

Topological Evolution of Networks: Case Studies in the US Airlines and Language Wikipedias

by

Gergana Assenova Bounova

B.S., Theoretical Mathematics, Massachusetts Institute of Technology (2003)
B.S., Aeronautics & Astronautics, Massachusetts Institute of Technology (2003)
S.M., Aeronautics & Astronautics, Massachusetts Institute of Technology (2005)

Submitted to the Department of Aeronautics and Astronautics
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2009

©2009 Gergana A. Bounova, All rights reserved.

The author hereby grants to MIT the permission to reproduce and to distribute publicly paper and electronic copies of this thesis document in whole or in part.

Author
Department of Aeronautics and Astronautics
February 27, 2009

Certified by
Prof. Olivier L. de Weck
Associate Professor of Aeronautics and Astronautics and Engineering Systems
Thesis Supervisor

Certified by
Prof. Christopher L. Magee
Professor of the Practice of Mechanical Engineering and Engineering Systems

Certified by
Dr. Daniel E. Whitney
Senior Research Scientist, Center for Technology, Policy and Industrial Development,
Senior Lecturer in Engineering Systems and Mechanical Engineering

Accepted by
Prof. David Darmofal
Associate Department Head
Chair, Committee on Graduate Students

Topological Evolution of Networks: Case Studies in the US Airlines and Language Wikipedias

by
Gergana Assenova Bounova

Submitted to the Department of Aeronautics and Astronautics
on February 27, 2009, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

This thesis examines the topology of engineering systems and how that topology changes over time. Topology refers to the relative arrangement and connectivity of the elements of a system. We review network theory relevant to topological evolution and use graph-theoretical methods to analyze real systems, represented as networks. Using existing graph generative models, we develop a profile of canonical graphs and tools to compare a real network to that profile. The developed metrics are used to track topology changes over the history of real networks.

This theoretical work is applied to two case studies. The first discusses the US airline industry in terms of routes. We study various airlines and segments of the industry statistically and find commonly occurring patterns. We show that there are topology transitions in the history of airlines in the period 1990-2007. Most airline networks have similar topology and historical patterns, with the exception of Southwest Airlines. We show mathematically that Southwest's topology is different. We propose two heuristic growth models, one featuring hub-seeding derived from the underlying patterns of evolution of JetBlue Airways and one featuring local interconnectedness, derived from the patterns of growth of Southwest. The two models match the topologies of these airlines better than canonical models over time. Results suggest that Southwest is becoming more centralized, closer to the hub-spoke topologies of other airlines.

Our second case study discusses the growth of language Wikipedia networks, where nodes are articles and hyperlinks are the connections between them. These knowledge networks are subject to different constraints than air transportation systems. The topology of these networks and their growth principles are completely different. Most Wikipedias studied grow by coalescence, with multiple disconnected thematic clusters of pages growing separately and over time, converging to a giant connected component via weak links. These topologies start out as simple trees, and coalesce into sparse hierarchical structures with random interlinking. One striking exception is the history of the Chinese Wikipedia, which grows fully connected from its inception. We discuss these patterns of growth comparatively across Wikipedias, and in general, compared to airline networks.

Our work suggests that complex engineering systems are hybrids of pure canonical forms and that they undergo distinct phase transitions during their evolution. We find commonality among systems and uncover important differences by learning from the exceptions.

Thesis Supervisor: Prof. Olivier L. de Weck

Title: Associate Professor of Aeronautics and Astronautics and Engineering Systems

Acknowledgments

I chose to stay at MIT for a doctoral degree because of Professor Olivier de Weck. I recognized early that he would be a great advisor. I know my peers share this opinion. In 2006 we nominated him for the best graduate student advisor award (Frank E. Perkins award) at MIT, and he deservedly won. Oli, thank you for letting me pursue research that I care about, and for supporting me in often taking an unconventional route. I admire your energy, kindness and concern for the well-being of students.

I would like to thank the rest of my committee, Prof. Christopher Magee and Dr. Daniel Whitney for their help, guidance, criticism and encouragement. Especially, I am grateful to Dan Whitney, for many interesting discussions, help with research directions and multiple revisions on my thesis till the very end. Thank you.

Thanks to Dr. Philippe Bonnefoy for helpful feedback on airline networks, and multiple revisions on my thesis.

The de Weck group throughout the years and the 33-409 cluster have been an integral part of my graduate school experience. I enjoyed working and sharing experiences with Kalina Galabova, Erica Gralla, Chris Graff, Ariane Chepko, Nii Armar, Jaemyung Ahn, Ryan Boas and Wilfried Hofstetter. Many thanks to Wilfried and Luca Bertuccelli for helping with my defense presentation.

Some of the best moments at MIT I spent playing volleyball. I am grateful for the existence of the MIT Women's Volleyball Club. Thanks to Tony, Deb, Ellen, Rahmat, Paola and Tracy and a hundred other people...! It's been great playing with you and winning some tournaments!

Thanks to everyone at St Mary's and all my friends who have kept me sane and made me who I am - Jaisel, Nodari, Bistra, Baber, Thi and Francois, Frances, Angelita and Sophocles. Many thanks to Anna Custo, for her friendship, encouragement, and for lending me some time on the MGH supercluster when the motif search was taking months to compute.

Finally, I would like to thank my whole family, my parents, grandparents and brother. My father inspired me to become an engineer, and to be curious about how things work, and my mother taught me how to think unconventionally. I am lucky to have my brother, who has made significant contributions to this thesis. He rewrote the motif search algorithm and compiled it in *C*, which made it a hundred times faster. He also parsed the static Wikipedia dumps into network snapshots that I could analyze. Without him, the Wikipedia Chapter would not be there. Dimo, I am proud to be your sister.

This thesis is dedicated to my grandmother, Gina, the most hard-working and selfless person I know.

Contents

0	Introduction	23
0.1	Complex Systems Change	23
0.1.1	Growth and Complexity	23
0.1.2	Visualization of Complex Data	24
0.2	Complex Systems and Topology	24
0.3	Complex Systems and Topology Evolution	25
0.4	Airline Networks	26
0.5	Research Goals. Hypotheses	27
0.5.1	Network Topology and Topology Evolution	27
0.5.2	Airline Networks	28
0.5.3	Wikipedia	29
0.6	Thesis Roadmap	30
1	Network Theory Review	31
1.1	Network Theory Basics	31
1.1.1	Introduction	31
1.1.2	Types of simple graph representations	33
1.2	Statistics on Graphs	35
1.2.1	Size and density - nodes, edges, average degree	35
1.2.2	Node centrality measures	36
1.2.3	Degree distributions	37
1.2.4	Clustering coefficient	39
1.2.5	Degree correlation / Assortativity	40
1.2.6	The S-metric; scale free graphs	40
1.2.7	Distances - average path length, diameter	41
1.3	Modularity, Motifs, Coarse-graining	42
1.3.1	Modularity	42
1.3.2	Subgraphs. Motifs	45
1.3.3	Coarse-graining	49
1.4	Evolution of Networks	51
1.4.1	Random graph models	51
1.4.2	Node-degree centric models	52
1.4.3	Spatial distribution algorithms	55
1.4.4	Node-copying and function preservation	57
1.4.5	Module-copying and growth by accretion	58
1.4.6	Edge-centric models	59
1.5	Conclusion	61

2	Network Topology	63
2.1	Statistical Indicators for Topology	64
2.1.1	Canonical topologies	64
2.1.2	Topology spectrum	68
2.1.3	Statistics for JetBlue Airways, August 2007	71
2.1.4	Degree-Betweenness Relationship	72
2.1.5	JetBlue August 2007 and degree distributions	74
2.2	Graph Similarity (Hubs/Authorities Comparison to Canonical Networks)	77
2.2.1	Graph similarity basics	77
2.2.2	Graph similarity measure assessment	77
2.2.3	JetBlue 8/2007 graph similarity	80
2.3	Modularity and Topology (JetBlue 8/07)	81
2.3.1	Newman-Girvan: modularity using betweenness	81
2.3.2	The Newman eigenvector method	82
2.4	Motifs - The Building Blocks	83
2.5	Coarse-graining the JetBlue 8/2007 network	86
2.6	Conclusion and Discussion	89
3	Airline Networks	91
3.1	Introduction	91
3.1.1	Trends in the airline industry	91
3.1.2	The hub-spoke phenomenon	93
3.1.3	Literature on airline networks	94
3.1.4	Dynamics drivers in the airline industry	97
3.2	Airline Data	99
3.2.1	Data source and slicing	99
3.2.2	Down-selection of the data slices	103
3.3	Topology Measures	103
3.3.1	Statistical indicators for topology	103
3.3.2	Topology profiles	115
3.3.3	Degree distributions	118
3.3.4	Motif analysis	122
3.4	Conclusion	128
4	Evolution of US Airline Route Networks	129
4.1	Graph-theoretical statistics over time	129
4.2	Detecting changes in topology	135
4.3	Comparing topology to canonical networks over time	139
4.4	Topology-derived growth models	143
4.5	Conclusion	151
5	Evolution of Language Wikipedias	153
5.1	The Language Wikipedias: Early Growth	154
5.2	The Interlingua Wikipedia	156
5.3	The Esperanto Wikipedia	161
5.4	The Spanish Wikipedia	164
5.5	The Russian Wikipedia	168
5.6	The Chinese Wikipedia	170

5.7	Topology Comparison of Language Wikipedias. Conclusion	173
6	Conclusion	175
6.1	Contributions	176
6.1.1	Analyzing network topology	176
6.1.2	Studying statistically system topology over time	177
6.1.3	Airline Networks	178
6.1.4	Wikipedia	179
6.1.5	Computational contributions	179
6.2	Limitations	179
6.2.1	Data/Computation	179
6.2.2	Approach limitations	180
6.2.3	Airline analysis	181
6.3	Future Work	181
A	Additional Materials	185
A.1	Airlines	185
A.2	Language Wikipedias	188
A.2.1	Interlingua Wikipedia	188
A.2.2	The Simple English Wikipedia	188
A.2.3	French and Italian Wikipedias	190
A.3	Airlines Data Sample	192
A.4	JetBlue Airways 8/2007 airport codes	192
A.5	Airline industry data slices	194

List of Figures

0-1	US airline flights in August 2007. High-frequency departure routes are highlighted. Alaska, Hawaii and Puerto Rico / US Virgin Islands are not shown.	24
1-1	Bowtie graph with four example representations: visual, adjacency matrix ($n \times n$, $A_{ij} = 1$ if i and j are linked), adjacency list (list of all neighbors of every node) and edge list (list of pairs of nodes with their link weights). More details on graph representations are given in Table 1.1.2.	32
1-2	Routes of all US airlines during 1990 (BTS [1]), domestic and international. Continental US, Europe, South America, Alaska, Hawaii and Guam and Japan are easily distinguishable.	32
1-3	Examples of network types: undirected graph with a single type of nodes and links (upper left); undirected graph with discrete nodes and links (upper right), undirected graph with weighted nodes and links (lower left) and directed graph (lower right). Reproduced from [2].	34
1-4	Two representations of the JetBlue August 2007 network. Nodes are airports, edges are existing routes. (left) Airports are plotted at their latitude and longitude; (right) The same graph is plotted but with node locations determined by a spring energy algorithm by Kamada and Kawai [3].	35
1-5	Matrix dot plots of the adjacency matrices of Southwest Airlines and Continental Airlines for 8/2007. The Southwest network has 74 airports and 809 one-way flight segments; Continental has 90 airports and 207 one-way flight segments. Every point is filled if there is a connection between cells i and j . Rows and columns are ordered by increasing nodal degree (increasing left and up). Clearly, the Southwest network is denser, and the Continental network features very few high-degree nodes.	36
1-6	Node sizes represent degree on the left, betweenness on the right. Nodes 2 and 8 become more central if measured by betweenness.	37
1-7	Example of a high betweenness edge for the bowtie graph.	37
1-8	Histogram, probability degree distribution (pdf) and cumulative degree distribution (cdf) computed for the network of all airlines routes in 1990 visualized in Figure 1-2. Nodes are airports, links are existing routes, node degree is the number of incoming airport connections. This is clearly an exponential degree distribution, but it can be interpreted as having a power-law cutoff.	38
1-9	Figure 6 from Li et al [4]. Five networks having the same node degree distribution concerning Internet route structure. (a) Common node degree distribution (degree versus rank on log-log scale); (b) Network resulting from preferential attachment; (c) Random graph built with the same degree sequence; (d) Heuristically optimal topology; (e) Abilene-inspired topology; (f) Sub-optimally designed topology.	39

1-10	Illustrating the definition of motif: all S_1, S_2 , and S_3 are valid motifs (connected subgraphs) of G	46
1-11	Pseudocode for $FindSubgraphInstances(H, G)$ from Grochow et al [5]	47
1-12	Pseudocode for $IsomorphicExtensions(f, H, G)$ from Grochow et al [5]	47
1-13	Pseudocode for $SymmetryConditions$ from Grochow et al [5]	47
1-14	Example of three generalizations of a feed-forward (XYZ) loop (source: Figure 2c from Kashtan et al [6]).	48
1-15	Z-score of the 3-node motifs (triads) shown in various real networks (Figure 1 from [7]). The four profiles on the figure correspond to transcription networks (biology), signal transduction (biology), social networks and language (word occurrences in sentences). The profiles show that different networks can be distinguished based on respective motif significance.	49
1-16	Example from Itzkovitz et al 2005 - coarse-graining of an electronic circuit. Four levels of representation of the 8-bit counter electronic circuit. In the transistor level network, nodes represent transistor junctions. In the gate level, nodes are CGUs made of transistors, each representing a logic gate. Shown is the CGU that corresponds to a NAND gate. In the flip-flop level, nodes are either gates or a CGU made of gates that corresponds to a D-type flip-flop with an additional logic gate at its input. In the counter level, each node is a gate or a CGU of gates/flip-flops that corresponds to a counter subunit. Numbers of nodes P and edges E at each level are shown.	50
1-17	Exponential network snapshots example, $t=100, 500, 1000$. The plots are created with NetDraw using the Kamada-Kawai spring energy minimization algorithm [3].	53
1-18	BA graph snapshots at $N=100, 500$ and 1000 nodes ($t=N$). The graph are plotted using the Kamada-Kawai spring energy minimization algorithm [3].	54
1-19	Master equation method with $a=2, m=1$. Graphs plotted using a spring energy minimization algorithm.	54
1-20	Master equation method with $a=2, m=2$. Graphs plotted using a spring energy minimization algorithm.	55
1-21	Fabrikant model with $N=100$, and $\alpha=1, 1.5$ and 20 respectively.	55
1-22	Newman-Gastner model for $\beta = 0.1, \beta = 0.9$, and $\beta = 1.1000$ nodes.	56
1-23	Optimizing telescope array configurations for cable length (cost) and uv density (performance). Left to right: 27, 60 and 99 nodes. Three configurations show minimum cable, nadir point and minimum uv metric designs from top to bottom (Bounova [8]).	57
1-24	Figure 2 from (Sole 2002) Growing network by duplication of nodes. First (a) duplication occurs after randomly selecting a node (arrow). The links from the newly created node (white) now can experience deletion (b) and new links can be created (c); these events occur with probabilities δ and α , respectively.	58
1-25	Different components of the change request network coalescing together over the course of nine years. (Left) Year 7; (right) Year 9.	59
1-26	Figure 1 from Dodds et al [9]. Illustrating the distances in a hierarchy - to the lowest common ancestor from each node (d_i and d_j) and from the lowest common ancestor to the top (D_{ij}).	60
1-27	Hierarchy patterns. Figure 2 from Dodds-Watts-Sabel [9].	61

2-1	Topology families in order of increasing graph density ($m/(n(n-1)/2)$): lines, circles, stars, balanced trees, BA/s-max, hierarchical trees with interlinking, lattices and random graphs.	66
2-2	Edge-to-node ratio for the topologies in Figure 2-1. As n increases the y-axis scale increases as $O(n)$, as random graphs and complete graphs grow in edges with $O(n^2)$. Most real networks (certainly airlines and Wikipedias studied here) have edge-to-node ratio between 2 and 10 (marked by circle on the plot).	66
2-3	Topology profile for an Erdős-Rényi random graph. The graphs compared to are: line, ring, star, binary tree, tertiary tree, Newman-Gastner graphs with $\alpha=0.1, 0.5, 0.9$, preferential attachment graph, s-max graph, hierarchical binary tree, hierarchical tertiary tree, Dodds-Watts-Sabel graphs with varying parameters, lattices (triangular, square and hexagonal), random graph with the same degree distribution, a random modular graph and an Erdős-Rényi graph with the same density.	69
2-4	Topology profile for a BA graph. The graphs compared to are: line, ring, star, binary tree, tertiary tree, Newman-Gastner graphs with $\alpha=0.1, 0.5, 0.9$, preferential attachment graph, s-max graph, hierarchical binary tree, hierarchical tertiary tree, Dodds-Watts-Sabel graphs with varying parameters, lattices (triangular, square and hexagonal), random graph with the same degree distribution, a random modular graph and an Erdős-Rényi graph with the same density.	70
2-5	Topology profile for the JetBlue 8/07 network. The closest match is to a preferential attachment (BA) graph.	72
2-6	JetBlue 8/2007 routes: geographic representation by airport latitude/longitude, and representation of node locations by minimizing spring energy. JFK and BOS are the two largest hubs.	72
2-7	Betweenness versus degree for all nodes of three types of graphs: (left) Erdős-Rényi graph with 200 nodes and $p=0.2$; (middle) random modular graph, with 200 nodes, 4 clusters, and $p=0.2$ general density; (right) BA graphs with 200 nodes.	73
2-8	Unweighted and weighted cumulative degree distribution for JetBlue 8/07. Both slopes close to -1, which means that $\gamma \approx -2$ for the frequency degree distribution.	75
2-9	Rewiring JetBlue 8/2007 from minimum $r=-0.667$ (left) to maximum $r=-0.542$ (right).	75
2-10	Betweenness versus degree for all JetBlue airports of August 2007. JFK and BOS stand out as both high-degree and high-betweenness. Orlando (MCO) emerges as a third important hub. This profile is most similar to the BA profile on Figure 2-7, confirming the similarity to BA graphs from the topology profile (Figure 3-23).	76
2-11	BA graph, random edge removal at every step. The four metrics from this section, sum of all entries of the similarity matrix (X), maximum entry, sum of the best diagonal (best matching node sequence), and the Euclidean distance between the diagonals compared. Each metric is used to compare consecutive graph snapshots as well as comparing every new graph to the original. Direction of high similarity is shown by the arrows.	79
2-12	Graph similarity score of JetBlue 8/2007 and the following canonical networks: star/circle (0.7948), binary tree (0.9539), tertiary tree (0.9625), itself (0.9986), random graph with the same degree distribution (0.9607), the s-max graph based on JetBlue's degree sequence (0.9949), triangular lattice (0.9024), square lattice (0.90199), hexagonal lattice (0.9196), BA graph (0.8827), random modular graph (0.9526), and an Erdős-Rényi graph (0.9261), all generate with the same characteristics (51 nodes, same density/degree sequence if relevant)	80

2-13	Modularizing JetBlue 8/2007 using the Newman-Girvan algorithm: up to 21 components, reflecting maximum Q value. The node locations are plotted using a spring energy algorithm.	82
2-14	Newman eigenvector method modularization of JetBlue 8/2007.	83
2-15	Motif significance profile for JetBlue 8/07. There are 16 motifs with Z score > 0.05 (shown in Table 2.4). Our motif finding algorithm identified 105 occurring motifs. This figure shows the statistical significance (Z-score - equation 2.2) of these motifs compared to a 100-random graph ensemble with the same degree distribution. The highest significance motifs are labeled, some with corresponding indices from Table 2.4.	86
2-16	Collapsing all stars from the JetBlue 8/2007 topology emphasizes the bi-partite subgraphs. Major bi-partite subgraphs are circled.	87
2-17	The coarse-grained JetBlue graph, showing motifs 4 (once), 6 (twice) and 8 (twice) from Table 2.4, and the star motif around JFK	88
2-18	JetBlue 8/07 A320 network (left) and Embraer 190 network (right).	89
3-1	US Airline Net profit since 1947, yearly. Plot from Hansman [10]. Additional study by Sgouridis [11]	92
3-2	Passenger traffic and GDP in the US over time. Plot from Hansman [10].	92
3-3	Increasing traffic world-wide 1970-2008 (Hansman [10]).	93
3-4	Emergence of secondary airports (Bonney [12]).	94
3-5	Trends in aircraft size, average number of seats per departure, 1990-2007. Figure from Hansman [10].	98
3-6	The airline networks by aircraft type from Bonney [12].	99
3-7	US airline data slices by aircraft type: regional jets and turboprops (left), narrow-body aircraft (middle) and wide-body aircraft (right).	100
3-8	US airlines dataset sliced by distance: (left) all flights under 500 miles, and (right) all flights above 1000 miles.	100
3-9	Slicing data (all US airlines) by departure frequency - top x flights for the entire period 1/1990-8/2007: top 80, top 60, top 40 and top 20 most frequent flights.	101
3-10	Slices by seat capacity offered (all US airlines): top 80, top 60, top 40 and top 20. A flight leg is in top $x\%$ if it has at least $x/100$ fraction of the seat capacity of the busiest flight in the network.	102
3-11	Total number of departures plotted against number of airports for all data slices of the entire airline dataset. Statistics are for January 1990 and August 2007 for almost all slices. Full nomenclature of data slices is given in Table A.5.	104
3-12	Total number of departures versus number of airports, for individual airlines only. Three groups of data points emerge: low-cost carriers, legacy carriers in 1990, in 2007, and Southwest Airlines is an outlier.	105
3-13	Total seat-miles for single airlines only.	106
3-14	Total number of passengers carried versus number of airports. 1/1990 and 8/2007, all data slices.	107
3-15	Passengers carried on US airlines - in 1990 and in 2007.	107
3-16	Number of edges (OD pairs served) versus number of nodes for all data slices. Text on figure is omitted for clarity. Zoomed plot shows single airline data points only.	108
3-17	Edge to node ratio for all major airlines including JetBlue and Southwest as low-cost representatives.	109
3-18	S-max for single airlines only.	110

3-19	Average path length versus number of nodes - all data slices. The average path length versus size seems to follow a log-like curve. The high path length slices are all short-distance or small-aircraft networks, so naturally, they do not have longer cross cutting flights which shorten the path lengths. Examples are flights with less than 100 seats, or under 500 miles.	111
3-20	Airline diameters versus number of airports. All airlines have diameters between 3 and 5 - which is very small, and mostly unexpected for Southwest which does from a diameter of 4 to 3 from 1/1990 to 8/2007.	111
3-21	C3 (traffic share in top 3 airports) index versus number of airports. All data slices.	114
3-22	Traffic share in top 3 airports versus number of airports for single airlines only.	114
3-23	JetBlue 8/07 topology profile, best match is a BA graph.	116
3-24	Southwest 8/07 topology profile, best match is a hierarchical tertiary tree graph.	116
3-25	Continental 8/07 topology profile, best match is a hierarchical tertiary tree graph, next BA graph.	117
3-26	Wide-body jets 8/07 topology profile, best match is a hierarchical tertiary tree graph.	117
3-27	Top 50 flights by departure topology profile, best match is a binary tree.	117
3-28	Top 50 flights by seat capacity topology profile, best match is a BA graph.	118
3-29	All airlines 8/07 weighted degree distribution. The weights are number of departures from a node. Top airports (steepest slope) are JFK New York, Atlanta, Chicago, Philadelphia, Denver, Charlotte, Los Angeles, San Francisco, Boston, Orlando, Salt Lake City, Seattle, Las Vegas, Phoenix, La Guardia, Detroit, Houston, Dallas, Minneapolis, and Covington, KY.	119
3-30	Unweighted (simple graph) degree distribution of the entire US airline system. Plots from left to right are: frequency distribution (or probability density plot), cumulative degree distribution (cumulative density plot) and rank plot ($k \sim \log d_k$).	119
3-31	Betweenness versus number of connections (degree) for all flights, US airlines 8/07. Atlanta is the most connected airport, but Anchorage has the highest betweenness. The linear-like betweenness-degree relationship is broken here, which means that the network is more modular and balanced rather than centralized with 1 or 2 dominating hubs.	120
3-32	Betweenness versus number of connections (degree) for all long haul flights (> 500 mi), 8/07. In long-haul flights, the centralized pattern is more present. Las Vegas is the most connected and most “intermediate” airport. Anchorage still has high betweenness but not that many long-haul flights.	120
3-33	Southwest log-log plot of cumulative degree distribution and rank plot. Exponential with a steep power-law-like cutoff.	121
3-34	Continental log-log plot of cumulative degree distribution and rank plot. The three airports that stand out in the distribution are Newark, Houston and Cleveland. The kink in the distribution means that there are more flights out of these hubs than predicted by an exponential cut-off.	121
3-35	Wide-body jets log-log plot of cumulative degree distribution and rank plot. A clear exponential distribution.	122
3-36	Top 50 seats flights log-log plot of cumulative degree distribution and rank plot. Power law distribution.	122
3-37	Predominant motifs in most airline (hub-spoke) topologies: stars, base-triangles, and bi-partite graphs.	123
3-38	Continental Airlines Z-score profile. About 125 motifs are found, of which 6-7 have a Z-score that stands out. They fall in the families described in Figure 3-37.	124

3-39	Wide-body jets data slice Z-score profile. From about 85 motifs found, 12 have a positive Z-score. All of them also fall in the families found in Figure 3-37.	125
3-40	Z-scores for all motifs found in the top 50 by seat capacity flights networks. Most Z scores are below and close to zero. The only significant motifs correspond to stars. .	126
3-41	Southwest motifs z-score profile. No significant (high Z-score) motifs.	127
3-42	Southwest top 60 significant motifs.	127
3-43	Southwest concentrates capacity in 8/2007. Line thickness is proportional to the number of seats offered on that leg.	128
4-1	Low-cost airlines growth in number of airports, over the period 1/1990-8/2007. Airlines plotted are JetBlue (starts 2/2000), Southwest, Airtran (starts 10/1994), ATA, Frontier (starts 7/1994), Spirit (starts 7/1992) and USA3000 (starts 12/2001)	130
4-2	Top eight airlines (by passengers carried) and low-cost airlines growth in number of airports, over the period 1/1990-8/2007. Airlines plotted are JetBlue, Southwest, Airtran, ATA, Frontier, Spirit, USA3000, Alaska, America West, American Airlines, United, Northwest, Delta, US Airways and Continental Airlines.	131
4-3	Top eight airlines and low-cost airlines growth in total number of seats offered monthly, over the period 1/1990-8/2007.	132
4-4	Low-cost airlines growth in total number of seats offered monthly, over the period 1/1990-8/2007. Notable events are the September 2001 spike and Airtran's merger with Valuate, beginning operations in September 1998.	132
4-5	Top eight airlines growth in total number of passengers offered monthly, over the period 1/1990-8/2007.	133
4-6	Low-cost airlines growth in total number of passengers offered monthly, over the period 1/1990-8/2007.	133
4-7	Top eight airlines growth in total number of minutes flown monthly, over the period 1/1990-8/2007.	134
4-8	Average path length over time for all low-cost carriers: JetBlue, Southwest, Airtran, ATA, Frontier, Spirit, USA3000, Alaska, America West	135
4-9	Example of tracking a random graph topology with the two similarity measures. The zoomed plot shows the graph similarity variation at a magnified scale.	136
4-10	ATA topology changes month to month, from 1/1990 to 8/2007. Graph similarity and topology vector similarity measures.	137
4-11	Plotting the two similarity measures, graph similarity (black) and topology vector (red) over time (JetBlue monthly index from 1 to 91).	137
4-12	Southwest topology changes over 212 months from 1/1990 to 8/2007.	138
4-13	Continental topology changes over 212 months from 1/1990 to 8/2007.	138
4-14	Wide-body jet network topology changes over 212 months from 1/1990 to 8/2007. .	139
4-15	JetBlue topological vector distance to canonical networks. Time period of 91 months from 2/2000 to 8/2007. Two transitions around months 22 and 70.	140
4-16	JetBlue months 23, 70 and 91	140
4-17	Southwest topological comparison to canonical network over the period 1/1990-8/2007. Weak topological transitions at month 45 and 150. 217 months including the 5 1980s snapshots: 7/1980, 2/1982, 2/1983, 1/1984 and 4/1988.	141
4-18	Southwest months 24, 96 and 212	142

4-19	Wide-body jets topological vector similarity to canonical network. 212 months; (top) all matches, (bottom) closest matches only - BA graph (BA), tertiary tree (ttree), binary tree (btree), Newman-Gastner graph (ng5), hierarchy with random interlinking (DWS R).	142
4-20	Continental Airlines topological vector similarity to canonical network. 212 months; (top) all matches, (bottom) closest matches only - tertiary tree (ttree), BA graph (BA), binary tree (btree), random modular graph (mod), Newman-Gastner graph (ng5).	143
4-21	Pseudocode for the hub seeding model.	145
4-22	Topology over time comparison: JetBlue topology comparisons with the new hub seeding model (red circles). The hub seeding model outperforms the rest of the canonical topologies for most data points, and consistently for last 15 months. The top plot shows the topology vector distance to the real graph; the bottom plot shows the topology rank based on that distance.	146
4-23	Topology over time comparison: Airtran topology comparisons with the new hub seeding model (red circles). The top plot shows the topology vector distance to the real graph; the bottom plot shows the topology rank based on that distance.	147
4-24	Topology over time comparison: Continental's topology best matches over time including the new hub seeding model (red circles).	148
4-25	Pseudocode for the Southwest expansion model.	149
4-26	Topology over time comparison for Southwest Airlines. Comparing the "Southwest model" to canonical networks. The bottom plot shows the best topologies only. Includes the 5 months from the 1980s. No major topology transitions are seen between the 1980s and 1990s growth.	150
4-27	Southwest top 80 topology evolution: 1/1990-8/2007. Hub seeding model clearly outperforms the rest as expected, in the last 60 months. The bottom plot shows topology rank (rather than vector distance) - which canonical graph (or model) is number 1, 2,... 24.	151
5-1	Number of articles versus days: first year growth in number of articles of the French, Italian, Spanish, Russian, Chinese, Esperanto and Interlingua Wikipedias. The Esperanto Wikipedia grows the fastest, though is currently outgrown by all the active language Wikipedias. All other Wikipedias see very slow growth in the first 200 days. The Spanish Wikipedia actually gets its first edge on day 285, and has few disconnected pages prior to that, but was plotted from 285 to 650 days to show comparative growth.	155
5-2	Average number of hyperlinks per article for the first year of the French, Italian, Spanish, Russian, Chinese, Esperanto and Interlingua Wikipedias. All Wikipedias are normalized to the same date (as if we they started to grow together). For example, the Spanish Wikipedia does not see growth until day 285, so its timeline on this figure is day 285 to day 650.	156
5-3	The first year of the Interlingua Wikipedia: total number of articles and hyperlinks and number of articles and hyperlinks in the giant connected component.	157
5-4	(left) Day 75 of the Interlingua Wikipedia. The star cluster is centered around the topic "languages". The two smaller components in the middle contain articles about space/astronomy and sciences. (right) Network modules of the giant component of day 75 (of the network on the left).	158
5-5	Topology profile for day 75 of the Interlingua Wikipedia.	158

5-6	Significant motifs for day 75 of the Interlingua Wikipedia: stars and base-triangle.	159
5-7	Topology evolution of the Interlingua Wikipedia. Days 2 to 75. Three major transitions	160
5-8	Days 20, 30, 45 and 75 of the Interlingua Wikipedia	160
5-9	The 200 days of the Esperanto Wikipedia: total number of articles and hyperlinks and number of articles and hyperlinks in the giant connected component. The giant component is small compared to the overall size in the first 100 days and then catches up and becomes 80-90% of all nodes.	161
5-10	Day 100 of the Esperanto Wikipedia: largely disconnected, with the main component centered around the word "encyclopedia" and featuring various topics. Other connected components include topics such as "Internet terminology" (ex: server, unicode), "Beijing places of interest and culture", names of geographic regions, lots of Chinese provinces. (right) Day 100 modules by the Newman eigenvector method.	162
5-11	Topology profile for day 100 of the Esperanto Wikipedia.	162
5-12	Significant motifs for day 100 of the Esperanto Wikipedia: stars and base-triangle, with Z-scores 0.121, 0.102, 0.195, 0.447, 0.01, 0.817 respectively	163
5-13	Topology evolution of the Esperanto Wikipedia. Days 21 to 115. Three transitions.	163
5-14	Days 40, 70, and 100 of the Esperanto Wikipedia	164
5-15	The first year of the Spanish Wikipedia: total number of articles and hyperlinks and number of articles and hyperlinks in the giant connected component. The giant component is small compared to the overall size in the first 200 days and then catches up and becomes 80-90% of all nodes. This transition happens during days 250-260.	165
5-16	(left) Day 235 of the Spanish Wikipedia. Different components include topics such as electro-technics, software, continents and Earth/planetary terms; (right) Modularization of the giant component of day 235 of the Spanish Wikipedia.	166
5-17	Topology profile of day 235 of the Spanish Wikipedia	166
5-18	Topology evolution of the Spanish Wikipedia, days 104:250	167
5-19	Days 124, 164, 204 and 250 of the Spanish Wikipedia	167
5-20	The first year of the Russian Wikipedia (days 8 to 356): total number of articles and hyperlinks and number of articles and hyperlinks in the giant connected component. The giant component becomes the majority of the network around day 275. About four major jumps in growth can be detected.	168
5-21	Day 180 of the Russian Wikipedia and its giant component modularized; The topics of separate components are 1/compounds, amino-acids and other molecules (this is the component that will outgrow the giant component on day 190) and 2/around the main page sciences and disciplines of study, such as chemistry, history, and ecology. It is interesting that the wiki main page is not in the giant component at that time. The most central article in the giant component is "Russia".	168
5-22	Topology profile of day 180 of the Russian Wikipedia. Best match are hierarchies, core-periphery and random-interdivisional and random graphs.	169
5-23	Topology evolution of the Russian Wikipedia, days 51 to 253. The first 50 days were too small to analyze (essentially two edges). A single component dominates the network, until day 190, when a star component centered around the topic "compounds/molecules" takes over.	170
5-24	Snapshots of the Russian Wikipedia history: days 180, 200, 220 and 250	170

5-25	The first year of the Chinese Wikipedia (days 5 to 325): total number of articles and hyperlinks and number of articles and hyperlinks in the giant connected component. Unlike all other Wikipedias, this one stays largely connected: the giant component contains almost all nodes. Another novelty is the large density, i.e. many more links per node on average.	171
5-26	Day 100 of the Chinese Wikipedia (giant component), modularized using the Newman-Girvan algorithm.	172
5-27	Significant motifs for day 100 of the Chinese Wikipedia: stars with Z-scores 0.166 and 0.338 respectively.	172
5-28	Topology evolution of the Chinese Wikipedia, days 5:158. One major transition from a pure-star to a multiple-star topology, at day 22.	173
5-29	Days 21, 22, 25 and 150 and 200 of the Chinese Wikipedia.	173
6-1	(Left) Lufthansa Airlines world routes 2006, modularized using the Newman-Girvan algorithm and (right) giant component of day 75 of the Interlingua Wikipedia, modularized using the Newman eigenvector algorithm.	176
6-2	Network Evolution Analysis Process	178
A-1	Number of airports and number of OD pairs monthly for all US airlines reporting to the BTS [1]. In August 2001, the DOT proposes the addition of military, cargo and charter flights to the reports, hence the jump in the data (source: http://www.bts.gov/publications/federal_register/2001/html/bts_20010828.html) . .	185
A-2	Total number of seats offered versus number of destinations. All data slices. 1990 and 2007.	186
A-3	Average path lengths for the eight major airlines and all low-cost airlines discussed in Chapter 4, 1/1990-8/2007	186
A-4	Passengers carried for the eight major airlines and all low-cost airlines discussed in Chapter 4, plotted for the month of August only, yearly from 8/1990 to 8/2007. . . .	187
A-5	ATA Airlines topological similarity to canonical topologies over time, 212 months, 1/1990-8/2007	187
A-6	Graph-theoretical metrics for the Interlingua Wikipedia: degree correlation, average path length, diameter and s-max measure.	188
A-7	Nodes and edges in the history of the Simple English Wikipedia from day 550 to day 730.	188
A-8	Day 730 of the Simple English Wikipedia. The fully connected cluster in the middle consists of the months of the year (i.e. the pages for January, February, etc.). . . .	189
A-9	Topology profile for day 730 of the Simple English Wikipedia.	189
A-10	Simple English Wikipedia significant motifs for day 730 (2 years)	189
A-11	Canonical networks comparison for the Simple English Wikipedia, days 550 - 730 . .	190
A-12	Growth of the French and the Italian Wikipedias	190
A-13	Days 250 and 365 of the French Wikipedia	191
A-14	Days 470 and 500 of the Italian Wikipedia	191

List of Tables

1.1	Simple graph representation data structures (N-number of nodes, M-number of edges) and the corresponding examples for the bowtie graph.	34
1.2	Node centrality measures and examples for the bowtie graph (see Figure 1-1).	36
2.1	Canonical network definitions	65
2.2	Statistics for some canonical graphs - m/n , r , C , l and d	67
2.3	Algorithm complexity for the components of the topology vector: density, clustering coefficient, degree correlation, s/s -max and diameter. The number of nodes is n , the number of edges m and the average nodal degree is k	71
2.4	JetBlue 8/2007 statistics for weighted (by number of departures monthly) and un-weighted route versions. Note that the original directed network is not connected.	71
2.5	Comparing canonical topologies using the "best-matching-sequence" diagonal sum measure of the similarity matrix. All graphs have 100 nodes. The graphs compared are star, binary tree, tertiary tree, triangular, square, hexagonal lattice, BA graph, random modular graph and an Erdős-Rényi graph.	80
2.6	List of motifs with positive Z-score for the JetBlue 8/07 route network.	85
3.1	Aircraft Types according to BTS [1] classification.	102
3.2	Correlation table of graph-theoretic "hub" metrics and the industry-developed. All airlines and data slices used for correlations.	113
3.3	Correlation table of graph-theoretic "hub" metrics and the industry-developed. Only single airline data points used for correlations.	113
3.4	Table of relative topology profile correlations. JetBlue, Continental, the wide body jets and the top 50 flights by seat capacity tend to be the most similar.	118
3.5	Continental Airlines (8/07) significant motifs statistics.	124
3.6	Wide-body jets (8/07) significant motifs statistics.	125
3.7	Top 50 by seat capacity significant motifs.	126
4.1	Low-cost carriers average growth rate, in terms of number of new destinations per month, 1990-2007.	130
4.2	Legacy carriers average growth rate, in terms of number of new destinations per month, 1990-2007.	131
4.3	Outline of the hub seeding growth model.	144
4.4	Counting frequencies of events according to the hub seeding model in various airlines.	145
4.5	Outline of the Southwest expansion model.	149
5.1	Table of Language Wikipedias analyzed in this thesis. The rank is by size among all Wikipedias. The code is used for the web address of the Wikipedia. The depth column is a rough indicator of how frequently the Wikipedia is updated.	154

A.1	One line from the US BTS [1] data. A JetBlue (B6) flight from Boston to Austin in Jan 2007, 30 departures.	192
A.2	All data slices of the US airlines dataset.	194

Chapter 0

Introduction

0.1 Complex Systems Change

0.1.1 Growth and Complexity

In August 2007, there were 850000 commercial and civil aircraft departures in the US [1]. More than 800000 of these flights offered more than 100 seats and about 700000 were jet-powered aircraft. These figures describe a massive flow of people across and beyond the continental US in only a single month. Despite recessions in this decade, projections are for steadily growing air traffic which will require an increased capacity and create infrastructural pressure on the US national airspace system. Though the imminent challenges such as airport congestion and air traffic control capacity-related problems are well-understood, the emergent problems associated with system growth cannot be foreseen. Are there ultimate physical limits to air transportation growth and is the system growing slowly enough, that with time we will be able to handle this technological challenge? How does technology infusion affect the system interactivity and performance over time? Neither technological advances, nor the economy can be forecasted very well - but growth can also be studied historically.

Apart from air transportation, the same challenges of growth and complexity affect many technological systems. With the explosion of Internet users in developing countries like China, and especially of mobile Internet users, it is unclear whether the communication protocols and designed-to-bandwidth routers will handle the massive flow of information robustly. Backbone cable failures (ex: Yellow Sea and Mediterranean) have already caused massive outages in entire countries. The challenge is in handling both growing information flow and increasing interactivity (more connections per node). The same can be argued for projects with long life-cycles which have to handle innovation and upgrades continuously. How changes propagate through the system can determine its future design and whether it can meet performance targets [13].

This thesis studies the larger patterns of growth in the evolution of technical systems, and especially non-dimensional patterns, to find out whether these patterns exist independent of system size. We combine metrics from the literature and develop new tools to analyze airline networks, which are networks of airports connected by flights. While the models and tools are informed and shaped by airline industry data, they are not exclusively applicable to airline networks growth. To argue generalizability, we also briefly look at growth of several language Wikipedias.

0.1.2 Visualization of Complex Data

In mathematics, a problem is said to be half-way solved if it is stated and represented in mathematical terms properly. For an engineer, visualizing a problem or a piece of data in a certain way can often be enough to diagnose a problem and propose a solution. In representing large systems as networks, the challenges are staggering, because it is very hard to put all aspects of the data on one chart. For example, the goal of Figure 0-1 is to show the top routes in the US by number of departures. Chicago-New York and BOS-NY-DC are the most frequented routes. This figure fails to show the type of aircraft flying on these routes, the number of passengers being hauled across and it does not tell what airline flies where. Other than the airspace operations, there are other layers of information not represented here, such as ground operations, crew rotations, airport infrastructure, flight corridors, and airspace restrictions. All of these will need to be considered to fully understand the limits of growth properly.

