwise) with the possibility that multiple types can be simultaneously expressed. The same dataset was coded by two experts for cross-validation and found to have very little inter-coder disagreement [132]. Occasionally, a single player sends several consecutive lines in the chat application. We make the assumption that these consecutive lines constitute a single coherent message which is a single *turn* in dialogue and recompute the expression of CC dialogue types in a turn-by-turn basis (see Fig. 4.1 for an illustrated explanation). 135 student teams were considered for our analysis of which 83 attempted the ‘Small Pyramids’ task and 52 did the ‘Warehouse’ task. The breakdown of their CPS outcomes is given in Table 4.2.

TABLE 4.2: Breakdown of CPS outcomes.

	Small Pyramids	Warehouse	Total
Successful	45	26	71
Unsuccessful	38	26	64
Total	83	52	135

4.4 Methods

In this section, we describe in detail the tools and techniques we employed in this chapter. We first describe the procedures we used to establish significant correlations between CC dialogue types and from them construct inferred transition networks. Next, we introduce the minimum preserved-reachability graph filtering algorithm used to simplify the inferred networks. Lastly, we mention a node centrality measure we call functional betweenness.

Player	Message	Dialogue Types											
A	The black one leh?;	0	0	0	...	0	0	0					
B	mine from the top is 20;	0	0	1	...	0	0	1					
B	the black is 20;	0	0	0	...	1	0	1					
A	The red box.;	0	0	0	...	0	0	0					
A	wait mine no numbers ;	0	1	0	...	0	0	1					
A	mine say fill in the red box.;	0	0	0	...	0	0	1					
B	4;	0	0	1	...	0	0	0					
B	write 4;	0	1	0	...	0	0	0					

Player	Dialogue Types											
A	0	0	0	...	0	0	0					
B	0	0	1	...	1	0	1					
A	0	1	0	...	0	0	1					
B	0	1	1	...	0	0	0					

FIGURE 4.1: Conversion of line-based coding data to turn-based. Top: Messages are read by human experts to pick out the presence of CC dialogue types. Bottom: Consecutive lines of the same players are condensed into a single vector. A dialogue type is coded ‘1’ if it appears once or more than once over the lines condensed.

4.4.1 Network Inference from Dialogue Data

In constructing the transition networks of dialogue types we consider, for simplicity, only the consecutive transitions between types. We define $d_i(m, t) = \{0, 1\}$ as the state (present or otherwise) of dialogue type i at turn t of team m . A transition from i to j is counted if $d_i(m, t) = 1$ and $d_j(m, t + 1) = 1$. Since dialogue progresses in a turn-by-turn basis in this experiment, we believe that the consideration of only such first-order correlations to be adequate in capturing the significant sequences. Then, $N(i, j)$ gives the total number of such transitions across all turns t and all teams m . We also define $p(j|i)$ to be the probability that $d_j(t + 1) = 1$ given that $d_i(t) = 1$ i.e. at the preceding turn. To say that a transition from i to j is significant, we intuitively require that

$$\frac{p(j|i)}{p(j|\neg i)} > 1, \quad (4.1)$$

i.e., j happens more frequently when conditioned by i at a preceding turn than the other way. In epidemiology, the left-hand side expression in Eq. 4.1 is known as the relative risk RR and is bounded on $[0, \infty)$. To circumvent the issues arising with infinite RR , we introduce a pseudocount of 1 to all $N(\neg i, j)$ i.e. the minimum value of $N(\neg i, j)$ is

1. It is understood as the ratio of the probability (i.e. risk) of contracting a disease under some type of exposure to the probability of contracting the disease without said exposure. $RR > 1$ indicates that i increases the probability of producing j (a positive correlation); $RR \approx 1$ means that they are independent; and $RR < 1$ points to a negative correlation.

When noise is present in the readings of d_i and d_j , the distribution of $\log RR$ is approximately normal [136] (see simulated distribution in Fig. 4.2). Its standard deviation (i.e. the uncertainty of $\log RR$), derived using the Delta method [137], is

$$\sigma = \sqrt{\left(\frac{1}{N(i, j)} + \frac{1}{N(\neg i, j)}\right) - \left(\frac{1}{N(i)} + \frac{1}{N(\neg i)}\right)}. \quad (4.2)$$

We reproduce the derivation of Eq. 4.2 in the Appendix. To be confident that an observed RR is significantly above the null value of 1, we want the bulk of the distribution $\mathcal{N}(\log RR, \sigma^2)$ to lie above 0. For a given RR , its z -score is simply

$$z = \frac{\log RR - \log 1}{\sigma}. \quad (4.3)$$

From this we can easily obtain the single-tailed p -value from standard conversion tables. If we want, for instance, to be 95% confident that the observed RR lies above 1, then we require that $p < 1 - 0.95$ and working backwards, we get $z > 1.645$.

Having obtained the transition frequencies $N(i, j)$ of type pairs, their relative risk RR , and their p -value significance, we can proceed to combine these quantities to define the transition edges in the networks. For a transition with frequency $N(i, j)$, we write its weighted transition frequency as

$$N_{wt}(i, j) = (RR - 1)N(i, j). \quad (4.4)$$

In this weighted count, transitions that are only marginally significant $RR = 1 + \epsilon$ are diminished close to 0 while significant transitions with $RR \gg 1$ are amplified. Inhibitory transitions ($RR < 1$) will have negative N_{wt} which we discard. We also do not consider edges if their p -value exceeds a threshold $p_{threshold} = 0.05$. Using these weighted frequencies as edge weights, we construct directed transition networks of dialogue types, where each node represents a CC type. The heaviest weighted edges in the networks represent dialogue type correlations that are dually *frequent* and *significant*.

We tested the sensitivity of the transition networks under changes in $p_{threshold}$. Different dialogue transition networks for the successful teams were generated with $p_{threshold}$ ranging from 0.10 to 0.01 (see Table 4.3). We found that lowering $p_{threshold}$ (i.e. tightening the significance condition) to $p_{threshold} = 0.01$ resulted in a MPRG-filtered network (see

FIGURE 4.2: Simulated $\log \widehat{RR}$ distribution. We simulate the distribution of $\log \widehat{RR}$ (with the hat over RR to denote a synthetically generated value) with parameters $N(i, j) = Np_{11} = 75, N(i, \neg j) = Np_{12} = 75, N(\neg i, j) = Np_{21} = 100, N(\neg i, \neg j) = Np_{22} = 150$ with an ensemble size $N_{ens} = 15000$. $\widehat{p}_{ij}$ s are drawn from the Poisson distributions $\text{Pois}(\lambda = p_{ij})$. The histogram gives the density distribution of this simulated $\log \widehat{RR}$. The green line is the normal probability density function with $\mu = \log RR$ and σ calculated according to Eq. 4.2. The inset shows the same plots with a logarithmic vertical scale.

TABLE 4.3: Network response to changes in $p_{threshold}$. The quantity u measures the similarity (i.e. proportion of edges shared) between a MPRG-filtered network with a given $p_{threshold}$ and the MPRG-filtered network at the default $p_{threshold} = 0.05$.

$p_{threshold}$	0.10	0.07	0.05	0.03	0.01
u	1.00	1.00	1.00	0.98	0.96

next section) that still shared 96% of its edges with the default-value $p_{threshold} = 0.05$ network. Raising $p_{threshold}$ (i.e. relaxing the condition) produced identical networks. Evidently, the filtered transition networks do not change much with respect to changes in the artificially imposed significance cut-off.

4.4.2 Minimum Preserved-Reachability Graph

Even with all negative and statistically insignificant ($RR < 1$) edges excluded, the resulting networks are still far too dense and hard for us to interpret visually. We introduce a *minimum preserved-reachability graph* (MPRG) filtering technique that (i) retains, as much as possible, the most heavily weighted edges; (ii) allows the weaker edges to be filtered out; and (iii) results in a filtered network with the same *reachability* as the original network. That is, if a directed path exists between nodes i and j in the

original network, i must be also reachable from j in the filtered network, even if not along the original shortest directed path.

Like the MST we described in Chapter 2, the MPRG is grown iteratively starting with an unconnected network that contains only vertices. Edges from the original network are added to this empty network in order of descending weight under the condition that an edge (i, j) is only admitted if j cannot be reached from i following existing paths in the growing network. We illustrate this procedure with a simple example in Fig. 4.3.

FIGURE 4.3: Minimum preserved-reachability graph filtering. The original network consists of both the solid and dashed directed edges. In growing the MPRG iteratively, the edge with weight $w = 10$ is first included followed by the one with $w = 8$. The $w = 6$ edge will not be admitted because the vertices it links are already joined by the path that contains the $w = 8$ and $w = 10$ edges. We find that the MPRG filters out also the $w = 4$ and $w = 1$ edges.

In the context of analyzing dialogue dynamics, the MPRG holds a number of advantages over other filtering methods we mentioned in Chapter 2. First, the transitive reduction [81], which finds the minimum number of edges to preserve the reachability of a network, was developed primarily for use with unweighted graphs. The widely-used MST, which we applied in the previous chapter, cannot guarantee that reachability is preserved. Moreover, it automatically disallows all loops in the filtered graphs. We believe that in certain processes, like dialogue, loops can represent important self-reinforcing dynamics and should not be categorically eliminated.

4.4.3 Betweenness and Functional Betweenness

We have introduced the concept of betweenness centrality earlier in Chapter 2. The unnormalized betweenness of a vertex v is given as

$$B(v) = \sum_{i,j} \frac{\psi(i,j|v)}{\psi(i,j)} \quad (4.5)$$

where $\psi(i, j|v)$ is the number of equal-length shortest paths connecting vertices i and j that contains the vertex v and $\psi(i, j)$ is the total number of equal-length shortest paths between the i, j pair. The implicit assumptions made in computing $B(v)$ are that (i) paths linking all possible pairs in the network are equally important (the summation goes through all pairs without attaching any weight coefficients) and (ii) flows between vertices always take the shortest network paths. We consider this a measure of *structural* betweenness in the network. For unweighted networks, path lengths are measured by simply counting the number of ‘hops’ needed to get from one node to another. In weighted networks, where high edge weights represent strong connections, the determination of path lengths is not so straightforward. One way to do so is to consider that the length of some path h to be the sum of *inverse* edge weights it contains i.e. $l_{wt}(h) = \sum_{e \in h} w(e)^{-1}$. This is in line with the intuition that one wants to keep the number of hops minimized while preferentially taking paths connected with strong edges when searching for the weighted shortest paths.

However, in many real-world functional networks, edges between vertices represent merely a probabilistic link. For instance, in the transition networks of dialogue types in this chapter, there is no guarantee that the presence of dialogue i will trigger dialogue j even if they are connected by an edge. Consequently, there is also no certainty that dialogue types are always *functionally* connected via the shortest paths in the network. Fig. 4.4 explains this claim.

We therefore propose a measure of *functional* betweenness that accounts for the actual observed dynamics of the system that underlies the network. Let us first define that $\zeta = (\zeta_1, \zeta_2, \dots, \zeta_{t_{max}})$ with each ζ_t being the list of variables (dialogue types in our case) present at a turn or time t in an observed instance of the system (in our case, records of some particular team). We connect the types in consecutive turns i.e. ζ_t and ζ_{t+1} only if we have judged from the overall data that the transition between them is significant (i.e. $p \leq p_{threshold}$) (see Fig. 4.4). If given some pair i and j , the shortest functional path is the one that joins i_t to $j_{t+\delta t}$ via a path that minimizes δt (see Fig. 4.4 again). The intermediate types connecting i_t and $j_{t+\delta t}$ are those *functionally between* i and j in this observed instance of the system. From the records of a single team $\zeta^{(m)}$, we define the betweenness of a vertex v to be

$$B_{func}^{(m)}(v) = \sum_{i,j} \frac{\Psi(i, j|v)}{\Psi(i, j)} \quad (4.6)$$

where the capitalized function Ψ measures the quantities of ψ on the shortest *functional* paths. The global functional betweenness is then simply the sum of the quantities across all teams $B_{func}(v) = \sum_m B_{func}^{(m)}(v)$.

FIGURE 4.4: Illustration of functional betweenness. In the transition network on the left, the shortest path from B to A is via C (highlighted in blue). However, it is possible that when we examine the sequential expression (activation) of variables from a portion of the data that we do not observe this (B, C, A) sequence. On the right, we illustrate a hypothetical record of expressed variables in which the shortest functional path from B to A is via D and then C . Note that we only join consecutively activated variables if a connection between them exists on the network on the left. This figure also helps to illustrate spontaneity and generativity. A spontaneous instance of a dialogue type has no incoming transitions. D in ζ_2 and A in ζ_3 are two such examples. Generative instances are those with an outgoing transition like B in ζ_2 and C in ζ_1 .

4.5 Results and Discussion

4.5.1 Transition Networks of CC Dialogue Types

In Fig. 4.5, we show the MPRGs of the successful teams, unsuccessful teams, and their overlapped graphs. The last visualization succinctly summarizes the similarities and differences in the dialogue dynamics of the two groups. One of the most differentiated edges between the two groups is the bi-directional transition between $10:CPD_4$ and $9:CPD_3$ which is much more heavily weighted in successful teams. Another transition important to successful teams is the bi-directional transition between $4:DPD_2$ and $13:PIDMGQF$. It appears that $9:CPD_3$ is a transitional sink for successful teams that attracts many incoming edges. The unsuccessful teams dominate transitions among nodes such as $6:DPDA$ to $18:PIDA_2$, $18:PIDA_2$ to $2:RFD_3$, and $17:PID_4$ to $20:PIDA_4$.

The unfiltered neighbourhoods of $9:CPD_3$ and $6:DPDA_1$ are shown in Fig. 4.6A and B. Transitions associated with $9:CPD_3$ are mainly from successful teams as already seen in Fig. 4.5. The $9:CPD_3$ to $15:PID_2$ and to $12:CPD_6$ transitions, though primarily appearing in unsuccessful teams, are relatively weak and infrequent. Unsuccessful teams dominate all transitions in and out of the $6:DPDA_1$ node.