Understanding how the entire industry changes over time requires a multi-faceted approach. While this thesis uses graphics heavily to represent results, the need for unified, possibly interactive visualization tools is clearly recognized.

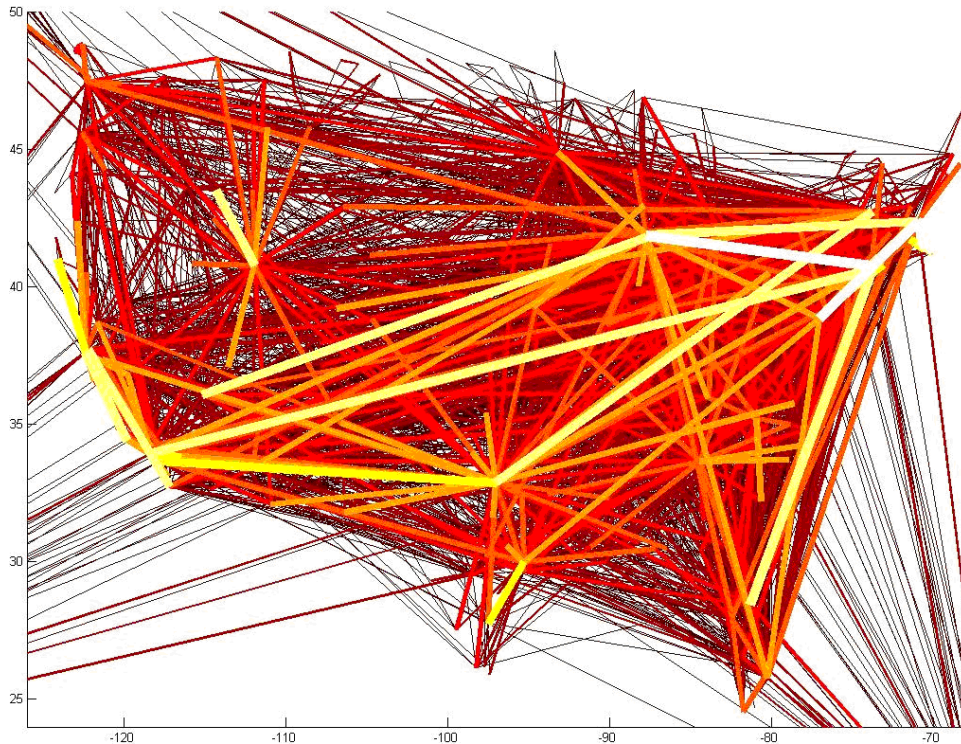


Figure 0-1: US airline flights in August 2007. High-frequency departure routes are highlighted. Alaska, Hawaii and Puerto Rico / US Virgin Islands are not shown.

0.2 Complex Systems and Topology

Network representation is one way to visualize systems of different size and type. The essence of a network is its elements and how they are connected. Network *topology* is the term that describes the layout and organization of the network. Conceptually, a topology is non-dimensional. For example,

star formations are star formations regardless of the number of spokes - there are many spokes and one hub. Topology is studied because it puts seemingly different networks under the same umbrella of "physical" laws and tools for analysis. Early research associated topologies with degree distributions (distributions of connections per node) [14]. More recent papers discuss frequently occurring patterns [7], network modularity [15] and node roles/classification [16]. In this thesis we review the topology models in the literature and analyze real systems against the spectrum of existing models, as well as against a set of non-dimensional metrics. We create a comprehensive view of network topology, in a novel and useful way.

This idea of comprehensive view of topology was deemed important because there are cases in the literature where single metrics were shown to be insufficient in understanding the structure of the system [4] compared to in-depth study and mapping of the nodes and connections. In our topology profile, we combine metrics capturing different aspects of the topology: reachability, degree correlations, clustering, density and scale-freeness. This is shown to be useful in distinguishing between different airline topologies.

0.3 Complex Systems and Topology Evolution

Human-made systems evolve, with and without human intervention, under physical, environmental and market forces. Any system humans build and operate is limited in growth by the resources of our planet, and our ability to create and sustain wealth, using these resources. The airline system is a great example of a growing technological system pushing against environmental and physical (capacity, financial) constraints. Understand the growth patterns in order to manage the systems we operate better is crucial for our ability to sustain and improve our standard of living.

The value in studying topology evolution across different systems is in the potential of finding universal principles and in developing multi-disciplinary tools, such as machine learning in biology and network theory in communications. In this thesis we apply network theory algorithms to study the route evolution of airlines, with the goal to study patterns more comprehensively than previous studies. From an airline expert's point of view, this approach is unusual. It is the applied mathematician's search for patterns in the real world. It is interesting to correlate the two sets of knowledge and use the mathematical description to assess the performance of a given airline. We devise two growth algorithms derived from findings of topology patterns over time, and show that they match airline topology better than canonical topologies from the literature. We claim that these two models capture mechanisms of growth for JetBlue Airlines and Southwest Airlines.

An empirical question about topology is whether it changes over time, or whether it stays constant with size. In other words, are there phases of growth a system goes through, with clear topology transitions, or not? Also, are these phenomena system-dependent, or are they detectable across systems? In this thesis we attack this question by studying empirically the topology over time of various airline networks and some Wikipedia networks. We find examples of both topology transitions and constant topology histories, and also examples of similar topology histories for different networks. We use the similarity in patterns over time to devise topology-derived growth models. However, the issue of identifying specific domain-dependent drivers of network growth and evolution and triggers of phase transitions is left for future work.

0.4 Airline Networks

There is a wealth of literature on airline networks - studying topology of airline routes and structure of operations from many angles. In Chapter 3 we review the most relevant work, divided in three areas: i/ studies on airlines using pure network theory only (without domain knowledge), ii/ studies in air transportation management that use network theory tools and iii/ econometric work on airline structure, based on profit and cost-maximizing models. This thesis falls in the first category of studies.

Research on airline networks using network theory only is popular due to the availability of data, and the easy representation of airport systems as networks. Guimera et al [17] study the global airport structure from the point of view of node centrality, and community structure. The authors consider the world-wide airport network, taking one year worth's data (OAG) and represent it as an almost symmetrical graph (adjacency matrix). They find that the path lengths are small for the size of the graph, so it is a small-world. They use centrality measures to classify airports according to their "role" in the network - global hubs, connectors, regional hubs, peripheral airports and ultra-peripheral airports. These ideas are further developed by the same authors in [16] where the premise is that within-module properties of systems represented as networks are different from global properties. Using a simulated annealing algorithm, they identify modules in various networks in biology, air transportation and communications (the Internet). Then they define node roles based on nodal connectivity within the modules and across modules. Finally they look at the connectivity profile for a network between different nodes of roles classes (hubs to hubs, peripheral to ultraperipheral etc) estimated against a random background.

In air transportation management, there are two interesting pieces of relevant work relevant, by Bonnefoy and Wojahn.

Bonnefoy [12] studies scaling mechanisms by which the airline industry has met growing demand in the past and is expected to do so in the future. He shows that the National Airspace System is not scale-free from the point of view of network theory, due to capacity constraints at major airports. He shows that the system has evolved to grow via multi-airport systems in metropolitan areas and if those are modeled as aggregated nodes, the entire systems does become scale-free (measured by the degree distribution). He performs in-depth case studies of various multi-airport systems and studies how they develop to provide recommendations for future airport infrastructure management.

Wojahn [18] studies the airline industry as a whole, discussing carrier statistics around the world. He uses "measures of network structure" to describe transformations in the industry, and finds that more and more, airlines adopt a hub-spoke model. The author analyzes the hub-spoke model versus the point-to-point model using cost and profit equations as a function of influence of travel time, flight frequency. He finds that the hub-spoke model is optimal if the passengers' valuation on flight frequency is high and of travel time is low. The number of hubs is also considered - where the result is that if congestion and slot restriction are in place, a multi-hub network is more profitable for he airline, compared to a single-hub network or a point-to-point network. Finally, Wojahn creates an asymmetric demand model, to reflect the fact that cities have different characteristics and different demand for travel. With the assumption in the model that spokes are connected to a single-hub and that hubs are fully connected, he finds that the cost-maximizing structure is a mixture of a point-to-point and single-hub networks.

0.5 Research Goals. Hypotheses

The focus of this thesis falls in three general areas:

- i (network theory) understanding network topology and topology evolution;
- ii (airline networks) applying topology ideas to analyze US airline routes in the period 1990-2007;
- iii (general tools) applying the network topology evolution tools to other growing networks, such as language Wikipedias to detect potential patterns of interest.

0.5.1 Network Topology and Topology Evolution

Hypotheses

- Non-dimensional graph theoretical metrics can uncover significant structure of real systems independent of size.
- There are clearly defined phases of evolution. And our proposed metrics will be able to detect these stages of evolution such as hub growth, new hub seeding, increased cross-linking, system capacitation (increasing capacity of selected "high profit" existing nodes and edges without significantly changing topology), as well as the impact of major external events such as bankruptcies and mergers.

Research Approach

Topology: combine existing metrics non-dimensionally into a vector profile, assess comparison measures; review and order "canonical networks" based on generative models in the literature; create topology profile based on "canonical networks"; search for frequently recurring motifs.

Topology evolution: use the topology metrics to plot time-varying comparisons for real data topologies; propose better-fitting models using the underlying motif structure, detect phase transitions in system evolution.

Rationale

Network topology describes how the elements of a network are arranged and connected. We aim to find out whether it can be quantified using simple non-dimensional metrics. The advantage of non-dimensional metrics is that they can characterize network topologies independent of network size, i.e. the number of nodes and edges, and can therefore be a basis for comparative studies of the same network at different times of its life or for comparison between different networks at the same time. This involves studying canonical topologies from the literature, such as simple idealized topologies (stars, trees) to constructed and optimized examples (hierarchies with randomized components) to topologies corresponding to growth models such as preferential attachment and random graphs, filling the spectrum between "hub-spoke" and "point-to-point". Then a network is said to exhibit a certain topology if it has similarities to a certain canonical topology or class of topologies. This will aid in describing different complex systems using similar tools.

With the concept of topology clarified, and a set of validated topology comparison metrics, we plan to study how topology changes over time. This means that for every time-tagged snapshot of real data (for example, a monthly instance of an airline route network), we will compare the point to all canonical topologies and track that comparison over time. The goal is to find whether topology goes through major phase transitions as a function of growth, and whether there are phases or stages of growth, general or specific to real systems (ex: the airline routes). It is likely that if patterns

of growth are discovered they will be different for the airlines, which are transportation systems and Wikipedia, which is a knowledge network. These differences can then be further explained by the presence of certain flow constraints that may only exist in physical systems that transport matter, energy and information, but that are irrelevant in knowledge networks where it is possible to instantly jump from one node to another in the network without violating any laws of physics.

To summarize, the goals for this research question is to i) distill a set of valid non-dimensional topology metrics, and assess metrics from the literature; ii) create a set of canonical topologies from simple and regular to random with as much continuity as possible between topology classes (ex: trees and hierarchies, random graphs with various properties); iii) plot the topology spectrum of system examples and study topology evolution looking for transitions, or stages of growth or decline.

0.5.2 Airline Networks

Hypotheses

- There are distinct stages in airline growth and evolution.
- Simple custom models based on underlying network motifs, combined with modest domain knowledge can perform better than canonical graph models in comparing network topology evolution.
- Southwest is topologically distinct from other airlines; this anomaly may be directly linked to its profitability but linking network evolution explicitly to competitive pressures and economics is beyond the scope of this thesis.
- Simple metrics are not enough to explain the network structure of Southwest.
- Southwest is likely to evolve into a more "conventional" network over time.
- The legacy carriers and the so-called "low cost" airlines can clearly be distinguished in terms of their network characteristics.
- Most airlines exhibit a dominant topology that takes advantage of efficiencies gained by hub operations - Southwest airlines is likely to be an exception.

Research Approach

- Split airline dataset into aircraft type, short and long haul networks. Extract single airlines as well, in particular, the top eight carriers (American, Continental, Delta, Northwest, United, US Airways, America West and Alaska Airlines), and a set of low-cost carriers such as JetBlue, Southwest, Frontier, Spirit, ATA, and Airtran. Analyze these datasets statistically, with both graph-theoretical and industry metrics, and plot their metrics altogether.
- Study recurrent patterns in a smaller set of data slices, including JetBlue and Southwest.
- Analyze the topologies over time (1990-2007) for the same small set of data slices and extract patterns of growth.
- Devise custom models from the underlying patterns and test them, using only the system's initial conditions and assumed internal growth parameters/probabilities to guide system growth and evolution. Test whether such models perform better than classical nodal degree based models such as preferential attachment.

Rationale

Airline networks in this study are modeled with airports as nodes and existing flights as links.

The entire set of routes in the United States flown by US airlines is difficult to analyze as a whole. Other than size and complexity, it is challenging to claim that the industry is evolving under

common economic or other principles. This is why the first step is to split the industry into network slices that make sense economically and technologically, i.e. by separate airline or by aircraft type, by distance flown, or by flight frequency. We are interested in comparing these slices statistically and concentrating on a few representative networks, such as JetBlue Airways, Southwest Airlines and Continental airlines. A natural question is whether legacy airlines and low-cost airlines have different topologies, and different recurrent patterns over time. Also, we discuss the relevance of network tools in the airline industry - and where the limitations of these tools are in explaining airline behavior.

Finally, we plan to tackle the question of Southwest Airlines as an outlier by all statistical measures. We discuss a custom growth model for Southwest.

0.5.3 Wikipedia

Hypotheses

- It is expected that Wikipedia network growth shows different patterns than the airline set because its growth is not subject to the same flow constraints and certainly does not have the same purpose - transportation.
- With a statistically large number of users (authors) differences in local structure between different languages will emerge.
- Also, while small and new Wikipedia networks initially look very different in terms of topology (since their growth is driven by only a few individuals with special interests and knowledge), over time as the number of nodes grows into the hundreds and the thousands these networks start to look more similar.
- Growth happens by coalescence: many smaller modules organized by topic grow separately and eventually connect. This is a similar pattern as has previously been observed in change propagation networks which can also be interpreted as a form of knowledge network.

Research Approach

- Wikipedia data is downloaded from data dumps regularly performed by Wikimedia, which contain the histories of all changes, edits, insertions, deletions. Only the first year is considered for analysis because these networks quickly grow too large to be handled computationally with the tools available.
- Each history is broken down into daily snapshots, which represent the network of the Wikipedia for that day. Given that different Wikipedias grow at different rates, for every language a suitable number of days is taken for analysis depending on the number of nodes and edges up to that day.
- The Wikipedias downselected for analysis are the Interlingua, the Esperanto, the Simple English, the Russia, Spanish, French, Italian and Chinese Wikipedias.

Rationale

It is our goal to develop general tools and concepts for network evolution analysis. The purpose of the Wikipedia analysis is to (i) demonstrate that the non-dimensional network topology metrics and methods are broadly applicable and (ii) to examine whether physical (transportation) networks and (virtual) knowledge networks behave differently. Suitable growing networks for benchmarking are different language Wikipedia datasets. These are not technological (transportation) networks, but rather "knowledge networks". They are governed by cognitive, rather than physical limits and may have different growth patterns. These are networks with articles as nodes and hyperlinks as edges. Hyperlinks are considered only as part of the text of the articles, so they are real cognitive

references between two topics. There are various auxiliary links associated with Wikipedia pages, which are not considered here. We plan is to compare a few language Wikipedias to each other and discuss the how fast they grow and what the patterns are. The same set of analysis tools will be used as for the airlines example.

0.6 Thesis Roadmap

This thesis is organized in 7 chapters. This Chapter, 0, presents the introduction. Chapter 1 reviews network theory relevant to the study of topology and its evolution. Chapter 2 uses the ideas from the literature review to develop tools to study network topology statistically. Chapter 3 applies the tools from Chapter 1 to our first case study - US airline networks. Chapter 4 discusses evolution of networks, tracking changes in topology and growth models using the first case study - the airlines. Chapter 5 aims to generalize the ideas and tools from previous chapters, by applying them to study the evolution of language Wikipedias. Chapter 6 contains the conclusion with contributions, limitations and ideas for future work.

Chapter 1

Network Theory Review

This chapter contains a review of network theory literature relevant to the study of network topology and its evolution. We discuss basic representation of systems using graphs, statistics on graphs, modularity, motif search and coarse graining. We use the algorithms and ideas developed in the literature to develop the discussion on how to analyze *topology* in Chapter 2. The most complete sources on this material are by Wasserman [19] and Newman [2].

The second half of this chapter presents models of network growth from the literature, including random graph models, node-degree based models, edge-centric models, as well as node-copying and module growth models inspired from biology and other disciplines. The most complete source on network evolution is by Dorogovtsev and Mendes [20].

1.1 Network Theory Basics

1.1.1 Introduction

Network theory is a modern branch of graph theory, concerned with statistics on practical instances of mathematical graphs. Graph theory itself started with Euler who solved a bridges crossing puzzle by clever representation [21]. This clever representation is essentially what a graph is - a collection of points in arbitrary (does not have to be metric) space and a set of links between them. Two of the most compact ways of representing graphs are an adjacency list (a list of nodes with their neighbors) and an edge list (a list of all links represented as pairs of nodes with their edge weights). Simplest to the eye is a visual representation which is suitable for smaller graphs and gets more challenging as their size grows. Figure 1-1 shows example representations of a six-node bowtie graph. Figure 1-2 shows a larger graph - the route network of all US airlines during 1990. 704 airports worldwide are plotted geographically, with 6935 flight legs between them. While some geographic patterns are distinguishable, such as flight patterns to Alaska, Hawaii, Europe and Asia, it is hard to learn much about this graph from the plot. Large graph visualization is the subject of research in many areas [22].

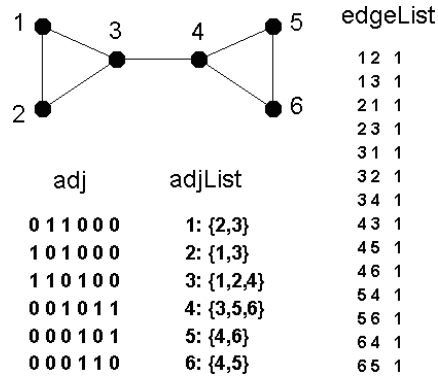


Figure 1-1: Bowtie graph with four example representations: visual, adjacency matrix ($n \times n$, $A_{ij} = 1$ if i and j are linked), adjacency list (list of all neighbors of every node) and edge list (list of pairs of nodes with their link weights). More details on graph representations are given in Table 1.1.2.

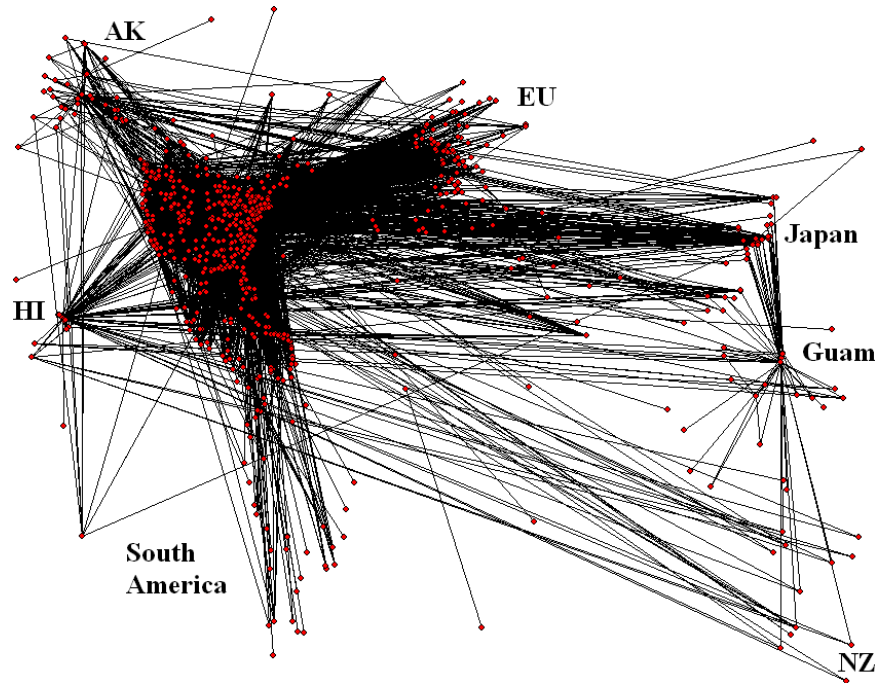


Figure 1-2: Routes of all US airlines during 1990 (BTS [1]), domestic and international. Continental US, Europe, South America, Alaska, Hawaii and Guam and Japan are easily distinguishable.

Statistics on networks first became interesting to social scientists who have used such approaches for more than 50 years to understand the dynamics of behavior in interconnected groups of people [19]. They often ask simple "centrality" questions to find out who is most influential, or what is the degree and profile of connectivity between the people in a group. They also study group structure and decomposition but usually in fairly small networks. More recently, the legacy of social science methods has been adopted by physicists who focus on statistical properties of very large graphs. This is driven by the recent availability of large datasets and the ability to collect, store and represent massive amounts of information. Typical networks of study can have hundreds and thousands

of nodes (ex: organizational network or protein interactions [23]) to a trillion nodes (www-graph 2008 [24]). Earlier work concentrated on similar centrality, robustness and sensitivity questions: How would the removal of a node affect the connectivity or flows in the entire network? What are the critical nodes to keep? Given the statistical implications of a million nodes, centrality is hard to pin down. There can be collections or clusters of important nodes. Statistics on nodes is only a step towards answering the questions about decomposition, structure, and function.

Recent research on structure and function includes work on community finding, modularity [25], biological applications of motifs [26], which are frequently occurring statistically significant subgraphs, and coarse-graining [27]. Modularity is of special interest in identifying communities in social networks and uncovering functional modules in technical networks. Using the network theory approach to modularity can be challenging in many engineering systems, where modularity is not exhibited in the connectivity in terms of cohesiveness, but with complexity built into the interfaces and not in the number of links.

The next question in this line of research is how certain network structure / topology came about - what are some generating principles reflecting function, purpose and the external environment. So far generative models have focused on linking node centrality metrics to the ability to receive new links, and some simple geometric models which optimize new link arrival based on Euclidean geometry.

Inevitably, this brings us to the Holy Grail of network theory - is it possible to understand, visualize, analyze and mathematically model the structure and dynamics of any graph? How to classify graphs according to structure and sub-graphs? And how is structure related to function? [2]

Understanding structure is challenging because the systems people plan, build, and deal with become more complex every day - either more intricately connected, more highly integrated in larger systems, and inside or just vastly larger, and with multiple scales of operations. Tools for systems engineering exist but not to deal with levels of complexity and emergence due to interconnectedness. Another problem is visualizing such complex and vast structures. Today, it is inconceivable to plot the network of hyperlinks of the World Wide Web or even of the Internet at the router level, which currently spans only about 30000 nodes [28], in a way that the structure is revealed. This is why complex systems research today is focusing on statistical, reductionist and bottom-up approaches to understand the topology and evolution of large systems.

This thesis shines some light into evolution of structure¹ of certain graphs describing engineering systems and knowledge networks.

1.1.2 Types of simple graph representations

A graph is an abstraction of a set of objects and their relations. The simplest graph representation of a network is a set of node² pairs. This description says nothing about how the nodes and links may be different in size, nature, color, name, location, capacity, nor does it specify the links as virtual, physical, temporal, constant, weighted, wireless, friendly or hostile. In real systems, nodes and links can be distinct; they can have different weights / sizes and be directional. This is shown

¹Structure and topology are used interchangeably here, though topology is the more precise definition mathematically. See Chapter 2

²The words *node/vertex*, and *edge/link* are used interchangeably.

in Figure 1-3. Various graph representation techniques are listed in Table 1.1.2. Most common is the adjacency matrix representation, which is a $N \times N$ matrix of 1s and 0s, with $A_{ij}=1$ if i and j are connected. In this work, the preferred representation is the edge list, a list of node pairs with their weights, because it allows a natural extension of node and link attributes in the data structure. An augmented edge list has rows with more entries than 3 that add relevant information about the nodes or the links. For example, [BOS LAX 31 A320] is an augmented edge which means that the flight from Boston to Los Angeles has 31 departures monthly, all with an A320 aircraft.

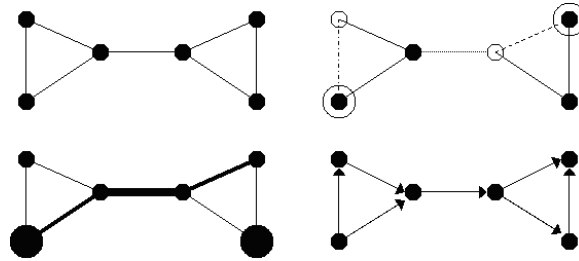


Figure 1-3: Examples of network types: undirected graph with a single type of nodes and links (upper left); undirected graph with discrete nodes and links (upper right), undirected graph with weighted nodes and links (lower left) and directed graph (lower right). Reproduced from [2].

Table 1.1: Simple graph representation data structures (N-number of nodes, M-number of edges) and the corresponding examples for the bowtie graph.

Data structure	Description	Bowtie example
Adjacency matrix	Matrix $A (N \times N)$ where $A(i, j) = 1$ if vertices i and j are adjacent, 0 otherwise.	<pre> 0 1 1 0 0 0 1 0 1 0 0 0 1 1 0 0 0 0 0 0 1 0 1 1 0 0 0 1 0 1 0 0 0 1 1 0 </pre>
Incidence matrix	Matrix $I (N \times M)$ where $I(i, j) = 1$ if vertex i is adjacent to edge j , 0 otherwise.	<pre> 1 1 0 0 0 0 0 1 0 1 0 0 0 0 0 1 1 1 0 0 0 0 0 0 1 1 1 0 0 0 0 0 1 0 1 0 0 0 0 0 1 1 </pre>
Adjacency list	List of adjacent nodes for every node: $\{v_i : [v_j \text{ if } A(v_i, v_j) = 1]\}$	$\{1 : [2, 3], 2 : [1, 3], 3 : [1, 2, 4], 4 : [3, 5, 6], 5 : [4, 6], 6 : [4, 5]\}$
One-line list	Compact representation of AdjList with neighbor list separated by 0s sequentially ordered by node index	2301301240356046056
Edge list	List of node pairs, usually including the strength of the relation, ex $[i, j, w_{ij}]$	$[[1, 2, 1], [2, 3, 1], [1, 3, 1], [3, 4, 1], [4, 5, 1], [4, 6, 1], [5, 6, 1]] + \text{symmetry}$
Graphic	Pictorial representation of nodes, links and their attributes.	Figure 1-1

The simplest graph representation is the most abstract and not always the most useful. Depending on the critical factors for a particular system, the graph description has to be augmented accordingly. Figure 1-4 shows two alternative representations of the JetBlue route network from August 2007. The left plot depicts airports at their respective latitude and longitude, which gives

a sense of the geographic presence of JetBlue Airways. The right plot shows the same graph, but node locations determined by a spring energy algorithm [3]. This representation shows the network connectivity patterns better, with the two major hubs, New York and Boston, and other interesting substructures. In this thesis, we use the spring energy plots primarily for visualization.

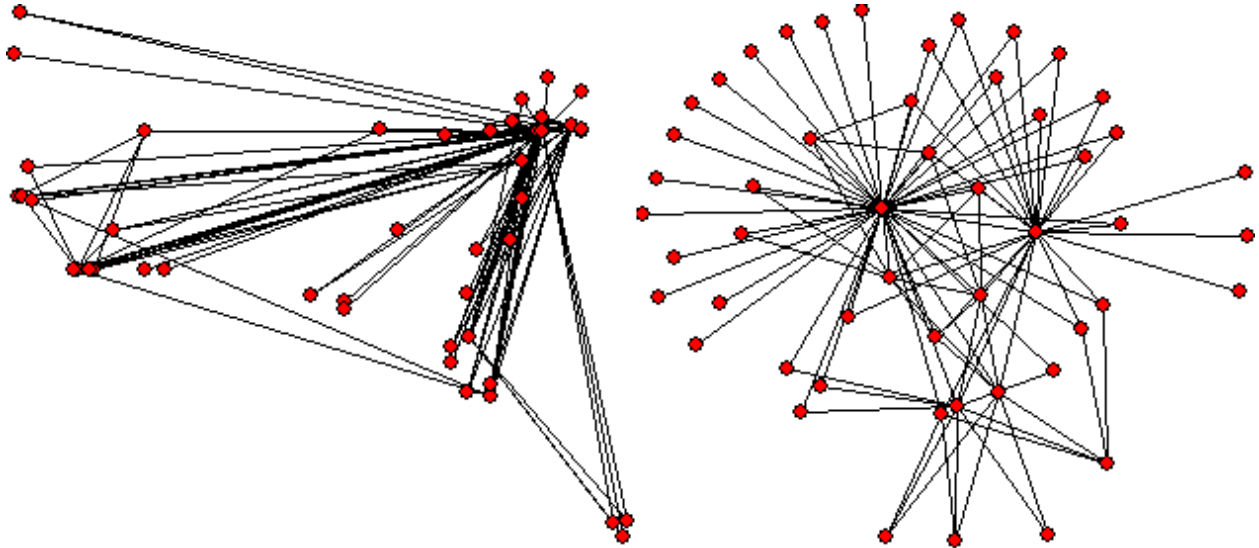


Figure 1-4: Two representations of the JetBlue August 2007 network. Nodes are airports, edges are existing routes. (left) Airports are plotted at their latitude and longitude; (right) The same graph is plotted but with node locations determined by a spring energy algorithm by Kamada and Kawai [3].

1.2 Statistics on Graphs

1.2.1 Size and density - nodes, edges, average degree

The first growth statistic to compute on any system is size. Size for networks is measured in terms of number of nodes and number of edges. Derived measures are the average degree, computed as the ratio of number of edges to the number of nodes, and the density, which is the number of edges, divided by the total possible number of edges ($\frac{n(n-1)}{2}$). Density is a useful indicator for topology, because it shows how relatively busy or interconnected the network is. The least dense (minimally-connected) graph is a tree with $n - 1$ edges, so the minimum possible density is $\frac{(n-1)}{n(n-1)/2} = 2/n$. Every dense network can be seen as built on top of a tree (a spanning tree). Most distribution networks where connection cost and distance are optimized are less dense with tree-like topologies. Figure 1-5 shows that Southwest Airlines is denser than Continental Airlines, i.e. has many more flights per origin-destination pairs.

Density is used in Chapter 2 in combination with other metric for comparing the network topology of graphs.

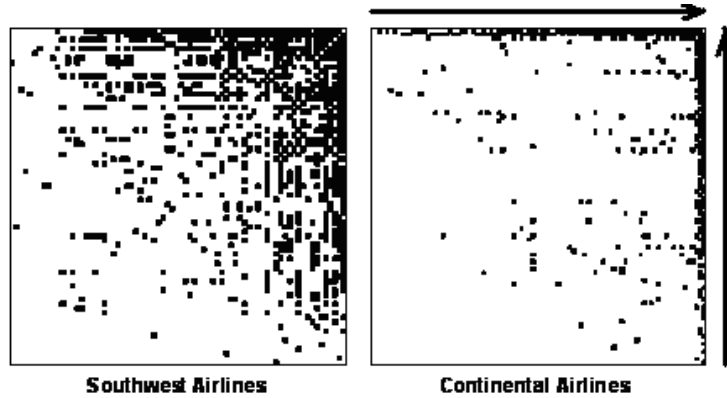


Figure 1-5: Matrix dot plots of the adjacency matrices of Southwest Airlines and Continental Airlines for 8/2007. The Southwest network has 74 airports and 809 one-way flight segments; Continental has 90 airports and 207 one-way flight segments. Every point is filled if there is a connection between cells i and j . Rows and columns are ordered by increasing nodal degree (increasing left and up). Clearly, the Southwest network is denser, and the Continental network features very few high-degree nodes.

1.2.2 Node centrality measures

The degree of a node equals the number of its links. If the graph is directed, then degree is the sum of the in-degree and the out-degree, which are the number of incoming and outgoing links. Derived from social science, the degree has been associated with the "importance" of the node, or its centrality with respect to other nodes in the network. High-degree nodes can be popular people, destinations or important multi-functional proteins. Degree is one of various centrality measures. Table 1.2.2 contains a summary of some node-centrality metrics.

Table 1.2: Node centrality measures and examples for the bowtie graph (see Figure 1-1).

Centrality Measure	Description	bowtie example
Degree centrality	number of links	1: 2, 2: 2, 3: 3, 4: 3, 5: 2, 6: 2
Closeness centrality	the average distance of a node to all other nodes (average shortest path)	1:2, 2:2, 3:1.4, 4:1.4, 5:2, 6: 2
Eigenvector centrality	$x_i = \frac{1}{\lambda} \sum_{j \in M(i)} x_j = \frac{1}{\lambda} \sum_{j=1}^N A_{i,j} x_j$, A - adj matrix, N - number of nodes, $M(i)$ - set of neighbor nodes of i , λ - largest eigenvalue. In vector notation: $Ax = \lambda x \Rightarrow$ eigenvector entries	Largest eigenvalue: $\lambda=2.4142$ Corresponding eigenvector: [0.3536, 0.3536, 0.5, 0.5, 0.3536, 0.3536]
Betweenness centrality	The number of shortest paths that go through i , weighted - $C_B(\nu) = \sum_{s \neq \nu \neq t \in S, s \neq t} \frac{\sigma_{st}(\nu)}{\sigma_{st}}$, σ_{st} is the number of shortest paths from s to t and $\sigma_{st}(\nu)$ is the number of shortest paths through ν	1:0, 2:0, 3:12, 4:12, 5:0, 6:0

In the example of the bowtie graph as seen in Table 1.2.2, nodes 3 and 4 (connected by the

middle edge) have the highest degree, eigenvector and betweenness centralities and the lowest closeness centrality, all of which imply that they are most central to the bowtie. This is not always the case. Various centrality measures do not always have the same message. An example is shown in Figure 1-6.

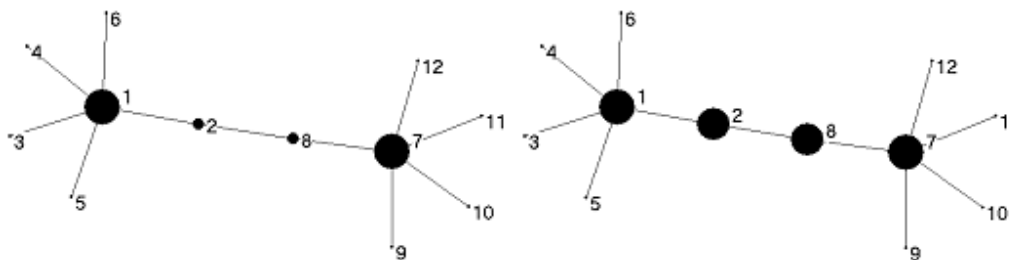


Figure 1-6: Node sizes represent degree on the left, betweenness on the right. Nodes 2 and 8 become more central if measured by betweenness.

Betweenness is a centrality measure that can be defined on both nodes and edges. It is the weighted number of shortest paths that go through a node or an edge (see description in Table 1.2.2). The same definition can be used for an edge, instead of a vertex, though edge betweenness is more complicated to calculate. Edge betweenness is used for graph partitioning in community finding [29]. An example of a high betweenness edge for the bowtie graph is shown in Figure 1-7.

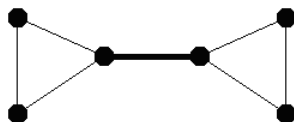


Figure 1-7: Example of a high betweenness edge for the bowtie graph.

Invariably high-betweenness nodes are not also of the highest degree since degree and betweenness are measures of different centrality. This is indicated by Guimera et al [17] who analyze world airports for centrality. They classify them as non-hubs: ultraperipheral nodes, peripheral nodes, nonhub connector nodes, nonhub kinless nodes, and hubs: provincial hubs, connector hubs, and kinless hubs, based on their connectivity within and outside their geographic region. It turns out that global connector hubs which are the most recognized airports in the world (such as NYC, Chicago, LAX, Frankfurt, London, Paris, Tokyo, Beijing) have the highest degrees, but not always the highest betweenness. For example, Anchorage, AK is one of the highest betweenness airports in the world. Obviously, Anchorage is a local hub, and due to the local landscape, weather and the remoteness of the region, flying with small airplanes is a common way to travel anywhere, based out of Anchorage. This explains the high number of shortest paths through Anchorage and its high betweenness.

1.2.3 Degree distributions

The degree distribution $P(k)$ is the frequency of nodes with degree k . The cumulative degree distribution $C(k)$ is the frequency of nodes with degree higher than degree k . While degree is node-centric, compiling the distribution of nodal degrees is used in network-wide analysis. Figure 1-8 shows the histogram, probability degree distribution and cumulative degree distribution for the 1990 airline network plotted in Figure 1-2. The x axis shows the degree, and the y axis the

frequency. This is an example of a fat-tail distribution with few high-degree nodes and many more low degree nodes. The fat tail results in an exponential for the cumulative distribution with a power-law drop off. The drop off signifies that beyond a certain degree there is no more growth. This is expected for airlines because airports do not have infinite landing capacity. The effective degree of an airport, in terms of number of landings per unit of time can be estimate by counting the number of runways and their capacity, and estimating what type of aircraft land there (because they have different landing distance and time separation requirements). This estimate does not say much about number of distinct connections to airports, but it can provide bounds for the cut-off of the weighted degree distribution.

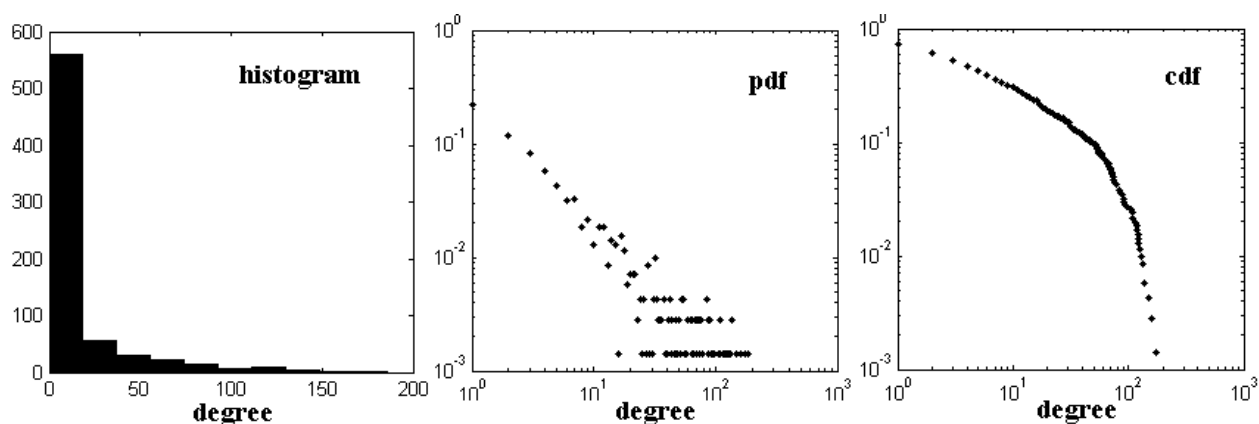


Figure 1-8: Histogram, probability degree distribution (pdf) and cumulative degree distribution (cdf) computed for the network of all airlines routes in 1990 visualized in Figure 1-2. Nodes are airports, links are existing routes, node degree is the number of incoming airport connections. This is clearly an exponential degree distribution, but it can be interpreted as having a power-law cutoff.

A lot of research has been done in linking degree distributions to network structure and more generally topology. For example, original papers [14] claimed that power-law degree distributions correspond to scale-free network topologies. First, that assumes a probability model behind the degree distribution of the degree sequence, i.e. $P[X = x] = f(x)$. Second, the power-law frequency-size distributions ($P[X > x] = cx^{-\alpha}$), which appear as a straight line on a log-log plot, were either measured or interpreted incorrectly. For example, all power laws were assumed to have been generated by preferential attachment [14]. Power laws are often mis-diagnosed because of missing or uncertain data, and wrong representation. Li et al [4] prove the ambiguities of distributions by showing that a single degree distribution can correspond to various types of network structures, i.e. there is a one-to-many mapping between degree distribution and topology.

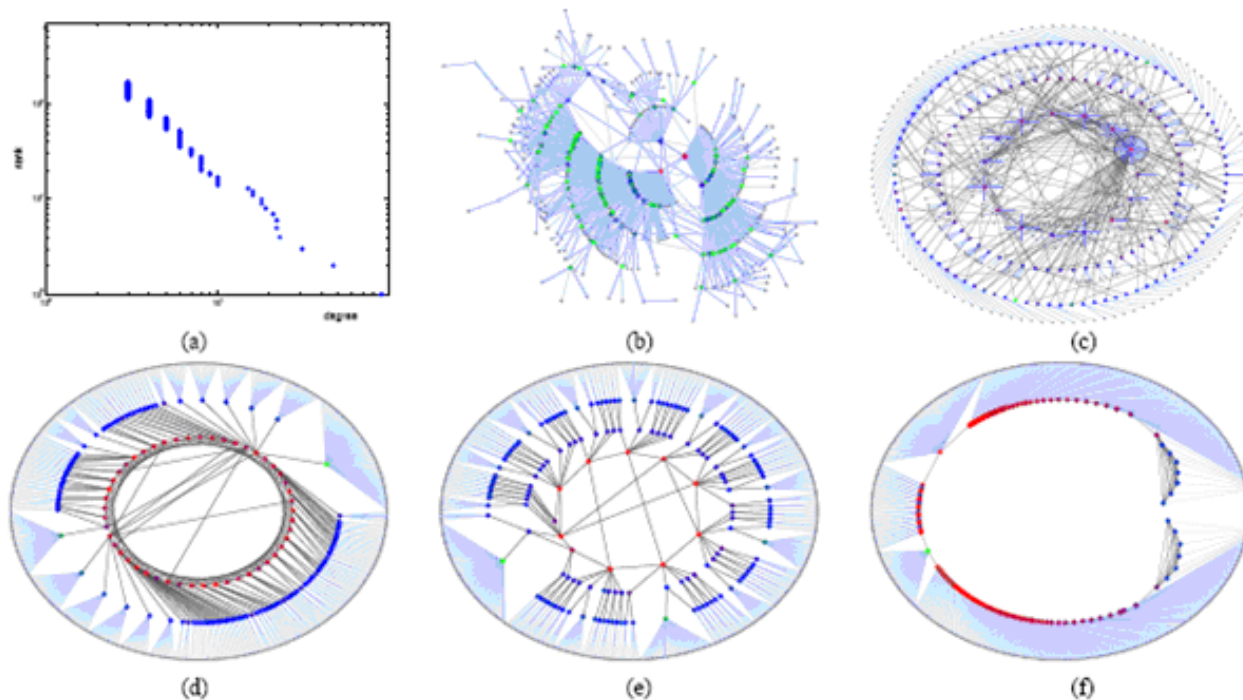


Figure 1-9: Figure 6 from Li et al [4]. Five networks having the same node degree distribution concerning Internet route structure. (a) Common node degree distribution (degree versus rank on log-log scale); (b) Network resulting from preferential attachment; (c) Random graph built with the same degree sequence; (d) Heuristically optimal topology; (e) Abilene-inspired topology; (f) Sub-optimally designed topology.

Figure 1-9 is from Li et al [4] and illustrates various network topologies corresponding to the same degree distribution. Only (b) is generated by preferential attachment. The most optimal network for the routing internet traffic is actually closer to the real topology-inspired design (e). It turns out that the topology in 1-9 (e) also corresponds to 1930 design of AT&T's long distance system [30].

In [31], there is an extensive discussion on representing data points stochastically and non-stochastically and how that leads to misconceptions in detecting power laws. The authors suggest using rank to determine whether a sequence has a scaling distribution. Their definition follows:

Definition: A finite sequence $y = (y_1, y_2, \dots, y_n)$ of real numbers, assumed without loss of generality always to be ordered such that $y_1 \geq y_2 \geq \dots \geq y_n$, is said to follow a power law or scaling relationship if $k = cy_k^{-\alpha}$, where k (by definition) is the rank, c is a constant and α is the scaling index.

The conclusion here is that degree distributions cannot tell the whole story about structure, and should be analyzed with caution. For example, detecting power law degree distributions or scaling distributions in the airline networks cannot imply necessarily that they are *scale-free* or grow by preferential attachment.

1.2.4 Clustering coefficient

Clustering coefficient as a graph measure comes from the social science literature [19]. The intent to answer the question: in what percentage of cases, a node's neighbors (friends) are also neighbors

(friends)? So the clustering coefficient, C , measures the local clustering in the graph. This can be computed by going through the neighbors of all nodes and checking whether they are connected as well. Another way to think about it is in terms of triangles in the graph. What is the percentage of triangles among all triples of nodes. A closed form way to compute that is shown in equation 1.1.

$$C = \frac{\# \text{ triangles}}{\# \text{ triples}} = \frac{\text{trace}(\text{adj}^3)/6}{\binom{n}{3}} = \frac{\text{trace}(\text{adj}^3)/6}{n(n-1)(n-2)/6} = \frac{\text{trace}(\text{adj}^3)}{n(n-1)(n-2)} \quad (1.1)$$

Clustering coefficient is one of the statistics used in combination with other measures to compare graph topologies in Chapter 2.

1.2.5 Degree correlation / Assortativity

Degree correlation is a measure of assortative mixing as defined in [15]. Are high-degree nodes connected mostly to other similar-degree nodes (assortative) or mostly to different-degree nodes (disassortative)? Networks with both characteristics exist and apparently assortativity implies certain structural characteristics. Studies have tried to relate degree distributions to degree correlation, but it turns out that with rewiring while keeping the degrees constant a wide band of degree correlations can be achieved, negative or positive [32].

The Pearson coefficient or degree correlation is defined [32] as

$$r = \frac{\sum_{i,j} (d_i - \bar{d}_i)(d_j - \bar{d}_j)}{\sqrt{\sum_{i,j} (d_i - \bar{d}_i)^2 \sum_{i,j} (d_j - \bar{d}_j)^2}} \quad (1.2)$$

where if m is the number of links in the network, the summation (i, j) is over all links, and

$$\bar{d}_i = \frac{1}{m} \sum_{i,j} d_i = \frac{1}{2m} \sum_k d_k^2 = \frac{\sum_k d_k^2}{\sum_k d_k} = \frac{n^{-1} \sum_k d_k^2}{n^{-1} \sum_k d_k} = \frac{\langle d^2 \rangle}{\langle d \rangle} \quad (1.3)$$

is the average degree of a node seen at the end of a randomly selected edge. Notice that $\bar{d}_i = \bar{d}_j$ by symmetry so the index can be dropped $\bar{d}_i = \bar{d}_j = \bar{d}$. It is essentially the covariance of the two distributions divided by the standard deviations. Positive degree correlation indicates uniform cores of connected nodes, “like prefers like”, whereas negative degree correlation indicates skewed topologies with high-degree nodes connecting to many low-degree nodes. The Pearson coefficient for the bowtie graph is $-1/6$.

The assortativity discussion is important in introducing the s -metric in the following section, which is widely used in our analysis. Related to the notion of how low and high-degree nodes connect, it measures how close a graph is, from its most scale-free counterpart, assuming the same degree distribution. Stars are scale-free graphs, and those commonly appear in the early stages of airline growth.

Degree correlation is another measure used in Chapter 2 to compare graph topologies.

1.2.6 The S-metric; scale free graphs

Given the studies using stochastic measures to claim many systems are “scale-free”, Li et al [31] attempt to demystify the “scale-free” term by showing the ambiguities of previous metrics and suggesting a new metric to quantify better what a scale-free graph is. The S-metric used together with a scalable degree distribution purportedly is a step towards a theory of scale-free graphs.

The S-metric measures the extent to which a graph g has a hub-like score and is maximized when high-degree nodes are connected to high-degree nodes.

$$s(g) = \sum_{(i,j) \in E} d_i d_j \quad (1.4)$$

where E is the edge set, so the S-metric is the sum of products of node degrees across every edge. There is a deterministic routine to construct a graph with maximum S given a degree sequence $\{d_i\}^{i=1 \dots n}$. The algorithm roughly works by connecting the edges with higher $d_i d_j$ products first. The main claim of Li et al [31] is that a graph is scale-free if its degree sequence is scalable, and $s(g)$ is maximal or close to the maximum value. According to the authors, high s-metric graphs with scalable degree sequences (as defined in 1.2.3) exhibit the scale-free properties discussed in the literature with previous metrics (m/n , r , C , l , d). They also argue that real systems (especially the router level of the Internet) are not scale-free, but scale-rich and that this is shown by the s-metric on a spectrum comparison with other random and scale-free graphs. Moreover, in the Internet case study, the low s-metric topology correlates with higher performance, different scales at different resolutions (scale-rich), compared to low-performance high s-metric scale-free (self-similar, strongly-invariant to scale) graphs, with the same degree distribution.

The bowtie graph is its own corresponding s-max graph, so the S-metric is 1.

The relation between r and s is studied by Whitney [32] who showed that if a graph is rewired to have max s then it will have max r as well, and similarly if it is rewired to have min s then it will have min r . The graph generated by rewiring for max r and the unique s -max graph do not have to be the same, but the max r and max s values are at their extremes. There is no algorithm for generating the min s other than rewiring [32]. Also, the normalized $s = (s - s_{min}) / (s_{max} - s_{min}) = (r - r_{min}) / (r_{max} - r_{min})$. So knowing r and its max and min allows one to calculate the normalized s , and vice versa.

1.2.7 Distances - average path length, diameter

Diameter is the longest shortest path between any two nodes in a network. In the pure graph sense, where edges have no associated weight, and there is no underlying Euclidean distance, it is the highest number of hops to be traveled between any two nodes. The diameter of the bowtie graph is 3. Diameter and path lengths are interesting to study because most real networks have surprisingly small diameters. Diameter or path length are studied together to classify networks as “small-worlds”. The idea comes from Stanley Milgram’s experiment in 1969 [33], who found that the average distance letters have to travel in a social network (which was not visible a priori) was 6, hence the phrase “six degrees of separation”.

The small-world phenomenon is discussed in one of the earliest papers in network theory by Watts and Strogatz in 1998 [34] who rewire randomly k -regular graphs and show that random interlinking makes the distances very small.

Transportation and communication networks are distribution networks so their function is to carry cargo, traffic, people, data across their structure where the nodes act as origin, destination and routing points. These networks are obviously set in metric spaces where computing and optimizing distance matters. Individual routing requires computing shortest paths defined in various ways (with or without attributes, edge-wise or Euclidean distances). Interestingly, by the definition of diameter above, the airlines are small-world (high clustering is also part of the definition). While

the networks have 50-200 nodes/airports typically, their diameter is between 3-6 on average. In another example, our computations show that the set of Internet routers in late December 2008 had 30584 nodes and a diameter of 13. For transportation networks that carry passengers small diameter is almost imperative - no one is like to book a flight of even twice the diameter of current airlines. Overall, technical networks tend to be centralized because there are many benefits in designing them that way.

In Chapter 2 diameter is combined with other metrics to compare real graph topology to canonical networks.

1.3 Modularity, Motifs, Coarse-graining

This section introduces tools to study graph structure from a more reductionist, bottom-up point of view, compared to the statistical measures presented in previous sections. Modularity is related to finding cohesive communities in the network, while motifs are frequently recurrent patterns that can be interpreted as building blocks. We use the modularization algorithms to understand the substructures of graphs visually, especially in the case of Wikipedia networks. Also, modularity is used in constructing random modular graphs with the same number of modules (see Chapter 2, Section 2.1.1). Motif finding is used to study common patterns in airline networks in Chapter 3, and to understand the underlying dynamics of these systems (Chapter 4).

1.3.1 Modularity

Clustering is a classical problem in graph theory and data analysis in general. Clustering algorithms can be hierarchical or partitional. Hierarchical clustering algorithms divide the network into two parts at every step creating a binary tree of subgraphs. Partitional clustering algorithms determine all clusters at once. In this thesis, we will introduce and use mostly partitional clustering because it is more suitable to understanding functional and structural modularity. Hierarchical clustering is better employed at analyzing social organizations or hierarchical structures in general, where there is a clear top-down or bottom-up organization of nodes with clear levels or roles.

Finding a set of communities based on connectivity only is not a trivial problem. The main issue is that if no number of prescribed modules is given, dividing into two, three, four or 10 ways has little meaning. Communities here are defined as groups of nodes more tightly connected within and with sparser connections to the outside. These can be sets of articles on a similar topic in Wikipedia, or a regional network for an airline, or a social unit formed by interest or else. This definition comes from social science [35].

In this section, we present three ways to address the classical problem of partitioning: i/ spectral partitioning, ii/betweenness-based (the highest betweenness edge is the cut), and iii/ an eigenvector approach.

Spectral partitioning[15]

Spectral partitioning is a classical way to divide a graph or a problem in two. It was first introduced by Fiedler [36]. The bisection is presented by vector v_i of $\pm s$, in which the sign of each entry indicates the membership of the node. If the edge weight is w_{ij} , then the bisection problem can be formulated as: $\min \sum_{i,j}(v_i - v_j)^2$, subject to the constraints $\sum_i v_i = 0$ and $\sum_i v_i^2 = 1$. The first condition ensures that the two partitions are of equal size (for even number of nodes).

The minimization will tend to assign equal signs to edges with large weights and different signs to “weaker” edges. The second constraint prevents the trivial solutions $v = 1, v = 0$. The minimization formulation can be recast as:

$$\min \frac{\sum_{ij} w_{ij} (v_i - v_j)^2}{\sum_i v_i^2}, \text{ such that } \sum_i v_i = 0 \quad (1.5)$$

The numerator can be expanded and rewritten as $\sum_{ij} w_{ij} (v_i - v_j)^2 = 2 \sum_{i=1}^N (\sum_{j=1}^N w_{ij}) v_i^2 - 2 \sum_{i=1}^N \sum_{j=1}^N w_{ij} v_i v_j = 2v^T L_G v$, where L_G is the Laplacian matrix, defined in equation 1.6

$$L_G(i, j) = \begin{cases} \sum_{j, j \neq i} w_{ij}, & i = j \\ -w_{ij}, & i \neq j \end{cases} \quad (1.6)$$

where w_{ij} is the weight of the edge between i and j . Using vector notation, one can obtain that the minimization in equation 1.5 is $2 \frac{v^T L_G v}{v^T v} = 2\lambda_2$ where λ_2 is the second smallest eigenvalue and the solution to the bisection is the second eigenvector, associated with λ_2 . Half of the entries of the second eigenvector are negative, the other half positive which defines the partition. This is necessary because $v_1^T v_2 = 0$.

The main disadvantage of spectral partitioning is that it is good for bisecting graphs, not so much for finding an arbitrary number of communities or identifying the most optimal number of communities. In general, knowing the number of communities is necessary.

Newman-Girvan algorithm [29]

Newman developed a set of community finding algorithms based entirely on connectivity but set to find *meaningful* communities while addressing the problem of what the right number of communities is. He applies the algorithms to various datasets from social to knowledge to biological networks.

The Newman-Girvan algorithm [29] was the first popular Newman algorithm and is based on betweenness. Betweenness is the node/edge centrality measure which reflects the number of shortest paths going through that node/edge. The outline of the algorithm is the following:

- i Calculate the betweenness for all edges in the network.
- ii Remove the edge with the highest betweenness.
- iii Recalculate betweennesses for all edges affected by the removal.
- iv Repeat from step 2 until no edges remain.
- v Quantify the strength of the communities found.

The metric used to measure the strength of communities, called Q , is computed as follows. Let $k \times k$ be a symmetric matrix e , whose element e_{ij} is the fraction of all edges in the network that link communities i and j . A row or a column in this matrix is $a_i = \sum_j e_{ij}$, which is the fraction of nodes connecting to community i . If every node is its own community then, $e_{ij} = a_{ij}$, (where A is the adjacency matrix). The modularity measure is defined as $Q = \sum_i (e_{ii} - a_i^2) = \text{Tr}(e) - \|e^2\|$, where $\|x\|$ is the sum of the elements of the matrix x . If the number of within community edges is no better than random, then Q tends to 0. If Q is close to the maximum, this indicated strong community structure. Improvements on this metric have been made by Hsieh [37].

The most computationally intensive part of the algorithm is re-calculating betweenness. The intuition behind this algorithm is to split the network at the highest-flow lowest connectivity cuts. This isn't always the same solution as the max-flow min-cut. In addition, the minimum cut solution merely counts edges, whereas a good division considers cuts with fewer edges than expected on average.

For the case of the bowtie graph, as seen in Figure 1-7, the highest betweenness edge is the middle edge. Once that edge is removed, the network remains disconnected into two equivalent components - triangles. Those cannot be partitioned further because all edges have equal betweenness by symmetry. So the partition is into the two 3-node triangles.

Newman eigenvector method [25]

Modularity semantic definition: The modularity is, up to a multiplicative constant, the number of edges falling within groups minus the expected number in an equivalent network with edges placed at random.

$$s_i = \begin{cases} 1 & i \in \text{group}_1 \\ -1 & i \in \text{group}_2 \end{cases} \quad (1.7)$$

$m = \frac{1}{2} \sum_i k_i$ is the number of edges, where k_i is the degree of the i^{th} node, A_{ij} is the (i, j) entry of the adjacency matrix, $\frac{k_i k_j}{2m}$ is the expected number of edges between i and j if placed at random.

The mathematical definition of modularity, using the expected degree at random is:

$$Q = \frac{1}{4m} \sum_{ij} (A_{ij} - \frac{k_i k_j}{2m}) s_i s_j = \frac{1}{4m} s^T B s \quad (1.8)$$

where s is the vector with \pm elements, B is the modularity matrix with entries $B_{ij} = A_{ij} - \frac{k_i k_j}{2m}$ and is real symmetric. Notice that the columns and rows of B sum up to zero, which means that $[1 \ 1 \ \dots \ 1]$ is an eigenvector. Let u_i be the normalized eigenvectors of B . Write s as a linear combination of u_i . So $s = \sum_{i=1}^n a_i u_i$, where $a_i = u_i^T \cdot s$. Then $Q = \sum_i a_i u_i^T B \sum_j a_j u_j = \sum_{i=1}^n (u_i^T \cdot s)^2 \beta_i$, where β_i is the eigenvalue of B corresponding to u_i^3 .

Let the eigenvalues of B be $\beta_1 \geq \beta_2 \geq \dots \beta_n$.

So if the definition of a module as having less number of links to the outside than expected is the norm, the modularity Q has to be maximized (for s_i and s_j having different signs we want $A_{ij} - k_i k_j / 2m$ to be very negative). This means choosing s so that the weight of the sum falls on the largest (most positive) eigenvalues. If s was unconstrained, it would be straightforward to set it to be proportional to u_1 (that would maximize the dot product at the largest eigenvalue). Unfortunately, that is not always possible, because s was chosen as a vector of $\pm 1s$. A simple effective solution (as in spectral partitioning) is to maximize only the term involving the largest eigenvalue while completely ignoring the others. In this case, since it is not possible to make s parallel to u_1 anyway, maximizing the largest term suffices. The greatest value of the coefficient $(u_1^T \cdot s)^2$ is achieved when the elements of s have the same signs as the elements of u_1 . So this is the gist of the algorithm: compute the eigenvector of the largest eigenvalue of the modularity matrix and divide all vertices in two parts according to the signs on the "largest" eigenvector. Interestingly, the magnitude of the eigenvector entries also matter - showing the strength of the belonging of a

³Note: $1/4m$ is an arbitrary factor.

node to its "sign" group. Entries close to zero show weak belonging to a particular group.

The adjacency matrix of the bowtie graph is given in Table 1.1.2, the degree sequence is [2 2 3 3 2 2] and the number of edges m is 7, so the modularity matrix can easily be computed. The largest eigenvalue is 1.7321 with corresponding eigenvector [-0.4440 -0.4440 -0.3251 0.3251 0.4440 0.4440], which splits the nodes in two groups (1,2,3) and (4,5,6) by sign, just as expected.

The Newman-Girvan and the Newman eigenvector algorithms give the same result for the bowtie graph, because this is a simple example, but they don't have to. The two algorithms reflect similar patterns but are based on different measurements in the graph. The Newman-Girvan counts shortest paths, so identifies communities by cohesiveness and remoteness, while the Newman eigenvector method identifies modules by whether they have more links inside the module than expected on average. The Newman-Girvan algorithm gives more precise results, but because of the expensive computation of edge betweenness is less practical for large graphs.

In Chapters 2, 3 and 5 we apply the Newman algorithms to study modularity in airline networks and Wikipedia.

1.3.2 Subgraphs. Motifs

Community finding is a reductionist, top-down approach to uncovering graph structure. A bottom-up view would be to look for the building blocks as they occur across the network and understand how they connect. Finding all instances of a given subgraph in a graph is a classical combinatorial problem, but unfortunately computationally hard to solve (NP). Most solutions involve exhaustive combinatorial search across the entire graph with some post-processing for symmetry elimination. As the size of the graph and the subgraph grow, this approach becomes impractical. A general question about what are all the significant subgraphs in any graph is particularly hard to solve.

One of the first detailed studies on motifs is by Milo et al [26], who study the frequency of simple 3- and 4-node directed motifs in biological networks and electronic circuits. They set the discussion of motifs as building blocks, core units behind important mechanisms and having a role in evolution. They discover frequent patterns such as the bi-fan and the feed-forward loop, especially important in biological systems. Their next study extends this concept to explore profiles of statistically significant 3-node and 4-nodes motifs for various systems. They use these profiles to compare across systems. Figure 1-15 shows an example. The motif profile is further developed by Kashtan et al [6] to introduce generalizable motifs. Generalizable motifs are a huge step in discovering internal structure patterns and classifying topology, compared to counting frequently repeating patterns. Later in this section, we discuss how we use the ideas from Kashtan et al [6] to do a more complete motif search, with larger motifs compared to Milo et al [26].

First we address the question of effective search for a single motif of any size. Grochow [5] gets closer to answering this question by ordering the search and doing symmetry breaking prior to exploring the entire graph. The inputs for the algorithm are a given subgraph and an original graph. The output is the set of instances of the subgraph in the entire graph.

Searching for any set of subgraphs is a general problem compared to searching for known structures, or motifs known to have a particular internal structure. Here we present how we address the general problem. Later on, in Chapters 2, 3 and 5, we perform motif search on some airline and Wikipedia datasets to study their topology.

Single motif finding

For clarification, motifs are (connected) frequently occurring subgraphs, which is where the name comes from - a recurrent pattern. A subgraph can be any set of connected node pairs that are a subset of the node pairs in the original graph. Figure 1-10 clarifies the definition of a motif.

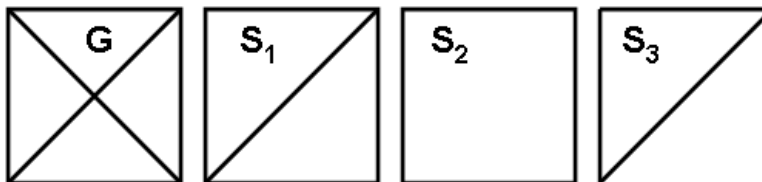


Figure 1-10: Illustrating the definition of motif: all S_1, S_2 , and S_3 are valid motifs (connected subgraphs) of G .

Grochow's single-motif-instances finding algorithm is presented below with one key modification⁴ which is related to our slightly different definition of a motif. In summary, the algorithm breaks down the motif-finding problem by first searching for a single given motif across the original graph, then using clever alignment of the nodal degrees at every matching and finally using labeling to eliminate additional search of symmetries to already queried subgraphs. The pseudocode is given in Figure 1-13.

Finding the set of isomorphisms of a given subgraph in a graph, is known to be NP-complete, though certain tricks can help computationally.

- i The first trick of Grochow's algorithm is mapping of the query network onto the graph, rather than enumerating all subgraphs of equal size and then testing for isomorphism.
- ii Symmetry breaking (not required for correctness)
- iii Isomorphism testing by aligning nodal degrees to ensure maximum mapping degrees of freedom.

The main algorithm routine is described in Figure 1-11. Let the original graph be G and the subgraph query be H . First the node sequence of G is ordered by degree and then nodes with equal degrees are ordered by increasing neighbor degrees. Once the degrees are ordered a partial map is created of a node h in H to a node g in G , such that there is no conflict between the degrees of h and g or the degrees of their neighbors (this property is termed " g in G supports h in H "). The partial map $f(h) = g$ is then extended in *IsomorphicExtensions*(f, H, G) (see Figure 1-12) which finds all full maps of $H \rightarrow G$ that involve h . *IsomorphicExtensions* works by recursively checking each neighbor of a newly added node to the map for conflicts. A conflict is present if node h in $D(f)$ (domain of f) has a neighbor h_n such that h_n is also in D , i.e. $f(h_n) = g_m$, where g_m is not a neighbor of $g = f(h)$. The second conflict condition in Figure 1-12 is incorrect for our definition of motif.

Symmetry enters the algorithm independently, only to speed up the search. Symmetry conditions ensure that in subgraphs where there are more than one isomorphism (different permutations of the nodes result in essentially the same graph), the algorithm maps and queries those only once. This is done by first finding all isomorphisms of the subgraph using *IsomorphicExtensions*(H, H),

⁴This condition does not present a conflict. It is perfectly fine for a neighbor node of $f(D)$ to not be a neighbor in D . In fact, keeping this condition restricts the algorithm from finding all valid subgraph instances, to finding all strict subgraphs (all edges of the subgraph nodes in the large graph, have to be in the motif also). It comes down to the definition of motif.

and then extracting equivalence classes of nodes (He) under the entire set of symmetries: groups of nodes that map inside the group under an isomorphism. The goal of *SymmetryConditions* is based on all isomorphisms and equivalence classes to compute a set of labeling conditions which define a unique node mapping order for each query subgraph.

```

FindSubgraphInstances(H,G):
  Finds all instances of query graph H in network G
  Start with an empty set of instances.
  Order the nodes of G by increasing degree and
  then by increasing neighbor degree sequence.
  For each node g of G
    For each node h of H such that g can support h
      Let f be the partial map associating f(h) = g.
      Find all isomorphic extensions of f
      i.e. call IsomorphicExtensions(f,H,G).
      Add the images of these maps to the set of all instances.
    Remove g from G.
  Return the set of all instances.

```

Figure 1-11: Pseudocode for *FindSubgraphInstances(H, G)* from Grochow et al [5]

```

IsomorphicExtensions(f,H,G):
  Finds all isomorphic extensions of partial map  $f: H \rightarrow G$ 
  Start with an empty list of isomorphisms.
  Let D be the domain of  $f$ .
  If  $D = H$ , return a list consisting solely of  $f$ . (Or write to disk.)
  Let m be the most constrained neighbor of any d in D
  (constrained by degree, neighbors mapped, etc.)
  For each neighbor n of f(D)
    If there is a neighbor d in D of m such that n is not neighbors with f(d),
or if there is a non-neighbor d in D of m such that n is neighbors with f(d)
    then continue with the next n.
    Otherwise, let  $f_0 = f$  on D, and  $f_0(m) = n$ .
    Find all isomorphic extensions of  $f_0$ .
    Append these maps to the list of isomorphisms.
  Return the list of isomorphisms.

```

Figure 1-12: Pseudocode for *IsomorphicExtensions(f, H, G)* from Grochow et al [5]

```

SymmetryConditions:
  Finds symmetry-breaking conditions for H given  $He, Aut(H)$ 
  Let M be an empty map from equivalence representatives to sets of conditions.
  For each n in  $He$ 
    Let C be an empty set of conditions.
     $n' \leftarrow n$ , and  $A \leftarrow Aut(H)$ .
    Do until  $|A| = 1$ :
      Add " $label(n') < \min\{label(m) | m \sim_A n' \text{ and } m \neq n'\}$ " to C.
       $A \leftarrow \{f \text{ in } A | f(n') = n'\}$ .
      Find the largest A-equivalence class E.
      Pick  $n'$  in E arbitrarily.
    Let  $M(n) = C$ .
  Return M.

```

Figure 1-13: Pseudocode for *SymmetryConditions* from Grochow et al [5]

Topologically Generalized Motifs

In biology and engineering there are examples where "the same" motifs do not have the same structure node-for-node. Function can be preserved under topologically similar structures. For example multiple-output motifs can have any number of "output" nodes. In general, motifs can be generalized by copying nodes that have the same "roles" and their links. This process is described in Kashtan et al [6]. The roles they refer to correspond to the equivalence classes described in [5].

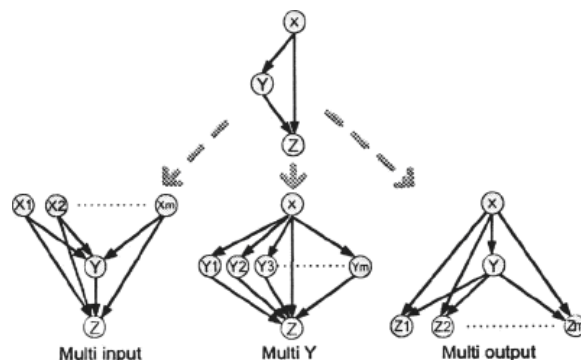


Figure 1-14: Example of three generalizations of a feed-forward (XYZ) loop (source: Figure 2c from Kashtan et al [6]).

Topological generalizations are found by first detecting all original single-role (roughly one node per equivalence class) motifs and then searching for instances of their extensions. These instances are compared to averaged occurrences in random graphs. Examples where generalizations are relevant in engineering are MIMO systems (Multiple-Input-Multiple-Output) systems which are in the same functional group and should be identified as such.

The tools described, finding all instances of a subgraph and finding all topological generalizations, are the base of the problem of motif finding. The last bit is to ensure that the occurrences found are not random. To claim that a motif is a significant pattern of the graph structure, it has to be tested against a null model: a random graph, with the same number of nodes and the same incoming and outgoing degree sequences. Statistically, one such random graph is not enough for a significance test. Usually, motif occurrences are tested against averages in an ensemble of random graphs. In Milo et al [7], a Z-score function is used as a significance test indicator which is computed as:

$$Z_i = \frac{N_{real_i} - \langle N_{rand_i} \rangle}{std(N_{rand_i})} \quad (1.9)$$

where N_{real_i} is the number of instances of motif i in the real network, $\langle N_{rand_i} \rangle$ is the ensemble average of motif occurrences and $std(N_{rand_i})$ is the standard deviation.

We implement both ideas of single motif search and topologically generalized motifs and find extensions of single motifs (i.e. motif families) in airline network. Further description is included in Chapter 2.

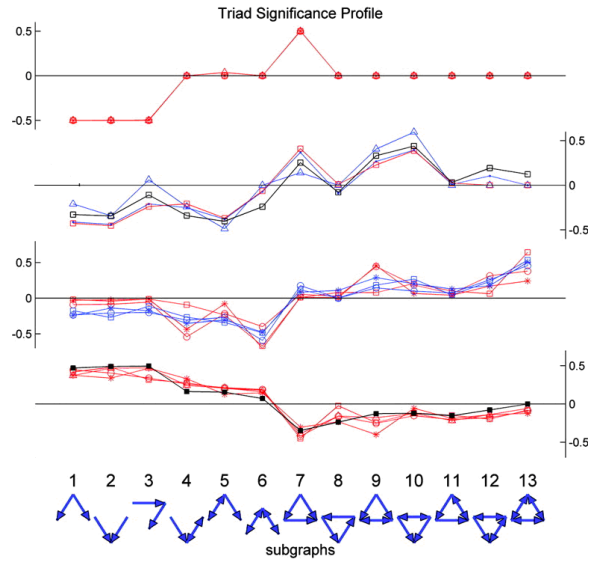


Figure 1-15: Z-score of the 3-node motifs (triads) shown in various real networks (Figure 1 from [7]). The four profiles on the figure correspond to transcription networks (biology), signal transduction (biology), social networks and language (word occurrences in sentences). The profiles show that different networks can be distinguished based on respective motif significance.

1.3.3 Coarse-graining

Coarse-graining is a natural continuation of motif analysis[27]. If the building blocks of the network are identified, the remaining question is how they fit together and how they comprise the network. For example, it is evident that triangles are the motifs of a triangular lattice and that they are stacked to form the entire structure. In the case of the lattice, a motif can be constructed of two, three or more adjacent triangles since those are also repeating patterns. We will define building blocks to be the smallest indivisible cohesive units of the network. More simply, motifs will be the smallest possible statistically significant repeating patterns. A coarse-grained view of a system is by definition a model where the fine detail is smoothed over or averaged out. A more precise way to define coarse-graining for this application is to replace a fine resolution view with a lower-resolution model. In the network sense, this means collapsing motifs to single nodes while keeping the links and considering the structure of the resulting supergraph.

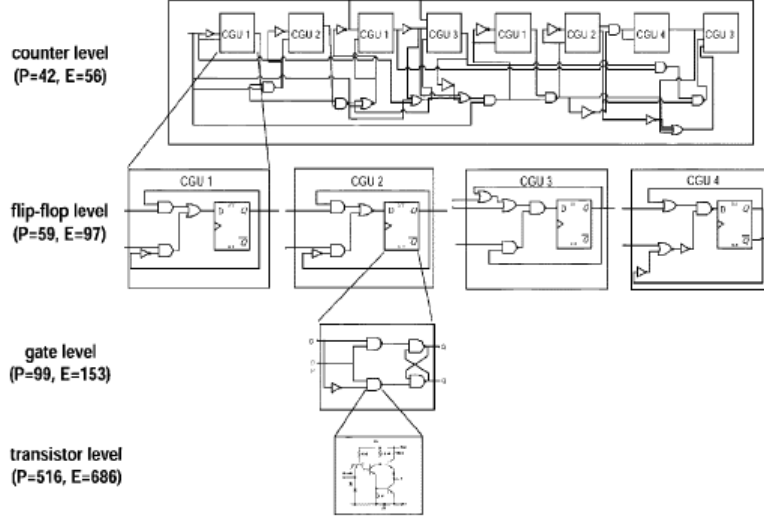


Figure 1-16: Example from Itzkovitz et al 2005 - coarse-graining of an electronic circuit. Four levels of representation of the 8-bit counter electronic circuit. In the transistor level network, nodes represent transistor junctions. In the gate level, nodes are CGUs made of transistors, each representing a logic gate. Shown is the CGU that corresponds to a NAND gate. In the flip-flop level, nodes are either gates or a CGU made of gates that corresponds to a D-type flip-flop with an additional logic gate at its input. In the counter level, each node is a gate or a CGU of gates/flip-flops that corresponds to a counter subunit. Numbers of nodes P and edges E at each level are shown.

In this thesis, we implement an algorithm developed by Itzkovitz et al [27]. In their work, the coarse-graining units (CGUs) are required to be 1/ as small as possible, 2/ as simple as possible and 3/ make the coarse-grained network as small as possible. The authors term these three properties, conciseness, simplicity and coverage. Each subgraph is treated like a black box with ports: incoming, outgoing and mixed (with both links to the outside and inside) and internal nodes. The number of ports ($H=I+O+2M$, I -incoming, O -outgoing, M -mixed) is calculated and added to the estimate of complexity. The lower H is, the more simple the CGU. Then for a given set of CGUs a scoring function is computed that reflects how good the coarse-graining is:

$$S = E_{covered} + \alpha \Delta P - \beta N - \gamma \sum_{i=1}^N T_i \quad (1.10)$$

where $E_{covered}$ is the number of edges covered by all CGUs, N is the number of CGUs, T_i is the number of internal nodes in the i^{th} CGU, ΔP is the difference between the number of nodes in the original network and the covered network ($\Delta P = P_{covered} - \sum_{i=1}^N n_i H_i$, $P_{covered}$ is the number of nodes covered by all CGUs, n_i is the occurrence of CGU i , and H_i is the number of ports of CGU i). So equation 1.10 becomes:

$$S = [E_{covered} + \alpha P_{covered}] - [\alpha \sum_{i=1}^N n_i H_i + \beta N + \gamma \sum_{i=1}^N T_i] \quad (1.11)$$

The first part of this function reflects coverage (edges and nodes), and the second part is related to conciseness and simplicity, in particular, minimizing the number of ports, the number of different CGUs (simplicity). The last term prevents the trivial solution of one complex CGU.

The function is optimized and the best set of CGUs is found with simulated annealing [38]. First all possible CGUs of sized 3-6 are detected. Their nodes are classified into *internal* (T), *incoming* (I), *outgoing* (O) and *mixed* (M). A candidate CGU is characterized by its adjacency matrix (or edgelist) and profile vector - (I,I,O,M,T). For the simulated annealing design vector, each CGU is assigned a spin ± 1 or 0/1 which signifies whether it's active or inactive in the current coarse-graining solution. Spins are perturbed randomly at every step, while the energy (S-scoring function) is annealed at a proper temperature cooling schedule.

One of the major challenges in implementing coarse-graining is to deal with motif overlaps. In a candidate set for coarse-graining, even among motifs of the same class (same pattern or topological class) there can be overlaps of edge and/or nodes which means that some motifs have to be eliminated from a feasible set. The rule adopted by Itzkovitz [27] is to duplicate an overlapping node in the coarse-grained network, only if this node receives only one input from neighboring coarse-graining units (CGUs). This ensures that the dynamic of individual CGUs is well-captured. Solution CGU sets that do not satisfy the single-input into overlapping node criterion are discarded and new solutions are sought. Overlapping edges are not allowed at all.

The algorithm can be repeated a few times to uncover several levels of structure. Though simple, coarse-graining does not always provide satisfactory results. The presumption is that a network is composed of relatively small (max 6-8 nodes) frequently recurring motifs, which is not the case for many systems. First, the recurring patterns could be of much bigger scale, or they could be expressed in some central plan (backbone) or in the mechanism of regulating different parts of the network. This is one of the reasons for which analyzing the static structure of a network is not sufficient to understand the significant patterns or mechanisms of its structure. That said, what the network does or represents is crucial for the interpretation of coarse-graining results. In a metabolic network, the low-level recurrent reaction mechanisms might be interesting, while in the case of airline routes, large-scale hub-feeding dynamics is where the story is.

In Chapter 2 we show the implementation of the coarse graining algorithm for JetBlue Airways and later on, in Chapter 4 use the underlying patterns to propose a growth model for hub-seeding type airline growth.

1.4 Evolution of Networks

The second cornerstone of this thesis is the study of growth or how network structure changes over time. The previous sections reviewed techniques related to structure. Here we review models related to growth, in particular random, node-degree centric, edge-centric, spatial distribution models, heuristically optimized topologies, and node- and module- copying, or *non-incremental* growth. One of the most complete references on network evolution is by Dorogovtsev and Mendes [20]. Chapter 3 contains references on network growth specific to the airlines.

1.4.1 Random graph models

Random graph models are the first "modern graph theory" developed by Paul Erdős and A. Rényi, to explain real-world networks [39]. The beauty of random graph mathematics comes from all the derivations possible due to averaging, such as mean degree, clustering coefficient and degree distribution closed form probability laws. The simplest ER (Erdős-Rényi) model says that for a set of nodes, edges arrive with probability 0.5 between any pair of nodes. Another simple version

of the model sets the maximum number of edges. The pseudo code below shows a typical random graph construction routine with a maximum number of edges m and a probability of attachment p .

```

While |E|<m:
  Pick random node  $n_1$ 
  Pick random node  $n_2 \neq n_1$ 
  If rand<p:
    Connect  $n_1$  and  $n_2$ 

```

Note that this routine is different from the classic Erdős-Rényi routine which goes through all possible edges and creates them with probability p . This results in $pn(n-1)/2$ mean number of edges, where n is the number of nodes.

Some easy derivations are the mean degree for the network: $z = np$, the clustering coefficient, $C = 0$, and the degree correlation $r = 0$.

1.4.2 Node-degree centric models

In these models, links arrive at the nodes with probability which is some function of the nodal degree.

Exponential model (Dorogovtsev [20])

Start with two nodes connected to each other twice $n(t=2) = 2$, $m(t=2) = 2$. At every step, add one new node and connect it randomly to any other existing node in the network. Thus, at every time step t , there are $n(t) = t$ nodes and $m(t) = t$ edges. The probability of an existing node receiving a new connection is $1/t$, because all nodes are equal in this model. The probability of a node s having degree k at time t is

$$p(k, s, t) = \frac{1}{t}p(k-1, s, t-1) + (1 - \frac{1}{t})p(k, s, t-1) \quad (1.12)$$

The initial and boundary conditions based on the attachment model are $p(k, s=1, 2, t=2) = \delta_{k,2}$, $p(k, s=t, t>2) = \delta_{k,1}$. This equation exhibits a common way to capture growth in the literature - by expressing the probability of a vertex (or a number of vertices) having certain degree at a certain time. This is probably due to the fact that changes in the degree distribution are most natural to track with adding and removing nodes, and also due to the strong focus on degree distributions [39][14]. Equation 1.12 can be used to derive the probability of a vertex having a certain degree at a certain time, using limits at infinity and approximations or assumptions. Define $P(k, t) = \frac{1}{t} \sum_{s=1}^t p(k, s, t)$ and apply $\sum_{s=1}^t$ to 1.12. In the limit of $t \rightarrow \infty$, the stationary form of this equation becomes $2P(k) - P(k-1) = \delta_{k,1}$. solution is the exponential form $P(k) = 2^{-k}$, which is why this growth model is called exponential. It is clear that this model generates trees, since nodes with degree bigger than k are never receive new links, as evident in Figure 1-17⁵.

⁵This figure and all subsequent figures of graph are created with Netdraw, which is a visualization program by Analytic Technologies.

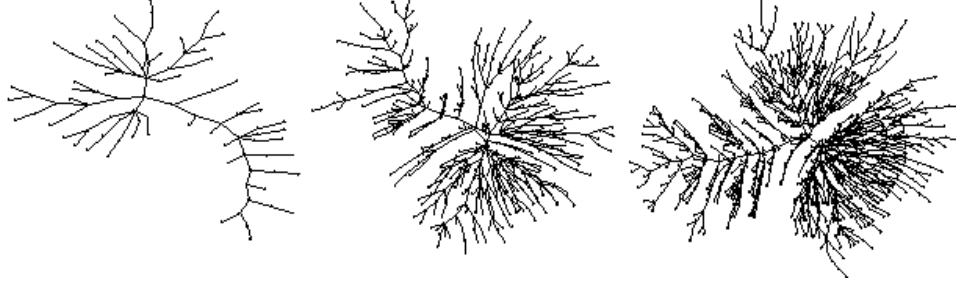


Figure 1-17: Exponential network snapshots example, $t=100,500,1000$. The plots are created with NetDraw using the Kamada-Kawai spring energy minimization algorithm [3].