FIGURE 4.5: MPRGs of successful and unsuccessful groups. (A) and (B) are the respective MPRGs of the successful and unsuccessful teams with $p_{threshold} = 0.05$. $N_{weighted}$ in each class are normalized by a factor T_{class}/T_{total} where the numerator is the total number of turns taken within each class and the denominator is the total across all classes. (C) shows the overlapped graph of (A) and (B). An additional cutoff is applied at $N_{weighted} = 5.5$ for all. Considering the T_{class}/T_{total} normalizing factor, transitions with $N_{weighted} = 5.5$ occurs in only $5.5/135 \approx 0.04$ of the teams and are relatively inconsequential. In the overlapped graph in (C), edges that belong to the filtered graphs of both classes are drawn in grey if their differences between classes are less than 0.3 of their mean. Otherwise, the edges are drawn in the color of the dominant class. Node sizes are proportionate to the summed frequencies of the corresponding dialogue types and the thickness of the arrows correlates with $N_{weighted}$. The dotted lines in each network gives the community partitions. 4, 4, and 3 communities are found respectively. Their modularities are: $Q_s = 0.11$, $Q_u = 0.46$, and $Q_{overlap} = 0.38$.

We believe that in successful teams, the 10:CPD4 (Solution Evaluation-Checking) to 9:CPD3 (Solution Evaluation-Acquiescence) transition indicates a logical sequence to

FIGURE 4.6: Unfiltered neighbourhoods of (A) *9:CPD3e* and (B) *6:DPDA1*. The immediate neighbourhood of these dialogue types are drawn here without filtering. The edge colours are interpreted identically to Fig. 4.5.

verify and then come to an agreement of a proposed solution. Conversely, unsuccessful teams are plagued with unproductive dialogue processes such as the transitions from *6:DPDA1* (Premature Closure) to itself or out towards *18:PIDA2* (Disaffective). Furthermore, we see evidence in students attempting to regulate unproductive dialogue types with *18:PIDA2* (Disaffective), *19:PIDA3* (Uncohesive-Task), and *20:PIDA4* (Uncohesive-Playful) having edges that lead to *2:RFD3* (Regulation) in unsuccessful teams. However, the lack of any outward transition from *2:RFD3* seem to suggest that there is no further progress into productive dialogue after such attempts at self-regulation.

4.5.2 Communities and Betweenness of CC Dialogue Types

Fig. 4.5 shows also the community partitions of dialogue type nodes found using the Louvain community detection method (see description of the algorithm in the Appendix of Chapter 3). The implementation here ignores the directionality of the edges. The communities are quite consistent whether partitioned on the successful, unsuccessful, or overlapped network. There is a peripheral community of nodes to the right-hand side of the networks consisting of *18:PIDA2*, *19:PIDA3*, etc. that is, in the unsuccessful and overlapped networks, connected to the rest of their respective graphs only via *6:DPDA1*. Another consistent community is the one found towards the bottom of the networks with high frequency dialogue types like *13:PIDMGQF*, *14:PIDMGRF*, etc. The top-left community is also interesting with relatively small (i.e. less frequent) nodes like *9:CPD3* but with many heavily weighted edges.

We also computed the figures for structural betweenness B_{struct} and functional betweenness B_{func} in the networks and report them in Table 4.4. In the successful transition network both measures agree in the high betweenness rankings of *13:PIDMGQF*, *10:CPD4*, and *7:CPD1*. However, there some large deviations like in the rankings of *9:CPD3* (higher in B_{struct}) and in *4:DPD2* (higher in B_{func}).

From the topology of the network, we can see that *9:CPD3* acts as the exclusive gateway for *5:DPD3* and as a possible transit point for *3:DPD1* and for *4:DPD2* to access the rest of the network, hence its relatively high B_{struct} score. However, its low B_{func} signifies that there are actually few functional sequences that went through it. In calculating B_{struct} , the paths between every node pair contribute equal weights to the score. Consider the long path (*5:DPD3, 9:CPD3, 10:CPD4, 11:CPD5, 12:CPD6*) and the shorter one (*5:DPD3, 9:CPD3, 10:CPD4*). Both path contribute a score of 1 in $B_{struct}(9)$. However, in actual observation the longer path is much less likely to occur than the shorter one. This is because the edges in the networks give only the probabilistic transitions as explained earlier in Section 4.4.3. Therefore, nodes that have high B_{struct} as a result of being between many far apart node pairs are unlikely to have also high B_{func} .

4:DPD2 is connected to the rest of the graph only via the (*13:PIDMGQF, 4:DPD2, 9:CPD3*) sequence. However, the alternate (*13:PIDMGQF, 14:PIDMGRF, 10:CPD4, 9:CPD3*) path is shorter ($l_{wt}(13, 4, 9) = 0.010$ vs $l_{wt}(13, 14, 10, 9) = 0.009$). Accordingly, the ‘winner-takes-all’ approach of B_{struct} never routes any shortest paths through *4:DPD2* and thus its B_{struct} score 0. In observed activations, however, some functional paths do go through *4:DPD2* since $B_{func}(4)$ is non-zero.

For unsuccessful teams, the *6:DPDA1* node achieves high B_{struct} score as all nodes in the right-hand side community of its network must use it to get to the rest of the network. The ranking of its B_{func} score is lower but still relatively high. The explanation we invoked earlier for *9:CPD3* in the successful network applies here as well.

We can also examine the *spontaneity* of a dialogue type. We define an instance of a dialogue type $d_k(t) = 1$ to be spontaneous if there is no significant in-transition into it from turn $t - 1$ (see Fig. 4.4 for examples). The spontaneity s is given as the proportion of k that spontaneously appears without in-transitions. We can similarly define the *generativity* g of a dialogue type as the proportion of k that resulted in at least one significant outward transition (see again Fig. 4.4). Fig. 4.7 and Table 4.5 shows that *6:DPDA1* and *12:CPD6* are more spontaneous (i.e. unpredictable) in successful teams while *5:DPD3* and *3:DPD1* are more unpredictable in unsuccessful teams. *6:DPDA1* and *18:PIDA2* are especially generative in unsuccessful teams. We compute as well the sequence-starting propensity ν which is the relative frequency that a dialogue type initiated a significant outward transition without a significant inward transition. The

TABLE 4.4: Betweenness scores of successful and unsuccessful teams. The scores are normalized by division with their respective maximums so that they lie on $[0,1]$.

Successful Teams	Rank	Type	B_{struct}	Type	B_{func}
	1	13:PIDMGQ	1.00	13:PIDMGQ	1.00
	2	10:CPD4	0.62	14:PIDMGRF	0.61
	3	7:CPD1	0.47	7:CPD1	0.29
	4	9:CPD3	0.40	10:CPD4	0.20
	5	16:PID3	0.37	16:PID3	0.11
	6	14:PIDMGRF	0.34	4:DPD2	0.09
	7	5:DPD3	0.28	3:DPD1	0.04
	8	8:CPD2	0.24	1:RFD2	0.03
	9	3:DPD1	0.24	8:CPD2	0.03
	10	0:RFD1	0.13	11:CPD5	0.02
	11	12:CPD6	0.12	9:CPD3	0.01
	12	11:CPD5	0.12	12:CPD6	0.01
	13	20:PIDA4	0.09	0:RFD1	0.01
	14	2:RFD3	0.09	15:PID2	0.01
	15	18:PIDA2	0.06	20:PIDA4	0.00
	16	17:PID4	0.05	19:PIDA3	0.00
	17	19:PIDA3	0.00	18:PIDA2	0.00
	18	15:PID2	0.00	17:PID4	0.00
	19	6:DPDA1	0.00	6:DPDA1	0.00
	20	4:DPD2	0.00	5:DPD3	0.00
	21	1:RFD2	0.00	2:RFD3	0.00

Unsuccessful Teams	Rank	Type	B_{struct}	Type	B_{func}
	1	6:DPDA1	1.00	14:PIDMGRF	1.00
	2	10:CPD4	0.99	13:PIDMGQ	0.79
	3	14:PIDMGRF	0.57	1:RFD2	0.33
	4	13:PIDMGQ	0.47	16:PID3	0.16
	5	18:PIDA2	0.39	6:DPDA1	0.14
	6	12:CPD6	0.37	10:CPD4	0.12
	7	15:PID2	0.35	4:DPD2	0.07
	8	16:PID3	0.21	17:PID4	0.06
	9	9:CPD3	0.21	2:RFD3	0.06
	10	8:CPD2	0.21	8:CPD2	0.05
	11	4:DPD2	0.20	7:CPD1	0.05
	12	0:RFD1	0.15	0:RFD1	0.05
	13	20:PIDA4	0.12	20:PIDA4	0.03
	14	2:RFD3	0.12	9:CPD3	0.02
	15	7:CPD1	0.06	19:PIDA3	0.01
	16	1:RFD2	0.04	18:PIDA2	0.00
	17	19:PIDA3	0.00	15:PID2	0.00
	18	17:PID4	0.00	12:CPD6	0.00
	19	11:CPD5	0.00	11:CPD5	0.00
	20	5:DPD3	0.00	5:DPD3	0.00
	21	3:DPD1	0.00	3:DPD1	0.00

sequence-terminating propensity ω describes the opposing property. *9:CPD3* tends more in both groups to end sequences than to start them.

FIGURE 4.7: Scatter plot of the differences in spontaneity s and generativity g . Dialogue types situated on the right side of the plot are more spontaneous in successful teams and those on the left in unsuccessful team. Types on the top of the graph are more generative in successful teams and vice versa.

That *9:CPD3* (Solution Evaluation-Acquiescence) has low functional betweenness does not contradict with the fact that it plays a key role as a transitional sink in the process of problem-solving. As seen in Table 4.5, *9:CPD3* is fairly low in generativity. This indicates that solution acquiescence is more often the *end-point* of a coherent sequence of functional activations as evident in its higher propensity to end a sequence than to start it.

4.5.3 Comparison with Time-Integrated Analysis

The time-integrated analysis of the dialogue data set treats each dialogue type independently and excludes the dynamical interactions between types over the course of a conversation. It relies, instead, on the aggregate frequencies of expressed dialogue types. Such a study has already been done in the earlier paper that introduced CC [132]. Here, we will repeat some of those results and comment on how they compare with our dynamical analysis.

Fig. 4.8 summarizes the relative rates of dialogue type production in the successful and unsuccessful teams. *6:DPDA1* is the most separated in its mean rates of production

TABLE 4.5: Spontaneity, generativity, sequence-starting propensity, and terminating propensity scores.

Type	s_s	s_u	g_s	g_u	ν_s	ν_u	ω_s	ω_u
0:RFD1	0.93	0.61	0.39	0.24	0.38	0.17	0.15	0.21
1:RFD2	0.61	0.31	0.45	0.43	0.26	0.14	0.17	0.36
2:RFD3	0.89	0.66	0.14	0.16	0.12	0.08	0.09	0.31
3:DPD1	0.83	0.83	0.27	0.14	0.11	0.55	0.29	0.03
4:DPD2	0.56	0.58	0.66	0.26	0.28	0.23	0.12	0.22
5:DPD3	0.49	0.87	0.19	0.19	0.22	0.55	0.20	0.03
6:DPDA1	0.94	0.32	0.06	0.76	0.06	0.18	0.06	0.14
7:CPD1	0.36	0.37	0.76	0.51	0.31	0.26	0.09	0.15
8:CPD2	0.59	0.48	0.15	0.46	0.10	0.18	0.41	0.28
9:CPD3	0.31	0.39	0.17	0.34	0.09	0.12	0.25	0.32
10:CPD4	0.23	0.28	0.84	0.75	0.17	0.14	0.09	0.13
11:CPD5	0.60	0.42	0.34	0.27	0.15	0.15	0.21	0.20
12:CPD6	0.73	0.50	0.51	0.30	0.32	0.00	0.16	0.00
13:PIDMGQ	0.34	0.66	0.86	0.88	0.16	0.40	0.09	0.05
14:PIDMGRF	0.04	0.02	0.63	0.57	0.02	0.01	0.36	0.42
15:PID2	0.73	0.63	0.20	0.35	0.13	0.27	0.21	0.22
16:PID3	0.53	0.71	0.36	0.55	0.19	0.39	0.34	0.13
17:PID4	0.49	0.40	0.59	0.60	0.24	0.22	0.16	0.21
18:PIDA2	0.73	0.64	0.17	0.45	0.09	0.22	0.19	0.20
19:PIDA3	0.86	0.67	0.31	0.26	0.31	0.02	0.09	0.21
20:PIDA4	0.40	0.12	0.60	0.76	0.20	0.08	0.20	0.20

being much higher for unsuccessful teams. This agrees with the network perspective that the *6:DPDA1* node is isolated in successful teams while in unsuccessful teams it is a highly generative hub. There are, however, many seemingly contradicting observations between the dynamical network picture and the time-integrated analysis. For example, *8:CPD2* is the second most differentiated in rate and appears more often in successful teams. However, transitions in and out of it are largely more significant only in unsuccessful teams. Also, the *9:CPD3* node which appears to be an important transition sink for successful teams is not very differentiated in its production rates.

It is reasonable that one assumes dialogue types that characterize successful teams in the time-integrated analysis (such as *8:CPD2* (Problem Analysis)) to generate other productive dialogue types. However, the lack of such an observation in our empirical results leads us to speculate that perhaps *8:CPD2* leads to other productive dialogue types only at higher-ordered (lag-2 or greater) correlations. For instance, if we discover a positive and significant lag-2 correlation between *8:CPD2* and *4:DPD2*, it could either mean that (i) *8:CPD2* led directly to the production of *4:DPD2* within the same player (somewhat like a player-internal process), effectively ignoring the team mate’s dialogue, or that (ii) there is a missing variable (i.e. CC competency) produced by the

FIGURE 4.8: Normalized production rates of dialogue types. For each team m , the rate of producing a dialogue type i is $R(i, m) = f(i, m)/T(m)$ i.e. the number of times i is produced divided by the number of dialogue turns. To get the normalized values, we take all $R(i, m)$ across all teams, center them, and divide them by the standard deviation. So $R_{Norm}(i, m) = (R(i, m) - \langle R(i) \rangle) / \text{std}(R(i))$. The means of the successful and unsuccessful $R_{Norm}(i)$ are plotted separately and ordered from left to right by the decreasing absolute difference. The top and bottom of the error bars gives the median values of the distribution of $R_{Norm}(i)$ above and below the mean respectively. Thus the range of the error bars covers half the distribution of $R_{Norm}(i)$ for each successful/unsuccessful class.

player’s partner in response to *8:CPD2*, and that this missing variable in turn causes the production of *4:DPD2*.