The preferential attachment model (Price [40], Barabasi [14])

Preferential attachment was first introduced by Price in 1965 who studied networks of scientific papers. The model has been re-introduced and become more popular since 1999 by Barabasi et al [14]. The cumulative advantage⁶ model has the same form of probabilistic equation as the exponential, except that the probability of attaching a new node to an existing node is proportional to the degree of the existing node. The initial and boundary conditions here are considered to be the same: $p(k, s = 1, 2, t = 2) = \delta_{k,2}$, $p(k, s = t, t > 2) = \delta_{k,1}$.

$$p(k, s, t + 1) = \frac{k - 1}{2t} p(k - 1, s, t) + \left(1 - \frac{k}{2t}\right) p(k, s, t) \quad (1.13)$$

Equation 1.13 reflects the fact that the probability of a node having degree k , at time step $t + 1$ is dependent on a node with degree $k - 1$ receiving a link at time t , with probability $\frac{k-1}{2t}$ or node with degree k at time t not receiving any links. The probability of receiving links, based on existing degree, reflects exactly the preferential attachment principle. The total number of nodes after t steps is t , and the total number of links is t , based on the Barabasi-Albert (BA) model. Then $2t$ is the sum of all degrees in the network⁷, and $\frac{k-1}{2t}$ is simply the fraction of (undirected) edges connected to a node of degree $k - 1$.

Applying $\sum_{s=1}^t$ to 1.13 to get the general distribution gives $P(k) - \frac{1}{2}((k - 1)P(k - 1) - kP(k)) - \delta_{k,1} = 0$ in the limit $t \rightarrow \infty$. The solution of this equation (can be verified by substitution) is $P(k) = \frac{4}{k(k+1)(k+2)}$, which means that $P(k) \sim k^{-3}$. A simulation of the BA model for the same graph at $t=100, 500$ and 1000 is shown in Figure 1-18.

⁶”Cumulative advantage” here is used interchangeably with ”preferential attachment”

⁷ $2\#edges = \sum_i d_i$ is an almost intuitive fact. Since every edge has two node ends, the sum of all degrees counts every edge twice.

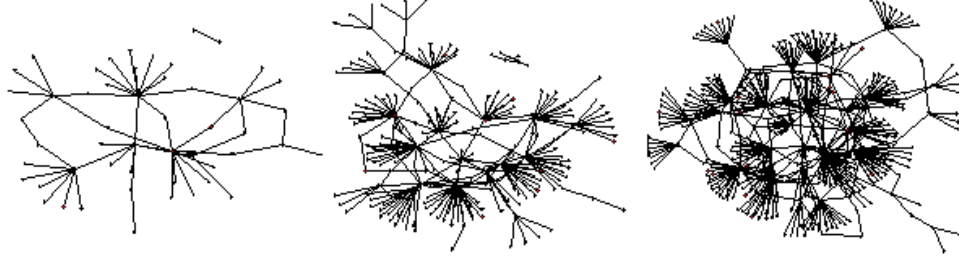


Figure 1-18: BA graph snapshots at $N=100$, 500 and 1000 nodes ($t=N$). The graph are plotted using the Kamada-Kawai spring energy minimization algorithm [3].

Master equation [20]

To generalize the previous two ideas, take a single arriving node at each time step (without loss of generality) and m new directed links simultaneously from unspecified origin (even from outside of the network). The target ends are distributed among vertices with probability proportional to the incoming degree of a node offset by a constant. Let $q(s)$ be the incoming degree of node s . Then node s receives new links proportionally to $q(s) + A$, where, $A = ma$, and a is a constant. Then the normalized probability of attachment is

$$\frac{q(s, t) + A}{mt + At} = \frac{q(s, t) + ma}{mt + mat} = \frac{q(s, t) + ma}{(1 + a)mt} \quad (1.14)$$

The probability that a vertex receives k new edges at a given point in time t is

$$P_s^{mk} = \binom{m}{k} \left[\frac{q(s, t) + am}{(1 + a)mt} \right]^k \left[1 - \frac{q(s, t) + am}{(1 + a)mt} \right]^{m-k} \quad (1.15)$$

Then the equivalent of equations 1.12 and 1.13 for the exponential and BA models is the following master equation:

$$p(q, s, t + 1) = \sum_{k=0}^m \binom{m}{k} \left[\frac{q(s, t) + am}{(1 + a)mt} \right]^k \left[1 - \frac{q(s, t) + am}{(1 + a)mt} \right]^{m-k} p(q - k, s, t) \quad (1.16)$$

The solution of this equation is $\bar{q}(s, t) + A = A(s/t)^{-\beta}$, where $\beta = 1/(1 + a)$.



Figure 1-19: Master equation method with $a=2$, $m=1$. Graphs plotted using a spring energy minimization algorithm.

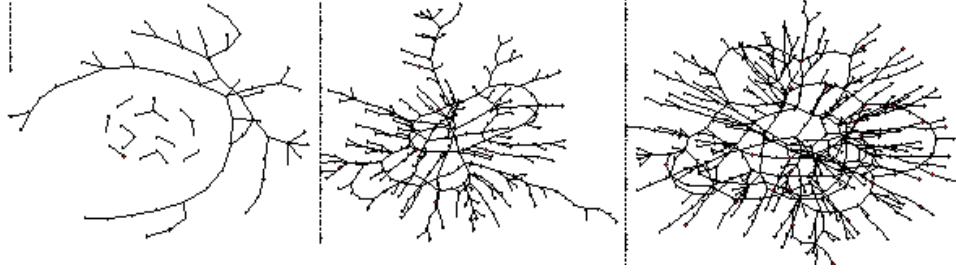


Figure 1-20: Master equation method with $a=2$, $m=2$. Graphs plotted using a spring energy minimization algorithm.

Figures 1-19 and 1-20 show that the master equation method is a generalized preferential attachment method. For $a = 1$, it should reproduce the BA model. For varying number of vertices m attached at each step, the existence of a global connected component is affected, as well as its degree distribution.

1.4.3 Spatial distribution algorithms

These models concentrate on positioning the nodes and then wiring them to optimize some objective.

Fabrikant model [41]

This is a tree building algorithm in which every new node arrives and attaches to one of the previous nodes, chosen with a certain objective function. Let h_j be some measure of centrality for node j , and α be some constant. The goal is to connect a new node i to an existing node j as close (in Euclidean sense) as possible, but which is also fairly central. Therefore, the metric is a weighted sum of closeness and centrality:

$$\min_{j < i} \alpha d_{ij} + h_j \quad (1.17)$$

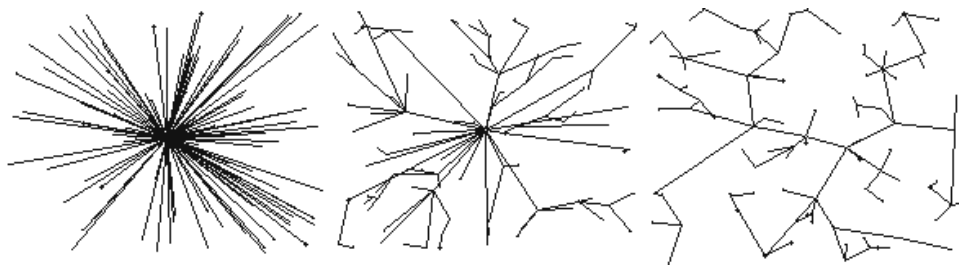


Figure 1-21: Fabrikant model with $N=100$, and $\alpha=1$, 1.5 and 20 respectively.

Figure 1-21 shows that for small α the closeness criterion is overpowered by centrality and every new node connects to the center. For higher values of α , local connectivity becomes more important. These kinds of considerations are relevant in energy, water and other distribution systems, where there is a source (and/or a sink), but there are also local distribution stations which are cheaper or easier to connect to logistically.

Newman-Gastner model [42]

This study, similarly to the Fabrikant work, concentrates on balancing traveling along the network versus Euclidean distance to the root. It is assumed that there is a source/sink type node in the system. The key definition is the *route factor*, expressed as: $q = \frac{1}{n} \sum_{i=1}^n \frac{l_{i0}}{d_{i0}}$, where l_{i0} is the distance from node i to the root along the edges of the network and d_{i0} is the Euclidean distance from i to 0. So for example a route factor of 2 means that the shortest path from a vertex to the root of the network is simply twice the Euclidean distance.

The authors examine a few subway and pipeline networks and find route factors often very close to 1, indicating that distance to the root is an important factor in their design. They propose an objective function for connecting new nodes similar to the Fabrikant model [41], but the results are different because the node locations are set prior to the design process. In this model, all nodes are given, (by some real problem setting) and connected starting at the root, one by one. A new node i is connected to node j which minimizes $w_{ij} = d_{ij} + \beta l_{j0}$, where l_{j0} as before is the distance from j to 0 along the network, and d_{ij} is the Euclidean distance between i and j . For $\beta = 0$, the new node always connects to the closest node (essentially a minimal spanning tree), while for large beta, the distance to the root matters more, so the new node will connect closer to the root, as shown on Figure 1-22.

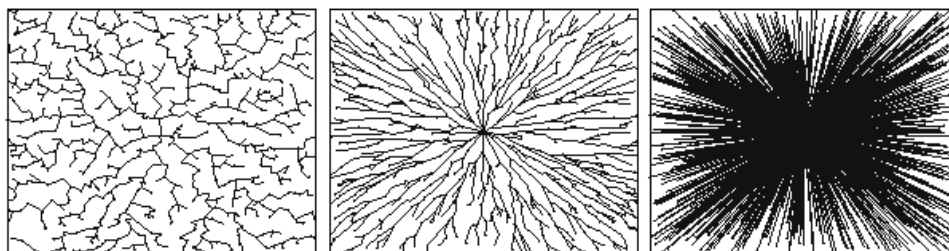


Figure 1-22: Newman-Gastner model for $\beta = 0.1$, $\beta = 0.9$, and $\beta = 1.1000$ nodes.

Heuristically-optimized configurations

In the models reviewed so far strictly graph statistics have been discussed for applications in biology, sociology, communications and engineering to "explain" the structure of systems without considering domain knowledge. As Li et al [31] argue, this is equivalent to simple (often inaccurate) curve matching and is far from the reality of the underlying architecture of the studied systems. Figure 6 from [4] (Figure 1-9) illustrates this point for the case of the Internet network at the router level. The degree distribution of the router-level network does not fix its structure, because there are many graphs with the same distribution, and the suggested mechanisms for generating scalable degree distributions suggest system architectures quite opposite from the real Internet architecture. The purely stochastic approach misses the cost, infrastructural constraints and the objectives/performance requirements of the system which drive the real architecture. Doyle et al argue that for the case of the Internet in [4].

Highly optimized tolerance is a concept termed by Doyle and his group at Caltech. In [4] Li et al show that the Internet, as an instance of an engineering system, grows as a heuristically optimal system, which means that its structure respects real design considerations, such as aggregating traffic (through high connectivity) and constraints, such as router bandwidth. They show that optimizing the topology for performance metrics and respecting constraints results in a topology a lot closer to the real one, compared to general graph growth methods.

Another example of a heuristically optimized engineering system for growth and robustness is our study on large telescope arrays [43][8]. Multiple antenna radio astronomy, also known as interferometry, uses many spread-out linked antennas to create an aperture equivalent to the aperture of a telescope with the diameter of the multi-antenna array. Usually, these arrays are spread-out along hundreds of kilometers and linked with expensive fiber. The design problem has naturally opposing objectives: to spread out the array which will benefit the uv distribution and to keep the distances between stations short to keep the length of expensive cable shorter hence cut cost. The Euclidean distance cost is a similar measure to the one used in (Gastner 2004). One of the main conclusions of this study is that different array topologies correspond to different balances of objectives. Optimizing for cable length produces Y-shaped arrays, while optimizing for coverage produces circular arrays. The algorithm is an adapted genetic algorithm with seeding geometries, so these geometries are not provable to the absolute answers to the optimization problem. Due to the non-linearity of the performance objective function, slightly non-regular geometries, with protruding branches are favored.

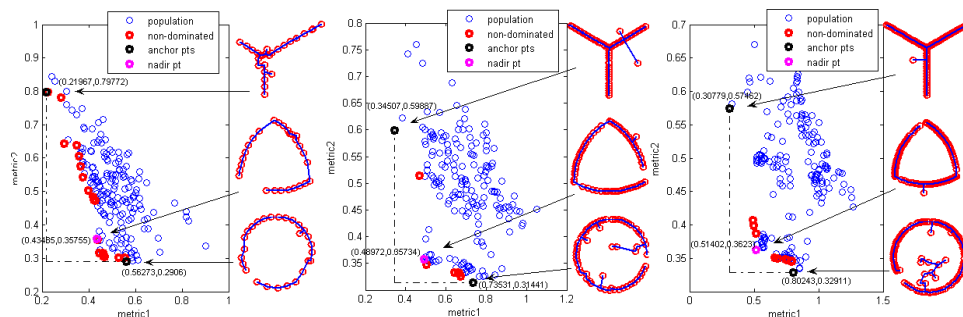


Figure 1-23: Optimizing telescope array configurations for cable length (cost) and uv density (performance). Left to right: 27, 60 and 99 nodes. Three configurations show minimum cable, nadir point and minimum uv metric designs from top to bottom (Bounova [8]).

1.4.4 Node-copying and function preservation

Node-copying is inspired by gene duplication and whole-genome duplication computational studies of various genomes (Vazquez [44]). Whole genome duplication has been proved to have occurred in various species. Duplication of a given area of the genome means that a nucleotide sequence repeats (is inserted twice) which could influence the proteome, i.e. producing copies of the same protein. The main idea is that a new copy of a protein will interact with the same proteins as the original, and eventually will lose/gain interactions based on mutations or selective pressure. This simple principle has been proposed as a network growth algorithm with fitting the probabilities of deleting and adding interactions. This is called a (DD) duplication-divergence model. The purpose of these models was to explore robustness of protein networks created by node-copying or duplication and to match the theory to concurrent studies in complex networks [34][14].

The divergence-duplication model can be summarized as follows:

- i *Duplication*: A node is a protein. A node i is selected at random. A new node i' is created and all neighbors of i are linked to i' as well. With some probability p , a link between i and i' is established as well, for self-interacting proteins.
- ii *Divergence*: For every node j linked to i and i' pick one the links at random and remove it with probability q .

The algorithm starts with two nodes connected to each other and continues with one step duplication and then one step divergence. On evolutionary scale, this is a fair assumption, because divergence is presumed to happen almost right after duplication, so naturally, it can be decoupled with the next duplication step.

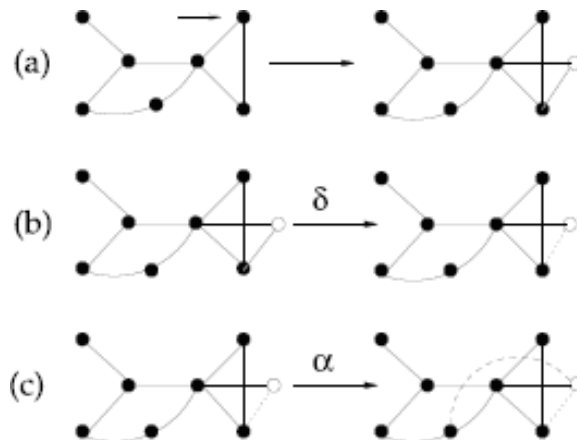


Figure 1-24: Figure 2 from (Sole 2002) Growing network by duplication of nodes. First (a) duplication occurs after randomly selecting a node (arrow). The links from the newly created node (white) now can experience deletion (b) and new links can be created (c); these events occur with probabilities δ and α , respectively.

The duplication-divergence model is similar to the node-degree centric models - pertinent to making asymptotic approximations. In particular, the authors of this model (Vazquez [44]) investigate the asymptotics of the average nodal degree, which they call the average connectivity.

By the algorithm, the change in average degree can be calculated as:

$$\langle k \rangle_{N+1} = \frac{(N)\langle k \rangle_N + 2p + (1 - 2q)\langle k \rangle_N}{N + 1} \quad (1.18)$$

which in the limit of large N has a stationary value for $k : \rightarrow k_\infty = \frac{2p}{1-2q} + O(N^{1-2q})$. The average degree is finite for $q > 1/2$ and grows exponentially with N , for $q < 1/2$ (as N^{1-2q}). Such conclusions can be derived for higher moments of the degree distribution. A similar model is presented by Sole et al [45].

The copying of links by functional association is a powerful idea. It is the same principle as of the equivalence classes for topological generalization. Interestingly, we adopt a similar approach of copying links of geographically similar nodes (airports) to create local structure in Chapter 4 in a growth model for Southwest airlines. Similarly, the other model we look at, derived from JetBlue's patterns of growth, features dual service to the same cities from two distinct hubs.

1.4.5 Module-copying and growth by accretion

A natural generalization to copying a node is to copy a larger subgraph, perhaps a "module", or a set of connected nodes with a particular uniting function. Then the links of this module would be preserved in the first step (copies of links of corresponding copied nodes) and then lost probabilistically over time. While this model may find its analogies in gene duplication, where the entire regulatory network is duplicated and its function diverges over time, it is valuable from

a completely different point of view. Growth node-by-node or edge-by-edge is a good first-order model. However, it is quite frequent to see disconnected networks become a larger connected whole, over time, starting with largely isolated nodes or clusters. This idea is presented in Watts's "Six Degrees" [46], describing the forming of social links between geographically separated communities over time. We recently found evidence of that type of growth in the evolution of change requests of a radar system [13], as well as in studying the evolution of Wikipedia in this thesis.

Giffin [13] studied changes proposed for the engineering design of a radar system, documented and tracked over time. Every time a new change request is made (a new node in the network) it is linked to prior changes as a child of a node (resulting from), a sibling of a node (related to, and on the same level). This resulted in 41000-node graph over 9 years, whose evolution was mapped day-by-day. The analysis showed many disconnected small components forming at the beginning and coalescing into larger components over time, in ways that could not be foreseen at the beginning. The incremental growth models discussed so far would predict that the network grows out of a single original change request. The data points to the contrary, large clusters in this system, form from smaller clusters, originating from completely unrelated original nodes. Figure 1-25 shows two snapshots of a connected component in year 7 and year 9 of the development of the program.

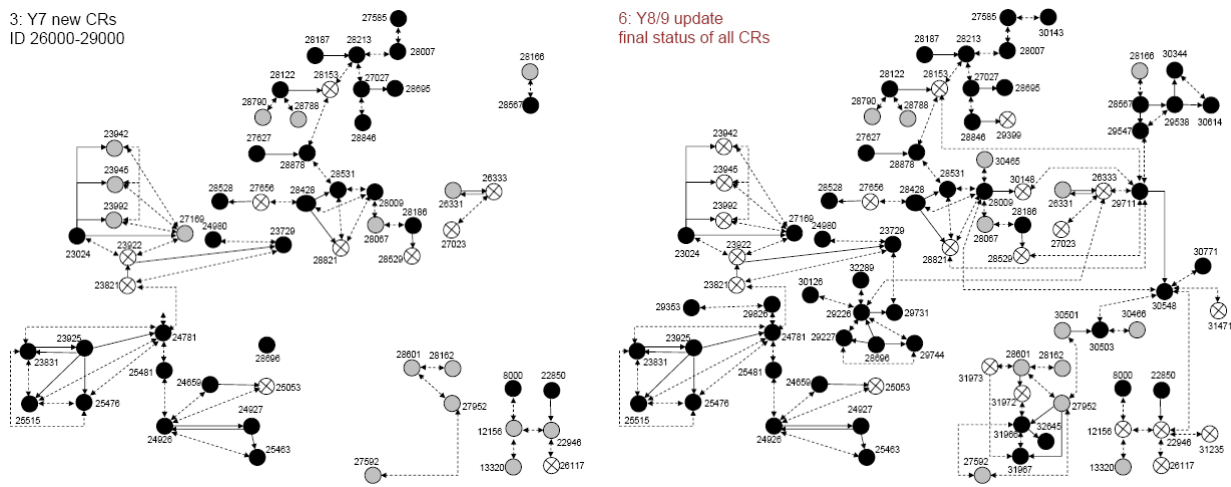


Figure 1-25: Different components of the change request network coalescing together over the course of nine years. (Left) Year 7; (right) Year 9.

As pointed out, this mode of growth is discussed later in the thesis, in the context of Wikipedia. We also discuss the implications of systems growing disconnected versus connected from the start and analyze examples from both technical (airlines) and social systems.

1.4.6 Edge-centric models

A consideration completely ignored so far is the pure propensity for a new edge forming, given a backbone network. This assumes that all nodes are already present in the network, because an edge forms due to a demand between two nodes to be connected. Simplistically, an airline will offer a new flight A-B if there is large population in A that wants to travel to B. There can be many reasons for an edge to form - demand at the nodes, demand and capacity at neighboring nodes, distance, etc. Note that the underlying network is already connected. This is still considered growth and arrives at a particular structure that can be analyzed at its end point. Airlines also connect directly

airports that they already serve, other than expanding to new airports.

Other than growth, edge-centric models based on rewiring were first introduced by Watts [34], who rewired k -regular graphs to create small worlds (networks with short diameter and high clustering) as cited in Section 1.2.7.

The Dodds-Watts-Sabel [9] model presents an edge-based growth model for hierarchies. The algorithm starts with building a pure tree with branching ratio b and depth L . Then extra edges are attached depending on the standing of corresponding nodes in the hierarchy. The probability of two nodes being connected (on top of the hierarchy backbone) is based on the depth level of their lowest common ancestor a_{ij} (D_{ij}) and their own depth below a_{ij} (d_i and d_j). To set aside cases where nodes are connected by default, it is required that $d_i + d_j \geq 2$. The organizational distance between two nodes is $x_{ij} = (d_i^2 + d_j^2 - 2)^{1/2}$. The assumption is that the larger the organizational distance, the less likely is that two nodes will associate with each other. Also, the probability of two nodes associating decreases with increasing depth level of their lowest common ancestor, which reflects the tendency of only higher rank nodes to associate. Figure 1-26 illustrates the model.

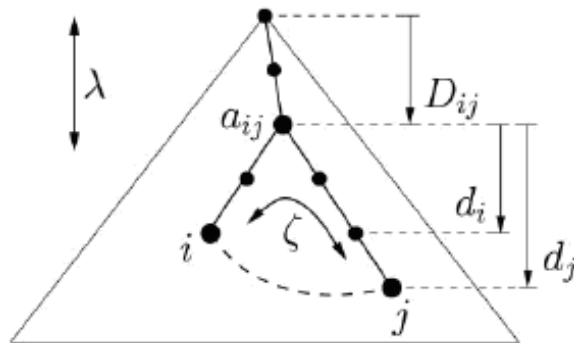


Figure 1-26: Figure 1 from Dodds et al [9]. Illustrating the distances in a hierarchy - to the lowest common ancestor from each node (d_i and d_j) and from the lowest common ancestor to the top (D_{ij}).

In the model, $P(i, j) \propto e^{-D_{ij}/\lambda} e^{-x_{ij}/\xi}$ is the probability that i and j will connect. The authors claim that varying the parameters λ and ξ gives rise to different network architectures. When $(\lambda, \xi) \rightarrow (\infty, \infty)$, $P(i, j)$ is nearly 1 and any pair of nodes can be connected, the network becomes random-like. When $(\lambda, \xi) \rightarrow (\infty, 0)$, the probability depends exclusively on the organizational distance between the two nodes, hence local teams are favored. When $(\lambda, \xi) \rightarrow (0, \infty)$, the probability depends on the depth of the lowest common ancestor, hence nodes associated with the top node in the hierarchy form a team and other divisions below them remain randomly connected. The authors call this configuration random interdivisional. Finally, when $(\lambda, \xi) \rightarrow (0, 0)$, both factors play a role - and links are exclusively added among the subordinates of the top node. This pattern can be seen in Figure 1-27 reproduced from [9].

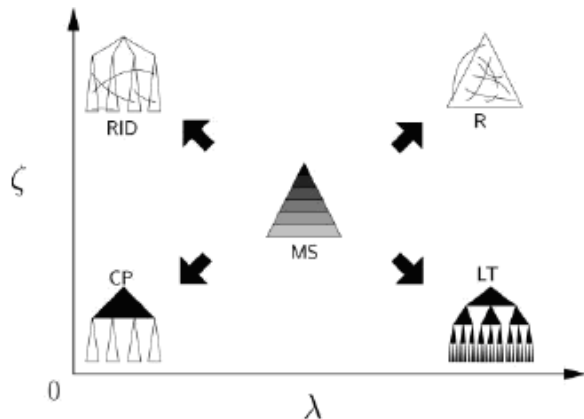


Figure 1-27: Hierarchy patterns. Figure 2 from Dodds-Watts-Sabel [9].

Interestingly, in the middle of the (λ, ξ) space, they identify "multiscale" networks where both the local teams and the global random ties are present.

We will use this model to create a range of "random hierarchies" as a family of topologies in Chapter 2. It turns out that many real networks resemble the Dodds-Watts-Sabel graphs statistically.

1.5 Conclusion

In this chapter, we reviewed network theory relevant to the i/ analysis of network topology and ii/ study of evolution of topology, or growth models for systems represented as graphs. With these ideas from the literature, we will develop the tools to analyze network topology in Chapter 2, and extend them to study growth in Chapter 4.

Five of the statistical metrics described in this chapter, density, clustering coefficient, degree correlation, the S-metric and diameter are combined and non-dimensionalized in a *topology* vector used to compare topologies in Chapter 2. The motif finding algorithm is used to uncover recurrent patterns in airline networks in Chapter 3. Degree distributions are computed for the airlines and discussed in Chapter 2 and Chapter 3. Most of the techniques described in this chapter appear implicitly, either in the computation or the visual representation of networks throughout this thesis.

Chapter 2

Network Topology

Topology is a field of mathematics which deals with properties of spaces that are preserved under continuous deformations. Intuitively, if an object is stretched, twisted, translated or deformed (not transformed) in some continuous way, topologically speaking, it is still the same object with the same properties. Topology studies properties of the space the objects are described in, what its rules are, such as connectivity. This definition of topology is not the one used in the title of this chapter, but should help to understand the intuition behind the narrower term *network topology*. Network topology describes how the elements of a network are arranged and connected. The definition implies that there are ways to arrange connected objects that can be described in some way. The common network topology classes, that come from electrical engineering (communications) are rings, meshes, stars, lines, buses, trees, regular (every node has the same number of connections) and complete graphs (fully connected, $(n - 1)$ -regular). These are very *pure* descriptions, and they don't tend to occur in reality, except for in relatively small simple designed systems. Later on in this chapter, these are described and analyzed as canonical topologies. Most systems represented as networks are *messy*, sometimes canonical-like, and sometimes random-looking, sometimes hybrids of canonical systems. This chapter describes tools for analyzing real topologies, statistically and in relation to canonical topologies.

The challenge in qualifying a topology or assigning it to a class of topologies, is in choosing the right metric. There isn't one perfect measure though there have been candidates - degree distributions among some of the more popular. One difficulty lies in the fact that the popular statistical measures such as diameter, clustering coefficient, edge to node ratio are very one-dimensional. One way to overcome that is to combine the non-dimensional metrics into a *topology vector* that can be compared across graphs and used to create a *topology profile*. This idea is presented in the chapter and discussed in the context of the route network of JetBlue 8/07. Obviously, the problem there is continuity - what is a topology class between two canonical topologies and what does this vector space look like?

A completely different approach is the reductionist way - looking at topology as composition of simple, well-defined elements. We discuss and compare the Newman modularization algorithms, followed by motif search. Looking for underlying patterns in the network, and extracting how they connect, is a more detailed magnifying glass approach to identify the backbone of the structure. Motif finding is presented in this chapter with our implementation of two ideas from the literature.

All of the above techniques are exemplified with the route network of JetBlue Airways of August

2007.

Aside from this graph theoretical discussion of topology, for a real system, there are many more factors that describe it. For the applications we are concerned with, the topological description has to be augmented by geometry, demand, energy considerations, legacy, and strategy. These are necessary to understand the annotated topology of a real system, though outside the scope of this thesis. We discuss airports geographically, and some airline strategy, but do not incorporate these factors in our models.

2.1 Statistical Indicators for Topology

First we discuss network topology in view of network density, clustering, assortativity and reachability. The statistical measures reflecting these characteristics are introduced in detail in Section 1.2. We concentrate on the following: edge-to-node ratio (m/n), clustering coefficient (C), degree correlation (r) average path length, (l), and network diameter (d). First, we present and discuss the statistics of various *canonical* networks, as waypoints for general networks, which are far from canonical, but might be interpreted as or generated from hybridized canonical topologies. Then we show how to use these statistics to compare a real topology to a *spectrum* of canonical topologies.

Degree distributions were reviewed in Section 1.2.3. Here, we discuss how they can be used in conjunction with other metrics to shed light on network topology. First, we discuss degree versus betweenness centrality, a topic first explored in the context of airlines by Guimera et al [17]. Then, for the example of JetBlue, we show the implications of the betweenness-degree relationship and discuss degree distributions and degree correlation elasticity, a topic explored by Whitney [32].

Throughout the section, the ideas are illustrated with the example of the route network of JetBlue Airways of August, 2007.

2.1.1 Canonical topologies

The term *canonical* here is used in the sense of "the simplest and most significant form possible without loss of generality"¹, that is - a simple generalization for a family of forms that is regular (non-random) and easy to comprehend. The number of canonical network types is arbitrary, so here we discuss only a limited set of simple regular forms: stars, loops, lattices and trees; and random-regular forms such as: random graphs, preferential attachment graphs, and random-modular graphs. All types are defined in Table 2.1.1.

¹By Webster's Online, third adjective interpretation.

Table 2.1: Canonical network definitions

Simple graphs	
→ Line graph	A linearly connected set of n nodes with $n - 1$ links between them.
→ Circle (or loop)	A graph, with n nodes and n links, connected in one cycle, every node with degree 2. The smallest circle graph has 3 nodes.
→ Star	A graph with n nodes, one central node connected radially to all $n - 1$ others. The central node has degree $n - 1$, all other nodes have degree 1.
Trees	Canonical trees in this thesis are trees with branching ratio b and depth (number of levels) L . Trees have no clustering (clustering coefficient of zero), $n - 1$ edges for n nodes, and a maximum diameter of $n - 1$.
Scale-free/Scale-rich	Graphs designed to particular scale, either in degree distribution or sub-structure patterns.
→ Preferential attachment graph	Graph in which the probability of adding links to a node is proportional to the nodal degree (number of neighbors), hence the name "preferential", which reflects preference to attach to higher degree nodes.
→ S-max graph	A unique graph for which $\sum_{(i,j) \in E} d_i d_j$ is maximized for a given degree sequence $\{d_1, d_2, \dots, d_n\}$.
(Random) Hierarchies	In this thesis, hierarchies refer to balanced trees with horizontal cross-linking along the same level and random interlinking across levels.
→ Hierarchical binary / tertiary trees	Balanced trees with horizontal cross links at all levels except the leaves.
→ DWS graphs (core-periphery, local team, random-interdivisional, and random)	Hierarchical binary / tertiary trees with random interlinking with several varieties dependent on distance between the two nodes and their common distance from the top of the tree 1.4.6.
Lattices	Lattices in this thesis are defined as geometric point arrangements in 2D, which can be formed using only one basis of vectors, by taking all linear combinations of the vectors with integer coefficients. Obviously, there are infinitely many lattices, so we only pick the set of three types that satisfy the following rotation symmetries: the symmetry moves a lattice point to a succession of other lattice points, generating a regular polygon in the same plane ² . This is the formal way of saying that the lattices considered here are triangular, rectangular and hexagonal.
Random graphs	Random graphs are also canonical structures in the sense that pure randomness allows analytical (via statistics) solutions, i.e. many properties are solvable in the limit of large graph size.
→ Random with the same degree distribution	Random graph in which every node has preset number of links (according to a given degree sequence) but the edges are distributed randomly. Self-loops and double edges sometimes occur in these random graphs.
→ Random modular	In this type of graph, the nodes are separated in modules and edges are assigned with greater probability inside the module, and lower probability across modules.
→ Erdős-Rényi	Graphs, which are constructed by flipping a coin with probability p to decide whether there is an edge between any two nodes.

Figure 2-1 shows the families of topologies presented above in order of "topological continuity". Though one cannot claim that these topologies continuously transform from one to another (under some invariance principle), the order of topology families in Figure 2-1 reflects a direction of increasing density and some form of notional complexity. On the far left are simple graphs, from the simplest trees, to balanced trees, evolving into trees with particular scalar structure. Mid-way are the Newman-Gastner type topologies, preferential-attachment graphs, the s-max graph, or heuristically optimized topologies. These are still very sparse, often still trees, or with few loops

²This is the Crystallographic Restriction Theorem.

and generally low clustering. The next notional jump is from pure hierarchies (balanced trees) to hierarchies with horizontal cross links, or random cross links, as in the Dodds-Watts-Sabel models [9]. These topologies are denser than trees, they have many loops and have an underlying *lattice*. That's why next on the line are lattices and one can add lattices with random cross links, or randomly rewired (as studied by Watts et al [34]). Finally, we have the random graphs of all varieties, random with the same degree distribution, with an imposed modular structure, and Erdős-Rényi random graphs with various densities. What is not shown on the far right is the densest possible graph - the complete graph. The complete graph is just a $(n-1)$ -regular graph, with a diameter of 1, average path length of 1 ($l = d$ addressed below), clustering coefficient of 1 and infinite degree correlation (same as any k -regular graph). Figure 2-2 shows the corresponding densities for the canonical topologies, marking most real systems with edge-to-node ratios between 2 and 10, which correspond to random hierarchies on this scale.

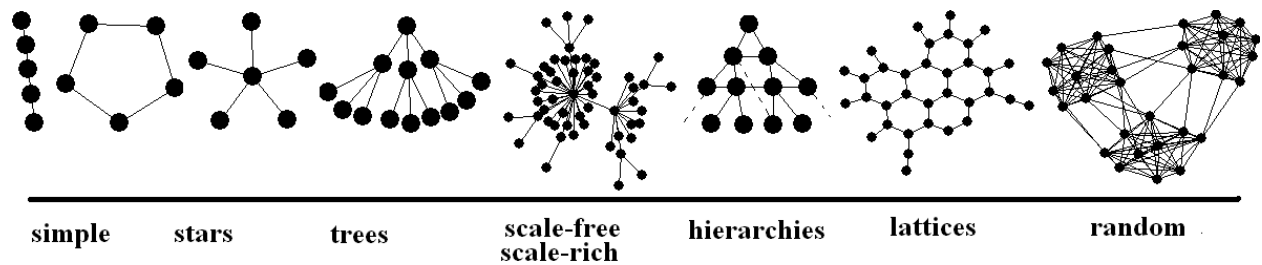


Figure 2-1: Topology families in order of increasing graph density ($m/(n(n-1)/2)$): lines, circles, stars, balanced trees, BA/s-max, hierarchical trees with interlinking, lattices and random graphs.

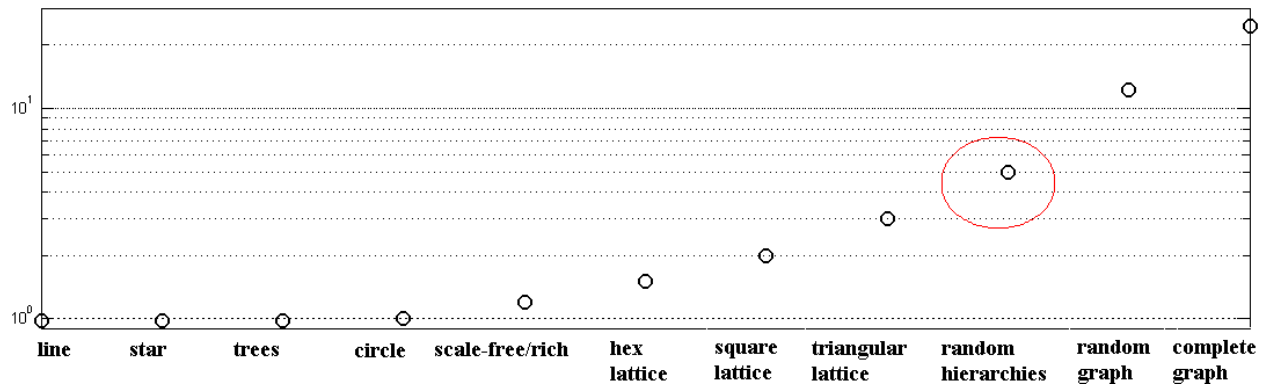


Figure 2-2: Edge-to-node ratio for the topologies in Figure 2-1. As n increases the y-axis scale increases as $O(n)$, as random graphs and complete graphs grow in edges with $O(n^2)$. Most real networks (certainly airlines and Wikipedias studied here) have edge-to-node ratio between 2 and 10 (marked by circle on the plot).

Table 2.1.1 shows a summary of statistical values for edge-to-node ratio, m/n , clustering coefficient, C , degree correlation, r , average path length, l , and diameter, d for the canonical networks defined above. Most values are derived analytically (star, circle and lattices), though there are some values that are computed as ensemble averages. The degree correlation for binary tree and the degree correlation and clustering coefficient for hierarchical trees (balanced trees with horizontal cross links) are cited from [32] and [47].

Table 2.2: Statistics for some canonical graphs - m/n , r , C , l and d .

Graph	m/n	r	C	l	d
Star	$1-1/n$	-1	0	$2(1-1/n)$	2
Circle	1	0/0	0 ($n>3$)	$(n+1)/4$ if n -odd; $n^2/4(n-1)$ if n -even	$\lfloor n/2 \rfloor$
tree(b,L)	$1-1/n$	-1/3 (binary tree)	0	$O(\log(n))?$	$\leq n-1$
BA	$O(\text{const})$	likely <0	low, ~ 0	$\alpha > 3, l = \ln(n),$ $\alpha = 3, l = \log(n)/\log\log(n),$ $2 < \alpha < 3,$ $l=1/2+2/(3-\alpha)$	$O(\log(n))$ for sparse graphs, $\log(n)/\log\log(n)$ asymptotically
Tree w/ horizontal cross linking	2.5 (at ∞)	-1/5 (at ∞)	11/28	probably $O(n)$	$2L-1, L$ - number of levels
triangular lattice	3	1/2	1	$\frac{2(n+1)(2n+1)}{n-1}$ for $n \rightarrow \infty$	$2(\sqrt{n} - 1)$ for a square lattice
square lattice	2	2/3	0	$\frac{2(n+1)(2n+1)}{3(n-1)}$ for $n \rightarrow \infty$	$2(\sqrt{n} - 1)$ for a square lattice
hexagonal lattice	1.5	0 (at ∞)	0	$\frac{3(n+1)(2n+1)}{n-1}$ for $n \rightarrow \infty$	$n/2-3$ for a square lattice
ER	$p(n-1)/2 = O(n)$	0	0	$\log(n)/\log(m/n)$	unknown, possibly $O(\log(n))^3$

To discuss the statistics, first we will show that if two metrics are close to each other, or asymptotically equal, that does not mean the two corresponding graphs are. Consider the following Lemma.

Lemma: If $l = d$, where d is the diameter of a network with n nodes and l is the average path-length, then the network is a complete graph.

Proof: By definition, l is the average path length, so $l = \frac{\sum_{i<j} p_{ij}}{n(n-1)/2}$, where p is the length of the shortest path between nodes i and j , and the sum is over all ordered pairs of nodes (i, j) , assuming the graph is undirected. Since the diameter is the longest shortest path, $p_{ij} \leq d$ for every (i, j) .

Then $l = \frac{\sum_{i<j} p_{ij}}{n(n-1)/2} \leq l = \frac{\sum_{i<j} d}{n(n-1)/2} = dl = \frac{\sum_{i<j} 1}{n(n-1)/2} = dl = \frac{n(n-1)/2}{n(n-1)/2} = d$, where equality is reached only if $p_{ij} = d$ for every i and every j . Therefore, if the average path length is equal to the diameter, then for this graph, all path lengths are equal, to each other and hence to the diameter. The question is what type of graph has all the same path lengths. An obvious answer is a complete graph. All paths in a complete graph equal 1, and the diameter is 1. It is easy to see that it is not possible for a graph to have equal path lengths of size greater than 1. Suppose there is such a graph and the path between i and j consists of $k > 1$ edges. Then along that path, there are nodes which have path lengths of size 1 from i and j , and all path lengths are not equal.

Given this Lemma, it appears that while some metrics can be very close in comparison, the networks they represent do not have to be. Though one might argue that infinite is tricky to consider, an infinite star satisfies the condition for a complete graph ($l = d$). Then consider a very large finite star for which $l \approx d$. The proximity of these metrics means nothing as shown by the Lemma. A star is a tree graph with the minimum possible density that ensures connectivity ($m = n - 1$), while a complete graph is maximally dense with $m = n(n - 1)/2$.

All degree correlations, path-lengths, and diameters are computed analytically for the circle graph, and all lattices. The first thing to notice about Table 2.1.1 is that the measures can either be a function of n , and grow with n , (sub)linearly ($\log(n)$) or superlinearly or be fairly constant and insensitive to the number of nodes. Discovering such networks was the beginning of the small-world hypothesis which identifies cases in which the average path-length or the diameter of the network remains small ($\log(n)$) regardless of the size of the network and despite significant clustering [34]. Lattices are on the opposite side of the spectrum as they spread out with size and their path lengths are longer. The small-world graphs (in Figure 2-1) are the stars and the variations of random graphs: Erdős-Rényi, random modular, and the preferential attachment graph.

Another point of comparison is the number of edges. The random networks are a lot denser than the trees and the lattices, which have $O(n)$ edges. More edges means more opportunities for clustering and short-circuiting which explains the lower diameters. The other interesting measure is the degree correlation which varies a lot across different networks and which has to be combined with the degree distribution to be interpreted properly. The degree correlation exhibits elasticity (changes under degree-preserving random rewiring) which is dependent upon the variation on the degree sequence. A highly peaked degree sequence results in a negative degree correlation with low elasticity (since there are few nodes with high degrees, and few possibilities for rewiring). A more uniform degree sequence, results in a positive degree correlation and higher elasticity. These phenomena are explored in detail with examples of real systems in [32].

Given the statistical measures of canonical networks we can look at the same measures for real networks and based on their statistical profile, derive possible similarities in network topology to canonical networks, their combinations or derivatives. First we present an approach to combine the topologies above and their metrics into a profile against which a real network can be measured. Then, to shed light on these ideas, we discuss the routes structure of JetBlue 8/07.

2.1.2 Topology spectrum

As demonstrated in the discussion of statistical measures, one metric is not enough to make a conclusion about the topology of a network. One straightforward but effective way to compare topologies using only the statistics is to match vectors of multiple metrics. A sum of squares distance between two "topology vectors", for example, can be used to tell how *far* two topologies are from each other. For this to work, the vector entries have to be non-dimensional, properly scaled and reflect widely different properties of the graph. We propose the following derivatives of m/n , C , r , d , and $s/smax$ some of which were discussed above.

Density:	Substitute edge to node ratio (m/n) with density $\frac{m}{n(n-1)/2} = \frac{2m}{n(n-1)}$ which is non-dimensional.
Clustering coefficient:	C is non-dimensional already, so is used as is.
Degree correlation:	r is a measure between -1 and 1, so it is scaled to fit between 0 and 1: $r \rightarrow 0.5r + 0.5$.
Scale-free index:	$s/smax$ is already between 0 and 1, so is used as is.
Diameter:	d is scaled by dividing it by the maximum possible diameter, $n - 1$: $d \rightarrow d/(n - 1)$.

The five-dimensional topology vector becomes $v = [\frac{2m}{n(n-1)}, C, \frac{r+1}{2}, \frac{s}{smax}, \frac{d}{n-1}]$. This vector can be used for relative comparison across networks. If two graphs G_1 and G_2 have vectors v_1 and v_2 , then the distance between them is $|v_1 - v_2|_2 = \sqrt{\sum_k (v_{1k} - v_{2k})^2}$. We call this distance the “topological similarity”. Zero distance means highest possible match, though it is not clear that also means the two graphs are isomorphic. This measure allows relative comparison along the line of canonical topologies, which we call the topology profile. Next follow some examples of how this idea can be used to analyze network topologies. We start by validating the topology similarity measure using canonical topologies to match graphs with already known topology.

Comparison to a canonical network would have greater value if that canonical network is designed to match the graph, in size, and other metrics if needed. The first requirement is to have the same number of nodes and edges (if the generation algorithm accounts for edges). In the case of a random graph, this could mean the same density, translated to a probability of attachment ($p = 2m/n(n-1)$). In the case of trees, the only thing fixed is the number of nodes. To create the s-max graph, the same number of nodes and the same degree distribution is needed. The generation algorithms for all of these graphs are reviewed in Chapter 1, Section 1.4. For the topology profile developed here, the list of canonical graphs is: line, circle, star, binary tree, tertiary tree, Newman-Gastner graphs with three α parameters (0.1,0.5 and 0.9), hierarchical trees (same as trees, but with cross-cutting links at every level), Dodds-Watts-Sabel graphs with four parameter conditions for λ and ξ , lattices, BA graph, random graph the same degree distribution, s-max graph and an Erdős-Rényi graph.

The arrangement of these graphs is according to increasing density and “complexity”, as discussed in the previous section. The vector distance, or topological similarity of an ER random graph ($p=0.5$) to the line of canonical topologies (from Figure 2-1) is shown in Figure 2-3.

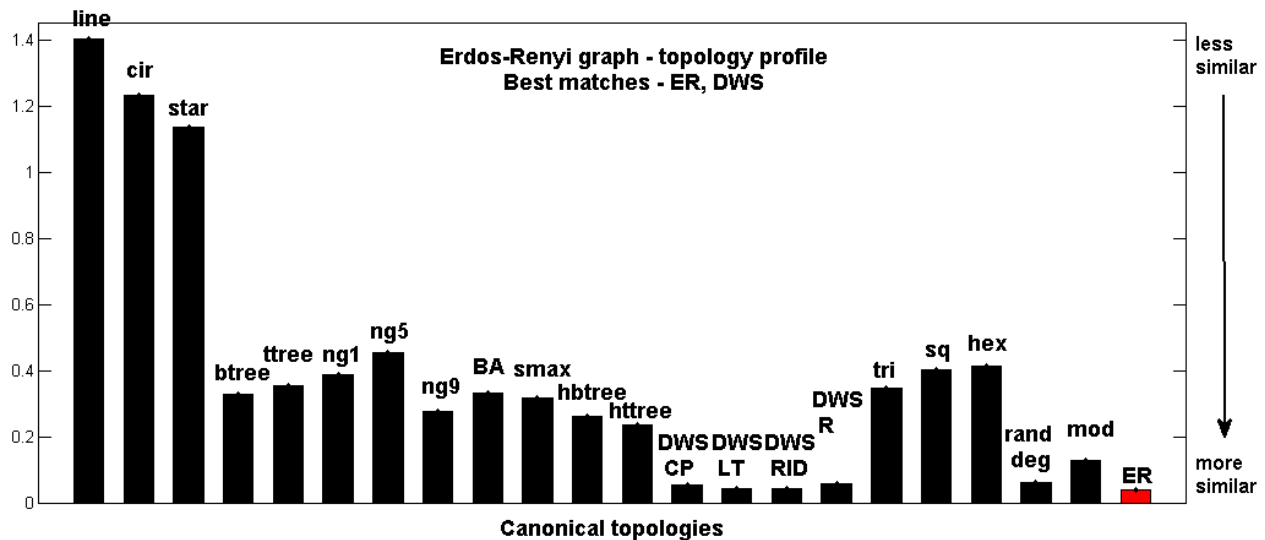


Figure 2-3: Topology profile for an Erdős-Rényi random graph. The graphs compared to are: line, ring, star, binary tree, tertiary tree, Newman-Gastner graphs with $\alpha=0.1, 0.5, 0.9$, preferential attachment graph, s-max graph, hierarchical binary tree, hierarchical tertiary tree, Dodds-Watts-Sabel graphs with varying parameters, lattices (triangular, square and hexagonal), random graph with the same degree distribution, a random modular graph and an Erdős-Rényi graph with the same density.

For the ER graphs, the best matches are other ER graphs (same number of nodes, same density) and the set of DWS graphs, or hierarchies with random interlinking. For very dense random graphs (for example $p=0.5$) there are many more edges on top of the hierarchical trees for the DWS graphs, so they look random, despite the underlying regular backbone. This explains their high similarity. From the other topologies, only the random graphs reflect density as well, so there are no other good matches.

The topology profile for a BA graph (preferential attachment) is shown in Figure 2-4. The best matches here are BA graphs, and trees, tertiary and then binary. The DWS hierarchies are a close match as well, because in this setting, they have very low density, so they are close to their underlying hierarchical trees. The BA graph is not very similar to its most scale-free corresponding graph (s-max). This finding has been discussed by Li et al [4] in the context of the Internet topology.

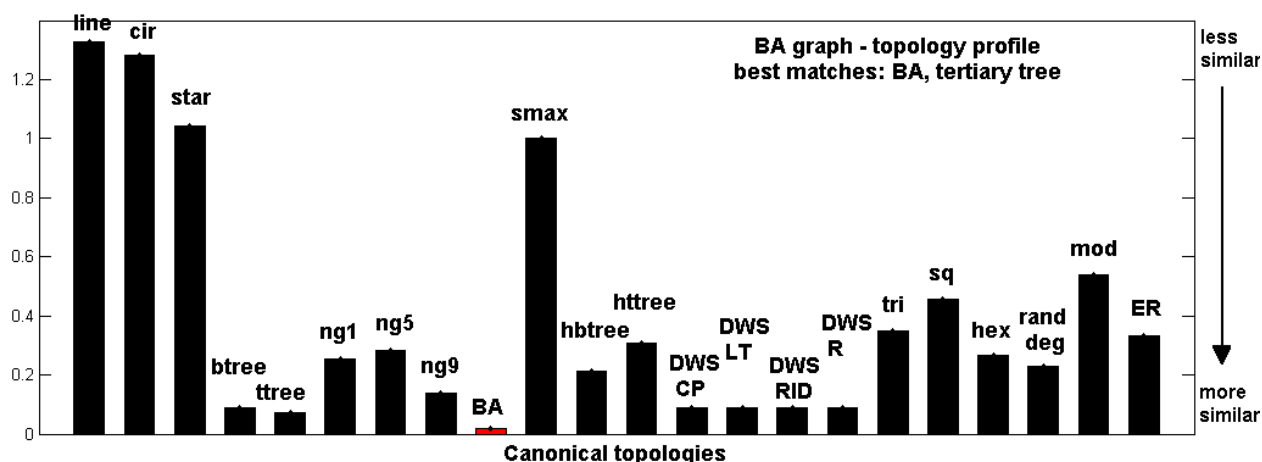


Figure 2-4: Topology profile for a BA graph. The graphs compared to are: line, ring, star, binary tree, tertiary tree, Newman-Gastner graphs with $\alpha=0.1, 0.5, 0.9$, preferential attachment graph, s-max graph, hierarchical binary tree, hierarchical tertiary tree, Dodds-Watts-Sabel graphs with varying parameters, lattices (triangular, square and hexagonal), random graph with the same degree distribution, a random modular graph and an Erdős-Rényi graph with the same density.

As the examples of topology profiles show, the purpose of the topology vector is to place a graph relatively on a scale of known topologies. Since equivalence to a topology or a graph is not pursued here, this is not a classification approach, but a reference scale measure. This profile, based on canonical topologies and topology vector distance, can spot interesting structures and substructures to study further, and also differentiate at a high-level between networks that fall in different regions of the profile. Results for airlines, using the topology profile will be discussed in Chapter 3. Chapter 4 uses the profile to compare real networks to canonical topologies over time - and thus give a relative sense of transitions and stable patterns in topology evolution.

Table 2.1.2 summarizes the difficulty of computing all the components in the topology vector. The hardest to compute are the s-max measure and the diameter. For dense graphs, the clustering coefficient and the degree correlation can be challenging too because they are quadratic functions of the number of edges. The clustering coefficient can also be seen as a function of the average degree (for a random graph, $k = m/n$). In practice, $s - max$ and d require most resources.

Table 2.3: Algorithm complexity for the components of the topology vector: density, clustering coefficient, degree correlation, s/s-max and diameter. The number of nodes is n , the number of edges m and the average nodal degree is k .

Component	Expression	Complexity
Density	$\frac{2m}{n(n-1)}$	$O(m)$
Clustering coefficient	C , equation 1.1	$O(n\frac{(m-1)(m-2)}{2})$
Degree correlation	r , equation 1.2	$O(m^2)$
S-max	$\frac{s}{s_{max}}$	$O(kn^2)$
Diameter	d	$O(n^2 \log n)$

2.1.3 Statistics for JetBlue Airways, August 2007

The example network in this chapter is the route network of JetBlue Airways for August 2007, as reported to the BTS [1]. In August 2007, JetBlue flew to 51 airports, with 100 point-to-point destination city pairs. All statistics for the network are shown in Table 2.1.3. The results for the unweighted and weighted route nets are different because the actual network is a directed and not a connected graph (following directional links, not all $i - j$ paths are possible). The one-way flights turn out to be the islands and Cape flights to Boston - Provincetown, Hyannis and Nantucket to BOS. This is probably because JetBlue outsources these flights to Cape Air. The weights are the number of departures monthly on a given leg.

Table 2.4: JetBlue 8/2007 statistics for weighted (by number of departures monthly) and unweighted route versions. Note that the original directed network is not connected.

JetBlue 7/2008 stats	undirected, unweighted	directed, weighted
m/n	1.96	273.2
r	-0.574	-0.542
C	0.373	0.294
l	2.0369	∞
d	3	∞

With a low edge-to-node ratio of about 2, JetBlue does not have a dense network. The mid-range negative degree correlation of -0.542 shows the tendency of nodes linking to nodes of different degree, so potentially high-degree to low-degree, i.e. hub-to-spoke type links. This is not a definite statement because a single degree correlation can correspond to many topologies. The skewness of the degree distribution affects the degree correlation elasticity. Coupled with the low network diameter and average path-length, it is evident that the network is star-like. These observations are confirmed by the topology profile shown in Figure 3-23. JetBlue is closest to a BA graph and trees, in the spectrum of topologies. In summary, the statistics and the topology profile comparison suggest a topology with many stars (short diameter), few hubs and many spoke flights. For reference, the actual route network of JetBlue 8/07 is plotted in Figure 2-6. The graphs with 51 nodes is small enough that the structure can be seen by eye. Figure 2-6 shows the JetBlue 8/2007 network with JFK (New York) as biggest hub, BOS (Boston) as secondary and Florida (MCO, PBI, FLL) forming a local pattern. Surprisingly, LGB (Long Beach) is small hub and LAX (Los Angeles) does not appear yet (service from JFK and BOS to LAX started May 21, 2008).

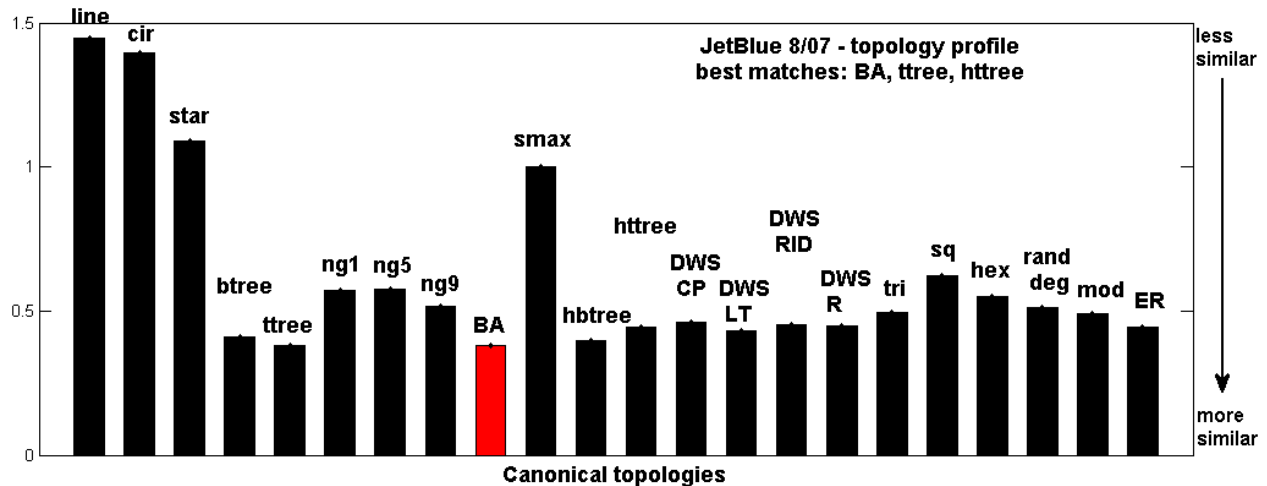


Figure 2-5: Topology profile for the JetBlue 8/07 network. The closest match is to a preferential attachment (BA) graph.

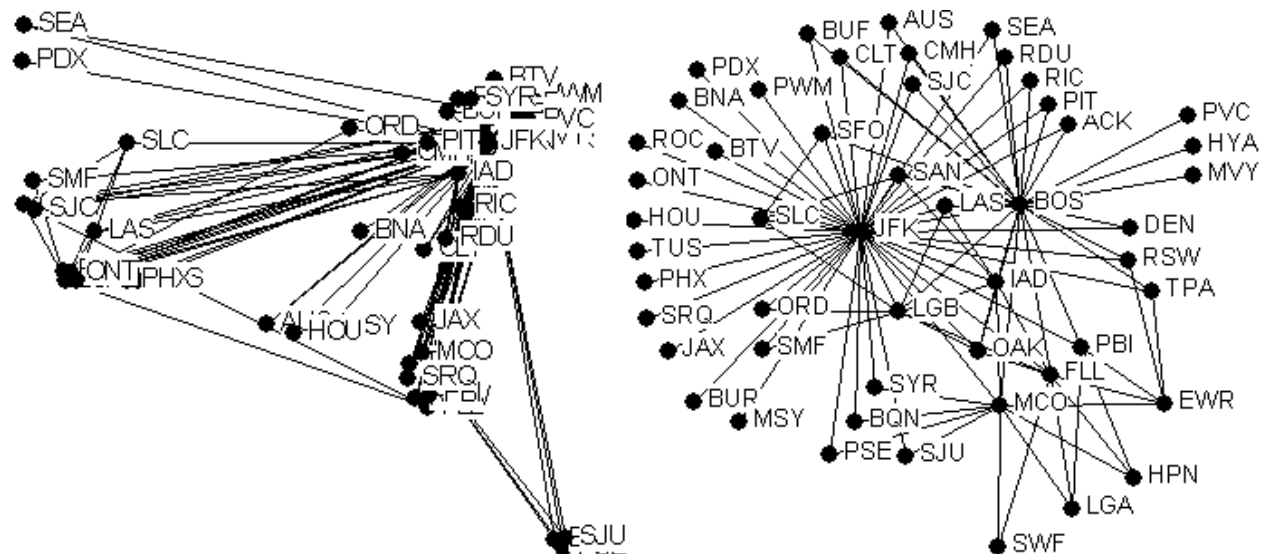


Figure 2-6: JetBlue 8/2007 routes: geographic representation by airport latitude/longitude, and representation of node locations by minimizing spring energy. JFK and BOS are the two largest hubs.

2.1.4 Degree-Betweenness Relationship

The relationship between nodal degree and nodal betweenness reveals more information about the graph compared to the degree distribution alone. With a simple mental exercise, one can prove that the two distributions are different yet proportional. Consider k -regular graphs. These are graphs for which every node has the same degree k . For example, a circle is 2-regular. An infinite square lattice is 4-regular. And a complete graph is $(n - 1)$ -regular where n is the number of nodes. Since a k -regular graph is symmetric with respect to all nodes, every node has the same degree, and the same betweenness measure, because all paths going through are symmetric.

The next level of regular graph is the "perfectly random" Erdős-Rényi graph. It is regular in expectation because all nodal degrees have the same expectation which is the mean degree. The betweenness-degree relationship of a random graph shows the null model relationship - that is all things equal, what is the connection between these two statistics. Intuitively, if a node has more links, then potentially more paths go through it, and hence a higher number of shortest paths. Though it is not always true that if a node has higher degree, it has higher betweenness, it certainly seems intuitive. This is exactly what the random graph results indicate. In Figure 2-7, the left-most plot shows the betweenness measure plotted versus the degree for all nodes of a random graph with 200 nodes and average degree of 39 ($p = 0.2$). The plot confirms that on average, higher degree implies higher betweenness. This signature-trend plot can be used as a null model to distinguish graphs from random (ER). Figure 2-7 also shows the relationship for modular random graph with 4 modules (in the middle) and a preferential attachment graph (right-most). In the case of the modular graph, the modules themselves are ER random graphs, connected with some probability. We expect that the nodes that connect modules will have higher betweenness since more cross-paths will be going through them. This is why the linear relationship is not as pronounced, which means that modularity affects the proportional betweenness-degree relationship. Finally, in the case of the hub-and-spoke graph, all the spokes of degree 1 will have 0 betweenness and all branch nodes of all degree will have low betweenness - so we expect to see a clustering of nodes around 0 (in the lower left corner of the plot). High-degree nodes are expected to have very high betweenness and really stand out together with a few peripheral hubs. The left-most plot in Figure 2-7 which shows this relationship for random preferential attachment graph with 200 nodes, confirms that.

In conclusion, for highly skewed degree distributions of graphs with strong hubs and many spokes, the betweenness versus degree relationship is expected to show this strong separation of nodes into hubs and spokes.

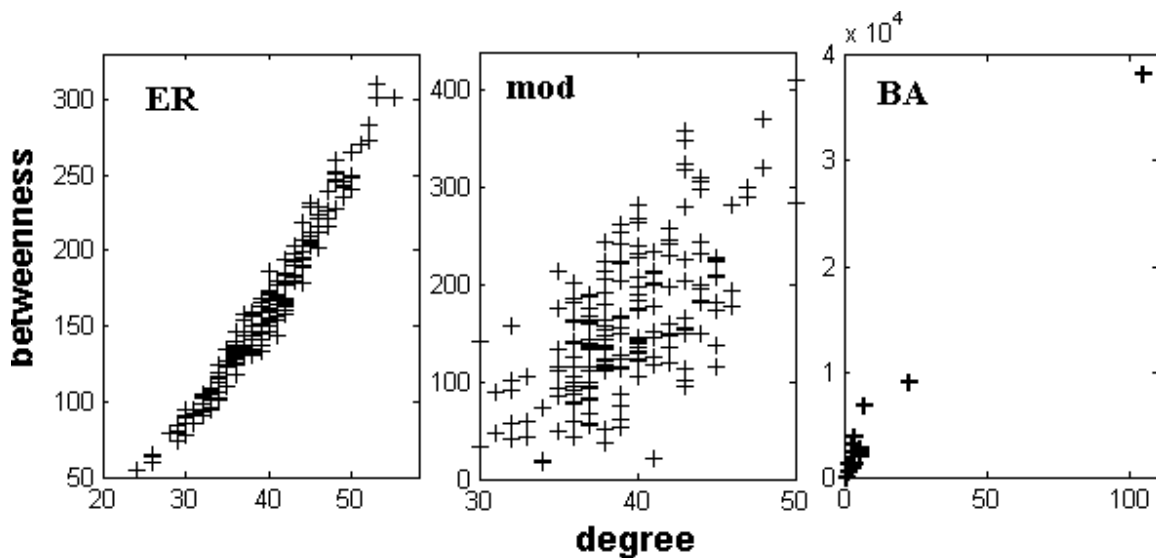


Figure 2-7: Betweenness versus degree for all nodes of three types of graphs: (left) Erdős-Rényi graph with 200 nodes and $p=0.2$; (middle) random modular graph, with 200 nodes, 4 clusters, and $p=0.2$ general density; (right) BA graphs with 200 nodes.

2.1.5 JetBlue August 2007 and degree distributions

The goal of this subsection is to show the degree distribution of JetBlue's route network, and discuss it alone, with betweenness and with degree correlation to find out what can be learned about JetBlue's topology.

The weights of edges for JetBlue are the number of monthly departures for every pair of airports. For example, if there is a daily flight between A and B, then the link weight will be 31 (31 departures in a 31-day month). So the weighted degree of a node is the sum of the weights of all links adjacent to that node. Counting the departures or another metric, gives a sense of "multiple edges" in the degree distribution and is a more accurate description of node centrality with respect to what the network actually does. What defines the operations of an airline over time is the number of flights the airline chooses to perform for any given city pair, as well as the seat capacity it offers. A hub in this sense is really defined by the total number of departures, rather than the total number of absolute links, and that results in different weighted and unweighted degree distributions [10][48]. This argument can be made for other transportation or resource distribution systems.

Mathematically, the resulting degree distribution exponent (if it exists, the distribution may not be a power-law) will be different, and will affect growth, depending on the model (ex: preferential attachment). If however, the flight frequency is proportional to the link connectivity, the degree (frequency) distributions will look the same. We find that in the case of JetBlue, the two distributions are different, but not drastically.

Figure 2-8 shows both the unweighted and weighted cumulative degree distributions for JetBlue. The overall weighted distribution is best matched by a slow exponential, with exponent -0.005. Only the tail fits a power law. The power law exponent is the slope of the tail plotted on a log-log scale. If the power law is defined as $P(k) \sim k^{-\gamma}$, then $\log P(k) \sim -\gamma \log k$, therefore $-\gamma$ is the slope of $\log P(k)$ versus $\log k$. If γ is the exponent of the probability density, $\gamma - 1$ is the exponent of the cumulative distribution, by integration. In Figure 2-8, the tail has a slope of -1.046, which means that $\gamma = 2.046$.

As for exponential fits, the unweighted distribution has a much faster decaying exponential (an exponent of -0.6 compared to -0.008) than the weighted distribution. That means that fewer airports have proportionally higher number of connections, and relatively more airports have a higher number of departures overall. The tail also matches a power law, with a similar but lower exponent of -1.072, which means that $\gamma = 2.072$, and that distribution does fall off faster. In other words, in airline terms, there are airports with low connectivity (number of connecting destinations), but still high traffic share (in total number of departures monthly).

As seen for the both the weighted and unweighted distributions, the tail follows the same power law with $\gamma \approx 2$. A "regular" preferential attachment process results in a power law with exponent 3 (see Chapter 1, Section 1.4.2). Typical real networks have exponents between 2 and 3 [2].

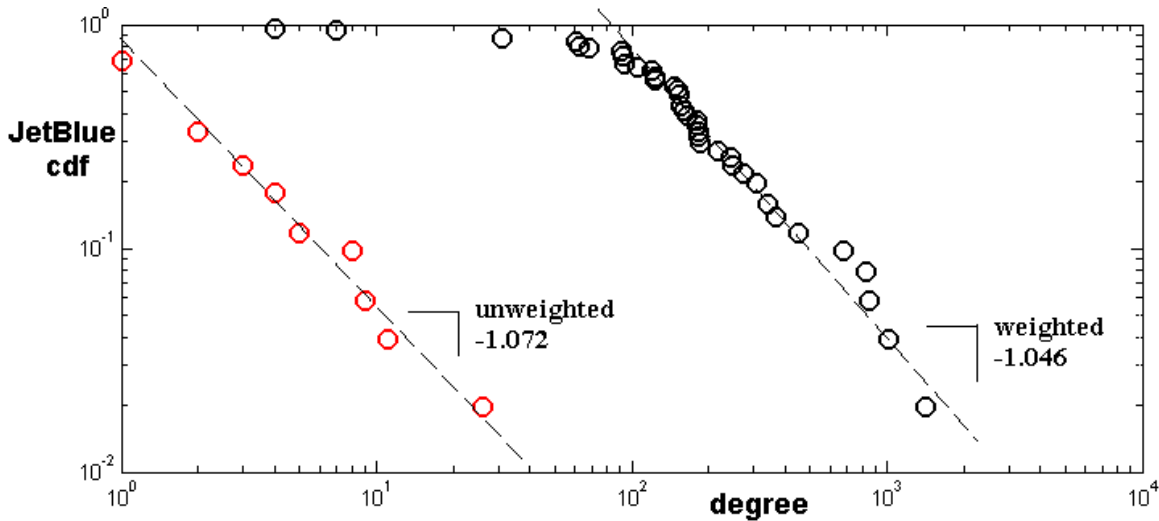


Figure 2-8: Unweighted and weighted cumulative degree distribution for JetBlue 8/07. Both slopes close to -1, which means that $\gamma \approx -2$ for the frequency degree distribution.

Since a given degree distribution together with a degree correlation can correspond to many graph instances, of varying topologies, one way to explore the realm of topologies is to rewire the network, while preserving the degree distribution, and find the maximum possible span of degree correlations. Some networks can be rewired to have a wide band of correlations, from negative to positive, while others can be fairly inelastic [32]. Figure 2-9 shows the spring-energy plots of the rewired JetBlue networks for maximum correlation of -0.542^3 and minimum correlation of -0.667^4 (the actual r is -0.574). The minimum correlation found also corresponds to the s -maximum graph for the JetBlue 8/07 network.

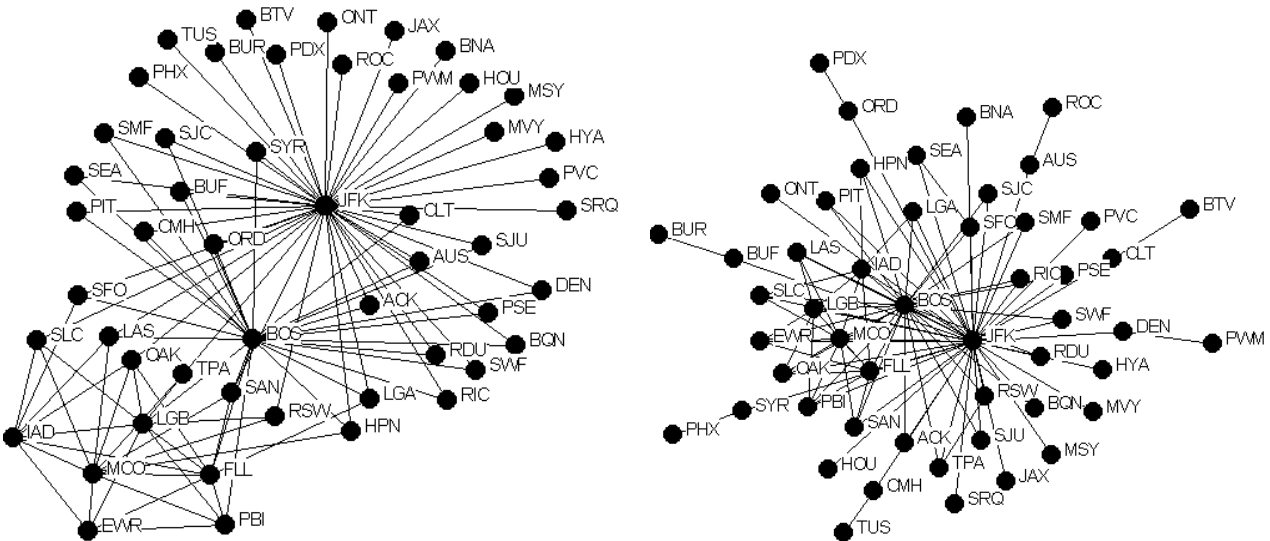


Figure 2-9: Rewiring JetBlue 8/2007 from minimum $r = -0.667$ (left) to maximum $r = -0.542$ (right).

Rewiring with preserving degree distribution and minimizing/maximizing the degree corre-

³This is the absolute maximum, corresponding to the s -max graph.

⁴This is a heuristic result, since there is no known algorithm for minimum r .

lation shows that the JetBlue route network can be rewired only within a very narrow band ($r_{min} = -0.667 < r_{actual} = -0.574 < r_{max} = -0.542$), hence has a fairly inelastic negative degree correlation.

Figure 2-10 shows the betweenness versus nodal degree for JetBlue Airways 8/2007. Every point is an airport, with coordinates its betweenness and degree in the network. Two airports stand out - JFK (New York) and BOS (Boston). As seen in Figure 2-6 these are the two major hubs, routing traffic to many spokes and other secondary hubs. So their high-betweenness is no surprise. In the cluster of low-degree, low-betweenness airports (relatively) Orlando stands out (MCO) as one of the bigger secondary hubs in JetBlue’s Florida operations. The disparity between JFK, BOS and the rest matches the betweenness-degree relationship seen in the case of the BA graph in Figure 2-7. This, together with JetBlue’s topology profile showing highest match to the BA graph, confirms the emerging hypothesis that JetBlue operates a hub-spoke, highly centralized network.

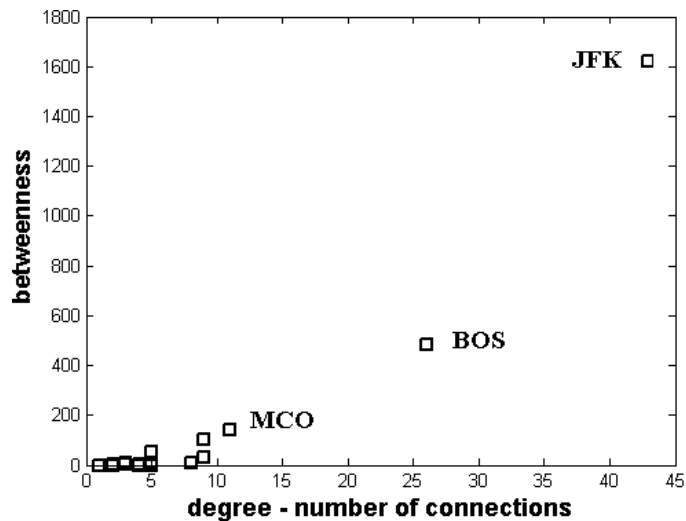


Figure 2-10: Betweenness versus degree for all JetBlue airports of August 2007. JFK and BOS stand out as both high-degree and high-betweenness. Orlando (MCO) emerges as a third important hub. This profile is most similar to the BA profile on Figure 2-7, confirming the similarity to BA graphs from the topology profile (Figure 3-23).

In this section, we showed the graph-theoretic statistics discussed in Chapter 1 can be used to extract information about a network. We developed the concept of a *topology profile*, and explicitly ordered a set of canonical topologies as a base of comparison for a real network. We combined non-dimensional metrics, such as density, clustering coefficient, scaled degree correlation, and normalized diameter to create a five-dimensional *topology vector* that can be used for general comparison between graphs. These ideas were applied to the route network of JetBlue airlines of August 2007.

In the final subsection we discussed the use of degree distributions combined with other measures to study topology. Using the same example, we found that JetBlue has a highly-skewed power-law degree distribution, with fairly inelastic degree correlation. This coupled with the betweenness versus degree results of outstanding hubs, confirms that JetBlue’s routes have a BA-like topology, with strong hubs and many spokes. Note that all of these conclusions were made without “looking” at the network, so its size is irrelevant.

The next section presents extensions on the graph similarity measure by Blondel et al [49] that can be used to detect changes in topology.

2.2 Graph Similarity (Hubs/Authorities Comparison to Canonical Networks)

2.2.1 Graph similarity basics

So far we have discussed canonical topologies as reference topologies, but not actively compared them to a real-world graph topology. Blondel et al [49] developed an algorithm based on Kleinberg’s hubs and authorities method [50] which finds a set of good hubs and a set of good authorities on a given web query. For example, if the query is ”university”, pages like MIT and Harvard’s homepages are good authorities on the query, and pages that point to them are good hubs. In general, good hubs point to good authorities, and good authorities are those that point to good hubs. This separation is meaningful in annotated graphs (with associated node semantics) with explicit directionality. Kleinberg proposes an iterative method that assigns *hub scores* and *authority scores* using flow equations:

$$\begin{cases} h_j \leftarrow \sum_{i:(j,i) \in E} a_i \\ a_j \leftarrow \sum_{i:(i,j) \in E} h_i \end{cases} \quad (2.1)$$

Equation 2.1 can be written iteratively as $\begin{bmatrix} h \\ a \end{bmatrix}_{k+1} = \begin{bmatrix} 0 & B \\ B^T & 0 \end{bmatrix} \begin{bmatrix} h \\ a \end{bmatrix}_k$, $k = 0, 1, \dots$, where B is the adjacency matrix of the graph. In compact form, this can be written as $x_{k+1} = Mx_k$. The matrix M is symmetric and non-negative. For hub and authority scoring only the relative scores are interesting, therefore they are computed as the normalized vector sequence: $z_0 = x_0 > 0$, $z_{k+1} = \frac{Mz_k}{\|Mz_k\|_2}$, $k = 0, 1, \dots$. This sequence does not always converge depending on the matrix M and the initial vector, z_0 , the first of which is given and the second not obvious to pick. In fact, usually the sequence oscillates between two limits. Blondel et al show that it is sufficient to pick $z_0 = 1$, which converges to the even limit at infinity, and is among all limits the one with the largest 1-norm. This limit is the formal definition of the similarity matrix between two graphs. The normalized vector equation above can be re-written in matrix form, as $X_{k+1} = BX_kA^T + B^TX_kA$, where A is the adjacency matrix of one graph, and B the adjacency matrix of the other. The resulting similarity matrix X has a score at every entry (i, j) showing relatively how similar, nodes i and j are, compared to other pairs of nodes.

2.2.2 Graph similarity measure assessment

While this metric was developed to aid search in a large hyperlink graph, we will discuss its relevance to comparing topologies, with the hope that it can track almost continuously discrete changes in topology and be able to place a real network topology in a neighborhood of canonical topologies.

First, imagine that the same network is monitored over time, i.e. nodes are compared to themselves, but at different times. There may be addition of new nodes or new edges. We would like a single number metric which:

- Is maximum when the graph is compared to itself.
- Changes little when small changes occur in the graph.

First, if we are comparing an instance of the same network, to track how it changes we are interested in checking how the similarity of nodes to themselves changes over time, even if i does

not remain most similar to i and so on. This is a different question from the one which nodes are similar to which in two instances of the network or in two different networks. Note that the similarity matrix only indicates how nodes in one graph match the nodes in the other, *relatively*, but not how *well* they match. Given that higher entry value $X(i, j)$ means better match relatively, here we use this fact to assess 4-5 metrics derived from the similarity matrix.

- i **Sum of all entries:** $s = \sum_{i,j} X(i, j)$, $i = 1 \dots n$, $j = 1 \dots m$, where n is the number of nodes in the first graph and m in the number of nodes in the second graph, and X is the similarity matrix.
- ii **Maximum entry:** $\max(\max(X))$ which is the maximum across rows and columns of the matrix X .
- iii **Sum of the X entries** corresponding to the **best matching sequence** of nodes from the two graphs. This is not always the diagonal of the matrix, since nodes are not always ordered by best matching sequence. The best matching sequence is found by consecutively identifying the maximum (left) entry and matching the $i \rightarrow j$, for the given maximum value $X(i, j)$.
- iv Numerical experiments with these metrics showed that among the entries along the best diagonal, what gets preserved more than the total magnitude, is the profile of the sequence of entry values. This is why we also suggest taking the **sum of squares of differences of two consecutive diagonal profiles**. Notice that this measure is useful only in tracking the topology changes across multiple stages, rather than point comparison of two graphs.

To test the continuity and overall behavior of all these metrics, we experiment with two canonical graphs: a preferential attachment graph and an Erdős-Rényi graph with initial size 100 nodes and for the ER graph, $p=0.2$ (chosen because it results in a connected fairly sparse random graph, which alleviates flow computations). A random edge is removed from the graph at every step and the resulting graph is compared to the preceding instance until all edges are removed.

Figure 2-11 shows the results for the BA graph. In each case, each snapshot is compared to the previous instance and to the original. The four metrics consistently show the same behavior - stability until the topology breaks, and then noisy behavior. The sum of all entries ($\text{sum}(X)$) falls consistently with every lost edge, while the sum of entries along the best "diagonal" falls off only when compared to the original graph.

The consecutive graph instances show stable diagonal sums, which is an indication of minute changes in topology given small changes in the graph. The max entry metric increases as the graph disintegrates, which does not make sense, because we expect high entry score to mean high similarity. This indicates that the maximum entry is not a good metric to use. Out of the four metrics, the sum of diagonals satisfies the two desired conditions above the best, because it reflects small changes with small deviations and does show differences in topology in a continuous way. Obviously, these are experimental conclusions, rather than rigorous, but will be useful in tracking changes in topology in Chapter 4.

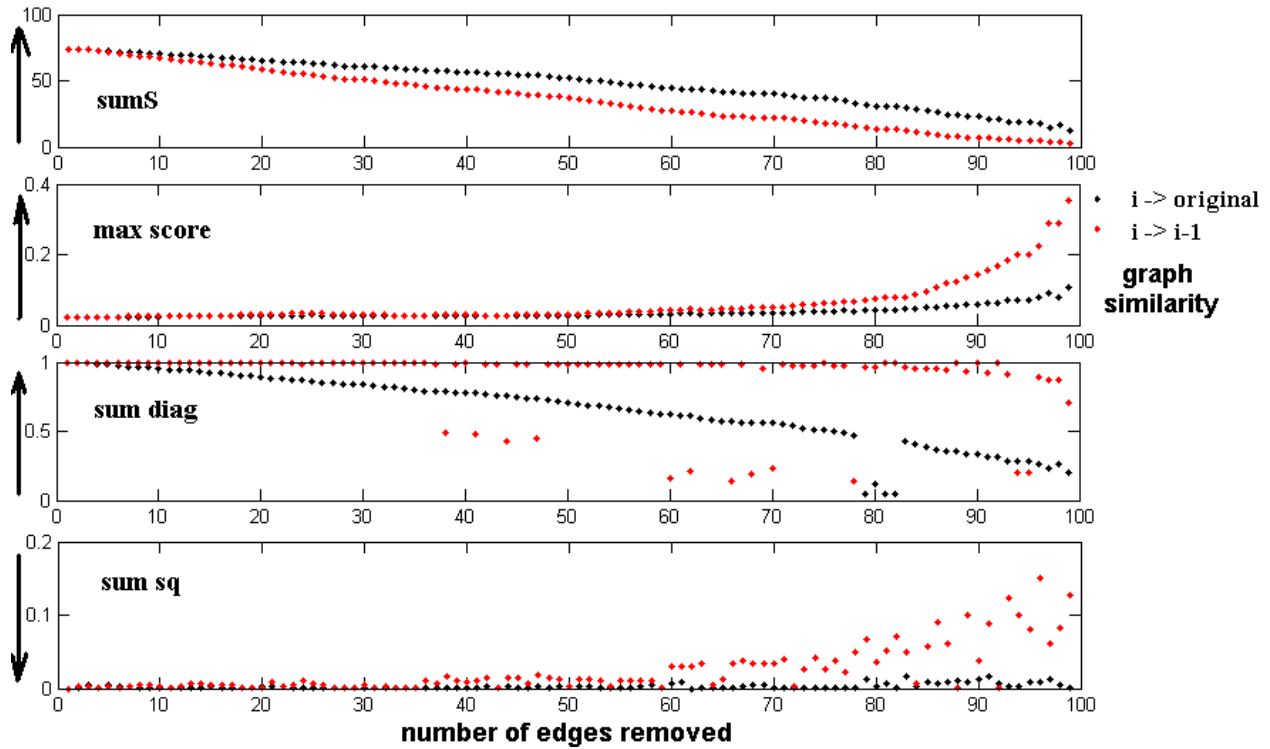


Figure 2-11: BA graph, random edge removal at every step. The four metrics from this section, sum of all entries of the similarity matrix (X), maximum entry, sum of the best diagonal (best matching node sequence), and the Euclidean distance between the diagonals compared. Each metric is used to compare consecutive graph snapshots as well as comparing every new graph to the original. Direction of high similarity is shown by the arrows.