4.5.4 Subgroups in Successful and Unsuccessful Teams

For each of the successful and unsuccessful student groups, we performed hierarchical clustering with Ward’s criterion (see procedure in the Appendix). The dendrograms obtained are given in Figs. 4.9A and B. They clearly show that the successful and unsuccessful teams can each be divided into 2 major sub-clusters (S-1, S-2 and U-1, U-2). Moreover, in both cases, the basis of discrimination can be partially accounted for in the dialogue durations of each subgroup as illustrated in Figs. 4.10B and C with the S-1 and U-1 groups being more verbose than their counterparts.

The overlapped MPRGs for subgroups S-1 and S-2 in Fig. 4.11 shows that the transitions between *10:CPD4* (Solution Evaluation-Checking) and *11:CPD5 Solution* (Evaluation-Critique) in both directions being dominated by group S-1 (i.e. the more verbose group). It seems that this back-and-forth of checking and critique may account for the longer dialogue durations of S-1. Two of the transitions most dominated by S-2 (successful concise

FIGURE 4.9: Hierarchical clustering dendrograms of **(A)** successful teams and **(B)** unsuccessful teams. The teams are reordered to show clearly the hierarchical structure. The vertical axis represents the ward distance between two clusters before merger. In both dendrograms, we see that there are two major clusters in each.

FIGURE 4.10: Histogram of dialogue durations of **(A)** successful vs. unsuccessful teams, **(B)** subgroups S-1 vs. S-2, and **(C)** subgroups U-1 vs. U-2. The histogram of all distribution pairs are plotted with identical bins with the overlapped regions shaded in an intermediate color.

subgroup) are from *10:CPD₄* (Solution Evaluation-Checking) and from *4:DPD₂* (Solution Generation-Concrete) both towards *9:CPD₃* (Solution Evaluation-Acquiescence). The quick paths to acquiescence may explain how these teams can so quickly finish their tasks.

Between S-1 (successful verbose subgroup) and U-1 (unsuccessful verbose subgroup), as drawn in Fig. 4.12, the *10:CPD₄* (Solution Evaluation-Checking) to *9:CPD₃* (Solution Evaluation-Acquiescence) path and the *11:CPD₅* (Solution Evaluation-Critique) to *10:CPD₄* (Solution Evaluation-Checking) path are more apparent in S-1. Interestingly, the *10:CPD₄* (Solution Evaluation-Checking) to *11:CPD₅* (Solution Evaluation-Critique) transition is significant and frequent in both groups but the reverse transition

FIGURE 4.11: Overlapped filtered transition graphs of S-1 and S-2 subgroups. A cutoff is applied at $w = 5.0$. Grey edges denote important transitions shared by both groups as in Fig. 4.5.

is only significant in S-1. It is possible that the ability to put to side an initial criticism and to continue with productive dialogue separates the successful sub-group S-1 from the unsuccessful sub-group U-1. The graphs for other sub-group combinations are given in the Appendix.

FIGURE 4.12: Overlapped filtered transition graphs of S-1 and U-1 subgroups. A cutoff is applied at $w = 6.0$. Grey edges denote important transitions shared by both groups as in Fig. 4.5.

4.6 Chapter Discussions Conclusions

In the work described in this chapter, we again highlighted the usefulness of network inference and analysis to visualize and understand dynamical processes in complex systems. In particular, we successfully used network inference, filtering, and analysis techniques to discern the key differences in dialogue dynamics among groups and subgroups of successful and unsuccessful problem-solving teams. We established dialogue transitions into solution acquiescence as a key indicator of problem-solving success in this data set. On the contrary, unsuccessful teams have highly generative premature closure dialogues leading to a number of undesirable dialogue types like those being disaffective or uncohesive. We further validated the usefulness of such network representations by showing that it can highlight the divergent dynamics among subgroups of similar-outcome teams. Using hierarchical clustering, we discovered that successful and unsuccessful student groups can be further split into two subgroups each. For instance, we found two major subgroups among successful student teams characterized by their verbosity — one group typically taking longer durations at task-completion than the other. Tellingly, the network comparison shows that verbose teams have a greater tendency to engage in criticism and discussion of a proposed solution whereas the less verbose teams are more likely to immediately acquiesce.

We developed two novel methodologies in network filtering and network analysis in this project. We introduced the *minimum preserved-reachability graph* (MPRG) which we think of as a natural extension of the MST for directed networks. In many processes, such as dialogue, loops may indicate important self-reinforcing processes that should not be categorically ignored like they are in the MST. MPRG retains a denser amount of information than the MST and is especially appropriate for small directed graphs like those in this chapter. We also proposed a measure of *functional betweenness* that, in our opinion, is more indicative of the amount of flow going through a node (dialogue type in our case) in an empirical system. We also pointed out that the traditional structural betweenness does not always agree with functional betweenness on how they rank node centralities. We revealed that the implicit assumptions in computing structural betweenness that flows always preferentially select the shortest path and that all paths, no matter its length, contribute equally to the structural betweenness score are not realistic. Functional betweenness, we think, is a more realistic measure. The analytical methods used here represent a step up in sophistication over the procedures described in Chapter 3 in which we restricted ourselves mainly to an ‘eyeball’ analysis of the filtered network of teaching practices.

In this work we focused on the excitatory correlations in dialogue transition. It is also possible to investigate the negative (i.e. inhibitory) correlations in the data by looking at

transition pairs with $RR < 1$. However, these two types of analysis have to be carried out separately. In the current representation of the transition networks, the highly weighted edges are those that have both high positive correlations and high transition frequencies (see Eq. 4.4). A network representation that shows the inhibitory relationships would require a different definition of edge weight — perhaps RR^{-1} so that strongest edges indicate the most negative correlations.

It also is hoped that the findings here can be eventually put to practical applications by education practitioners. For example, we have discovered that some unsuccessful teams tended to reach an impasse upon the critique of a solution while successful teams were able to put aside the disagreement and continue with productive dialogue. In essence there is a ‘gap’ in the transition network of the unsuccessful teams that impeded their success. This indicates that it might be fruitful for educators to focus on teaching students how to positively handle disagreements during collaborative tasks and bridge this ‘gap’ in the network.

Appendix

A4.1 CPS Task: Small Pyramids

In the ‘Small Pyramids’ task, student pairs have to collaborate to figure out the rules of a number pyramid. Each player has access to different pieces of information and must collaborate to solve the problem. One player can key in numbers at the bottom of the pyramid while the other can see the result at the top of the pyramid. Three sub-tasks have to be completed to consider the team successful: (i) to find the number in the black box when the red box is ‘6’; (ii) to find the number in the black box when the red box is ‘20’ (the players can only key in numbers less than ‘10’, so this requires a joint understanding of the number pyramid); and (iii) to solve a modified pyramid with fewer tiers. A screenshot of the task is shown in Fig. A4.1. Refer to [135] for more details.

Figure A4.1: Screenshot of the ‘Small Pyramids’ task. Top: View from player A. Bottom: View from player B.

A4.2 CPS Task: Warehouse

In the ‘Warehouse’ task, players combine to ‘secure’ a warehouse by correctly positioning security cameras to cover the warehouse using as few cameras as possible. Tall objects

in the warehouse limit the line of sight of certain camera positions. Similarly, players must also complete three sub-tasks to be considered successful. A screenshot of the task is shown in Fig. A4.2. Refer to [135] for more details.

Figure A4.2: Screenshot of the ‘Warehouse’ task. Top: View from player A. Bottom: View from player B.

A4.3 Derivation of RR Standard Deviation

The textbook derivation of the standard deviation of relative risk is reproduced here. We begin with a hypothetical cause i (the exposure) and effect j (the disease) and their joint probabilities under all permutation in the table below.

Counts		Effect j	
		Yes	No
Cause i	Yes	$N(i, j)/N = p_{11}$	$N(i, \neg j)/N = p_{12}$
	No	$N(\neg i, j)/N = p_{21}$	$N(\neg i, \neg j)/N = p_{22}$

N is the sum of all permutations. As defined, the relative risk is

$$RR = \frac{p(j|i)}{p(j|\neg i)} = \frac{\frac{p_{11}}{p_{11} + p_{12}}}{\frac{p_{21}}{p_{21} + p_{22}}} = \frac{p_{11} \cdot (p_{21} + p_{22})}{p_{21} \cdot (p_{11} + p_{12})}. \quad (4.7)$$

According to the multivariable delta method [137], the variance of $y = f(p_{11}, p_{12}, p_{21}, p_{22})$ is

$$\text{Var}(y) \approx \nabla f(p_{11}, p_{12}, p_{21}, p_{22}) \cdot \text{Cov}(p_{11}, p_{12}, p_{21}, p_{22}) \cdot \nabla f(p_{11}, p_{12}, p_{21}, p_{22})^T. \quad (4.8)$$

To estimate the variance of $\log(RR)$, we let f be

$$f = \log(RR) = [\log(p_{11}) + \log(p_{21} + p_{22}) - \log(p_{21}) - \log(p_{11} + p_{12})]. \quad (4.9)$$

Then, ∇f is given as

$$\nabla = \left(\frac{\partial f}{\partial p_{11}}, \frac{\partial f}{\partial p_{12}}, \frac{\partial f}{\partial p_{21}}, \frac{\partial f}{\partial p_{22}} \right) \quad (4.10)$$

$$= \left(\frac{p_{12}}{p_{11}^2 + p_{11}p_{12}}, -\frac{1}{p_{11} + p_{12}}, -\frac{p_{22}}{p_{21}^2 + p_{21}p_{22}}, \frac{1}{p_{21} + p_{22}} \right). \quad (4.11)$$

Also, the variance covariance matrix for a 4-category multinomial distribution is

$$\Sigma = \frac{1}{n} \begin{pmatrix} (1 - p_{11})p_{11} & -p_{11}p_{12} & -p_{11}p_{21} & -p_{11}p_{22} \\ -p_{11}p_{12} & (1 - p_{12})p_{12} & -p_{12}p_{21} & -p_{12}p_{22} \\ -p_{11}p_{21} & -p_{12}p_{21} & (1 - p_{21})p_{21} & -p_{21}p_{22} \\ -p_{11}p_{22} & -p_{12}p_{22} & -p_{21}p_{22} & (1 - p_{22})p_{22} \end{pmatrix}. \quad (4.12)$$

Performing the $\nabla f \Sigma \nabla f^T$ multiplication yields

$$\text{Var}(y) \approx \frac{1}{N} \times \left[-\frac{1}{p_{11} + p_{12}} + \frac{1}{p_{21}} - \frac{1}{p_{21} + p_{22}} + \frac{1}{p_{11}} \right]. \quad (4.13)$$

Recalling that $N(i, j)/N = p_{11}$, $N(i, \neg j)/N = p_{12}$, etc. and rearranging the expression, we retrieve Eq. 4.2.

A4.4 Hierarchical Clustering of CPS Teams

Let us make clear that we are applying hierarchical clustering to *data*, specifically vectors of data, instead of applying it to nodes in a network like is done in [20], for example. In our implementation of hierarchical clustering, we first, from the vector of dialogue type frequencies for team m (i.e. how many times is dialogue type i expressed in team m

given n total types)

$$\mathbf{f}^{(m)} = (f_1^{(m)}, f_2^{(m)}, \dots, f_n^{(m)}), \quad (4.14)$$

define a mean normalized rank vector $\mathbf{r}^{(m)}$ where each component i is calculated using

$$r_i^{(m)} = \frac{1}{2}(p(f_i^{(m)} \geq f_i) + p(f_i^{(m)} > f_i)) \quad (4.15)$$

where f_n denotes the frequencies of dialogue type n over all teams so $p(f_i^{(m)} \geq f_i)$ gives the proportion of f_i that is smaller than $f_i^{(m)}$. We use Ward’s criterion [138] which seeks to minimize the total within-cluster variances. Given 2 clusters X_A and X_B , the distance between them is the increase in the total within-cluster variance in the event that they are merged

$$\begin{aligned} d(A, B) = & \sum_{i \in X_A \cup X_B} \|\mathbf{r}^{(i)} - \overline{\mathbf{r}(X_A \cup X_B)}\|^2 \\ & - \sum_{i \in X_A} \|\mathbf{r}^{(i)} - \overline{\mathbf{r}(X_A)}\|^2 \\ & - \sum_{i \in X_B} \|\mathbf{r}^{(i)} - \overline{\mathbf{r}(X_B)}\|^2. \end{aligned} \quad (4.16)$$

The first term is the within-cluster variance of the merged set of X_A and X_B while the two terms subtracted are the within-cluster variances of X_A and X_B respectively.

The algorithm starts with every sample m in their own cluster X_m . It then finds the closest pair of clusters according to the distance criteria set out above and merges them, creating a new cluster. The search and merge operations are performed iteratively until all samples belong to the same cluster.

A4.5 Overlapped Sub-Group MPRG

Figure A4.3: Overlapped filtered transition graphs of S-1 and U-2 subgroups. A cutoff is applied at $w = 8.0$. U-2 transitions are in general much stronger than S-1. There is a very significant switching between *10:CPD4* (Solution Evaluation-Checking) and *11:CPD5* Solution (Evaluation-Critique) in U-2 teams.

Figure A4.4: Overlapped filtered transition graphs of S-2 and U-1 subgroups. A cutoff is applied at $w = 6.0$. The successful and less verbose sub-group transitions strongly into *9:CPD3* (Solution Evaluation-Acquiescence) from *10:CPD4* (Solution Evaluation-Checking) and *4:DPD2* (Solution Generation-Concrete).

Figure A4.5: Overlapped filtered transition graphs of S-2 and U-2 subgroups. Comparing both less-verbose groups, we see again that successful groups transition into *9:CPD3* (Solution Evaluation-Acquiescence).

Figure A4.6: Overlapped filtered transition graphs of U-1 and U-2 subgroups. The more verbose unsuccessful group tended to switch between *10:CPD4* (Solution Evaluation-Checking) and *11:CPD5* Solution (Evaluation-Critique) more often.

Chapter 5

Functional Shortcuts in Language Co-occurrence Networks

5.1 Introduction

In this chapter, we deal with data from the system of language — a complex system of a very different nature compared to those in the previous two chapters. Specifically, we are interested to know if the syntactic organization of natural language can be automatically detected from language networks created from data of real language use. To do so, we turn to the automatic detection of motifs in these networks. The motifs we define in this chapter are different from the sub-graph motifs of Milo *et al.* we described in Chapter 2 and from the sequential motifs of Chapter 3. This project was initiated by Prof. Luke Kang-Kwong, the doctoral co-supervisor of the thesis author, who hypothesized that constructions of Construction Grammar can be detected with a data-driven approach. This results of this work has been written up, sent for publication, and is currently (January 2017) under review. Most passages in this chapter are lifted verbatim from the paper.

The pioneering linguist de Saussure defined language as “a system of interdependent terms in which the value of each term results solely from the simultaneous presence of the others” [139]. Under this definition, it becomes reasonable to employ the framework of complex networks in linguistic studies. As we have elaborated in Chapter 2, the network approach is especially appropriate in analyzing the complex relationships among components in a complex system. In the case of language, networks relate words or other linguistic components within the context of semantic, syntactic, co-occurring, or other types of relationships. Semantic networks have been shown to possess small-world and scale-free properties [140, 141]. The syntactic relationship between linguistic

components has also been examined by building co-occurring word networks [142, 143] and with networks of syntactic dependencies [144, 145]. Such networks also display the same small-world and scale-free properties.

The works cited above have largely concentrated on the macro-structural properties of language networks. These global approaches, however, are unable to detect grammatical/syntactic structures in the networks. For instance, it has been found that measures like mean path length [146], mean degree [146, 147], and mean clustering coefficient [146, 147] are not significantly different in syntactic networks and non-syntactic networks derived from scrambled text. In fact, Zipf’s Law, the scale-free distribution of word frequencies [148] said to be responsible for the scale-free topology of language networks, can be derived from various language models that do not consider syntax at all [149–151]. An exception is noted in [147] which introduced a selectivity measure that distinguished shuffled text from real ones. Given that the existence of grammatical patterns is universal in all human languages, one expects that these patterns must manifest in localized regularities in the language network as detectable structures (motifs). In fact, since the network approach should be applicable for all languages, motif analysis, which we will describe later in detail as the unsupervised detection of such structural regularities in language networks, becomes a generalized process to characterize the grammar of any language.

Ultimately, the organizing principle of language exist to serve the purpose of efficient communication. Human languages are vastly more complex than any other forms of animal communication [152–155]. Tellingly, the vastness of the lexicon and the recursive application of structural patterns allow language an almost infinite creative potential in producing meaningful utterances [156, 157]. Despite the complexity and variety of language, the processes of creating and understanding expressions are accomplished remarkably rapidly [142]. This efficiency can be partially attributed to the small-world organization in both semantic and syntactic networks [140, 142, 158–160]. However, since we know that even non-syntactic networks are small-world as described in [146] and later in the results, we propose that the navigation of language networks is further optimized in the local structures of the network.

In this work, we adopt an unsupervised learning algorithm [161] to recursively extract motifs from word-to-word networks of English corpora. These motifs are then re-embedded back into the network as coarse-grained network shortcuts. We show that motif density is drastically different between syntactic networks and non-syntactic networks. We also give evidence that network motifs function efficiently as navigational shortcuts. Finally, we examine how low-level motifs are organized around stop words while high-level motifs correlate with more abstract structures.

5.2 Syntactic Network of Languages

There are two general approaches in constructing the syntactic networks of languages. The first is with a syntactic dependency network [144, 145, 158] obtained usually from corpora that have been manually annotated with dependency trees [162]. The second is through means of a co-occurrence network [143, 146] which models the linear ordering of words (or other linguistic units) [163] in a corpus. We focus on the latter technique because the former requires expert guidance in generating the dependency trees whether it is done manually or automatically (algorithms for parsing require supervised training data). It also precludes the analysis of very large or new corpora, esoteric languages with no existing training sets, as well as languages with unknown grammars.

The most straightforward method of constructing co-occurrence networks is to simply draw an edge (directed or undirected) for all pairs of neighbouring words (distance $d = 1$) in a corpus. With this, we risk missing out on interactions that take place at longer distances (i.e. $d \geq 2$). Fig. 5.1A illustrates a co-occurrence network converted from a small toy corpus of the following 4 sentences:

1. Elephants drink a lot of water.
2. He owns a lot of risky equities.
3. The students have a lot of homework.
4. The baby cries a lot in the morning.

The simplicity of constructing co-occurrence networks means that it can be applied for any language. Even artificial ones can be cast on this network as long as the language is ‘tokenizable’, that is, the corpus can be split into tokens of linguistic units and expressed linearly with tokens arranged one after another. However, keeping only short-range correlations while discarding the long-range ones produces an over-generalized network. In the toy corpus, the words *elephants* and *equities* do not appear in the same sentence nor should they do so in normal English usage. Yet, in the network in Fig. 5.1A, one can easily navigate along an unrestricted path p_{ur} from *elephants* to *equities* and produce a grammatical but unlikely phrase *elephants drink a lot of risky equities*. Clearly, not all possible paths in a co-occurrence network are accessed, or even allowed, in regular language use. In the later sections of this chapter, mentions of language/linguistic networks will refer to such co-occurrence networks.

To overcome this over-generalization, Solan *et al.* [161] proposed what they call a pseudograph which limits the number of allowable paths in the network. Instead of having

FIGURE 5.1: Toy co-occurrence network and pseudograph. (A) shows the co-occurrence network based on Sentences 1 to 4. We draw a directed edge between all adjacent word pairs. The path from *elephants* to *equities* is highlighted in red. The pseudograph of the same toy corpus is shown in (B). Each sentence is represented by a different colored path on the network. The box highlights the motif *a lot of*. It is marked by a fan-in of edges at the start (indicated by the sharp drop in P_L) and a fan-out at *of* (indicated by the sharp drop of P_R). A new node for the significant motif is created in (C) and the edges are routed through it. (D) identifies an equivalence class within a window $L_w = 4$.

edges that merely join pairs of words, edges in the pseudograph span multiple nodes (i.e. words). Multiple edges between nodes and loops are also allowed. In [161], each edge in the pseudograph connects all the words in a sentence in a linear order. Fig. 5.1B shows the pseudograph of the same toy corpus. Since allowable paths must trace the available edges, a restricted path p_r that leads from *elephants* to *equities* does not exist. Although the pseudograph does not lend itself well to traditional complex network analysis, the same work uses such a network to detect ordered patterns they called MEX motifs. These motifs appear as coherent bundles of edges spanning short sequences of words and are bookended at both ends with the convergence (fan-in) and divergence (fan-out) of edges. In the example in Fig. 5.1B, the sub-sequence *a lot of* represents such a motif. The procedure of motif detection is described in detail in Section 5.3.1.

The motif is then embedded back into the corpus as a single linguistic unit (i.e. a single node, see Fig. 5.1C), acting essentially as a shortcut by reducing network distances between words on opposite sides of the condensed phrase.

It is reasonable that linguistic networks contain shortcuts if we adopt the perspective that generating and deciphering language is an exercise in network navigation. When we speak, or listen to, a sentence we are probably subconsciously tracing a path from one linguistic item to another in the linguistic network. Yet, the speed at which humans process language leads us to believe that this network is not navigated at a purely word-to-word level as this would entail longer path lengths (≈ 20 words, the average length of a sentence) than can be stored in our working memories of 7 ± 2 objects [164, 165]. Shortcuts would allow us to traverse multiple nodes at a time, shortening the effective path lengths and utilizing less cognitive effort. We will demonstrate in this chapter that the motifs detected are not just effective but also efficient network shortcuts.

5.3 Methods

5.3.1 Motif Extraction Algorithm

The Motif EXtraction (MEX) algorithm was developed by Solan *et al.* [161] to automatically extract patterns and syntax from language corpora. Consider a corpus of N_{sents} sentences of varying lengths totaling N_{toks} tokens made up of N_{words} unique words. It is useful to visualize this corpus as a pseudograph where each word is represented by a node and each sentence is a directed edge or path going through multiple nodes. As described, motifs are coherent sub-sequences of nodes where a number of edges bundle up and are bookended with fan-ins and fan-outs of edges.

We define a path in this graph as $(e_i; e_j) = (e_{i+1}, e_{i+2}, \dots, e_j)$ where each e_k represents some word node w_l . The *extension probabilities* of such a path are

$$P_R(e_i; e_j) = \frac{l(e_i; e_j)}{l(e_i; e_{j-1})}, j > i \quad (5.1)$$

and

$$P_L(e_j; e_i) = \frac{l(e_i; e_j)}{l(e_{i+1}; e_j)}, j > i. \quad (5.2)$$

The function $l(e_i; e_j)$ returns the number of edges that traverse the sequence $(e_i, e_{i+1}, \dots, e_j)$. Therefore, the equation for $P_R(e_i; e_j)$ describes the probability that a sub-path $(e_i, e_{i+1}, \dots, e_{j-1})$ is found to extend rightward (or forward) to word e_j . Similarly, $P_L(e_j; e_i)$ measures the leftward (or backward) extension probability i.e. the probability that a sub-path

$(e_{i+1}, e_{i+2}, \dots, e_j)$ is preceded by word e_i . $P_R(e_i; e_i) = P_L(e_i; e_i)$ is simply the occurrence probability of e_i . As illustrated in Fig. 5.1B, the start and end of a significant motif are marked by sudden drops in their rightward and leftward extension probabilities. The *drop ratios* are defined as

$$D_R(e_i; e_j) = \frac{P_R(e_i; e_j)}{P_R(e_i; e_{j-1})}, j > i \quad (5.3)$$

and

$$D_L(e_j; e_i) = \frac{P_L(e_j; e_i)}{P_L(e_j; e_{i+1})}, j > i. \quad (5.4)$$

When $D < \eta$, the drop threshold parameter, we consider the sub-sequence a significant motif. A new pattern node $e_{patt} = [e_{i+1}, e_{i+2}, \dots, e_j]$ which contains the condensed sequence is created. Paths that previously went through $(e_{i+1}, e_{i+2}, \dots, e_j)$ are now routed through the e_{patt} pattern ‘supernode’ like in Fig. 5.1C. When there are overlapping significant sub-sequences, we prioritize the one with greater significance based on the smallest combined binomial probability $B_{comb} = B_R + B_L$ where

$$B_R(e_i; e_j) = \sum_{x=0}^{l(e_i; e_j)} \text{Binom}(l(e_i; e_{j-1}), x, \eta P(e_i; e_{j-1})), j > i \quad (5.5)$$

and

$$B_L(e_j; e_i) = \sum_{x=0}^{l(e_j; e_i)} \text{Binom}(l(e_{i+1}; e_j), x, \eta P(e_{i+1}; e_j)), j > i. \quad (5.6)$$

MEX can also accommodate *equivalence classes* of linguistic units that share the same context. Fig. 5.1D shows an example of an equivalence class contained within a context window of length $L_w = 4$. The slot for the equivalence class can exist at any position of a context window except at the edges. The equivalence class is then merged into an equivalence class ‘supernode’.

In our implementation of MEX, we consider the network formed from the original corpus of terminal words to be the 0-th level graph ($G^{(0)} = (V^{(0)}, E^{(0)})$). From this, we scan the graph first for all possible level 1 equivalence classes (i.e. the generalization step) and embed them to $G^{(0)}$, obtaining $G^{(1)-OnlyEC}$. We then seek for pattern candidates from all possible sub-sequences in $G^{(1)-OnlyEC}$ and rank the candidates by B_{comb} . The patterns are embedded into the network starting from the most significant and the process is repeated until no more $V_P^{(1)}$ nodes are found (i.e. the embedding step). This culminates in the level 1 network $G^{(1)}$. This ordering of steps enables the equivalence class detection and pattern detection procedures to bootstrap each other to create more equivalence classes and pattern motifs. The algorithm then moves on to construct $G^{(2)}$ and so on until no more new objects are identified.

5.3.2 Equal-Cost Null Model for Network Shortcuts

As MEX is performed on a linguistic network, new nodes acting as shortcuts are created and added to the network. This, obviously, has the effect of shrinking distances between word pairs. To determine if MEX-detected shortcuts are efficient in shrinking distances, we compare them against a null model where shortcuts are randomly created. Consider that when $G^{(k)}$ is derived from $G^{(k-1)}$ using MEX, we are adding $N(V_P^{(k)})$ nodes and $N(E_P^{(k)})$ edges to the network. We count only pattern nodes and edges because equivalence classes do not decrease network lengths. We hypothesize that these new network objects are created in the mental lexicon at the cost of

$$\Gamma(G^{(k)}, G^{(k-1)}) = \Gamma_{Node} N(V_P^{(k)}) + (1 - \Gamma_{Node}) N(E_P^{(k)}). \quad (5.7)$$

with $\Gamma_{Node}/(1 - \Gamma_{Node})$ being the cost ratio of node formation to link formation i.e. $\Gamma_{Edge} = 1 - \Gamma_{Node}$. We let $G^{(k)\dagger}$ be the null model. Starting with $G^{(0)\dagger} = G^{(0)}$, we randomly select sub-sequences in the network to be condensed into shortcuts in $G^{(0)\dagger}$ to obtain $G^{(1)\dagger}$. To ensure fairness, we set $\Gamma(G^{(k)\dagger}, G^{(k-1)\dagger}) = \Gamma(G^{(k)}, G^{(k-1)})$. Subsequent null model networks $G^{(k)\dagger}$ are derived from $G^{(k-1)\dagger}$.

5.3.3 Null Rate of Pattern Statistics

The null statistics of level k pattern motifs are the properties of the ensemble of patterns in $V_P^{(k)}$ if, instead of being detected with the MEX criteria, they were randomly defined. To compute this, we take $G^{(k)-OnlyEC}$, which is $G^{(k)}$ before patterns are embedded using MEX, and create the ensemble of null model patterns $V_P^{(k)*}$ by randomly selecting sub-sequences in this network. The number of patterns in each ensemble are kept equal (i.e. $N(V_P^{(k)*}) = N(V_P^{(k)})$).

5.4 Results

5.4.1 Motif Detection in Language Networks

We applied the Motif EXtraction algorithm (MEX) on three English corpora approximately 10^6 words in length. They are: (i) the Uppsala Student English Corpus (USEC) (1007839 tokens, 22471 words) [166] of essays collected from 1999 to 2001 from students at Uppsala University in Sweden taking English as a second language which represents a learner's corpus; (ii) the Brown Corpus (BC) (988331 tokens, 41018 words) [167] which

is compiled from samples of mainly professionally written American English text compiled in the 1960s to represent English usage at a high level of sophistication; and (iii) a single-author corpus (SAC) (757542 tokens, 14456 words) of Jane Austen’s writings freely available from Project Gutenberg [168] to represent also high proficiency but from a single user’s perspective. The drop threshold parameter is set $\eta = 0.65$ and the context window length is $L_w = 6$ following [161] which selected these criteria based on the optimal trade-off between precision and recall when tested against a corpus of context-free grammar.