The figures above and the discussion show that the graph similarity ideas by Blondel et al [49] can be used to track changes in topology. However, these metrics are not reliable in comparing one topology to another. For example, a star and a circle (closed loop) are equivalent according the similarity matrix and any of the four measures, and a square lattice is more similar to a triangular lattice of the same size, rather than itself, though in the same order of magnitude. The results show that the similarity measure really is about matching the flow conditions of different nodes. For example, all lattices are very similar to each other and to random graphs, because all nodes, either precisely, or on average, get equal flow through, so they are "equal" in a sense. The BA graph and the star/circle graphs stand out as not too similar to any of the other topologies.

Table 2.2.2 shows the numbers behind this discussion. The conclusion is that graph similarity, as defined here, cannot be used to compare topologies, or determining the topology of a given graph, singly, but can be used more effectively to track changes in topology over time, and potentially major transitions.

Table 2.5: Comparing canonical topologies using the "best-matching-sequence" diagonal sum measure of the similarity matrix. All graphs have 100 nodes. The graphs compared are star, binary tree, tertiary tree, triangular, square, hexagonal lattice, BA graph, random modular graph and an Erdős-Rényi graph.

	star/cir	bin tree	ter tree	tri lattice	sq lattice	hex lattice	BA	mod	ER
star/cir	1								
bin tree	0.7598	0.9964							
ter tree	0.7525	0.9712	0.9438						
tri lattice	0.9212	0.9055	0.8753	0.9997					
sq lattice	0.9275	0.9031	0.8719	0.9994	0.9991				
hex lattice	0.8066	0.9611	0.9166	0.9609	0.9578	1			
BA	0.7851	0.849994	0.7948	0.9147	0.8816	0.9406	0.9848		
mod	0.9885	0.8395	0.8258	0.9659	0.96996	0.8830	0.8512	1	
ER	0.9784	0.8662	0.8508	0.9776	0.9805	0.9065	0.8671	0.9982	1

2.2.3 JetBlue 8/2007 graph similarity

Comparing the JetBlue 8/07 topology to canonical topologies does not give too much information, as discussed in the previous section, because graph similarity does not give an absolute measure of topology. Figure 2-12 shows a comparison of the same topologies, as in Table 2.2.2, with the addition of JetBlue 8/07, a random graph with the same degree distribution as JB 8/07 and the s-max graph with the JetBlue degree distribution. Overall, the JetBlue-related graphs are similar to each other and more so the trees, than to lattices. Notice that these random graphs are derived from the same degree distribution, which means that the degree distribution for the case of JetBlue is such that it strongly determines flows in the network - and it is easy to identify similar nodes based on their degree (hence the high similarity to degree-derived graphs).

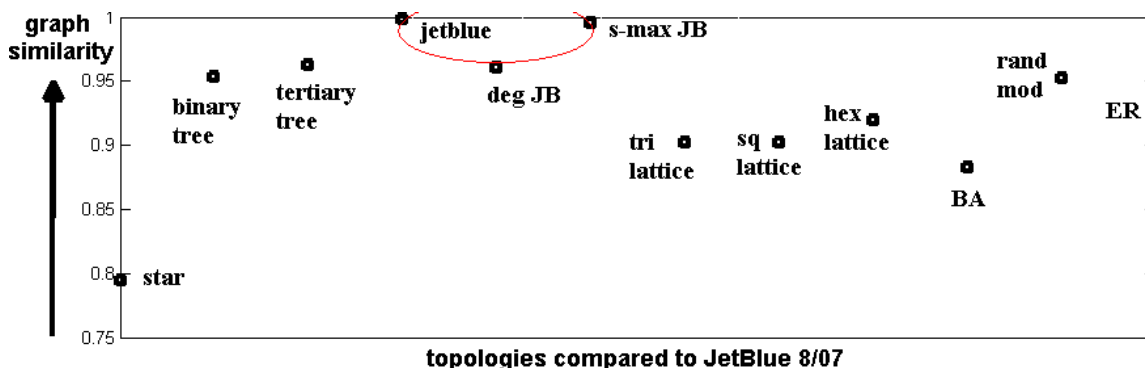


Figure 2-12: Graph similarity score of JetBlue 8/2007 and the following canonical networks: star/circle (0.7948), binary tree (0.9539), tertiary tree (0.9625), itself (0.9986), random graph with the same degree distribution (0.9607), the s-max graph based on JetBlue's degree sequence (0.9949), triangular lattice (0.9024), square lattice (0.9019), hexagonal lattice (0.9196), BA graph (0.8827), random modular graph (0.9526), and an Erdős-Rényi graph (0.9261), all generate with the same characteristics (51 nodes, same density/degree sequence if relevant)

This conclusion has already been supported by our results showing an i/ inelastic degree correlation, and ii/ betweenness-degree relationship with outstanding hubs.

2.3 Modularity and Topology (JetBlue 8/07)

Modularizing a graph in this thesis means detecting or identifying subgraphs that can be classified as modules or components. A module is a subgraph which is more cohesive internally than connected to other modules, or in a probabilistic sense, has more links inside the modules than expected on average [15]. Being modular is a quality that not all graphs have. Some cannot be divided into well-defined components (ex: star). Studying modularity is a higher level of analysis than studying graph statistics.

We will compare two algorithms in for the JetBlue 8/2007 route network - the Newman-Girvan algorithm based on betweenness and the Newman eigenvector method, based on eigenvector computation. There are improvements on the Newman-Girvan algorithms by Hsieh [37], but they are not reviewed here.

2.3.1 Newman-Girvan: modularity using betweenness

Figure 2-13 shows the JetBlue network modularized by cutting high-betweenness edges, with 21 components (a lot of these are single nodes), because it turns out that at 21 the algorithm has the highest Q metric (see Chapter 1). The first substructure the algorithm finds is the bi-partite-like subgraph of connections between Florida and the Caribbean (marked with blue circles on the plot). As it can be seen in Figure 2-13, the connectors are FLL (Fort Lauderdale), PBI (West Palm Beach) and MCO (Orlando). The three airports connected directly to JFK and to the bi-partite component are San Juan (SJU), Ponce (PSE) and Aguadilla (BQN), all in Puerto Rico. The secondary hub component around BOS is well-formed, with fork airports, connected with both BOS and JFK and with pure-BOS spokes, such as Hyannis and Martha's Vineyard. The "pure stars" around JFK and BOS, i.e. the spokes that have no other connections but BOS or JFK, are split into single-node components, which shows that the Newman-Girvan algorithm does not see stars as indivisible components. This is because the algorithm splits the network along the highest-betweenness edges, and all the spoke edges around the same hub have the same zero betweenness.

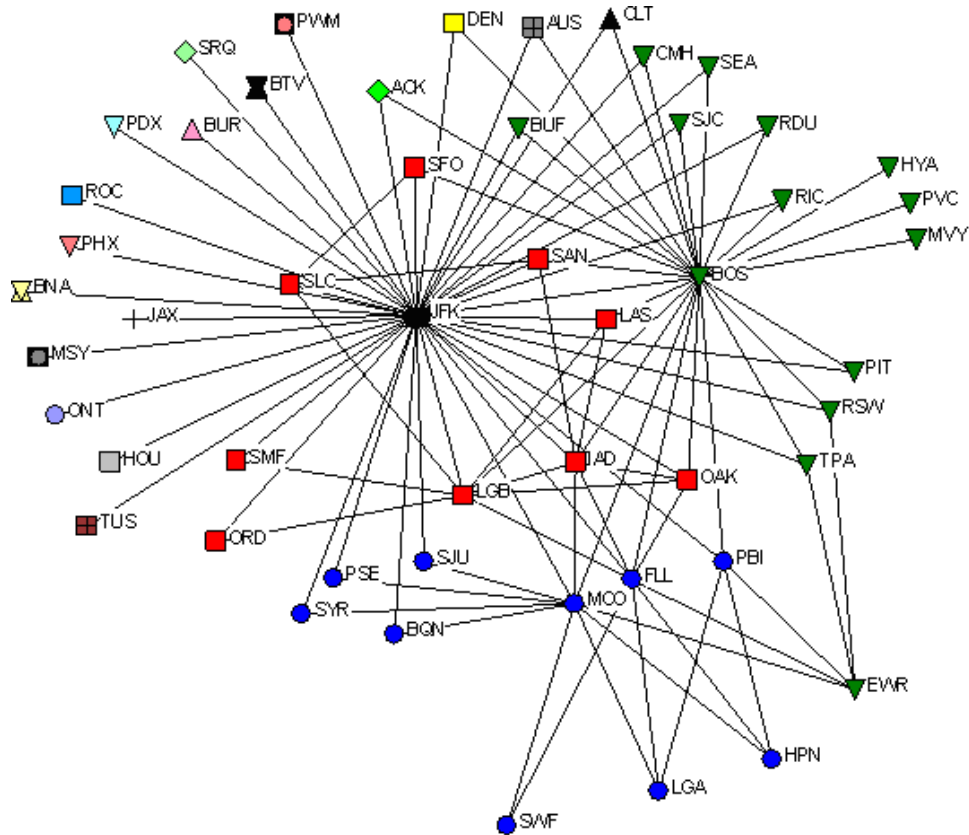


Figure 2-13: Modularizing JetBlue 8/2007 using the Newman-Girvan algorithm: up to 21 components, reflecting maximum Q value. The node locations are plotted using a spring energy algorithm.

The conclusion from the Newman-Girvan method is that the route network is composed of two stars, (JFK and BOS-based), one tight-knit, bi-partite Florida-Caribbean module and an internal well-cross connected module. The JFK-based star is not identified the algorithm, but we will designate it as a component anyway. The internal module is labeled with squares on Figure 2-13, and it involves Oakland (OAK), Long Beach (LGB), Salt Lake City (SLC), Las Vegas (LAS), San Francisco (SFO), Dulles (IAD), San Diego (SAN), Chicago (ORD) and Sacramento, CA (SMF). The Florida-Caribbean module is based out of Orlando (MCO), Fort Lauderdale (FLL), West Palm Beach (PBI), and is connected to New York state and Connecticut on one side (La Guardia, Newburgh, CT (SWF) and Westchester County, NY (HPN)) and through Orlando to San Juan, Puerto Rico (SJU), Aguadilla, PR (BQN), Ponce, PR (PSE).

2.3.2 The Newman eigenvector method

Results from the eigenvector method are similar to those of the Newman-Girvan algorithm, however there are no problems with identifying the stars (Figure 2-14). The spokes around JFK are all grouped into one component, so are most BOS-spokes. The Florida-Caribbean module is present, though with some minor differences. FLL is outside of the Florida-Caribbean group. To explain that better, one would have to look at the strength of the partitions, in particular the magnitude of the corresponding eigenvalues. In the case of JetBlue 8/07 there are no weak splits, in any of the five modularization steps. A weak split means that the entries of the eigenvector used to split the network, are close to zero, not very negative or positive. The strong split is an indication of

strong modularity, which could be quantified using the magnitudes of the eigenvector entries.

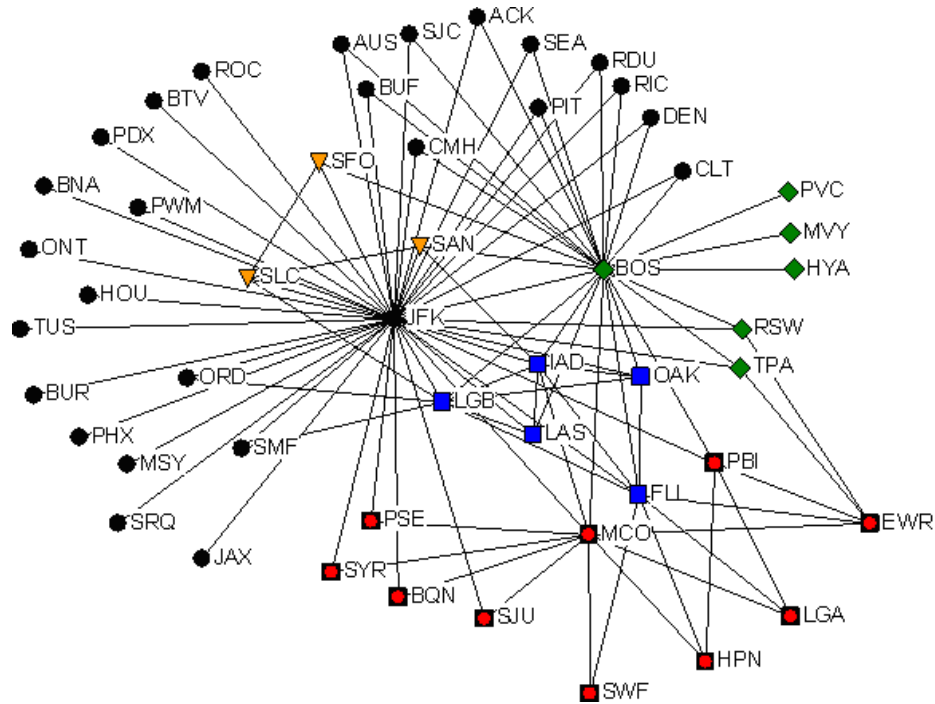


Figure 2-14: Newman eigenvector method modularization of JetBlue 8/2007.

For the JetBlue example, the Newman eigenvector method performs better because it identifies the stars around hubs as components, while the Newman-Girvan algorithm does not. This is natural, because spoke edges have zero betweenness.

2.4 Motifs - The Building Blocks

The motif search in this thesis applies only to unweighted connected graphs. Underlying patterns where edge weights are part of the recurrent motifs can be very useful, but very computationally intensive to find. Finding all possible motifs, as discussed in Chapter 1, is already combinatorially very hard. In Chapter 1 we reviewed two pieces of work on motif finding - searching for a single motif [5] and finding topologically generalized motifs [6]. In this work, we combine the above approaches with heuristics to find "most" frequent subgraphs.

The motif finding routine begins with n (number of nodes) breadth first searches starting at every node. Not every possibility is explored, but the BFS is performed with some degree of randomness. Then a special search is performed for stars, loops and cliques. Finally, the subgraphs found are tested for extensibility, i.e. used in search for topologically similar motifs, by copying nodes in the same equivalence class (see Chapter 1). The steps in the motif search are summarized below:

- i Set a maximum motif size n (number of nodes).
- ii Perform n Breadth-First-Searches starting from every node of the graph, with depth equal to the maximum motif size.
- iii Search for stars, loops and cliques of size 3 up to the maximum motif size.

- iv Search for generalizable motif structures based on the motifs so far, using the ideas from Kashtan et al [6].

At every step the search is capped by a pre-defined motif size, for example for motifs only of size 4,5 or 6 etc. Also, only motifs of size higher than or equal to 3 are considered, with the only 3-node (connected, undirected) motif being the triangle. A line of size three (a 3-node linear motif) is not interesting because it is present by nature of connectivity in every connected graph (except for a star). In terms of motif size, we found that the memory becomes insufficient above 9 nodes (10+) on an Intel processor, Pentium 4CPU, 3.60 GHz machine with 2GB of RAM, so all analysis is done with up to 9-node motifs (the maximum was mostly set to 6, and 7 for some smaller datasets).

Theoretically, motif size is important, and depending on the network size and growth patterns, larger motifs would be very interesting to analyze. The most obvious example is the one of large stars for hub-spoke topologies. There are star patterns in airline networks that have more than 9 nodes, and of course those should be identified as single patterns rather than a set of smaller stars, that share the hub. It would be interesting to combine the motif idea with the idea of modularity. If a system can arguably be modular, then recurring patterns should be sought within modules, not across them. If that is true, that would limit the motif size, and alleviate the computation. Reducing the problem to searching within modules has its caveats, for example, one might miss out on key loop motifs that connect all or some modules in the entire network.

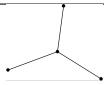

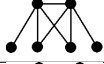




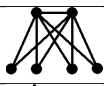



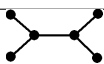
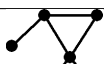



The set of motifs found has many copies and symmetries by virtue of the search procedure itself. Graph copies and graph symmetries are eliminated to present a set of distilled unique motifs. Also, due to the randomness in the BFS, the motif finding routine is repeated until no new motifs (by symmetry) are found. The resulting motifs are still not necessarily significant, but simply occurring in this particular graph. Occurrence or absence is a feature - for example, a tree will have no loops, and that should be evident⁵. To probe significance, we test the occurrence of the set of unique present motifs in an ensemble of random graphs built with the same number of nodes and same degree sequence. Such an ensemble is often used as a null model to test statistical significance of various metrics in random or real/constructed graphs. The larger the ensemble the statistically better is the null model. The frequency of motif occurrence is compared in the real case to the null model by computing a Z score as explained in Chapter 1.

$$Z_i = \frac{N_{reali} - \langle N_{randi} \rangle}{std(N_{randi})} \quad (2.2)$$

where N_{reali} is the number of instances of motif i in the real network, $\langle N_{randi} \rangle$ is the ensemble average of motif occurrences and $std(N_{randi})$ is the standard deviation. If Z is very large, then the number of occurrences in the actual network is much higher than the average in the ensemble of random graphs and this means that the given motif is significant in occurrence (absence corresponds to highly negative Z -score). Usually, a significance is declared if a Z score stands out in the background of all Z scores. Most insignificant scores are < 0.1 as seen in Table 2.4. All the motifs found for the JetBlue 8/07 route network are statistically more frequently occurring in the real network than in random graphs with the same degree distribution.

⁵We do not perform analysis on missing motifs, but only on statistically less or more present than average.

Table 2.6: List of motifs with positive Z-score for the JetBlue 8/07 route network.

#	Motif	Z score	Count	Mean	Std
1		0.104	15377	1807	391
2		0.292	139030	3886	1384
3		0.0914	231429	11619	7202
4		0.142	100927	2623	2079
5		0.119	320438	12060	7761
6		0.182	503677	14372	8037
7		0.0557	43551	4929	2076
8		0.199	7489	106	111
9		0.831	1029151	6923	3684
10		0.105	1307345	86675	34813
11		0.0842	347213	23520	11513
12		0.0500	396423	55689	20399
13		0.0711	900099	71097	34945
14		0.117	102660	3981	2524
15		0.176	7548	148	126
16		0.0693	236556	20146	9359

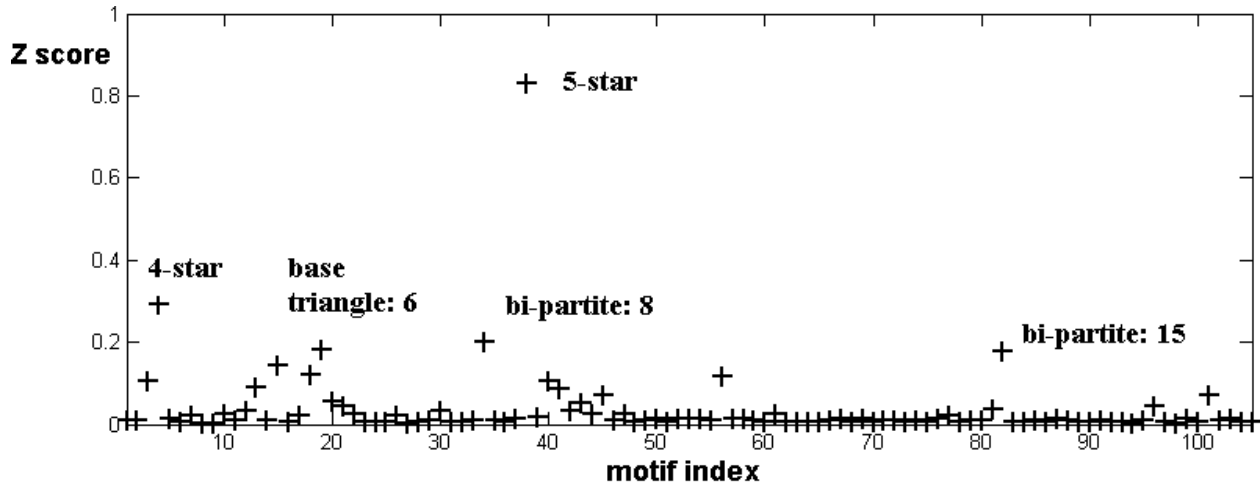


Figure 2-15: Motif significance profile for JetBlue 8/07. There are 16 motifs with Z score > 0.05 (shown in Table 2.4). Our motif finding algorithm identified 105 occurring motifs. This figure shows the statistical significance (Z-score - equation 2.2) of these motifs compared to a 100-random graph ensemble with the same degree distribution. The highest significance motifs are labeled, some with corresponding indices from Table 2.4.

Figure 2-15 and Table 2.4 show that the significant motifs fall into 2 classes: stars and bi-partite subgraphs. Other motifs are equally present in random graphs with the same degree sequence. Looking back at the route network of JetBlue 8/2007 on Figure 2-6, this is not a surprising result. The prevalent features of the network are the two stars (spoke formations) around JFK and BOS, and the bi-partite formation with the Florida equivalence class, FLL, MCO and PBI on one side and the Puerto Rico, and New York state destinations as the other equivalence class. The analysis in this chapter may not capture all the bi-partite subgraphs as motifs, because the motif size is capped at 6 nodes. However the pattern is obvious.

2.5 Coarse-graining the JetBlue 8/2007 network

The coarse-graining algorithm applied here was reviewed in Section 1.3.3. In this section, we discuss the process of coarse graining the JetBlue network and the lessons learned.

Since the star subgraph formations here are present with sizes much bigger than 6 nodes, we'll do a trick to coarse-grain the whole graph. At a rough approximation, if stars are considered a statistically significant subgraph (in an extended topology, they represent the same subgraph), they can all be "collapsed". This means that every star subgraph can be replaced with its hub node only. This is a special trick that only works with types of subgraphs that by virtue do not overlap with any other subgraph. Since their edges cannot also take part in other motifs, there is no optimization involved in selecting a "good" subset of motifs for coarse-graining. Figure 2-16 shows the resulting JetBlue network after collapsing all star motifs. Now the emerging patterns are even clearer. BOS and JFK are the equivalence class of a large bi-partite network, with BUF, PIT, SFO, ORD, SEA, SJC, DEN etc in the other equivalence class. This giant bi-partite graph is connected to the 6-node bi-partite graph of the Florida-Puerto Rico pattern. The two bi-partite subgraphs are connected with a few intermediate airports: EWR, TPA, RSW, PSE, SYR and BQN. There are a few shared nodes, such as IAD and LGB, with Long Beach probably starting off as a major hub like BOS and

JFK but becoming eventually more peripheral, as the network gets more weight on the East Coast.

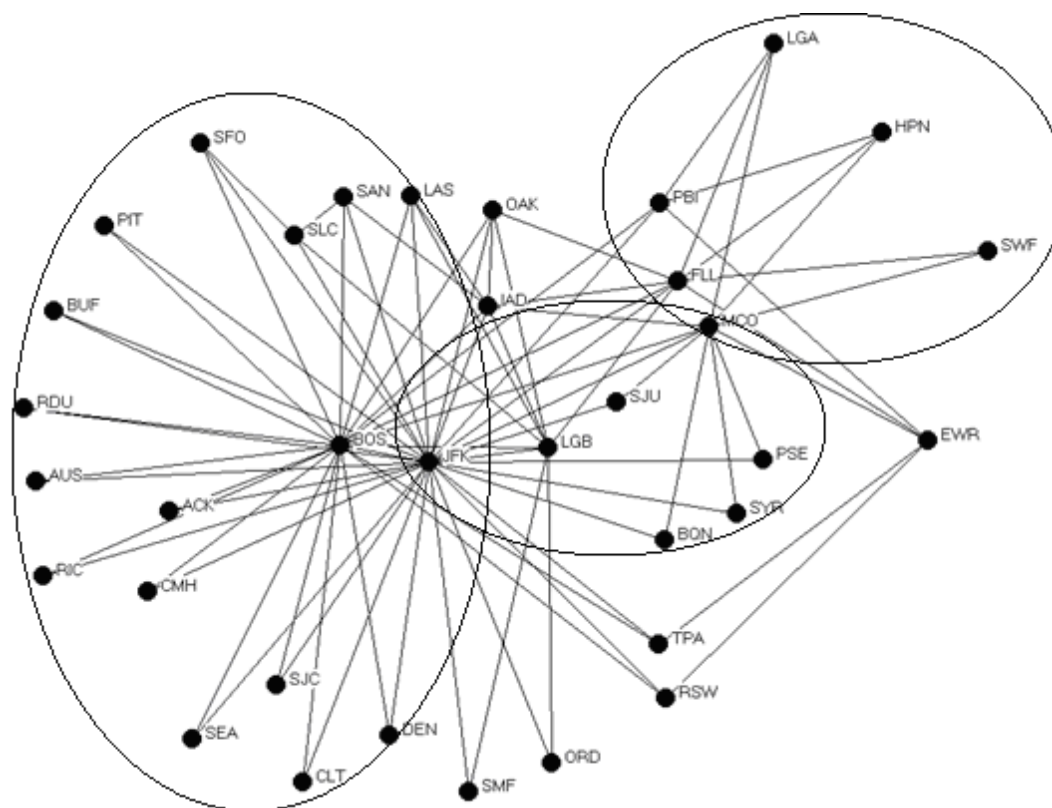


Figure 2-16: Collapsing all stars from the JetBlue 8/2007 topology emphasizes the bi-partite subgraphs. Major bi-partite subgraphs are circled.

The active motifs fed into the coarse-graining algorithm for the JetBlue example are the two classes found in Section 2.4, stars and bi-partite graphs. The solution contains a set of bi-partite and star graphs that cover part of the network. We already know what the solution should be, in view of Figure 2-16. The coarse-grained network should look like 2 or 3 large (6-20)-node bipartite graphs connected to each other in a circle with overlaid stars at the hubs. While this solution is easy to picture, it is hard to represent with a coarse-grained network and much harder to deduce just by looking at the result. For the sake of clarity, we show the motifs as found by the coarse graining algorithm without collapsing them (i.e. one level of coarse-graining only).

The coarse-graining of JetBlue 8/2007 was performed in two steps. Instead of searching solutions among all motif classes together, first all stars were collapsed (see Figure 2-16) and then the coarse-graining was done with bi-partite graphs only, in the star-free network. The "ideal" motif-finding and coarse-graining would include all motif classes together but also would be able to search motifs of any size, i.e. the 20-node stars around JFK and BOS would be easily identified. Actually, the two approaches give different solutions, depending on how the simulated annealing energy is computed. If number of CGU is maximized, the simulated annealing finds only stars and collapses them into few nodes, such as JFK and BOS. If motif size is an objective (i.e. more edges in the bi-partite subgraphs), then results are similar to what is shown in Figure 2-17.

By coarse-graining with all motifs, with motif size restricted at 5-6, we find that the solutions include many small motifs that should be combined into one. Collapsing all stars and then searching

for bi-partite graphs is a trick employed to help with this problem, but obviously may not be generalizable.

Figure 2-17 shows the coarse-grained JetBlue 8/2007 network. In summary, the JetBlue network is a connected core of bi-partite graphs with overlaid stars or spoke flights out of 3 or so hubs (a hub here is loosely defined as evidently having many connections, such as JFK, BOS).

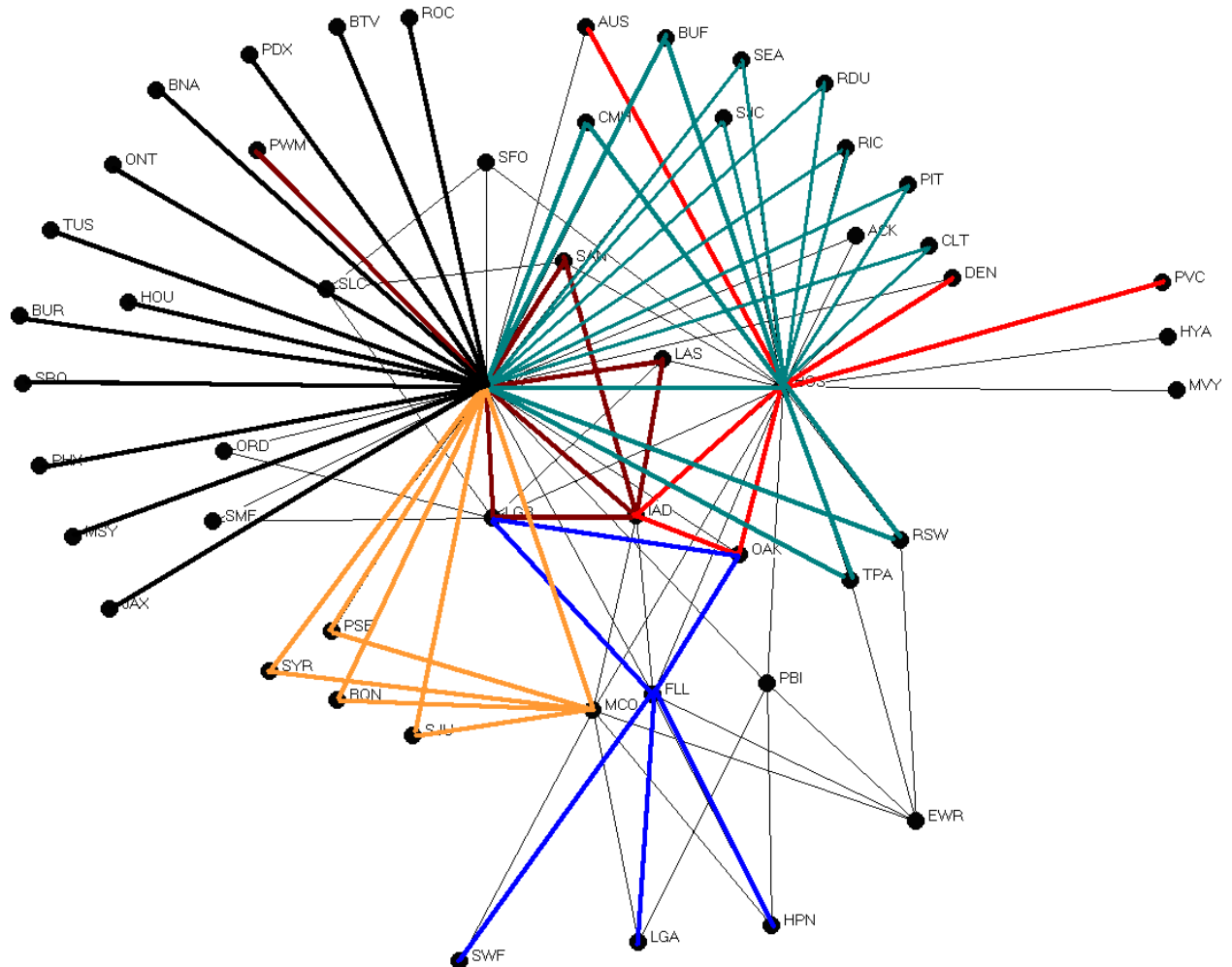


Figure 2-17: The coarse-grained JetBlue graph, showing motifs 4 (once), 6 (twice) and 8 (twice) from Table 2.4, and the star motif around JFK

Overall, the coarse-graining analysis complements well the motif search. In the case of JetBlue 8/07, the motifs are easy to visualize, as they connect to form the bigger network. The hub and spoke phenomenon is present with heavy hubs such as JFK and BOS, and lighter (peripheral) hubs, such as FLL, IAD and MCO. However there are other patterns as well, such as the dual servicing of airports, based at two major hubs, or what we called a bipartite motif. The example is the JFK-BOS based dual servicing of PSE, SYR, BQN and SJU (motif 8).

In cases where the network is not easy to visualize either due to size or complexity, it is possible that the coarse-graining will contribute more to uncovering the network topology and quantifying the stability of the backbone topology over time. Something interesting to notice is that motifs

found span the country geographically, and are not necessarily correlated with geographical distance. This is partly due to the fact that JetBlue has been flying regional jets only in the past two years and has had a A320-fleet from inception. If we separate the network by aircraft type we find that both the A320 and the Embraer 190 networks are based off of JFK and BOS and both feature the basic star and bi-partite graph motifs (see Figure 2-18). So they are built on top of each other, rather than composed separately. In fact, there are legs on which JetBlue flies both aircraft at the same time. Thus aircraft type is not an apparent factor affecting network topology.

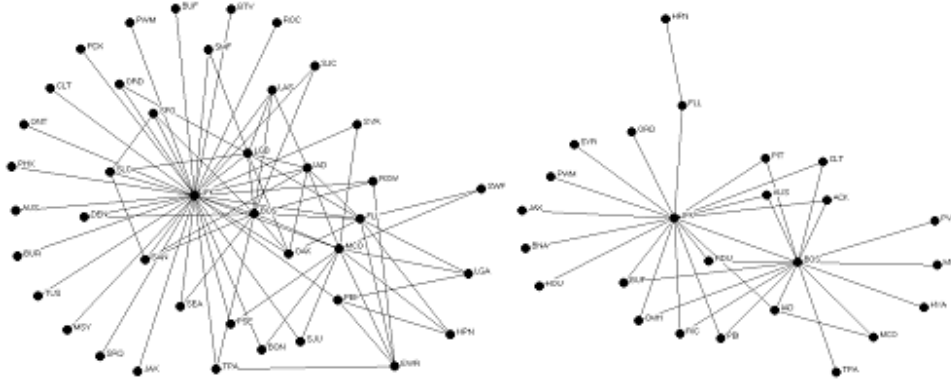


Figure 2-18: JetBlue 8/07 A320 network (left) and Embraer 190 network (right).

2.6 Conclusion and Discussion

In this chapter we discussed what network topology means and presented various ways to analyze it, using the JetBlue 8/07 route network as an example. Starting with statistical measures, comparison to canonical networks, and degree distributions, we showed that JetBlue 8/07 can be viewed as a hub and spoke network mathematically, from different angles. This was confirmed by its skewed power-law degree distribution, inelastic degree correlation, and topological similarity to a preferential attachment graph. In the later sections of the chapter, we looked into the topology in finer detail, by searching for significant motifs and how they connect to form the backbone of the route system. We found that the building blocks are more than hubs and spokes, but also bi-partite subgraphs, or dual service patterns from major hubs, which maybe be considered redundant in a pure hub-spoke (star) topology.

In summary, we showed that unraveling the structure of a complex network needs a multifaceted point of view. There are some measures that give a very good sense of the global picture, such as the s/s_{max} metric and the topology profile, but ultimately, the motifs finding, coupled with modularity and coarse-graining provide a more complete picture. What is missing from this discussion is how the uncovered structures evolved. In Chapter 4 we plan to use the techniques described here to analyze the topology of the airline dataset over time. Before that though, we look at the static picture in Chapter 3 - what can the measures used in this chapter tell us about the topology of major US airlines and slices of the industry in 2007.

Chapter 3

Airline Networks

This chapter presents the first case study - US airline networks. The chapter discusses what can be learned from graph-theoretic techniques about the topology of the airline routes. Trends in the industry are reviewed as well as previous literature on using network theory in airline research. We describe the data, and how to separate it into various sub-networks, by industry segment and by airline. Our analysis section includes general statistics, discussion of graph-theoretical versus industry metrics, then topological profiles of airlines, degree distributions and finally motif search in down-selected networks. Results show that all single airlines analyzed have common statistics and significant motifs, while the structure of Southwest Airlines is mathematically different.

3.1 Introduction

3.1.1 Trends in the airline industry

Studies of airline economics, route optimization or strategy, usually mention the Airline Deregulation Act of 1978, after which airlines could decide their own routes, pricing and seating capacity offered on destinations (Wojahn [18]).

Furthermore, studies (Wojahn [18], Hendricks [51]) claim that deregulation caused airlines to transition to a hub-spoke structure, which translates to concentrating flight path structure and seat capacity both in space (geographically) and in time (schedule-wise) at a few airports, called hubs. Deregulation also brought on more competition, so fares also fell and quality of service for the passenger greatly improved.

The “opposite” of the hub-spoke model is the point-to-point route structure, in which an airline flies from all its origins to all its destinations directly. The canonical versions of these models are a star and a complete graph. A point-to-point model is extreme and does not occur often, except in some special cases, ex: Go!, a Hawaiian airline operated an almost complete (except for one leg) graph route structure between the Hawaiian islands.

Though the hub-spoke model is claimed to be economically intuitive (Wojahn [18]) and mathematically sound (Hendricks [51]), hub-spoke airlines have not been largely profitable. And while the route structure is probably not the sole reason (competition is a key factor), it is interesting to note that the only profitable US airline (first quarter loss in 9/08) is Southwest, which operates a notably non-hub-spoke route system. Results in this chapter show that Southwest’s structure is not hub-spoke using various statistical measures. History shows that the airline industry operates in an unstable sinusoidal-like profit-loss regime, as seen in Figure 3-1. It is evident that the cyclical

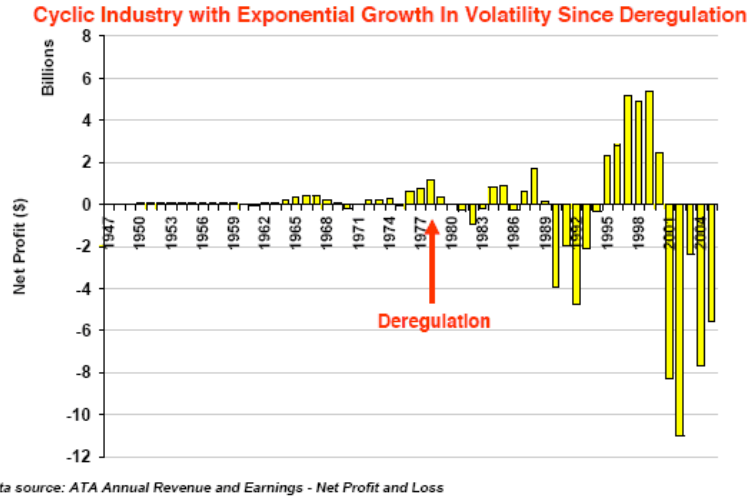


Figure 3-1: US Airline Net profit since 1947, yearly. Plot from Hansman [10]. Additional study by Sgouridis [11]

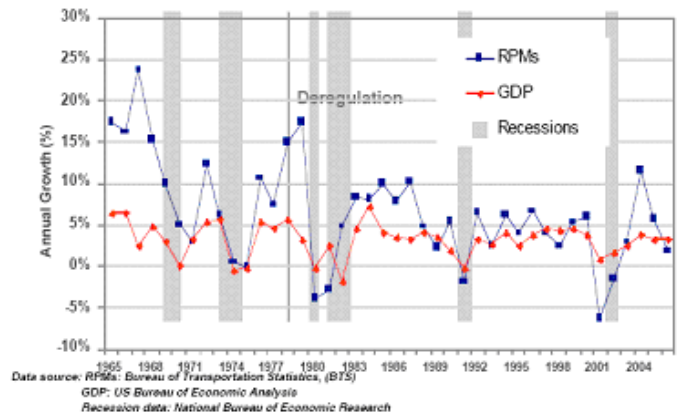


Figure 3-2: Passenger traffic and GDP in the US over time. Plot from Hansman [10].

nature of the industry was present before deregulation. Something else that history has proven is that passenger traffic is positively correlated with GDP (Figure 3-2).

Speculation about what causes the cycle centers on the lag between aircraft orders and supply. Airlines place orders for aircraft when they are profitable, but because of backlogs only receive the planes 2-4 years later, after their upturn, which creates overcapacity and reinforces a downturn. Also making predictions in terms of growing segments, both in the OD market and the passenger types, can be extremely hard. A few concepts such as the Boeing Sonic Cruiser did not see light. The toughest challenge by far nowadays has to do with congestion. The world is seeing a combination of increasing traffic, growing market-active population, especially in Asia, and also growing restrictions in terms of emissions and noise, environmental concerns, and finally increasing fuel cost. All of these require active remodeling of the industry from lighter, greener aircraft, to maybe novel route structures. Obviously, there are other operational and business concerns related to crew management, maintenance, gate operations, and labor union relations. There are attempts to develop a new system, or a new way to manage airline traffic over the United States - the Next Generation Air Transportation System [52]. Given the current "increasingly inefficient and

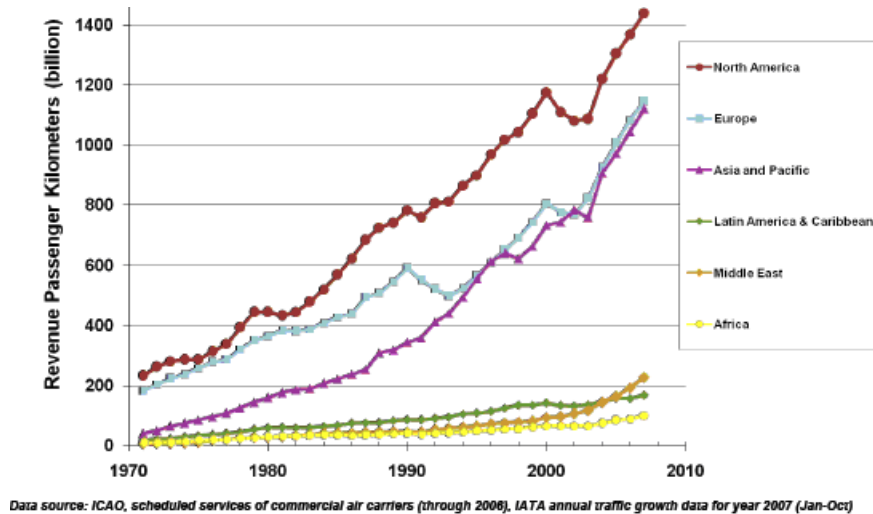


Figure 3-3: Increasing traffic world-wide 1970-2008 (Hansman [10]).

operationally obsolete” [53] approach to management, several agencies are working on NextGen, which is envisioned as a major redesign of the NAS management, with precision satellite navigation, digital networked communications, integrated aviation weather system and more.

As mentioned one of the reasons for a re-design is the increasing traffic (Figure 3-3), causing congestion and complexity in airport and runway operations.

Often problems in one airport propagate to the entire system and cause delays, stranded passengers and sky-rocketing costs for airlines holding aircraft on the ground or in the air accumulating delays. The JetBlue snowstorm havoc at JFK on February 18, 2007 propagated through the entire system, as crews could not get to their next rotation and resulted in the cancelation of hundreds of flights. Airlines have tried to deal with these issues by trying to match supply (seats, aircraft and scheduling) better to the market demand. This has resulted in smaller-size, more centralized networks, carrying more passengers in recent years. There has been a trend towards smaller aircraft, which keep delays, prices and also demand high. Authorities have tried to deal with congestion by imposing slot restrictions at airports to constrain the scheduling behavior of airlines by capping the total number of operations that can be performed at the airport. Finally, a system-wide effect of congestion and increasing traffic has been the emergence of secondary airports (Figure 3-4)[48][12].

The issues that the industry has to solve to operate in solvency are interesting to this research, because as traffic increases, the network grows, but it also restructures to deal with congestion and capacity constraints.

3.1.2 The hub-spoke phenomenon

After the deregulation act of 1978 it is assumed that the airline industry has transitioned from a point-to-point model to a hub-spoke model [18]. A pure point-to-point network corresponds to a complete graph, i.e. there is a direct flight between any two points in the network. A pure hub-spoke model is essentially a star, where every route has two hops, going through only one intermediate airport, the only one that is one connection away from everywhere else. Surprisingly both pure structures have occurred. There are point-to-point airline route structures in Hawaii for example, where distances between islands are small enough and there is enough demand that

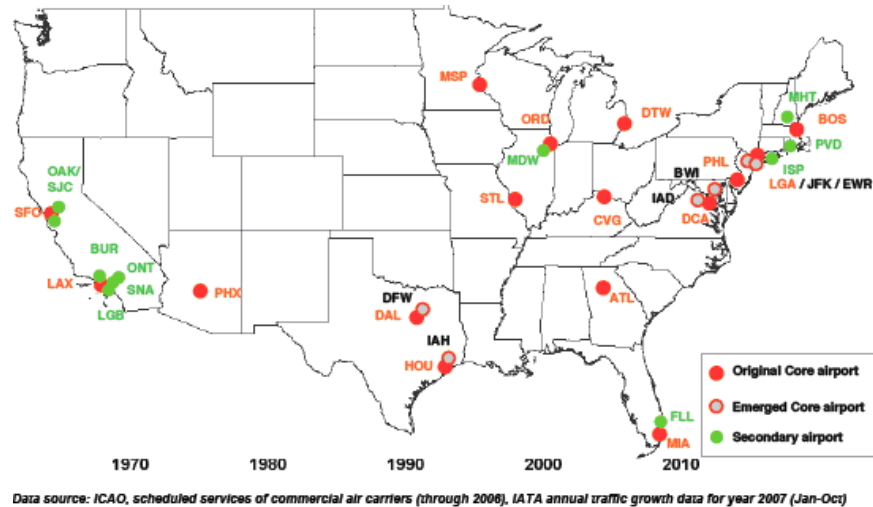


Figure 3-4: Emergence of secondary airports (Bonney [12]).

a point to point network makes sense, for a small number of destinations (order of 5 airports, ex: Mesa Airlines in 2006). Star networks are more common, and often mark the beginnings of an airline (ex: JetBlue).

The explanation for the shift to hub-spoke operations is savings due to economies of density and economies of scale. Economies of scale come from the fact that offering more seats per route (with higher capacity aircraft) drops the cost per seat, hence is cheaper to operate overall, given steady demand. If the passenger demand stays high because of the frequency of service, despite the inconvenience of flying non-direct - that is the reason for economies of density and makes the hub-spoke model worth it. For the airline, if flight frequencies are high, the cost of flight is lower, and the demand is still high. Economies of scale also figure in expansion. Presence in certain not so profitable markets can be beneficial because of the network effect of connecting these regions to more profitable OD markets where most of the revenue is made. Presence is often one of the key factors stimulating the market. Overall, hub and spoke configuration allows you to reach many more markets with the same number of seats.

With these considerations justifying the shift to hub-spoke structures, it increasingly important to understand why they are not solely enough to make airlines profitable, why the few airlines making profit (aka Southwest) are not hub-spoke, and how this model can perform under rising traffic and increasing congestion, as people fly more and more. Even though, recent fuel price rise indicated that the higher cost of travel may put a natural damping on this problem, the actual statistics show that air traffic is growing world-wide (Figure 3-3).

3.1.3 Literature on airline networks

There is a wealth of literature on airline networks, route structure and its evolution, and from all different angle of the problem. We review a set of papers that discuss network theory techniques, air transportation management ideas and econometrics.

“Pure network theory” studies

Guimera et al [17] study the global airport structure from the point of view of node centrality, and community structure. The authors consider the world-wide airport network, taking one year worth’s data (OAG) and represent it as an almost symmetrical graph (adjacency matrix). They find that the path lengths are small for the size of the graph, so it is a small-world. The degree distributions and betweenness distributions are power-law. Also, they plot degree versus betweenness and find that connected (by degree) airports are not necessarily the most central (by betweenness) airports. The prevalent example is Anchorage, which is a local hub in Alaska, but not a global hub. As an almost single port of entry to Alaska, ANC is the connection to all Alaskan airports, and as such, has high betweenness. The authors also use a simulated annealing routine to find network modules (subgraphs that are more connected within than to other modules), and discover that the network modules found coincide with geopolitical regions. This finding, as well as the distinction between highly connected and highly central nodes, inspires them to classify airports according to their “role” in the network - global hubs, connectors, regional hubs, peripheral airports and ultraperipheral airports. These ideas are the ground work for their Nature Physics paper in Jan 2007 [16].

In Guimera et al [16] the premise of the authors is that within-module properties of systems represented as networks are different from global properties. Using a simulated annealing algorithm, they identify modules in various networks in biology, air transportation and communications (the Internet). They define node roles based on nodal connectivity within the modules and across modules. For example, nodes that are well-connected across the board, are global hubs; nodes that are connected only well within their neighborhood are provincial hubs. Finally they look at the connectivity profile for a network between different nodes of roles classes (hubs to hubs, peripheral to ultraperipheral etc) estimated against a random background. The random background is usually an ensemble of random graphs created with the same global properties. So if certain connection types stand out, that is a signature of that particular network. The authors single out two classes of networks based on this connectivity profile, and based on their examples from air transportation, the Internet, metabolic and protein interaction networks. Air transportation and metabolism networks tend to form “hub oligarchies”, which means dense interconnectedness between large hubs (across and within modules) and lack or poor connectivity among hubs of lower rank, and among nodes of lower degree overall (these tend to connect to their closest local hub). The other type of networks look more like connected strings of stars, in which the hub-to-hub connectivity is lower, except for inter-module connectors. The former type of hub oligarchy has also been found in another study of Chinese airline networks [54].

Another “pure network theory” study is by [55]. The idea in this paper is the “hidden” inherent metric space behind a network. They assume that every network has a hidden connectivity or closeness, not apparent in the observable connectivity, which nodes are “aware of” when they route information or energy. The efficient routing of information is termed “navigability” and is measured in number of hops. The authors claim that these hidden metric spaces are an underlying link between function and structure of networks. They analyze navigability for a special class of networks they construct against degree distribution exponent and clustering. The degree distribution exponent is used to measure the number of large hubs in the network, whereas the clustering is tied to the hidden closeness of nodes. The special class of constructed networks is “small-world, scale-free, and with strong clustering” and is created using a modified preferential attachment routine, in which the preference depends both on nodal degree and on “hidden” distance. The authors claim that for airport networks the underlying metric space is Euclidean distance or geographic

closeness. This claim can be challenged, because it is not likely that small or large airports that are very close geographically will be connected. Flying in extremely small distances is not efficient, and is replaced by alternative transportation modes (car, train etc). On the other hand hubs do tend to connect to each other which justifies the other part of the idea. But again, when hubs are geographically very close, they will not be connected, ex: NYC airports. The final point of the paper is that routing happens in a zoom-out/zoom-in fashion as a route starting at small airport connects to progressively bigger hubs, and when in close proximity starts to zoom-in to smaller airports to reach another small destination. This is extended to Internet routing for example, without examination of the network, and without discussion of the BGP protocol.

Air transportation management studies

Lee and Kornfeld [56] examine the optimality of the hub-spoke network from the point of view of cost, demand and supply in terms of aircraft types. Using CPLEX they analyze a very small number of airports (7 cities) for a FEDEX cargo delivery example and find that the optimal network type varies from point-to-point to hub-spoke depending on aircraft types, distances between airports, and cargo demand. In some cases, a point-to-point network will be preferable. This result may not be scalable with increasing number of airports, as direct flights on the order of n^2 for large n are impractical.

Bonnefoy [12] studies scaling mechanisms by which the airline industry has met growing demand in the past and is expected to do so in the future. He shows that the National Airspace System is not scale-free from the point of view of network theory, due to capacity constraints at major airports. The system has evolved to grow via multi-airport systems in metropolitan areas and if those are modeled as aggregated nodes, the entire systems does become scale-free (by measure of degree distribution). He performs in-depth case studies of various multi-airport systems and studies how they develop to provide recommendations for future airport infrastructure management.

Wojahn [18] studies the airline industry as a whole, discussing carrier statistics around the world. He uses "measures of network structure" to describe transformations in the industry, and finds that more and more, airlines adopt a hub-spoke model. The author analyzes the hub-spoke model versus the point-to-point model using cost and profit equations as a function of influence of travel time, flight frequency. He finds that the hub-spoke model is optimal if the passengers' valuation on flight frequency is high and of travel time is low. The number of hubs is also considered - where the result is that if congestion and slot restrictions are in place, a multi-hub network is more profitable for he airline, compared to a single-hub network or a point-to-point network. Finally, Wojahn creates an asymmetric demand model, to reflect the fact that cities have different characteristics and different demand for travel. With the assumption in the model that spokes are connected to a single-hub and that hubs are fully connected, he finds that the cost-maximizing structure is a mixture of a point-to-point and single-hub networks.

Econometric models

Burghouwt's work [57] focuses on applying the Gini index to study airline configurations. The authors claim that the effect of deregulation will be different in Europe than the one observed in the United States and propose to use the Gini index to investigate how "centralized" an airline is, where n is the number of airports and y_i is the seats offered at airport i . They use the Network Concentration (NC) index which is the normalized Gini index divided by the maximum possible

Gini index ($NC = G/G_{max}$), which is $G_{max} = 1 - 2/n$, and corresponds to a single-hub network in which traffic is concentrated along one hub-spoke route only. The authors use the NC to analyze European airlines in the period 1990-1999, with discussion on European airspace regulation, and airline division into national carriers, regional airlines and low-cost airlines and the differences between them. Given the results, they separate airlines into four loose categories depending on whether they grow or decline in size and whether they increase their concentration or not. The categories are concentrated network builders (increasing concentration and size), deconcentrated network builders (decreasing concentration and increasing size), concentrated network rationalizers and deconcentrated network rationalizers (both size and concentration decreasing). The classification shows great diversity among airlines, and that low-cost carriers tend to decentralize more over time, because with lower fixed cost of a single flight, they can afford to fly more direct routes.

Real-world engineering systems are inevitably tied to economics because they operate under cost and schedule constraints and are built to deliver some function. Bhadra et al [58] present an example of an economic model of network growth in attempt to analyze the evolving US air traffic. They concentrate on modeling original-destination pairs (OD pairs) of airports and the demand for travel along those segments based on 10% of itineraries released data. They find that passengers, weighted average fare, average distance and types of air carriers empirically determine the itinerary choices. Notice that this is a different angle on growth - from the demand side, rather than from the airlines supply side.

The dependent factors listed above are chosen initially with the assumption that a binary choice of one itinerary over another will be linearly dependent on each factor (or a vector of such variables) - $Z_i = \alpha + \beta X_i$. A probability distribution function is described as: $P_i = F(\alpha + \beta X_i) = P(Z_i)$. They use a logistic distribution probability composed of these parameters to describe the probability of binary choice. They use the data to estimate parameters in a regression by maximizing the likelihood of certain itinerary choices and find that all initial factors considered have an effect in itinerary choices.

3.1.4 Dynamics drivers in the airline industry

As for any real system operating in an uncertain environment, it is hard to enumerate all the factors influencing the airline industry. Following is a list of the major ones, in view of the analysis performed later in the following sections.

Size of origin-destination market (number of stations/airports): Size is a critical factor in complexity of operations for an airline or any networked system. While topology has to be assessed independent of size, due to the physical limitations imposed by size, in the airline industry, the topology is influenced by size, as constraints or economies of scale take effect. Size in the airline industry can be measured in terms of OD pairs served, total number of seats offered, total number of passengers carried or total number of departures.

Historical background of the carrier and its network (strategic airline management): History, strategy, character and company philosophy are part of any business. For the airlines the company character not only plays in with the operations, but also the labor relations and passenger market attraction pool. For example, entertainment and leg room are JetBlue's selling points, and go along with the marketing line of "jetting" versus "flying".

Intercontinental/continental orientation: Not all US carriers fly internationally. In the

US international markets are carefully sliced. Continental has more of South America, United has the Japanese market, American more presence in Europe. We will not analyze route structure outside of the US, so this is outside of the scope of this thesis.

Fleet composition: The fleet plays a huge role in operations, stage lengths, expansion strategy, maintenance scheduling, pilot and crew training. Low-cost carriers tend to carry few types of aircraft to simplify their operations (ex: Southwest still flies only 737s, which are great for its overall mid-range flight legs). Choosing aircraft size can be traded with increasing flight frequency, which has an effect on networks. For example, using smaller aircraft more often on a popular route makes the load factor appear higher given the same demand. This is compared to using bigger aircraft less frequently (fewer flights per day). Figure 3-5 shows the trends of decreasing aircraft size since 1990. Also, consistent use of similar aircraft can decrease training costs and improve pilot flexibility. These effects in turn may limit the impact of disruptions due to congestion and weather delays.

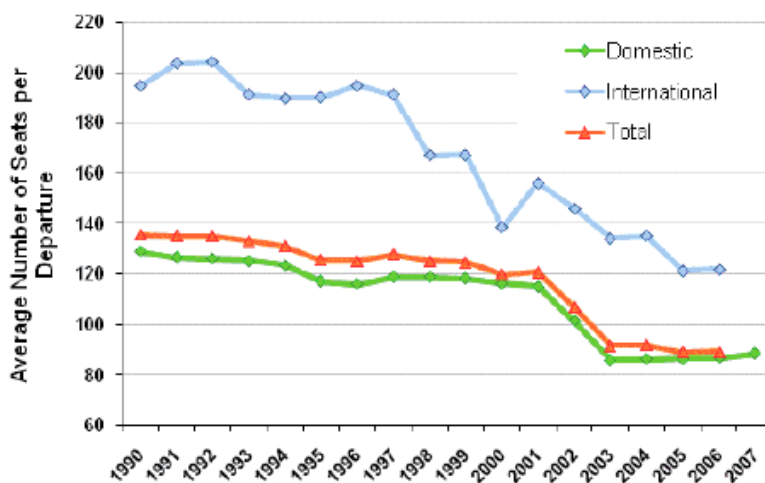


Figure 3-5: Trends in aircraft size, average number of seats per departure, 1990-2007. Figure from Hansman [10].

Hub capacity: Hubs are strategic in how an airline positions itself across the country. In the US, with the coast-to-coast travel demand, the ideal airline hub is a city like Chicago, which is closer to the middle, has its own high demand travel market and can be used to stage flights in waves. The wave model means that hub are fed from one coast at a certain hour, so within 1 hour, or 1.5 hours, the connections for the other coast can leave and arrive by evening/afternoon.

Average stage length: Stage length plays a role in network structure because it is tied with aircraft type (range) and can be dependent on a completely different market segment. Range is only a factor when it comes to regional jets and turboprops, because in the US the distances are small enough that a narrow-body jet will fly the same stage length as a wide-body jet. That said, it is unlikely that a 777 will be used for a BOS-JFK flight, and the demand economics will still be different. The JFK-LAX passenger pool is very different from the pool on a local flight in Florida or a flight from Salt Lake City to Seattle, for example. Airlines are well aware of this, which is why the route structure looks very different at different stage lengths (for example, all flights under 500 mi versus all flights above 1000 mi). This is also related to fleet composition (See Figure 3-6).

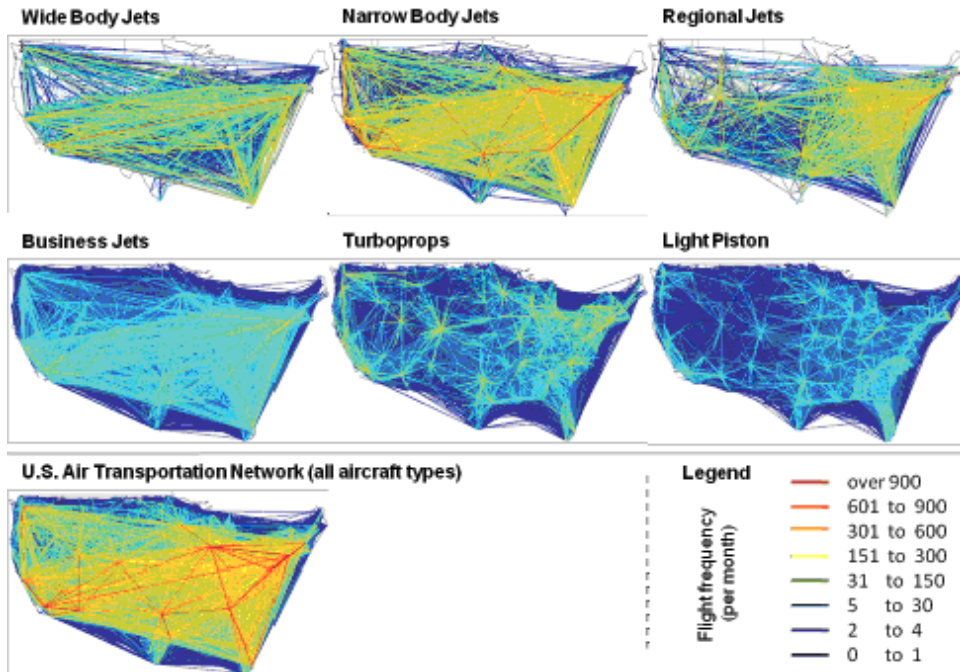


Figure 3-6: The airline networks by aircraft type from Bonnefoy [12].

3.2 Airline Data

3.2.1 Data source and slicing

The airline dataset in this thesis is publicly available by the US Bureau of Transportation Statistics [1], which collects statistical data for all transportation systems in the United States. US airlines are required to report monthly flight data, including origin-destination, seats offered, number of passengers, cargo in pounds, minutes in the air and leg distance in miles. They also report costs, broken into operational, labor and capital. A tenth of passenger fare/itinerary is also reported, which gives sense of market demand. The BTS statistics site contains various databases starting around 1990. A sample of the data is included in the Appendix, in Table A.3. In this thesis, we focus on the supply side, and analyze what the airlines put out in the market, in terms of flight frequency, OD (origin-destination) information, and seat capacity.

Given the different drivers in the airline industry and the richness of this dataset, understanding the topology of the entire industry, as opposed to of an individual airline, requires slicing the data carefully. As discussed earlier, major drivers in the industry are the size of the OD market, the historical background of the carrier, the size of its current operations, its fleet composition, inter-continental/continental orientation, the hub capacity which constrains its operations, the average stage length and of course, the strategic management of the airline. All of these factors account for different dynamics on different scales. And these would not necessarily be obvious in an aggregate analysis.

We address these questions by splitting the dataset and analyzing it separately, in 4 slices. We use the sliced networks as separate networks in our analysis. In particular, we will analyze some individual airlines, as well as aggregate networks, such as the wide-body jet network, and the airline

network of top 50% most frequent flights, by departures and by seat capacity.

Here we present four ways to split the industry by drivers, aircraft types, stage length, top percentage of flights in departure frequency and in seat capacity.

i **Slice by aircraft type.** In these networks the routes appear as edges only if they are flown by a certain type of aircraft, nodes are airports. For example, the right-most plot in Figure 3-7 shows only wide-body jet routes in the continental US. Other types plotted are regional jets and turbo-prop (left) and narrow body jets (middle).

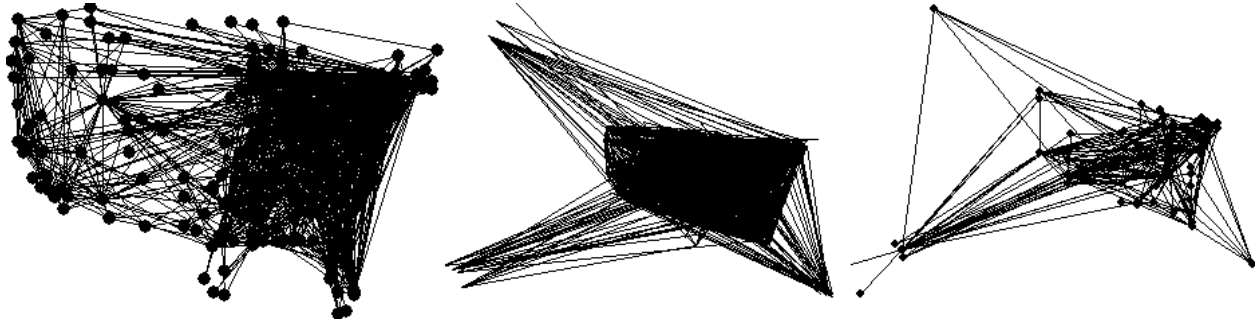


Figure 3-7: US airline data slices by aircraft type: regional jets and turboprops (left), narrow-body aircraft (middle) and wide-body aircraft (right).

ii **Split by distance flown.** The left plot in Figure 3-8 shows only flights under 500 miles, and right, flights longer than 1000 miles, nodes are airports. The left-most plot shows two disconnected components, because Alaska appears split from the continental US.

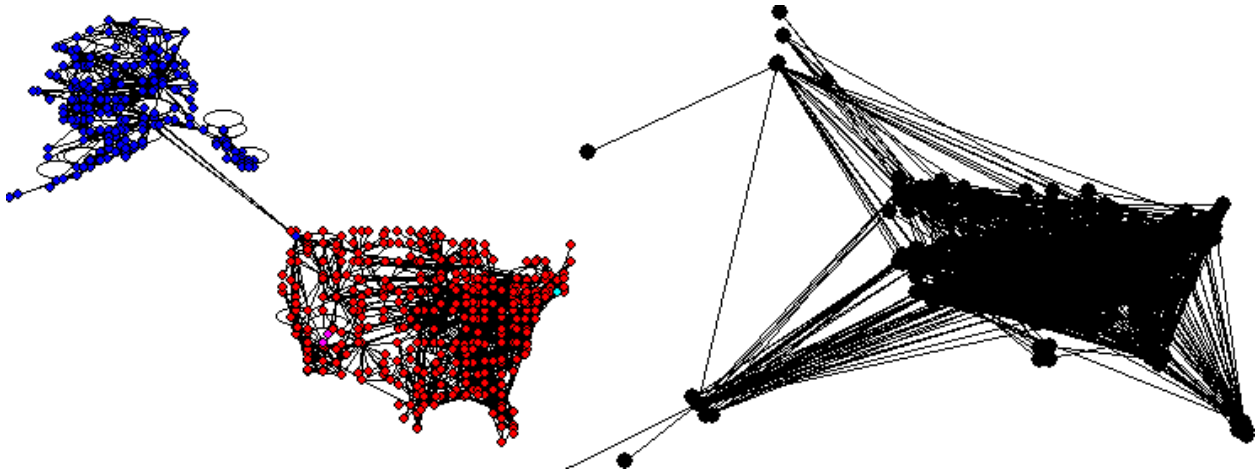


Figure 3-8: US airlines dataset sliced by distance: (left) all flights under 500 miles, and (right) all flights above 1000 miles.

iii **Split by flight frequency** in terms of number of departures. A leg/edge appears if its frequency is within $x\%$ of the most frequent leg, where x is 80%, 60%, 40% and 20%. So the right-bottom plot in Figure 3-9 shows only the top 20% of flights in the US, by number of departures.

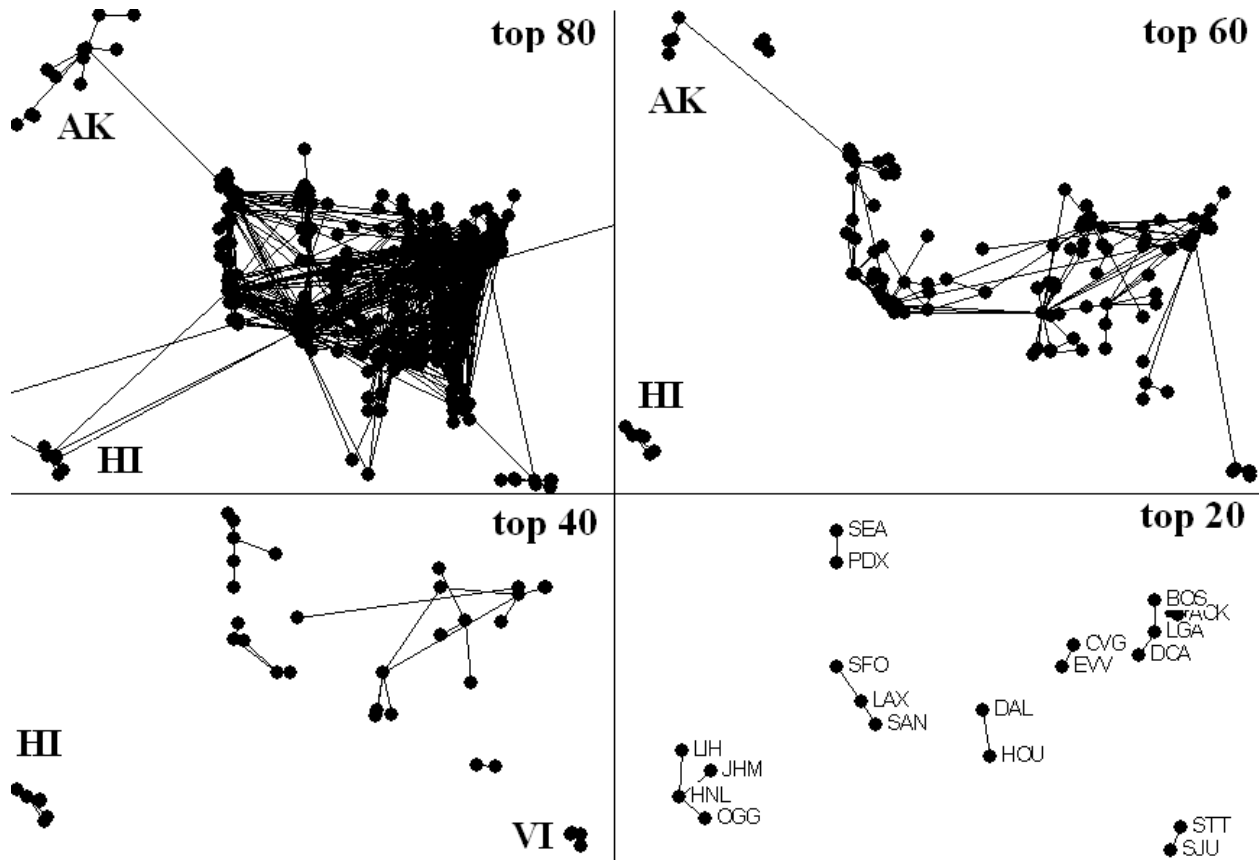


Figure 3-9: Slicing data (all US airlines) by departure frequency - top x flights for the entire period 1/1990-8/2007: top 80, top 60, top 40 and top 20 most frequent flights.

iv **Split by seat capacity.** This is the same as above, but the number of departures is replaced by seat capacity.

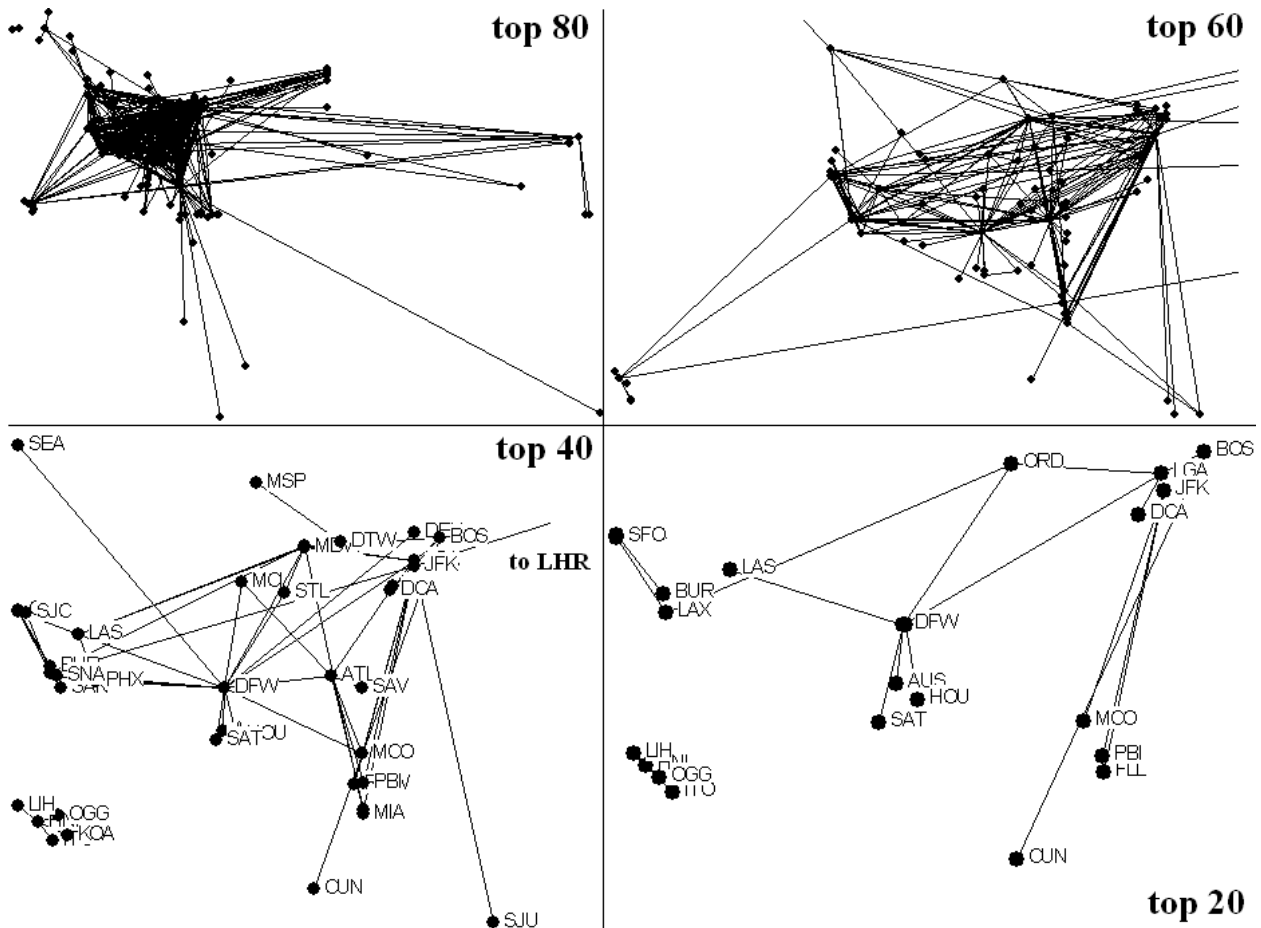


Figure 3-10: Slices by seat capacity offered (all US airlines): top 80, top 60, top 40 and top 20. A flight leg is in top $x\%$ if it has at least $x/100$ fraction of the seat capacity of the busiest flight in the network.

Aircraft type classification, referenced from US BTS data [1], is presented in Table 3.2.1. Using this classification, all jets are classes 6,7 and 8, for example. Types 0, 1, 2, and 3 are not considered at all in this analysis.

Table 3.1: Aircraft Types according to BTS [1] classification.

0:	Piston, 1-Engine/Combined Single Engine (Piston/Turbine)
1:	Piston, 2-Engine
2:	Piston, 3-Engine/4-Engine
3:	Helicopter/Stol
4:	Turbo-Prop, 1-Engine/2-Engine
5:	Turbo-Prop, 4-Engine
6:	Jet, 2-Engine
7:	Jet, 3-Engine
8:	Jet, 4-Engine/6-Engine

3.2.2 Down-selection of the data slices

The entire airline dataset is difficult to analyze not only because of its size (817 airports in 8/2007 and 5703 legs served) but also because of complexity in terms of factors driving its dynamics. The slicing in the previous section separates the data into segments by "industry driver", such as fleet composition and market demand for seats on certain destination pairs. Of these slices we pick the top 50 flight frequency and top 50 seat capacity routes, composed over time. This means that a flight leg is in the top 50 if its share is within 50% of the month's maximum during at least one month between 1990 and 2007. We also pick the wide body jet network, i.e. routes flown only by wide-body jets - because these present a separate "long-distance" passenger slice of the system. These three slices are chosen for in-depth analysis such as motif search and topology evolution. However, we will still draw examples from the entire industry or other slices if they are useful in our analysis.

All data slices that appear in tables and plots in Section 3.3.1 are shown in Table A.5.

Of the airlines, we will use the analysis of JetBlue Airways, done as example in Chapter 2, Southwest Airlines, as an outlier in the industry and Continental Airlines, as legacy carrier representative. Continental was chosen among the large carriers because its bankruptcy happened early in the dataset history (early 90s) and it has had steady network operations over the years compared to the other legacy carriers. The other major carriers (American, Delta, United, Northwest, US Airways, Alaska, America West) and some low-cost carriers (ATA, Spirit, Frontier, USA3000) appear in our overall discussion of statistics in the industry, in Section 3.3.1.

For the purposes of statistical analysis, all data slices and single airlines are shown, while the down-selected group is used for in-depth analysis, i.e. degree distribution plots and motif search.

3.3 Topology Measures

In this section, we begin with statistical measures for topology and discuss size/traffic industry measures such as number of departures and passengers carried, then industry-related hub measures and finally the graph theoretical measures such as diameter and degree correlation. These are calculated and discussed across the entire airline industry. We then move on to more in-depth analysis of the slices and airlines selected in the previous section. These networks (JetBlue, Southwest, Continental, the wide body jets, the top 50 seats capacity and departure frequency flights) are analyzed in terms of their topology profile, degree distributions and significant motifs.

3.3.1 Statistical indicators for topology

Traffic-related, industry-related size metrics

Number of departing flights

Figure 3-11 shows total number of departures against number of airports for a slice of the airline dataset. Individual airlines are in the down left corner, flying to less than 200 destinations and with less than 10000 departures. A linear mid-trend appears, with linear growth in departures as airport destinations grow. There are two outlier groups: the top 20, 40 and 50 routes which exclusively serve very few airports, but with extremely high frequency, hence they are in the upper left corner; and small aircraft networks, with under 100 seats, these flights occur between many small to many

other small airports across the country, but they don't carry many passengers and are infrequent, hence they are in the lower-center and lower-right part of the plot. These slices of the airline network are likely to have longer path lengths, and positive or zero degree correlation, because they are more spread out both geographically and in terms of nodal degree. An interesting point are the regional jets, low-center, right on the red line, with just above 200 airports and fewer than expected departures. Also, the entire airline industry (ALL 1/90,8/07) has doubled in destinations and grown almost 50% in departures.

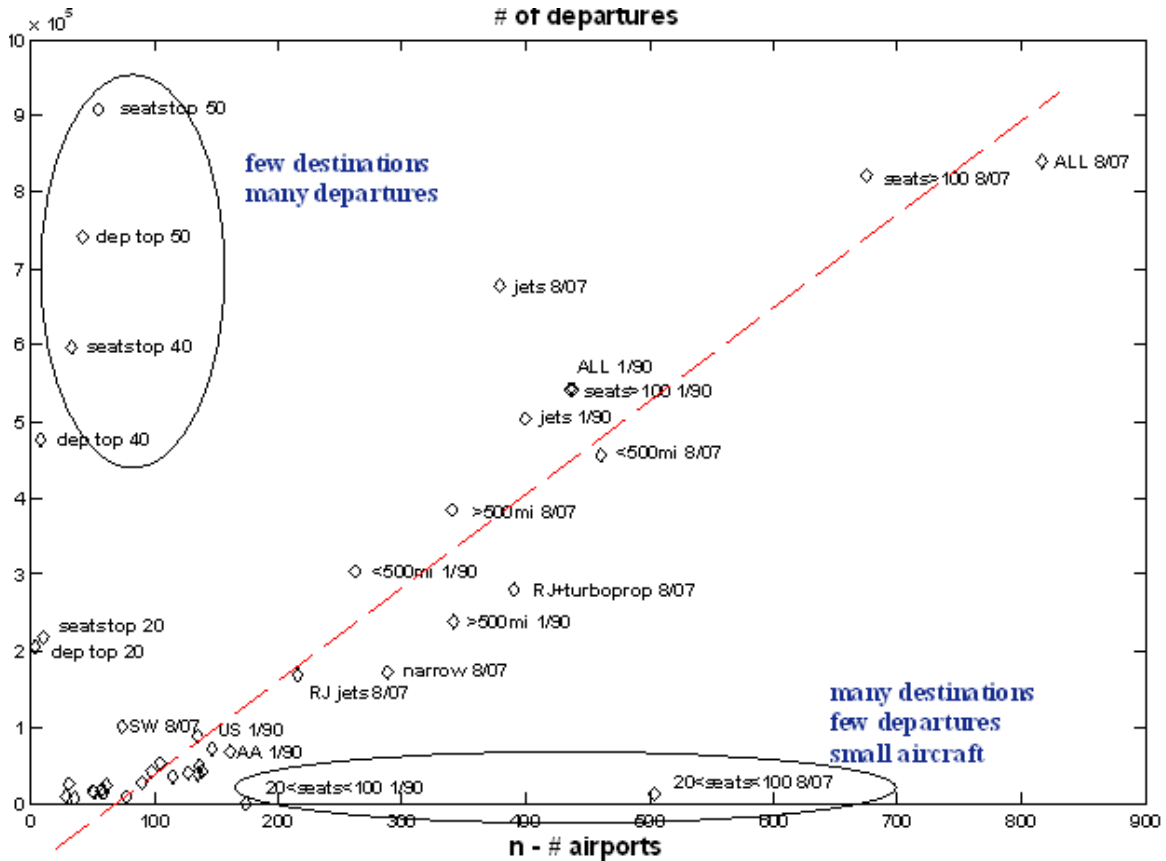


Figure 3-11: Total number of departures plotted against number of airports for all data slices of the entire airline dataset. Statistics are for January 1990 and August 2007 for almost all slices. Full nomenclature of data slices is given in Table A.5.

If we zoom into the individual airlines from the previous plot, we have Figure 3-14. Three groups emerge among the airlines - the "low-cost" carriers, Alaska, America West, Southwest 1/90 and JetBlue (and US Airways 8/07 severely downsized), the legacy carriers in 1990, and the legacy carriers in 2007 (United, Delta, American, Northwest and Continental). Southwest 8/07 is the outlier - with more departures than any other airline, and flying to a mid-range number of destinations. This means a higher-density network than any of the others. If Southwest packs more departures into its schedule monthly, that means that there are many more flights daily. A denser network with a higher number of OD pairs is one explanation - confirmed by the results in Figure 3-16 which plots edge to node ratio for all airlines. But another factor driving the departures is performing many departures of a single aircraft on many short-haul (1.5-2 hours) trip per day. The Southwest timetable gives some evidence of that. The frequent short haul flights indicate strong

local connectivity - a fact that we will use to suggest a model for Southwest's expansion since 1990.