MEX found the greatest number of patterns/motifs and equivalence classes (collectively termed objects) in the USEC. Fig. 5.2A details the results. For example, at level 1, we detected $(6466 + 10191)/22471 = 0.74$ as many objects as there were terminal words in the original corpus (i.e. the motif density). It is followed closely by the SAC with 0.64 while the BC lags substantially with 0.16. We define corpus compression by measuring the change in the number of tokens ΔN_{toks} remaining in the corpus at each embedding level. It is greatest for the USEC which at level 5 contains $794922/1007839 = 0.79$ as many tokens as it did at level 0 (i.e. the normalized token count). The SAC and the BC follow at 0.81 and 0.88 respectively. With other network measures (see Table 5.1), the BC also stands out with its very low clustering coefficient C and density ρ when compared to the other two real corpora (1.1 vs. 3.1 and 4.4, all $\times 10^{-2}$).

Corpus	$\rho \times 10^4$	$\langle k \rangle$	$C \times 10^2$	r	$\langle \min(d_{ur}) \rangle$	$\langle \min(d_r) \rangle$	$\langle \text{mean}(d_r) \rangle$
USEC	5.1	23.0	3.05	−0.21	3.08	8.97	11.03
SAC	9.7	28.1	4.41	−0.26	2.89	16.10	20.70
BC	2.3	19.2	1.10	−0.19	3.09	9.79	11.48
U-S	8.4	37.5	7.66	−0.24	2.96	8.37	10.57
U-PS	7.8	34.9	5.43	−0.24	2.85	8.71	10.62

TABLE 5.1: Table of network measures. USEC-S refers to the scrambled USE corpus and USEC-PS the POS-scrambled corpus. The measurements are density ρ , average degree $\langle k \rangle$, clustering coefficient C , assortativity r , average minimum unrestricted distances $\langle \min(d_{ur}) \rangle$, average minimum restricted distances $\langle \min(d_r) \rangle$, and average mean restricted distances $\langle \text{mean}(d_r) \rangle$. For U-S and U-PS, the values are given up to the precision not affected by fluctuations in the random scrambling.

We then tested MEX on artificially-generated corpora with little or no syntactic structures of real language. In Fig. 5.2B, the word order in each sentence of the USEC was shuffled to produce the *shuffled corpus* and the *POS-shuffled corpus* was generated by swapping words of the same Part-of-Speech (POS) category within the corpus. The POS categories for the USEC and SAC were determined using a Maximum Entropy tagging algorithm implemented through Python’s Natural Language Toolkit (NLTK) while the BC came with the POS tags included. As expected, only a negligible number of objects were identified (motif densities of 0.01 and 0.05 respectively compared to 0.79

FIGURE 5.2: Results of MEX embedding. The bar graphs in (A) and (B) show the motif densities for the real corpora (USEC, SAC, and BC) and the artificial corpora (USEC vs. shuffled and POS-shuffled) respectively. USE-S refers to the scrambled USE corpus and USE-PS the POS-scrambled corpus. The line plots in (A) and (B) chart the decrease in the number of tokens N_{toks}^k / N_{toks}^0 as more motifs are embedded in the network. (C) show how $\langle \text{mean}(d_r) \rangle$ decreases in the USEC and how it compares to null models set at different cost parameters Γ_{Node} . (D) shows relative the word frequencies $f(r)$ against rank r and highlights the stop word cutoff in each corpus.

of the original USEC) in the scrambled corpora. However, we noted that the scrambled corpora have remarkably higher C , ρ , and average degree $\langle k \rangle$ (Table 5.1). We also compared the in/out degree distributions k^{in}, k^{out} and in/out selectivity distributions e^{in}, e^{out} of the USEC in Fig. 5.3. As seen in Fig. 5.3(A) and (B), there is practically no difference in the degree distributions between the original corpus and its scrambled version. Selectivity of a node i , $e^{in/out}(i)$, is a measure given by $e^{in/out}(i) = \frac{s(i)}{k^{in/out}(i)}$ [147] where the strength $s(i)$ is the frequency of the word associated with the node i . As already known, the distributions of selectivity are distinctly different between real and scrambled corpora.

The MEX algorithm identified motifs from overlaps in the word-to-word pseudograph. The BC, being a corpus purposefully curated to include a diverse range of genres, subjects, and authorship to represent a wide range of American English usage, is designed to minimize the number of overlaps. Accordingly, the algorithm detected only a small number of motifs in this corpus. The USEC and the SAC, on the contrary, have much greater overlap densities. The USEC is built from students' essays written on a limited number of topics, with many essays per topic, guaranteeing content overlap. The

FIGURE 5.3: Cumulative distributions of degree and selectivity of the USEC. The in/out degree distributions are drawn in (A) and (B). The distribution of the original corpus (blue markers) follows the left-side vertical scale and the distribution of the scrambled corpus (red markers) is offset downwards by one decade and read against the right-side vertical scale. The dashed lines give the best-fit power law $y = kx^{-\beta}$ with β calculated using the method set out in [169]. Note that this are the exponents for cumulative distributions, so those of the actual distributions are $\gamma = \beta + 1$. Ferrer i Cancho and Solé argued that there could actually be two regimes in the distribution [142] but for simplicity we assume that the data can be fitted with a single power law. (C) and (D) give the cumulative distributions of inwards and outwards selectivity. The legend in subplot (B) applies for all subplots.

SAC, composed of only 8 books from Jane Austen, contains both content and stylistic overlaps. It is apparent that selecting a corpus with the appropriate overlap density is more crucial for MEX to detect motifs than it is to choose a large corpus. The network measures C , ρ , $\langle k \rangle$, and degree assortativity r proved to be unreliable in determining whether the word-to-word network contains MEX-detectable syntactic regularities. In real corpora, increases in C , ρ , and $\langle k \rangle$ seem to correlate loosely with a larger motif density. However, the correlation is completely reversed with the scrambled corpora in the sense that they register hardly any objects in MEX despite possessing higher C , ρ , and $\langle k \rangle$. This agrees with the results of [146] which found that $\langle \min(d_{ur}) \rangle$, $\langle k \rangle$, and C are slightly, but not significantly, higher for non-syntactic networks compared to their syntactic equivalents.

5.4.2 Network Distances

Network distances will invariably shrink when a sequence of words is condensed into a single-node shortcut as it does under MEX. We quantify the evolution of network distances under MEX and compare it with an *equal-cost null model* (see Section 5.3.2) to demonstrate that MEX motifs are efficient in shrinking network distances. We consider distances only along paths p_r already existing on the pseudograph. The reason is two-fold: (i) as stated, many of the unrestricted paths are not accessed, or even allowed, in regular language use and (ii) as shown in Table 5.1, unrestricted distances $\langle \min(d_{ur}) \rangle$ are much smaller ($\approx 1/3$) than observed distances $\langle \min(d_r) \rangle$ in the network.

The decrease of average mean network distances $\langle \text{mean}(d_r) \rangle$ are charted in Figs. 5.2C and A5.1(A-B) (see A5.1(C-E) in the Appendix for $\langle \min(d_r) \rangle$). At level 5, the maximal layer of embedding, $\langle \text{mean}(d_r) \rangle$ for the USEC, SAC, and BC are respectively $8.571/11.027 = 0.777$, $16.637/20.696 = 0.804$, and $9.936/11.481 = 0.865$ of the values of $\langle \text{mean}(d_r) \rangle$ at level 0 (i.e. the normalized distances). Regardless of the node-formation cost Γ_{Node} , the distances of the MEX-embedded networks are always smaller than the null models. This is especially true when Γ_{Node} of the null models are set larger than 0.5. Null model distances shrink and approach the MEX results when $\Gamma_{Node} < 0.5$. For $\langle \min(d_r) \rangle$, null models with $\Gamma_{Node} < 0.5$ sometimes even surpass MEX in distance reduction (e.g. in SAC and BC). Selecting a realistic Γ_{Node} is not trivial. In this, one may be guided by empirically calculating the ratio of new nodes to new edges created at each MEX level k and set $\Gamma_{Node}N(V_P^{(k)}) = \Gamma_{Edge}N(E_P^{(k)})$. This assumes that equal cost is expended in node and edge creation. This yields Γ_{Node} between 0.82 to 0.97 which is well within the range where MEX shrinks the network more than the null model in both $\langle \text{mean}(d_r) \rangle$ and $\langle \min(d_r) \rangle$. The choice of $\Gamma_{Node}/\Gamma_{Edge} > 0.5$ can also be motivated from neuropsychology — the creation of new memory that can be associated to previously acquired knowledge (i.e. edges) requires less effort than remembering a new and isolated piece of information (i.e. nodes) [170].

5.4.3 Taxonomy of MEX Patterns

We also attempted to create a taxonomy of MEX-detected patterns. We use the POS tags attached to each token to create *POS templates* of MEX patterns. For example, the phrase *the united states* would have the template *DT (Determiner) JJ (Adjective) NN (Noun)* as would the phrase *a short man*. Significant POS templates are singled out by how their standard scores (z-scores) measure against the null ensemble of patterns $V_P^{(k)*}$. The most significant POS templates at the first level is *DT (Determiner) NN (Noun) IN (Preposition)* (e.g. *a lot of*) with $z = 56.75$ and frequency $f = 371$. The next is *IN*

(Preposition) *DT* (Determiner) (e.g. *in the*) (see Table 5.2 for results of the USEC and Tables A5.1 and A5.2 for the SAC and the BC). Additionally, the 6466 MEX pattern motifs at level 1 of the USEC can be represented by 1190 POS templates whereas the null model required twice the number of template types (2290). The relatively small set of POS templates used suggests that motifs at this level are in part characterized by a small class of grammatical/syntactic regularities. Significant templates are also detected in $G^{(2)}$ such as $[[IN\ DT]\ JJ\ NN]$. There are no significant ($z > 3$) POS templates beyond $G^{(2)}$.

Rank	Template	Z	F	Example
Level 1 POS Templates: 6466 Motifs; 1190 Types; 2290 Null Types				
1	[DT NN IN]	56.75	371	[a lot of]
2	[IN DT]	41.70	434	[in the]
3	[DT VBZ]	34.00	81	[this is]
Level 2 POS Templates: 1662 Motifs; 1312 Types; 1442 Null Types				
1	[[IN DT] NN [IN DT]]	12.97	6	[[as a] {consequence,result} [of this]]
2	[[IN DT] JJ NN]	12.04	22	[[in the] first place]
3	[[IN DT] NN NN]	7.43	11	[[in the] right way]
Level 1 Stop Word Templates: 4038 Motifs; 83 Types; 98 Null Types				
1	_ the _ of _	39.77	227	[the age of]
2	_ a _ of _	26.54	110	[a lot of]
3	_ the _ to _	19.46	58	[the right to]
Level 2 Stop Word Templates: 353 Motifs; 24 Types; 38 Null Types				
1	_ of _ to _	7.88	3	[[in the] ages of {15,16...} to]
2	_ of _	3.81	47	[[as a] matter of fact]
3	_ to _ to _	3.37	2	[to {quick,young} to]

TABLE 5.2: Motif templates of the USEC. We present for levels 1 and 2 the top 3 ranked (in terms of Z-score) POS templates and stop word templates respectively. The Penn Treebank [171] POS tags are used here. Beyond level 2, there are no regularly used templates. For each level, we also gave the number of template types used by the motifs in the corpus and also the number of template types that appeared in randomly extracted motifs.

Classification templates for patterns can also be constructed from *stop words* [172] which in computational linguistics refer to very common closed-class words such as prepositions, conjunctions, and pronouns with little semantic value. While the determination of POS tags demands expert input, *stop word templates* can be constructed from a purely data-driven approach. The list of stop words is not universal and varies among natural language processing tools. We can, however, adopt an operational definition of stop words from empirical data. In Fig. 5.2D, we plotted the ranked relative word frequencies $f(r)$ of each corpus on log-log plots and observed that there are sudden drops in frequency after $r = 8, 4, 6$ for respectively the USEC, SAC, and BC. This boundary marks a conservative rank cutoff r_{cutoff} to define an operational list of stop words for a particular corpus. It is entirely possible that stop words (in the closed-class definition

sense) exist beyond these cutoffs but are not reflected in the trend of the data and thus cannot be unambiguously identified. For example, stop words like ‘the’ and ‘to’ are common for all three corpora and others like ‘is’ is USEC-specific.

We employ these stop words to construct *stop word templates*. Stop word templates are defined by replacing the non-stop words in a motif with a blank. For example, the stop word template of ‘the {reason,explanation} for’ is ‘_ the _ for _’ with the inclusion of blanks also at the front and back. Of the 6466 MEX patterns at level 1 of the USEC, 4038 contain at least one stop word in them and can be classified by a template (Table 5.2). At $G^{(1)}$ of the USEC, we find significant templates in ‘_ the _ of _’ (e.g. ‘the age of’) ($z = 39.77$, $f = 227$) and ‘_ a _ of _’ (e.g. ‘a lot of’) ($z = 39.77$, $f = 227$). Similar templates are also found in the other two corpora (see Tables A5.1 and A5.2). Beyond $G^{(1)}$, the significances and frequencies of stop word templates fall drastically with very few registering $z > 3$. This suggests that at $G^{(>1)}$, motifs stop being organized around these operational lists of stop words. We also observe that at the first level $G^{(1)}$, stop words appear in MEX-detected motifs 58 to 83% more frequently than by chance (see Table A5.3). The trend is inverted at higher levels $G^{(>1)}$ as the rates of stop word usage decrease below that of random noise (e.g. at level 2, 16 to 30% below null rate). Although, stop word templates cannot classify patterns as precisely as POS templates (fewer possible template types), they are nevertheless an unsupervised alternative.

5.5 Discussion

In linguistics, the theories of Construction Grammar (CxG) describe a family of grammar models in which constructions are the basic units of grammar [173–175]. Constructions, in this framework, are *pairings of form and function* [174, 175]. Form refers to the morphological and syntactic features of the construction while function specifies its semantic features [176]. Consider the idiom *kick the bucket* [174]. One cannot derive its meaning by considering its syntax (the ‘X does something to Y’ form) and its component words separately. In CxG, not only idiomatic expressions, but all constructions are said to be form-function pairings. There is also “a continuum from schematic complex constructions to substantive atomic constructions” [177]. Isolated words and complex grammatical arrangements are treated on equal footing in CxG.

One of the processes through which children acquire constructions is entrenchment [178]. Entrenchment is when a person performs a task successfully enough times, the way they perform this task becomes habitual and automatic. Similarly, MEX detect motifs in language pseudographs by searching for repeated sub-sequences. They are reminiscent of constructions in CxG. Take, for example, the generalized pattern $[i \{feel, think, believe\}$

that] found in the USEC. All three variants of the pattern are identical in their functions of expressing an opinion. Even though it shares an identical form with a pattern like [*i suggest that*], MEX considers them to be different entities due to their usage in different contexts. Such form-function pairings in MEX motifs are features found also in CxG. Of course, it would be presumptuous to claim that all motifs in MEX are constructions, or that all constructions in the corpus are detectable with MEX. Nevertheless, MEX remains a useful tool to quickly and crudely process a corpus for possible constructions.

Frequently used constructions become entrenched in the mental grammar [176]. These constructions become mental routines directly accessed without invoking higher schemas [176]. This is consistent with our earlier claim that the language network has to contain shortcuts to account for how rapidly we navigate it. In MEX, significant motifs are similarly entrenched in the network as condensed routines. As shown, motifs reduce observed network distances d_r effectively and efficiently. We believe that the creation of such shortcuts in the mental lexicon comes at the expense of using more memory to store the new linguistic units and the new associations that come with it. This cost, modelled in Γ , is offset by the reduction of network distances which we associate with an increase in language processing speed.

We can also speculate on the mechanisms that guide the formation of constructions by investigating the properties of MEX motifs. We showed that incidence rates of stop words in level-1 motifs are higher than expected from the frequencies of the stop words themselves. In syntactic networks of natural languages, stop words are hubs owing to their high usage frequency and their flexible combinatorial potential [159]. It has been shown that stop words only become hubs in the syntax network of a child after the *syntactic spurt* [179] stage of language acquisition [158]. If we assume that level-1 motifs correspond to constructions acquired in early stages of language development, this study lends credence to our observation that stop words are important building blocks of low-level motifs.

In linguistics, stop words are called function words [180]. They have little lexical meaning and, instead, serve to mediate and/or to emphasize the interaction between words. We believe that function words arise necessarily out of cognitive limitations. Before they undergo the syntactic spurt, toddlers have small one-word lexicons which can be easily searched through when producing or deciphering utterances. As the lexicons grow, and as the children start to produce 2-to-3 word phrases, the numbers of possible expressions explode combinatorially. The same linear search strategy becomes too slow for real-time linguistic interactions. Function words are constrained by the word category preceding them and they in turn constrain the word category following them. For example, a toddler may say ‘put table/bed/box’ to mean ‘put the toy on the table/bed’ or ‘in the

box’. With the introduction of a function word like *on*, after saying ‘put on’, it is much more probable to produce *table* or *bed* than *box*. As such, the function word *on* reduces the search space of the possible productions after it.

5.6 Chapter Conclusions

The work described in this chapter applied network techniques to investigate the organization of syntax in the co-occurrence pseudographs of English corpora. We showed that the co-occurrence networks of original texts and their scrambled equivalents are remarkably similar in many quantitative measures such as density ρ , assortativity r , and average path lengths $\langle \min(d_{ur}) \rangle$, suggesting that these measures do not capture the microscopic and mesoscopic structural regularities of grammatical syntax in the networks. The density of sequential motifs in these networks as detected by the MEX algorithm, though, are very different by almost two orders of magnitude — obviously, real networks have many more coherent motifs of language use than scrambled networks. Moreover, we demonstrate how these motifs act as functional shortcuts by shrinking word-to-word distances in the network effectively and efficiently, potentially explaining why we are able to generate and decipher language expressions so rapidly. We also suggested a link between the automatically discovered motifs and constructions in the theories of Construction Grammar and speculate on how these constructions are acquired through entrenchment and are necessary in our mental lexicon to speed up language processing capabilities.

Appendix

54.1 Decrease of Network Distances

Figure A5.1: Decrease of network distances. **A** and **B** shows the decrease of $\langle \text{mean}(d_r) \rangle$ in the SAC and BC and how they compare to null models set at different cost parameters. Sub plots **(C-E)** chart the decrease of $\langle \text{min}(d_r) \rangle$ and their null models for the USEC, SAC, and BC respectively.

A5.2 SAC Templates

Rank	Template	Z	F	Example
Level 1 POS Templates: 3522 Motifs; 647 Types; 1392 Null Types				
1	[DT NN IN]	57.58	301	[the {exigence,air,...} {for,of}]
2	[IN RB IN]	27.56	34	[as soon as]
3	[PRP VBD]	25.22	174	[it was]
Level 2 POS Templates: 781 Motifs; 665 Types; 718 Null Types				
1	[[IN DT] NN]	11.33	18	[[in the] world]
2	[[IN DT] JJ NN]	9.17	9	[[in the] first place]
3	[[IN DT] NN IN]	6.87	11	[[in the] habit of]
Level 1 Stop Word Templates: 1334 Motifs; 20 Types; 23 Null Types				
1	_ the _ of _	72.27	294	[the sight of]
2	_ to _ to _	9.59	15	[to speak to]
3	_ to _ of _	9.27	11	[to think of]
Level 2 Stop Word Templates: 122 Motifs; 5 Types; 11 Null Types				
1	_ of _	3.21	36	[[with a] look of]
2	_ and _ to _	0.86	1	-
3	_ the _	0.49	21	-

Table A5.1: Motif templates of the SAC. We present for levels 1 and 2 the top 3 ranked (in terms of Z-score) POS templates and stop word templates respectively in a similar manner to Table 5.2.

A5.3 BC Templates

Rank	Template	Z	F	Example
Level 1 POS Templates: 2472 Motifs; 762 Types; 1502 Null Types				
1	[AT NN IN]	85.09	319	[a number of]
2	[PPS BEZ]	31.95	42	[it is]
3	[IN AT]	30.40	158	[of the]
Level 2 POS Templates: 513 Motifs; 391 Types; 481 Null Types				
1	[IN NN [IN AT]]	26.39	12	[in spite [of the]]
2	[[IN AT] NN IN]	26.35	21	[[on the] basis of]
3	[[IN AT] NN [IN AT]]	12.02	8	[[at the] end [of the]]
Level 1 Stop Word Templates: 1625 Motifs; 44 Types; 44 Null Types				
1	_ the _ of _	66.66	307	[the number of]
2	_ a _ of _	33.17	75	[a number of]
3	_ in _ of _	16.92	14	[in terms of]
Level 2 Stop Word Templates: 154 Motifs; 12 Types; 15 Null Types				
1	_ of _	2.86	38	-
2	_ to _ the _	2.59	4	-
3	_ in _ of _	2.05	1	-

Table A5.2: Motif templates of the BC. We present for levels 1 and 2 the top 3 ranked (in terms of Z-score) POS templates and stop word templates respectively in a similar manner to Table 5.2. It uses a different list of POS tags obtainable from [32].

A5.4 Pattern Statistics

Lv	Cr	$\langle N_{len(P)=2} \rangle$	$\langle N_{len(P)>2} \rangle$	$F_{SWinP} \times 10$	$F_{CinP} \times 10$	$F_{PinP} \times 10$
1	U	44.3 (1.3 $\pm$ 0.3)	9.4 (6.3 $\pm$ 0.1)	2.9 (1.1 $\pm$ 0.0)	2.5 (1.9 $\pm$ 0.0)	-
	S	60.0 (10.0 $\pm$ 1.0)	8.3 (5.5 $\pm$ 0.3)	1.9 (0.7 $\pm$ 0.0)	2.4 (2.1 $\pm$ 0.0)	-
	B	98.5 (35.2 $\pm$ 2.4)	7.4 (5.3 $\pm$ 0.3)	3.3 (1.5 $\pm$ 0.1)	1.2 (1.1 $\pm$ 0.0)	-
2	U	22.8 (14.0 $\pm$ 0.3)	4.9 (3.6 $\pm$ 0.1)	0.7 (-0.3 $\pm$ 0.1)	2.3 (1.9 $\pm$ 0.0)	4.2 (1.9 $\pm$ 0.1)
	S	27.9 (19.0 $\pm$ 1.7)	4.8 (3.6 $\pm$ 0.0)	0.5 (-0.2 $\pm$ 0.0)	2.5 (2.3 $\pm$ 0.0)	4.3 (2.2 $\pm$ 0.1)
	B	37.5 (29.0 $\pm$ 1.7)	5.3 (4.2 $\pm$ 0.1)	1.0 (-0.2 $\pm$ 0.1)	1.9 (1.8 $\pm$ 0.0)	3.9 (2.5 $\pm$ 0.1)
3	U	9.3 (3.8 $\pm$ 1.0)	4.3 (3.2 $\pm$ 0.0)	0.6 (-0.5 $\pm$ 0.1)	2.0 (1.7 $\pm$ 0.0)	5.2 (2.9 $\pm$ 0.2)
	S	8.5 (2.2 $\pm$ 1.5)	3.7 (2.5 $\pm$ 0.1)	0.5 (-0.2 $\pm$ 0.1)	3.4 (3.2 $\pm$ 0.1)	4.7 (3.0 $\pm$ 0.2)
	B	17.5 (10.7 $\pm$ 1.3)	3.7 (2.5 $\pm$ 0.1)	1.0 (-0.1 $\pm$ 0.1)	2.6 (2.4 $\pm$ 0.1)	4.7 (3.4 $\pm$ 0.2)

Table 5.3: Table of pattern statistics. Here we show the properties of the patterns extracted by MEX for the first 3 levels. $\langle N_{len(P)=2} \rangle$ gives the mean occurrence frequency of length-2 patterns and $\langle N_{len(P)>2} \rangle$ is for patterns with lengths greater than 2. F_{SWinP} , F_{CinP} , and F_{PinP} are the proportions of objects in the patterns that are stop words, classes, and lower-level patterns respectively. The values are presented with the difference between the observed values and the null values in parentheses together with the error margin. For example, $\langle N_{len(P)=2} \rangle$ for the USEC at level 1 is 44.3 and the null model yields $44.3 - 1.3 = 43.0$ with an error margin of ± 0.3 .

Chapter 6

Conclusions

In this thesis we presented three examples of applying network construction and analysis techniques to interpret, represent, and reveal the internal organizations of real-world data. Network approaches, as explained, are versatile analytical tools that allow us to adopt a holistic perspective when looking at complex systems. Aspects of such systems like hierarchy, flows, communities, and causality can all be defined in terms of network topologies. In utilizing these techniques on empirical data, we hoped to be able to create networks that represent crude reconstructions of the underlying complex systems that produced the data and from the features of these networks glean useful insights into the dynamics and properties of these systems.

Chapter 3 detailed our first attempt at the network analysis of empirical data where we used directed networks to represent and understand the dynamical processes of teaching and learning. We inferred networks from high-dimensional, multi-scale, and time-resolved data of teaching practices in actual classroom settings. These networks represented important time-ordered transitions among different types of teaching practices. The method of network inference we developed in this chapter is unique as it takes into consideration varying characteristic durations of different teaching events to define event transitions. Moreover, to test such transitions for significance, we compared observed transition frequencies to synthetic ones using a Monte Carlo sampling method. Beyond its utility in Chapter 3's work, we believe that this 'first-activation' definition of transitions can be extended to the network analysis of other time-continuous processes that output such binary (events are either present or not) and discrete-time data like in high-throughput gene regulatory and other biological signalling experiments.

We also demonstrated in Chapter 3 how network filtering, specifically the MST and PMFG, can help to simplify extracted networks and highlight the most important connections in these systems. Such representations render the complex data of this project

much more accessible for analysis. From these reduced networks, we could easily identify the dominant transitions between teaching practices and also establish hubs of inward and outward transitions. Whereas the outward ‘source’ hubs corresponding with highly frequent teaching practices that exemplifies the transmissionist model of teaching, the inward ‘sink’ hubs we found, like students’ individual work, do not stand out in how often they appear during actual lessons and their special role can only be detected with this time-resolved network analysis.

Chapter 4 continued with the modelling of dynamical processes in the realm of education. In this chapter we inferred transition networks from the temporal production of ‘collective creativity’ dialogue types when students solve problems collaboratively. The analytical methods used here represent a step up in sophistication over the procedures described in Chapter 3 in which we restricted ourselves mainly to an ‘eyeball’ analysis of the filtered networks of teaching practices. Moreover, the data in this chapter contains an additional dimension of interest because it distinguishes between the data obtained from successful and unsuccessful problem-solving teams, allowing us to perform network-to-network comparisons of the different outcomes.

The inference methodology used in this chapter is different from Chapter 3 as we assumed that the process of dialogue alternates turn-by-turn rather than proceeding in a time-continuous fashion. Also, we adopted a novel topological filtering method termed the *minimum preserved-reachability graph* which we believe is more suited for filtering directed networks than the MST and PMFG used in Chapter 3. We also proposed a measure of *functional betweenness* that, in our opinion, is more indicative of the amount of actual flow going through a node (dialogue type in our case) in an real dynamical system.

The results showed that dialogue transitions into solution acquiescence is a key indicator of problem-solving success. On the contrary, unsuccessful teams have highly generative premature closure dialogues leading to more undesirable dialogue types. We also showed that traditional structural betweenness does not always agree with functional betweenness on how central certain nodes are. Particularly, the implicit assumptions that flows always preferentially select the shortest path and that all paths, no matter its length, contributes equally in computing structural betweenness are not always true in real systems. Functional betweenness, we think, is a more realistic measure.

We deal with a fundamentally very different system in Chapter 5 where we considered the problem of how we can examine the syntactic organization in natural language with networks. We showed that language networks are typified by the presence of coherent motifs which can be interpreted as regular patterns of language use and that real networks have many more such motifs than their randomized equivalents. These motifs act

as functional shortcuts by shrinking word-to-word distances in the network effectively and efficiently. It potentially explains why we are able to generate and decipher language expressions so rapidly. Motifs also seem to take on the properties of grammatical units called constructions as proposed by the theories of Construction Grammar. In this chapter we also speculated on the possible mechanisms behind the emergence of constructions based on the properties of the motifs we found. Potentially, motif detection can become a generalized process to characterize the grammar of any language, even unknown ones.

The network inference and analysis methods developed over the course of this thesis also have a number of applications outside those already described. In recent years, Internet memes [181] have rapidly gained popularity as a medium to transmit jokes, opinions, slogans, etc. Memes propagate among Internet users not only within platforms (e.g. Internet message boards like Reddit.com or social networking websites like Facebook) but also do so *across* platforms. It would be possible to track the temporal propagation of memes on these platforms and construct a functional network of meme (and by extension information) transmission between these platforms by using the network inference methods described in Chapter 3. This network would reveal, for example, which platforms are the most influential hubs and which ones have the highest betweenness scores which suggest the propensity to receive and pass on information. We may even use the motif discovery algorithms of Chapter 5 to find any patterned ‘transmission highways’ in the propagation of Internet memes. One can also imagine constructing functional gene regulatory networks from high-throughput data using the network inference techniques introduced in this thesis. This will probably require us to adopt the simplified picture that individual genes have Boolean states of being either ‘on’ or ‘off’ [30]. These dynamic sequences of gene activations can also be processed for motif-like activation sequences.

In conclusion, this thesis presented the methods we developed and the results we obtained from the application of network analysis to data of real-world systems. We defined new methods to infer transitions from time-continuous and turn-based processes in Chapters 3 and 4 respectively. Also, in Chapter 4, we introduced a novel method of network filtering procedure for directed networks and a definition of functional betweenness. Chapter 5 described the procedure to identify coherent motifs. We learnt in Chapters 3 and 4 that network analysis can reveal generative event hubs and motifs of event chains that cannot be easily seen with conventional methods and in Chapter 5 that coherent motifs in language networks are probably due to the syntactic organizations in language.

Bibliography

- [1] E. Bonabeau, Social insect colonies as complex adaptive systems. *Ecosystems* **1**, 437–443 (1998).
- [2] L. Steels, *International Conference on Parallel Problem Solving from Nature* (Springer, 2000), pp. 17–26.
- [3] R. J. Lempert, A new decision sciences for complex systems. *Proceedings of the National Academy of Sciences* **99**, 7309–7313 (2002).
- [4] R. Descartes, *A discourse on method: Meditations and principles* (transl. J. Veitch, JM Dent and Sons, 1637).
- [5] N. R. Franks, Army ants: a collective intelligence. *American Scientist* **77**, 138–145 (1989).
- [6] B. Hölldobler, E. O. Wilson, *The ants* (Harvard University Press, 1990).
- [7] S. Higashi, K. Yamauchi, Influence of a supercolonial ant *Formica* (*Formica*) *yessensis* Forel on the distribution of other ants in Ishikari Coast [Japan]. *Japanese Journal of Ecology* (1979).
- [8] M. Bollazzi, F. Roces, To build or not to build: circulating dry air organizes collective building for climate control in the leaf-cutting ant *acromyrmex ambiguus*. *Animal Behaviour* **74**, 1349–1355 (2007).
- [9] W. Weaver, Science and complexity. *American Scientist* **36**, 536–567 (1948).
- [10] M. Mitchell, Complex systems: Network thinking. *Artificial Intelligence* **17**, 1194–1212 (2006).
- [11] N. F. Johnson, *Two’s company, three is complexity: A simple guide to the science of all sciences* (Oneworld Publications Ltd, 2007).
- [12] M. J. Mauboussin, Revisiting market efficiency: The stock market as a complex adaptive system. *Journal of Applied Corporate Finance* **14**, 47–55 (2002).

- [13] J. H. Holland, *Hidden order: How adaptation builds complexity* (Basic Books, 1995).
- [14] M. Gell-man, Thinking about the future - the big picture (2005). Lecture by Murray Gell-man at The Frederick S. Pardee Center for the Study of the Longer-Range Future, Boston University [Accessed: 7/12/16].
- [15] D. J. Watts, S. H. Strogatz, Collective dynamics of ‘small-world’ networks. *nature* **393**, 440–442 (1998).
- [16] A.-L. Barabási, R. Albert, Emergence of scaling in random networks. *science* **286**, 509–512 (1999).
- [17] M. E. Newman, Scientific collaboration networks. i. network construction and fundamental results. *Physical review E* **64**, 016131 (2001).
- [18] M. E. Newman, Scientific collaboration networks. ii. shortest paths, weighted networks, and centrality. *Physical review E* **64**, 016132 (2001).
- [19] R. Albert, H. Jeong, A.-L. Barabási, Error and attack tolerance of complex networks. *Nature* **406**, 378–382 (2000).
- [20] R. N. Mantegna, Hierarchical structure in financial markets. *The European Physical Journal B-Condensed Matter and Complex Systems* **11**, 193–197 (1999).
- [21] J.-P. Onnela, A. Chakraborti, K. Kaski, J. Kertesz, A. Kanto, Asset trees and asset graphs in financial markets. *Physica Scripta* **2003**, 48 (2003).
- [22] T. Mizuno, H. Takayasu, M. Takayasu, Correlation networks among currencies. *Physica A: Statistical Mechanics and its Applications* **364**, 336–342 (2006).
- [23] J. Moody, Race, school integration, and friendship segregation in america. *American journal of Sociology* **107**, 679–716 (2001).
- [24] M. Girvan, M. E. Newman, Community structure in social and biological networks. *Proceedings of the national academy of sciences* **99**, 7821–7826 (2002).
- [25] W. W. Zachary, An information flow model for conflict and fission in small groups. *Journal of anthropological research* pp. 452–473 (1977).
- [26] O. Sporns, C. J. Honey, R. Kötter, Identification and classification of hubs in brain networks. *PloS one* **2**, e1049 (2007).
- [27] R. Milo, *et al.*, Network motifs: simple building blocks of complex networks. *Science* **298**, 824–827 (2002).

- [28] M. Faloutsos, P. Faloutsos, C. Faloutsos, On power-law relationships of the internet topology. *SIGCOMM Comput. Commun. Rev.* **29**, 251–262 (1999).
- [29] R. Guimera, S. Mossa, A. Turtshi, L. N. Amaral, The worldwide air transportation network: Anomalous centrality, community structure, and cities' global roles. *Proceedings of the National Academy of Sciences* **102**, 7794–7799 (2005).
- [30] S. A. Kauffman, *The origins of order: Self organization and selection in evolution* (Oxford University Press, USA, 1993).
- [31] S. Epskamp, *et al.*, qgraph: Network visualizations of relationships in psychometric data. *Journal of Statistical Software* **48**, 1–18 (2012).
- [32] S. A. Cheong, A. Nanetti, M. Philippov, Digital maps and automatic narratives for the interactive global histories. *The Asian review of World Histories* **4**, 83–123 (2016).
- [33] L. A. N. Amaral, A. Scala, M. Barthelemy, H. E. Stanley, Classes of small-world networks. *Proceedings of the national academy of sciences* **97**, 11149–11152 (2000).
- [34] M. E. Newman, The structure and function of complex networks. *SIAM review* **45**, 167–256 (2003).
- [35] M. Newman, *Networks: an introduction* (Oxford university press, 2010).
- [36] A.-L. Barabási, J. Frangos, *Linked: the new science of networks science of networks* (Basic Books, 2002).
- [37] D. J. Watts, *Six degrees: The science of a connected age* (WW Norton & Company, 2004).
- [38] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, D.-U. Hwang, Complex networks: Structure and dynamics. *Physics reports* **424**, 175–308 (2006).
- [39] E. Estrada, *The structure of complex networks: theory and applications* (Oxford University Press, 2012).
- [40] B. Bollobás, *Modern graph theory*, vol. 184 (Springer-Verlag, New York, 1998).
- [41] L. Euler, Solutio problematis ad geometriam situs pertinentis. *Commentarii academiae scientiarum Petropolitanae* **8**, 128–140 (1741).
- [42] E. Bullmore, O. Sporns, Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature Reviews Neuroscience* **10**, 186–198 (2009).

- [43] O. Sporns, D. R. Chialvo, M. Kaiser, C. C. Hilgetag, Organization, development and function of complex brain networks. *Trends in cognitive sciences* **8**, 418–425 (2004).
- [44] D. J. Watts, Networks, dynamics, and the small-world phenomenon. *American Journal of sociology* **105**, 493–527 (1999).
- [45] P. Sen, *et al.*, Small-world properties of the Indian railway network. *Physical Review E* **67**, 036106 (2003).
- [46] J. G. White, E. Southgate, J. N. Thomson, S. Brenner, The structure of the nervous system of the nematode *Caenorhabditis elegans*. *Philos Trans R Soc Lond B Biol Sci* **314**, 1–340 (1986).
- [47] A. Rapoport, W. J. Horvath, A study of a large sociogram. *Behavioral Science* **6**, 279–291 (1961).
- [48] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, B. Bhattacharjee, *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement* (ACM, 2007), pp. 29–42.
- [49] L. Egghe, R. Rousseau, *Introduction to informetrics: Quantitative methods in library, documentation and information science* (Elsevier Science Publishers, 1990).
- [50] B. A. Huberman, The laws of the web (2001).
- [51] S. L. Pimm, J. H. Lawton, J. E. Cohen, Food web patterns and their consequences. *Nature* **350**, 669–674 (1991).
- [52] A. Davis, B. B. Gardner, M. R. Gardner, *Deep South: A social anthropological study of caste and class* (Univ of South Carolina Press, 1941).
- [53] M. E. Newman, Coauthorship networks and patterns of scientific collaboration. *Proceedings of the national academy of sciences* **101**, 5200–5205 (2004).
- [54] W. Feller, *An introduction to probability theory and its applications: volume I*, vol. 3 (John Wiley & Sons London-New York-Sydney-Toronto, 1968).
- [55] J.-P. Onnela, A. Chakraborti, K. Kaski, J. Kertesz, Dynamic asset trees and Black Monday. *Physica A: Statistical Mechanics and its Applications* **324**, 247–252 (2003).
- [56] G. Bonanno, *et al.*, Networks of equities in financial markets. *The European Physical Journal B-Condensed Matter and Complex Systems* **38**, 363–371 (2004).

- [57] C. Spearman, The proof and measurement of association between two things. *The American journal of psychology* **15**, 72–101 (1904).
- [58] Z. Bin, H. Steve, A general framework for weighted gene co-expression network analysis. *Statistical applications in genetics and molecular biology* **4**, 1–45 (2005).
- [59] A. Barberán, S. T. Bates, E. O. Casamayor, N. Fierer, Using network analysis to explore co-occurrence patterns in soil microbial communities. *The ISME journal* **6**, 343–351 (2012).
- [60] S. Liang, S. Fuhrman, R. Somogyi, Reveal, a general reverse engineering algorithm for inference of genetic network architectures. *Pac. Symp. Biocomput.* **3**, 17–28 (1998).
- [61] L. Song, P. Langfelder, S. Horvath, Comparison of co-expression measures: mutual information, correlation, and model based indices. *BMC bioinformatics* **13**, 328 (2012).
- [62] L. Kullmann, J. Kertész, K. Kaski, Time-dependent cross-correlations between different stock returns: A directed network of influence. *Physical Review E* **66**, 026125 (2002).
- [63] M. McDonald, O. Suleman, S. Williams, S. Howison, N. F. Johnson, Detecting a currency's dominance or dependence using foreign exchange network trees. *Physical Review E* **72**, 046106 (2005).
- [64] C. W. Granger, Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society* pp. 424–438 (1969).
- [65] T. Schreiber, Measuring information transfer. *Physical review letters* **85**, 461 (2000).
- [66] A. Brovelli, *et al.*, Beta oscillations in a large-scale sensorimotor cortical network: directional influences revealed by Granger causality. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 9849–9854 (2004).
- [67] A. Roebroeck, E. Formisano, R. Goebel, Mapping directed influence over the brain using Granger causality and fmri. *Neuroimage* **25**, 230–242 (2005).
- [68] W. Liao, *et al.*, Small-world directed networks in the human brain: multivariate Granger causality analysis of resting-state fmri. *Neuroimage* **54**, 2683–2694 (2011).
- [69] R. Vicente, M. Wibral, M. Lindner, G. Pipa, Transfer entropy - a model-free measure of effective connectivity for the neurosciences. *Journal of computational neuroscience* **30**, 45–67 (2011).

- [70] K. J. Friston, L. Harrison, W. Penny, Dynamic causal modelling. *Neuroimage* **19**, 1273–1302 (2003).
- [71] S. Fortunato, Community detection in graphs. *Physics reports* **486**, 75–174 (2010).
- [72] V. M. Eguiluz, D. R. Chialvo, G. A. Cecchi, M. Baliki, A. V. Apkarian, Scale-free brain functional networks. *Physical review letters* **94**, 018102 (2005).
- [73] S. Allesina, A. Bodini, C. Bondavalli, Secondary extinctions in ecological networks: bottlenecks unveiled. *Ecological Modelling* **194**, 150–161 (2006).
- [74] F. Radicchi, J. J. Ramasco, S. Fortunato, Information filtering in complex weighted networks. *Physical Review E* **83**, 046101 (2011).
- [75] Z. Wu, L. A. Braunstein, S. Havlin, H. E. Stanley, Transport in weighted networks: partition into superhighways and roads. *Physical review letters* **96**, 148702 (2006).
- [76] R. Ferrer i Cancho, R. V. Solé, The small world of human language. *Proceedings of the Royal Society of London B: Biological Sciences* **268**, 2261–2265 (2001).
- [77] J. B. Kruskal, On the shortest spanning subtree of a graph and the traveling salesman problem. *Proceedings of the American Mathematical society* **7**, 48–50 (1956).
- [78] M. Tumminello, T. Aste, T. Di Matteo, R. N. Mantegna, A tool for filtering information in complex systems. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 10421–10426 (2005).
- [79] W. P. Goh, D. Kwek, D. Hogan, S. A. Cheong, Complex network analysis of teaching practices. *EPJ Data Science* **3**, 36 (2014).
- [80] M. Á. Serrano, M. Boguná, A. Vespignani, Extracting the multiscale backbone of complex weighted networks. *Proceedings of the national academy of sciences* **106**, 6483–6488 (2009).
- [81] A. V. Aho, M. R. Garey, J. D. Ullman, The transitive reduction of a directed graph. *SIAM Journal on Computing* **1**, 131–137 (1972).
- [82] J. R. Clough, J. Gollings, T. V. Loach, T. S. Evans, Transitive reduction of citation networks. *Journal of Complex Networks* **3**, 189–203 (2015).
- [83] J. Chalupa, P. L. Leath, G. R. Reich, Bootstrap percolation on a bethe lattice. *Journal of Physics C: Solid State Physics* **12**, L31 (1979).
- [84] S. Itzkovitz, *et al.*, Coarse-graining and self-dissimilarity of complex networks. *Physical Review E* **71**, 016127 (2005).

- [85] R. Guimera, L. A. N. Amaral, Functional cartography of complex metabolic networks. *Nature* **433**, 895–900 (2005).
- [86] A. Rapoport, Contribution to the theory of random and biased nets. *The bulletin of mathematical biophysics* **19**, 257–277 (1957).
- [87] P. Erdős, A. Rényi, On random graphs, i. *Publicationes Mathematicae (Debrecen)* **6**, 290–297 (1959).
- [88] S. Milgram, The small world problem. *Psychology today* **2**, 60–67 (1967).
- [89] J. Travers, S. Milgram, An experimental study of the small world problem. *Sociometry* pp. 425–443 (1969).
- [90] J. Guare, *Six degrees of separation: A play* (Vintage, 1990).
- [91] M. Meringer, E. W. Weisstein, Regular graph (2014). Visited on 10/12/16.
- [92] R. Albert, A.-L. Barabási, Statistical mechanics of complex networks. *Reviews of modern physics* **74**, 47 (2002).
- [93] B. Bollobás, O. Riordan, The diameter of a scale-free random graph. *Combinatorica* **24**, 5–34 (2004).
- [94] R. Cohen, S. Havlin, Scale-free networks are ultrasmall. *Physical review letters* **90**, 058701 (2003).
- [95] V. Latora, M. Marchiori, Efficient behavior of small-world networks. *Physical review letters* **87**, 198701 (2001).
- [96] R. Albert, H. Jeong, A.-L. Barabási, Internet: Diameter of the World Wide Web. *Nature* **401**, 130–131 (1999).
- [97] D. S. Bassett, E. Bullmore, Small-world brain networks. *The neuroscientist* **12**, 512–523 (2006).
- [98] M. E. Newman, D. J. Watts, Scaling and percolation in the small-world network model. *Physical Review E* **60**, 7332 (1999).
- [99] D. J. d. S. Price, Networks of scientific papers. *Science* **149**, 510–515 (1965).
- [100] S. Wasserman, K. Faust, *Social network analysis: Methods and applications*, vol. 8 (Cambridge university press, 1994).
- [101] E. Ravasz, A.-L. Barabási, Hierarchical organization in complex networks. *Physical Review E* **67**, 026112 (2003).

- [102] M. E. Newman, Assortative mixing in networks. *Physical review letters* **89**, 208701 (2002).
- [103] M. McPherson, L. Smith-Lovin, J. M. Cook, Birds of a feather: Homophily in social networks. *Annual review of sociology* pp. 415–444 (2001).
- [104] M. E. Newman, M. Girvan, Finding and evaluating community structure in networks. *Physical review E* **69**, 026113 (2004).
- [105] M. E. Newman, Modularity and community structure in networks. *Proceedings of the national academy of sciences* **103**, 8577–8582 (2006).
- [106] M. E. Newman, Fast algorithm for detecting community structure in networks. *Physical review E* **69**, 066133 (2004).
- [107] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, D. Parisi, Defining and identifying communities in networks. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 2658–2663 (2004).
- [108] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, E. Lefebvre, Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment* **2008**, P10008 (2008).
- [109] G. Palla, I. Derényi, I. Farkas, T. Vicsek, Uncovering the overlapping community structure of complex networks in nature and society. *Nature* **435**, 814–818 (2005).
- [110] A. Lancichinetti, S. Fortunato, J. Kertész, Detecting the overlapping and hierarchical community structure in complex networks. *New Journal of Physics* **11**, 033015 (2009).
- [111] B. S. Everitt, Unresolved problems in cluster analysis. *Biometrics* pp. 169–181 (1979).
- [112] O. Sporns, R. Kötter, Motifs in brain networks. *PLoS Biol* **2**, e369 (2004).
- [113] J.-P. Onnela, J. Saramäki, J. Kertész, K. Kaski, Intensity and coherence of motifs in weighted complex networks. *Physical Review E* **71**, 065103 (2005).
- [114] A. Barrat, M. Barthélemy, R. Pastor-Satorras, A. Vespignani, The architecture of complex weighted networks. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 3747–3752 (2004).
- [115] P. Bonacich, Power and centrality: A family of measures. *American journal of sociology* pp. 1170–1182 (1987).

- [116] L. C. Freeman, A set of measures of centrality based on betweenness. *Sociometry* pp. 35–41 (1977).
- [117] N. Mercer, The seeds of time: Why classroom dialogue needs a temporal analysis. *The Journal of the Learning Sciences* **17**, 33–59 (2008).
- [118] D. Hogan, P. Towndrow, D. Kwek, M. Chan, Final report of the core 2 research programme, *Tech. rep.*, National Institute of Education (2013).
- [119] A. Luke, P. Freebody, C. Cazden, A. Lin, Singapore pedagogy coding scheme, *Tech. rep.*, National Institute of Education (2004).
- [120] A. Luke, C. Cazden, A. Lin, P. Freebody, The Singapore classroom coding system, *Tech. rep.*, National Institute of Education (2004).
- [121] J. Hattie, *Visible learning: A synthesis of over 800 meta-analyses relating to achievement* (Routledge, London, 2013).
- [122] D. Hogan, *et al.*, *Globalization and the Singapore Curriculum: From Policy to Classroom*, Z. Deng, S. Gopinathan, C. K.-E. Lee, eds. (Springer Science & Business Media, Singapore, 2013), pp. 121–150.
- [123] O. Kwon, J.-S. Yang, Information flow between stock indices. *EPL (Europhysics Letters)* **82**, 68003 (2008).
- [124] C. J. Honey, R. Kötter, M. Breakspear, O. Sporns, Network structure of cerebral cortex shapes functional connectivity on multiple time scales. *Proceedings of the National Academy of Sciences* **104**, 10240–10245 (2007).
- [125] Y. Hu, Efficient, high-quality force-directed graph drawing. *Mathematica Journal* **10**, 37–71 (2005).
- [126] D. R. Olson, N. Torrance, *The handbook of education and human development: New models of learning, teaching and schooling* (Blackwell, Cambridge, 1996), pp. 9–27.
- [127] V. Richardson, The role of attitudes and beliefs in learning to teach. *Handbook of research on teacher education* **2**, 102–119 (1996).
- [128] R. K. Sawyer, *The Cambridge handbook of the learning sciences*, R. K. Sawyer, ed. (Cambridge University Press, New York, 2008), vol. 2.
- [129] D. L. Ball, *et al.*, *Mathematical proficiency for all students: Toward a strategic research and development program in mathematics education* (Rand Corporation, Santa Monica, 2003).

- [130] OECD, PISA 2012 assessment and analytical framework, *Tech. rep.*, OECD (2012).
- [131] F. Christie, *Classroom discourse analysis: A functional perspective* (Bloomsbury Publishing, London, 2005).
- [132] J. P.-L. Tan, I. S. Caleon, C. R. Jonathan, E. Koh, A dialogic framework for assessing collective creativity in computer-supported collaborative problem-solving tasks. *Research and Practice in Technology Enhanced Learning* **9**, 411–437 (2014).
- [133] D. J. Treffinger, S. G. Isaksen, K. B. Stead-Dorval, *Creative problem solving: An introduction* (Prufrock Press Inc., 2006).
- [134] R. W. Weisberg, *Creativity: Understanding innovation in problem solving, science, invention, and the arts* (John Wiley & Sons, 2006).
- [135] ATC21S, ATC21S Launch: Project Background (2012). Visited on 20/12/16.
- [136] N. P. Jewell, *Statistics for epidemiology* (CRC Press, 2003).
- [137] G. W. Oehlert, A note on the delta method. *The American Statistician* **46**, 27-29 (1992).
- [138] J. H. Ward Jr, Hierarchical grouping to optimize an objective function. *Journal of the American statistical association* **58**, 236–244 (1963).
- [139] F. De Saussure, *Course in general linguistics (translated by W. Baskin)* (Columbia University Press, 2011).
- [140] M. Sigman, G. A. Cecchi, Global organization of the Wordnet lexicon. *Proceedings of the National Academy of Sciences* **99**, 1742–1747 (2002).
- [141] A. E. Motter, A. P. de Moura, Y.-C. Lai, P. Dasgupta, Topology of the conceptual network of language. *Physical Review E* **65**, 065102 (2002).
- [142] R. Ferrer i Cancho, R. V. Solé, The small world of human language. *Proceedings of the Royal Society of London B: Biological Sciences* **268**, 2261–2265 (2001).
- [143] A. Masucci, G. Rodgers, Network properties of written human language. *Physical Review E* **74**, 026102 (2006).
- [144] R. Ferrer i Cancho, R. V. Solé, R. Köhler, Patterns in syntactic dependency networks. *Physical Review E* **69**, 051915 (2004).
- [145] H. Liu, The complexity of Chinese syntactic dependency networks. *Physica A: Statistical Mechanics and its Applications* **387**, 3048–3058 (2008).

- [146] H. Liu, F. Hu, What role does syntax play in a language network? *EPL (Europhysics Letters)* **83**, 18002 (2008).
- [147] A. Masucci, G. Rodgers, Differences between normal and shuffled texts: structural properties of weighted networks. *Advances in Complex Systems* **12**, 113–129 (2009).
- [148] G. K. Zipf, *Human behavior and the principle of least effort: An introduction to human ecology* (Addison-Wesley, Reading, MA, 1949).
- [149] S. N. Dorogovtsev, J. F. F. Mendes, Language as an evolving word web. *Proceedings of the Royal Society of London B: Biological Sciences* **268**, 2603–2606 (2001).
- [150] R. Ferrer i Cancho, R. V. Solé, Least effort and the origins of scaling in human language. *Proceedings of the National Academy of Sciences* **100**, 788–791 (2003).
- [151] B. Corominas-Murtra, R. Hanel, S. Thurner, Understanding scaling through history-dependent processes with collapsing sample space. *Proceedings of the National Academy of Sciences* **112**, 5348–5353 (2015).
- [152] R. Solé, *Phase Transitions*, Primers in Complex Systems (Princeton University Press, 2011).
- [153] J. Maynard Smith, E. Szathmary, The major evolutionary transitions. *Nature* **374**, 227–232 (1995).
- [154] D. Bickerton, *Language and human behavior*. (University of Washington Press, 1995).
- [155] C. F. Hockett, The origin of speech. *Scientific American* **203**, 88–96 (1960).
- [156] N. Chomsky, *Language and mind* (Cambridge University Press, 2006), third edn.
- [157] M. D. Hauser, *The evolution of communication* (MIT press, 1996).
- [158] B. Corominas-Murtra, S. Valverde, R. V. Solé, The ontogeny of scale-free syntax networks: phase transitions in early language acquisition. *Advances in Complex Systems* **12**, 371–392 (2009).
- [159] R. V. Solé, B. Corominas-Murtra, S. Valverde, L. Steels, Language networks: Their structure, function, and evolution. *Complexity* **15**, 20–26 (2010).
- [160] J. Borge-Holthoefer, A. Arenas, Semantic networks: structure and dynamics. *Entropy* **12**, 1264–1302 (2010).

- [161] Z. Solan, D. Horn, E. Ruppín, S. Edelman, Unsupervised learning of natural languages. *Proceedings of the National Academy of Sciences* **102**, 11629–11634 (2005).
- [162] R. Hudson, *An introduction to word grammar* (Cambridge University Press, 2010).
- [163] L. Tesnière, *Eléments de syntaxe structurale* (Klincksieck, Paris, 1959).
- [164] G. A. Miller, The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review* **63**, 81–97 (1956).
- [165] M. H. Christiansen, N. Chater, The now-or-never bottleneck: A fundamental constraint on language. *Behavioral and Brain Sciences* **39**, e62 (2016).
- [166] M. W. Axelsson, USE - the Uppsala Student English corpus: an instrument for needs analysis. *ICAME Journal* **24**, 155–157 (2000).
- [167] W. N. Francis, A standard corpus of edited present-day American English. *College English* **26**, 267–273 (1965).
- [168] M. Hart, *Project Gutenberg* (Project Gutenberg, 1971).
- [169] A. Clauset, C. R. Shalizi, M. E. Newman, Power-law distributions in empirical data. *SIAM review* **51**, 661–703 (2009).
- [170] J. R. Anderson, G. H. Bower, *Human associative memory* (Psychology Press, 2014).
- [171] M. P. Marcus, M. A. Marcinkiewicz, B. Santorini, Building a large annotated corpus of English: The penn treebank. *Computational linguistics* **19**, 313–330 (1993).
- [172] S. Lawrence, C. L. Giles, Searching the World Wide Web. *Science* **280**, 98–100 (1998).
- [173] A. E. Goldberg, Constructions: a new theoretical approach to language. *Trends in cognitive sciences* **7**, 219–224 (2003).
- [174] C. J. Fillmore, *Annual Meeting of the Berkeley Linguistics Society* (1988), vol. 14, pp. 35–55.
- [175] W. Croft, *Radical construction grammar: Syntactic theory in typological perspective* (Oxford University Press, 2001).
- [176] H. Diessel, *The acquisition of complex sentences* (Cambridge University Press, 2004).

-
- [177] W. Croft, Construction grammar. *The Oxford handbook of cognitive linguistics* pp. 463–508 (2007).
- [178] M. Tomasello, Acquiring linguistic constructions. *Handbook of child psychology* (2006).
- [179] A. Radford, *Syntactic theory and the acquisition of English syntax: The nature of early child grammars of English* (Blackwell, 1990).
- [180] C. C. Fries, *The structure of English: An introduction to the construction of English sentences* (Longman, 1973).
- [181] L. K. Börzsei, Makes a meme instead: A concise history of internet memes. *New Media Studies Magazine* (2013).