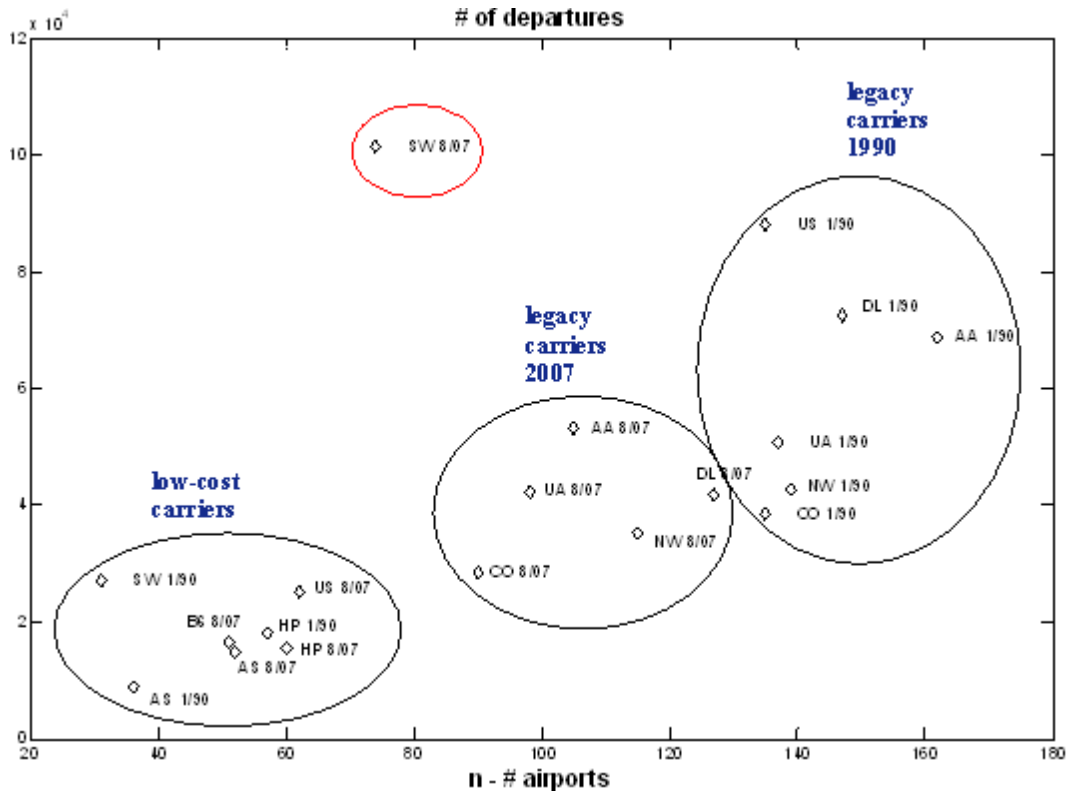


Figure 3-12: Total number of departures versus number of airports, for individual airlines only. Three groups of data points emerge: low-cost carriers, legacy carriers in 1990, in 2007, and Southwest Airlines is an outlier.

Total number of seats and seat-miles offered

Total number of seats offered is related to the number of departures, but not linearly since a departure can be performed by many different types of aircraft. In the case of Southwest for example, this is the same measure, just scaled, since the airline only flies 737s. Figure 3-15 shows the total number of seats and largely resembles the departures plot. It is interesting that the top 40 and top 50 flights by seat capacity (for all 1990-2007) by far outweigh the entire industry in 8/07. Also, wide body jets are barely present in this plot, and regional jets are dwarfed by narrow-body jet flights. Wide-body jets in US airline fleets are mostly used for international flights, and some coast-to-coast flights (757, 777). Narrow-body jets dominate the airspace, especially with airlines such as Southwest flying only 737s and dominating the industry by capacity.

Zooming on the airlines seats offered only, the same picture as with the departures emerges. Southwest Airlines is the major supplier in the industry and the fastest growing airline. Alaska Airlines also sees growth, but on much smaller scale. US Airways downsizes the most of all, and a month after the 8/07 data point it actually merges with America West. These events are not considered as they occur after the dates analyzed. The seat-mile picture is the same as the total seats and total departures, except that American Airlines outflies Southwest in mileage (Figure 3-13).

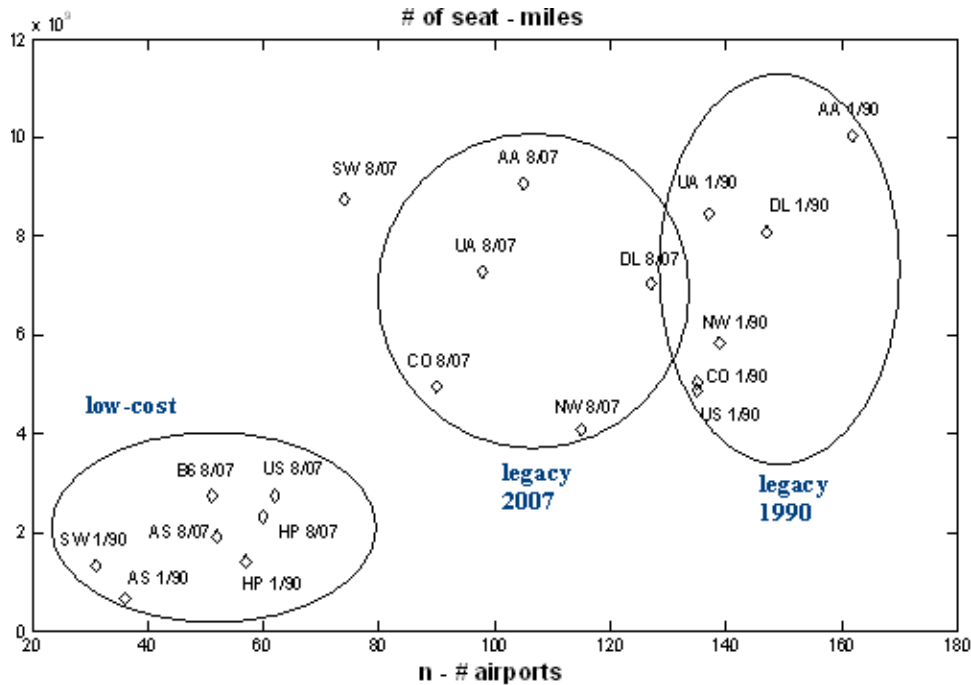


Figure 3-13: Total seat-miles for single airlines only.

The story of departures, seats and seat-miles is almost the same - for all airlines, except for Southwest and Alaska, the number of departures per month, number of seats offered and seat-miles per month has decreased (in some cases drastically). Only Southwest and Alaska have grown from 1/90 to 8/07. Alaska (as well as America West) is somewhat special and while in the top 10, they differ from the top 6 network carriers. Alaska is thought to have lower costs than most network carriers and has a niche advantage with its strong position in the Pacific Northwest and West Coast market.

Regarding size from an industry perspective then, the industry has definitely downsized, for the sake of efficiency, profitability or dealing with bankruptcy.

Passenger enplanements

Passenger enplanements are the total number of passengers boarded a given airline/flight or fleet for a given time period. In this case, these are the total number of passengers flown for the data slices we analyze, for 1/1990 and 8/2007.

According to the numbers on Figure 3-14, most passengers travel on jets and fly more than 500 miles. Despite the downsizing seen in large carriers, even short haul (< 500 miles) flight passengers in 2007 are more than the long haul passengers in 1990. More people are flying, and the airlines seem to be getting more efficient at carrying the higher load factor. The zoomed plot of airlines only is shown in Figure 3-15. The same patterns as with the number of departures and seats offered are seen, except that the legacy carriers in 2007 are carrying more passengers, even though they are flying less.

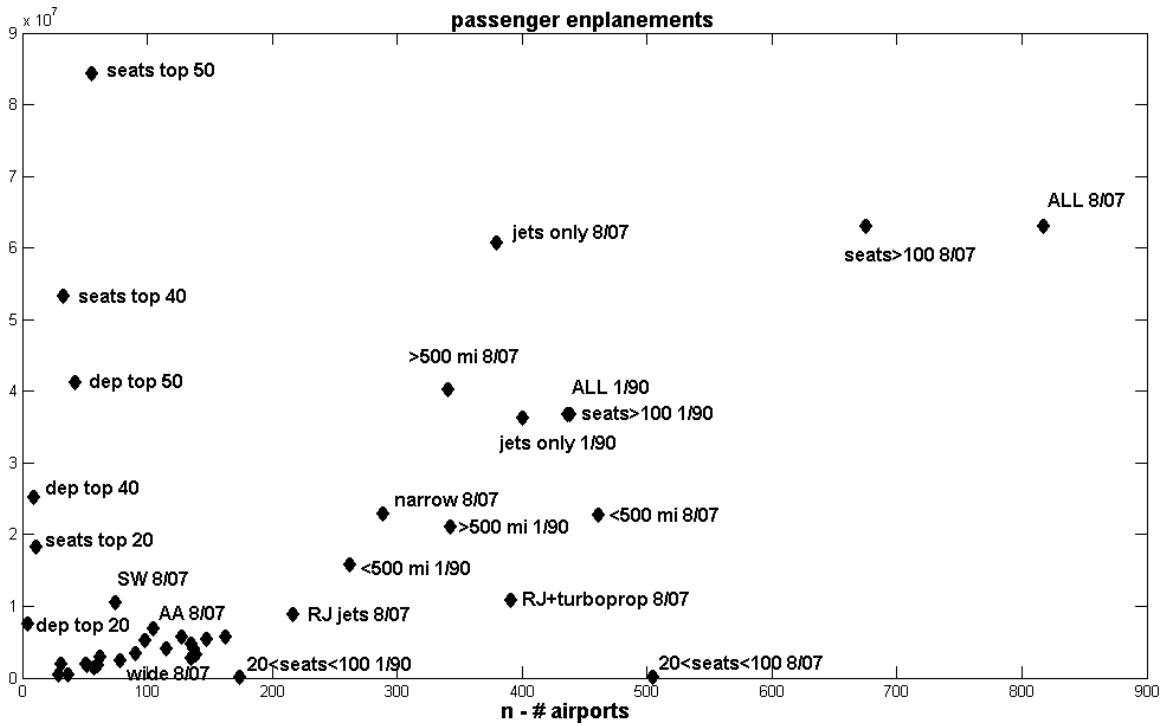


Figure 3-14: Total number of passengers carried versus number of airports. 1/1990 and 8/2007, all data slices.

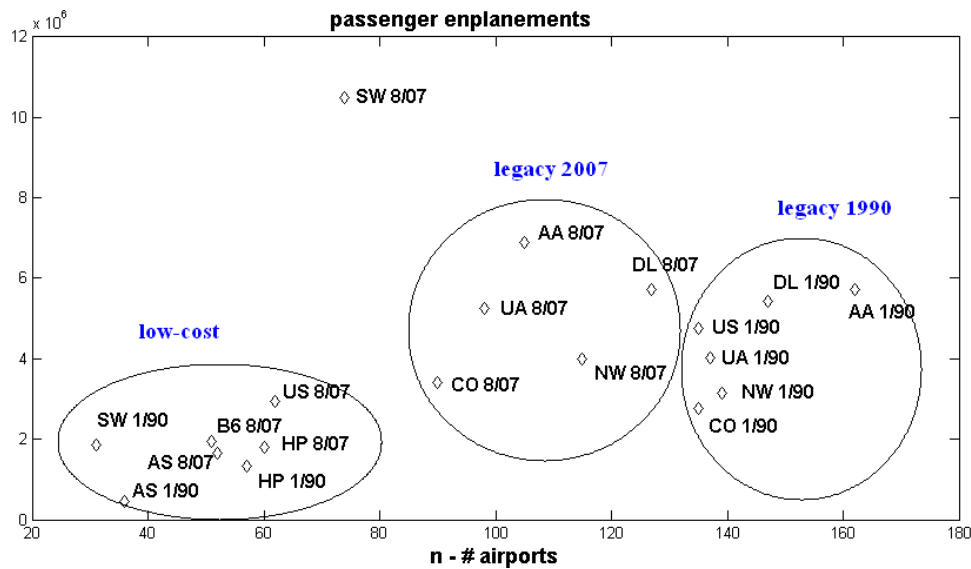


Figure 3-15: Passengers carried on US airlines - in 1990 and in 2007.

Graph-theoretic statistical metrics

Size and density

The metrics discussed here are number of nodes (airports), number of edges, edge to node ratio, average path length, diameter, and s-max metric.

There is almost a linear-like relationship between number of nodes and number of edges (see Figure 3-16). A node here is an airport, and an edge connects a pair of airports, if there is a flight between them. Figure 3-16 shows that there is linear-like growth in density in airline networks. For the entire dataset, we measured an edge to node ratio of 6.6 while for the single airlines only the edge to node ratio is 3.8 (zoomed plot on Figure 3-17). If we look at the airlines only, we see the same separation in low-cost, and legacy airlines in 1990 and 2007, with Southwest Airlines, as the inevitable outlier.

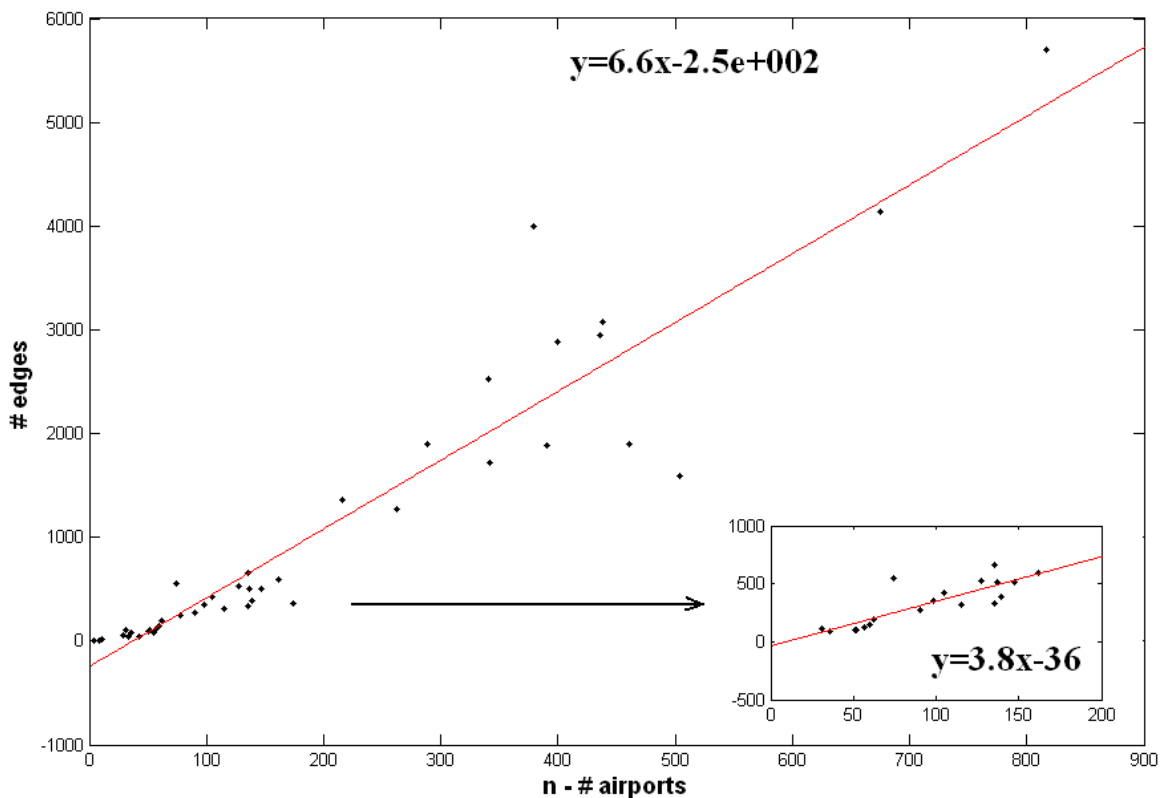


Figure 3-16: Number of edges (OD pairs served) versus number of nodes for all data slices. Text on figure is omitted for clarity. Zoomed plot shows single airline data points only.

Southwest is most dense network by far, which is the first hint that Southwest is different in topology. Even though those are not marked on Figure 3-16, the jet and long haul slices are most dense among the industry. Most airlines have $m/n < 5$ which is quite sparse.

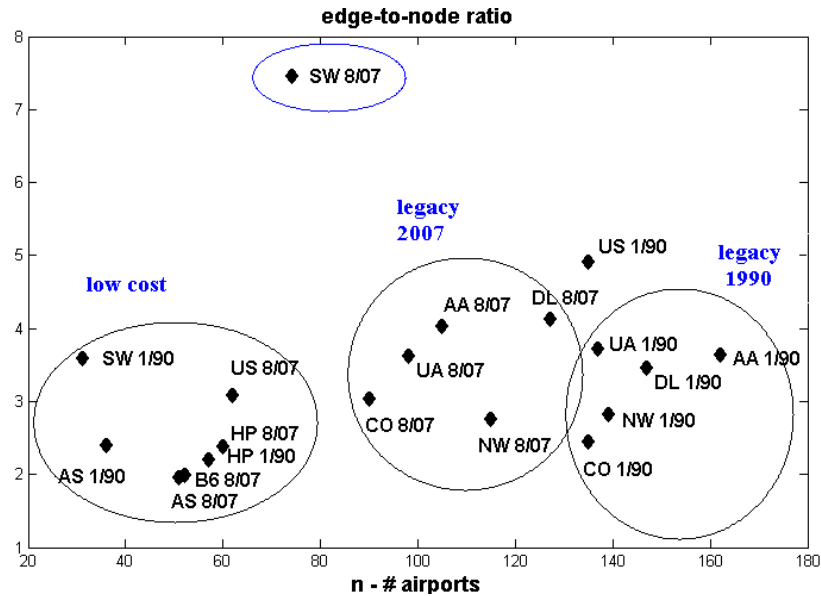


Figure 3-17: Edge to node ratio for all major airlines including JetBlue and Southwest as low-cost representatives.

S-max metric

As a reminder, the s-max metric is the s-metric of the graph, divided by the s-metric of the corresponding s-max graph. The s-metric is the sum of products of degrees across all edges of the graph (see equation 1.4). Plotting the s-max metric (degree of "scale-freeness") for the airlines only shows the same pattern of low-cost, and 1990/2007 legacy separation, as seen before. It is interesting that the legacy airlines, with downsizing are carrying more passengers and becoming more scale-free. That means that more hubs are connected to hubs, rather than solely to spokes. Also, US Airways and America West in 8/07 lead in the s-max scale, right before merging officially a month later.

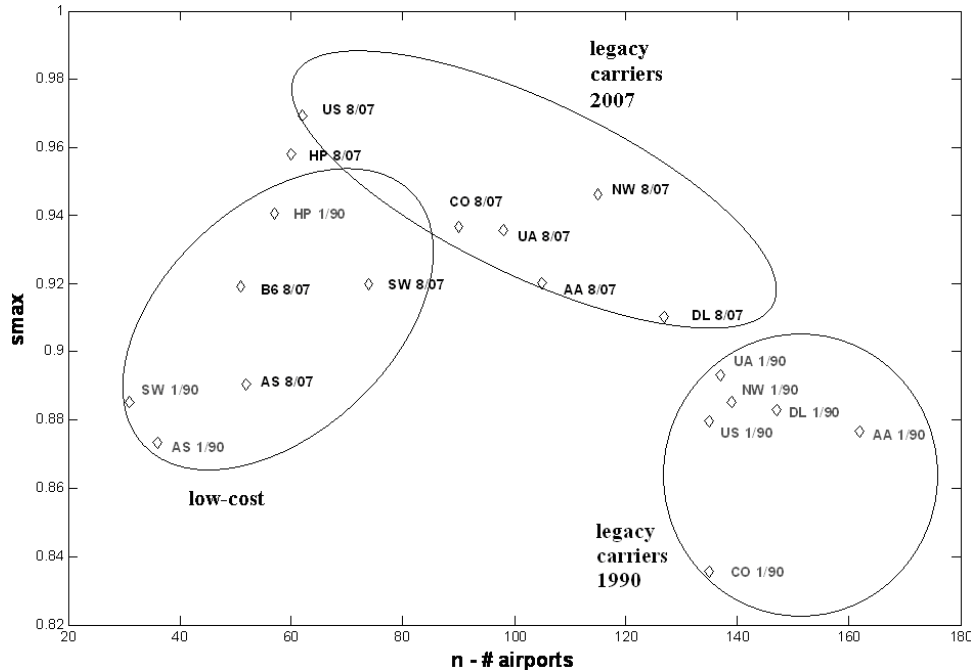


Figure 3-18: S-max for single airlines only.

Also, surprisingly, Southwest does not stand out here. It is not the most decentralized, but hanging mid-way. Alaska, as well as most airlines in 1/90 tend to have less scale-free networks. In terms of the scale of "scale-freeness", the s-metric tends to vary between 0.5,0.6-0.9 overall. In this range, anything above 0.8, and especially 0.9 can be considered "close" to its s-max configuration. A purely random graph (ER, $p=0.5$) will have s/s_{max} of about 0.5, so for airlines, which are very centralized and sparse networks, 0.8 should be average, and above that - scale-free. All the legacy carriers in 2007 with s-max above 0.9 can be considered "scale-free".

Diameter and average path length

Since the diameter is a very discrete measure, and tends to be small across all airlines and data slices, we also plotted the average path length, which is the average shortest path (in number of hops) across all shortest paths between all destinations in the network¹. The average shortest path, called the average path length, is a more continuous measure of the span of the network. Figure 3-23 shows the average path length plotted versus network size. Most data points lie on a log-like curve except for short haul and small aircraft slices. Because of their short physical range, these networks are bound to have long path lengths and diameters. Above the curve are also top 20, 40 and 50 data slices. These are extremely small in number of airports, and very tree-like as graphs. This is what explains their relative longer path lengths. The $\log(n)$ curve is not too surprising as it has been shown that preferential attachment graphs (which are sparse) have a diameter which scales as $\log(n)/\log(\log(n))$ [59], and for types of general random graphs this has been shown in the 1980s [60].

¹Note that shortest path in the graph-theoretical sense does not coincide with an itinerary an airline / agent will put together based on the route network. However, given the few hubs and the sparsity of these networks, these will often also be the same.

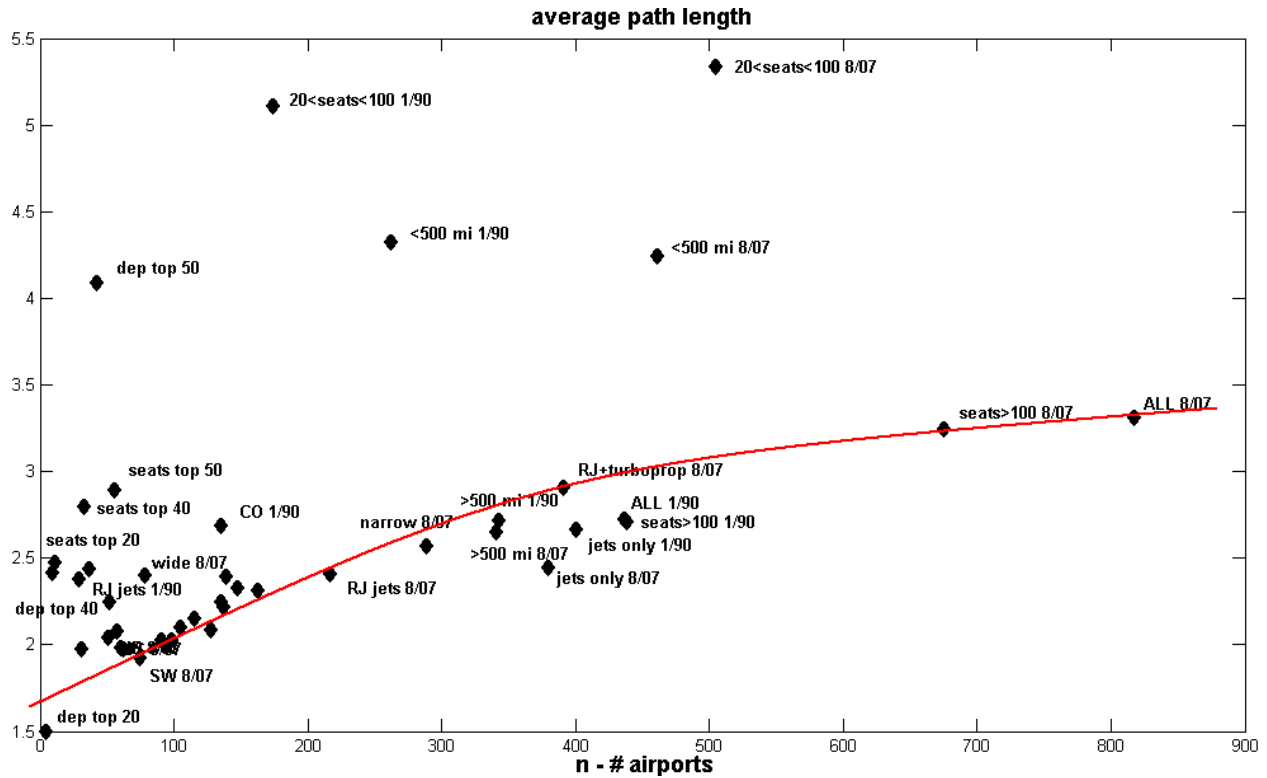


Figure 3-19: Average path length versus number of nodes - all data slices. The average path length versus size seems to follow a log-like curve. The high path length slices are all short-distance or small-aircraft networks, so naturally, they do not have longer cross cutting flights which shorten the path lengths. Examples are flights with less than 100 seats, or under 500 miles.

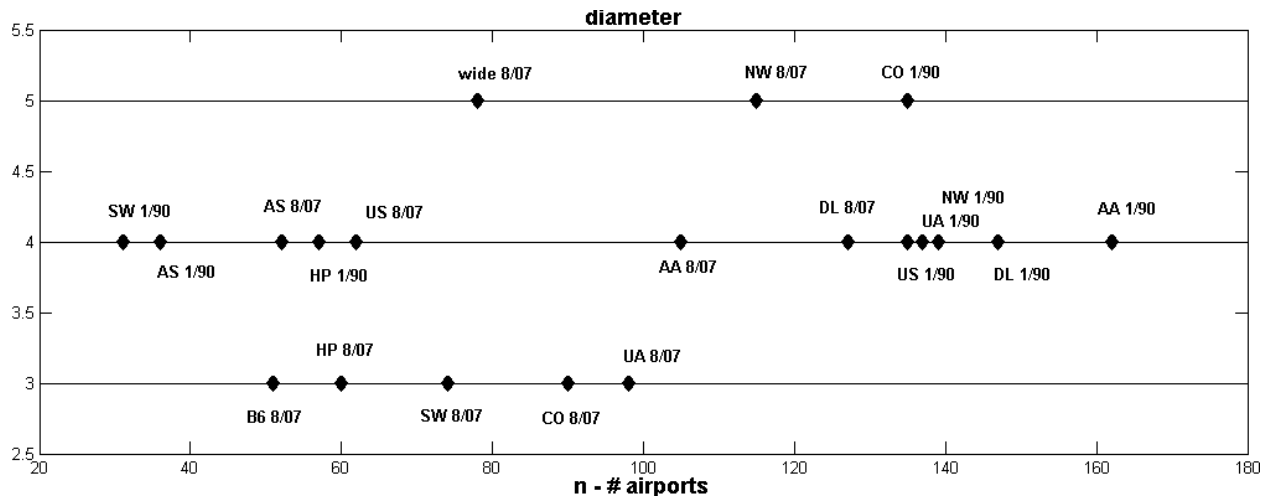


Figure 3-20: Airline diameters versus number of airports. All airlines have diameters between 3 and 5 - which is very small, and mostly unexpected for Southwest which does from a diameter of 4 to 3 from 1/1990 to 8/2007.

In terms of diameter, there are three values among single airlines - 3, 4 and 5. Diameter of 3 have JetBlue, America West 8/07, Southwest 8/07, Continental 8/07 and United Airlines 8/07.

The only two airlines with diameter of 5 are Northwest 8/07 and Continental 1/90. The conclusion here is that as far as span goes, these networks do not vary much.

For all data slices diameter varies more - with really high values for the short-haul and small aircraft networks. As explained earlier, due to the small local reach and lack of large network-wide hubs, these will naturally have large span overall. The diameter across the entire airline dataset for 1/90 is 6 and for 8/07 is 7.

So far we have found that size and traffic-related airline statistics and graph theoretical statistics agree on separating airlines into three groups: low-cost carriers, legacy carriers of 1990 and legacy carriers of 2007. Also most of them single out Southwest Airlines, which means that we expect there to be general topology similarities among all airlines, closer similarities among low-cost carriers versus legacy and, finally, Southwest is bound to be very different.

Industry-related hub metrics [18]

The "hub" metrics presented here are concentration measures. They present various ways to compute how centralized or concentrated the network is in space and in time. Wojahn [18] presents some strict definitions of what properties concentration measures should have, such as anonymity principle (invariance to relabeling) and scale independence. Here we will review briefly each concentration measure and show some correlations between measures and one interesting plot.

Hubbing indices are based on traffic share at an airport: s_i , is the share of traffic at airport i , during a specified period of time, such that $\sum_i s_i = 1$ (the sum of all shares is one).

- i **m -airport concentration** ratio C_m : the sum of traffic shares at the m biggest airports: $\sum_i^m s_i$. The number m is arbitrary, Wojahn [18] uses $m=3$. For example, for $m=2$, the top 2 JetBlue airports by traffic share would be JFK and Boston.
- ii **McShan-Windle index** M measures the percentage of traffic concentrated at the largest 3% of airports: $M(s) = \sum_{i=1}^{\lfloor m \rfloor} s_i + (m - \lfloor m \rfloor)s_{\lfloor m \rfloor+1}$, where $m = 0.03n$, and $\lfloor m \rfloor$ is the largest integer not greater than m . The added fraction ensures continuity in taking the top 3 percent.
- iii **Herfindahl index**: The sum of squares of all traffic shares - $H(s) = \sum_i^n s_i^2$. This measure does capture connectivity despite the sum over all nodes, because smaller fractions squared will contribute less to the total sum, while big airports can dominate.
- iv **Generalized Entropy** (more sensitive to variations in the lower tail distribution), for $\alpha=1$ or 2. $GE(s, \alpha) = \frac{1}{\alpha^2 - \alpha} [\frac{1}{n} \sum_{i=1}^n (ns_i)^\alpha - 1]$. This is another type of sum of fractions. Depending on the alpha parameter, the sum can be made more or less sensitive to the difference scale fractions in the distribution of traffic in the network.
- v **Gini index**: $G(s) = \frac{1}{2n} \sum_{i=1}^n \sum_{j=1}^n |s_i - s_j|$. The Gini index is normalized, as network concentration (size factored out) by Burghouwt [57] by dividing by the maximum Gini index possibly for a given number of airports $G_{max}(n) = 1 - 2/n$. The Gini index is used in economics to measure equality in income distribution. If all (traffic) shares are equal (imagine a pure point-to-point network) then the Gini index is 0 and there is perfect equality. On the contrary, the more skewed the distribution is, the higher the sum of all absolute differences will be, hence G will be higher, signifying inequality, i.e. a more concentrated network.
- vi Ratio between connecting passengers and total enplanements.

The last is probably the best practical measure of "hubbyness", but harder to calculate than the rest. Only 10% of itineraries are published, so those can still be used to calculate the percentage

of passengers flying direct versus non-direct.

The point of all the industry-related indices is to find out how centralized, spatially and in time, the network is - how much operation does it concentrate in the hubs. It is interesting to see whether the industry metrics correlate well with the s-metric, for example, and how well. Table 3.3.1 shows correlations between graph-theoretic measures (degree correlation, r , diameter, d , and s_{max} measure, s/s_{max}) and the five concentration measures (NC is normalized from the Gini index). The correlations are computed for measure vectors across the entire industry, single airlines and data slices. The only stronger relationships to note are between s_{max} and C_m and the Herfindahl indices. Table 3.3.1 shows the same correlations but using only single airlines as data points. The pattern is completely different, which signifies that these metrics measure different (though related) aspects of the topology and cannot be analyzed in combination with each other.

Table 3.2: Correlation table of graph-theoretic "hub" metrics and the industry-developed. All airlines and data slices used for correlations.

	C_m	McShan-Windle	Herfindahl	Gen Entropy	Gini	NC
r	-0.3628	0.0555	-0.3841	0.0763	-0.0316	-0.0616
d	-0.4154	-0.0073	-0.3618	0.0764	-0.1310	-0.1496
s/s_{max}	0.4249	-0.0778	0.4206	-0.0522	0.0309	0.0565

Table 3.3: Correlation table of graph-theoretic "hub" metrics and the industry-developed. Only single airline data points used for correlations.

	C_m	McShan-Windle	Herfindahl	Gen Entropy	Gini	NC
r	-0.0865	-0.2289	-0.0324	-0.1912	-0.2145	-0.2068
d	-0.0697	0.2580	-0.1340	0.0028	0.0471	0.0393
s/s_{max}	0.0683	-0.0174	0.0350	0.1157	0.1526	0.1502

Figure 3-21 shows the C_3 measure - traffic share in top 3 airports - plotted versus size (in number of airports). There is an unmistakable trend of decreasing load on top airports with network size. Of course, this is because the fractions are an inverse function of size. But it is clear that the larger a network is, the less centralized it is. And this is the meaning of the plot - as an airline grows, by the C_m measure, it becomes less centralized in its original largest traffic share (per day/month) airports. This is due to airport capacity, but also network logistics becoming more fragile and complex with single large hubs. These are also the reasons for secondary hub emergence.

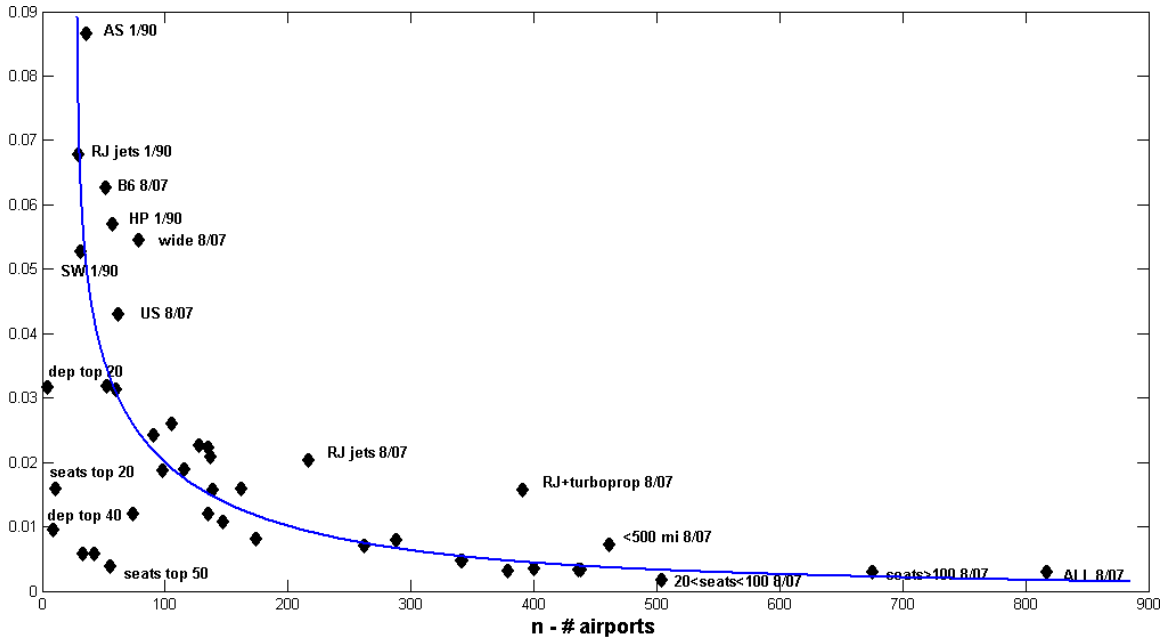


Figure 3-21: C3 (traffic share in top 3 airports) index versus number of airports. All data slices.

Figure 3-22 shows the same plot zoomed on the airlines only. Again, we see the 1990 and 2007 separation. As airlines downsized they concentrate traffic better. The low-cost carriers are most concentrated, with the exception of Southwest - which at a low number of destinations still distributes traffic more equally. On top is Alaska in 1/90, followed by JetBlue in 8/07.

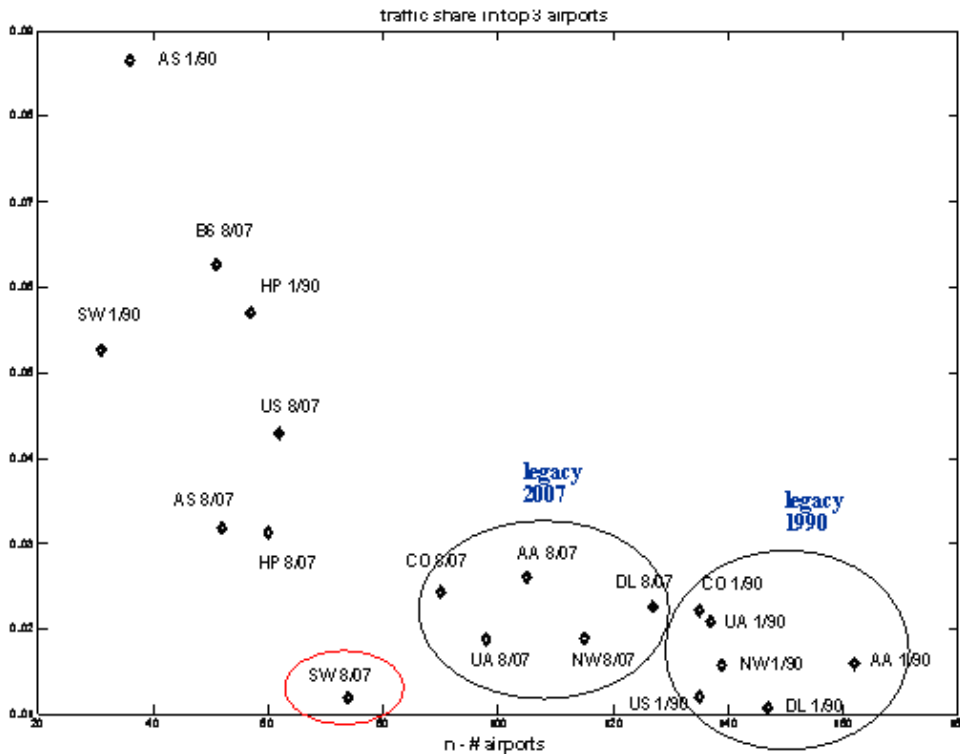


Figure 3-22: Traffic share in top 3 airports versus number of airports for single airlines only.

There are two main takeaways from the discussion of statistics in the airline industry, be it traffic-related, graph-theoretical or industry concentration-indices. First, the legacy carrier part of the industry does evolve together in all these measures. The separation of 1990 and 2007 of legacy carriers is present on all plots. It is usually related to downsizing overall for the legacy carriers, but also to carrying more passengers (on average) and becoming more concentrated (higher s_{max} and C_3). This means that the legacy carriers are making efforts to restructure after bankruptcy, by cutting routes and adding more flights to profitable destinations.

The other main point is the uniqueness of Southwest Airlines. With its explosive growth in passengers carried since 1990 and its unusually decentralized network, Southwest is interesting to watch and a must to analyze in more detail.

3.3.2 Topology profiles

The topology profile concept was described in Chapter 2, Section 2.1.2. The idea is to compare a topology to a spectrum of canonical topologies by the Euclidean distance of their topology vectors. The topology vector is composed of five non-dimensional metrics that describe different aspects of the network topology: density, clustering coefficient, degree correlations, s-max metric and diameter. In mathematical notation, $v = [\frac{2m}{n(n-1)}, C, \frac{r+1}{2}, s/s_{max}, \frac{d}{n-1}]$, where n is the number of nodes, m is the number of edges, C is the clustering coefficient, r is the degree correlation, s_{max} is the s metric of the s_{max} graph, and d is the diameter. The comparison is done, as shown in Chapter 2, Section 2.1.2, by calculating the Euclidean distance between the two topology vectors corresponding to the two different graphs. A smaller value (closer to 0) means greater similarity.

The canonical networks (described in detail in Chapter 2, Section 2.1.1 are plotted on Figure 2-1 in the following order: line graph, circle graph, star, binary tree, tertiary tree, Newman-Gastner graphs with $\alpha=0.1, 0.5, 0.9$, preferential attachment graph (BA), s-max graph, hierarchical binary tree, hierarchical tertiary tree, Dodds-Watts-Sabel graph with varying parameters λ and ξ ($(\epsilon, \epsilon), (\infty, \epsilon), (\epsilon, \infty), (\infty, \infty)$), triangular, square and hexagonal lattices, a random graph with the same degree distribution, a random graph with the same number of modules (as found by the Newman eigenvector method) and an Erdős-Rényi graph. As explained in Chapter 2, each canonical network is created with the same characteristics, density, degree distribution, etc.

Figures 3-23 through 3-28 show that all of these example airline networks differ largely from the simplest graphs, lines, circles and pure stars, as well as from their most scale-free corresponding graph. The profile analysis points to a few other patterns:

- **JetBlue** is closest to a preferential attachment type topology, next to pure and hierarchical trees.
- The **Southwest** topology is closest to a hierarchical tertiary tree, to hierarchies with inter-linking (all the Dodds-Watts-Sabel varieties) and to its corresponding random graphs. These are very interesting results, that will tie with the motif analysis later. They indicate an overall uniformity in the network (either regular or random) rather than a concentrated network or a hub-spoke topology.
- **Continental** Airlines is closest to trees, binary and tertiary, both pure and hierarchical, as well as to preferential attachment graphs. This is the perceived notion of a legacy carrier.
- Among the data slices, the story is the same for the **wide body jets** and the **top 50 flights** by **seat** capacity. As in the case of Continental and JetBlue, these topologies are close to preferential attachment graphs and trees, both binary and hierarchical. In the case of top

seat capacity flights this is especially true because the network is very sparse, so very tree like, and probably the preferential attachment mechanism is related to market demand at large-population destinations.

- The case of **top 50 flights by departures** is different, because the graph is practically a tree (with 3 extra edges that create loops), so this explains the profile similarity to a binary tree and to hierarchies (because in the DWS graphs, there are not that many extra edges to be added for interlinking).

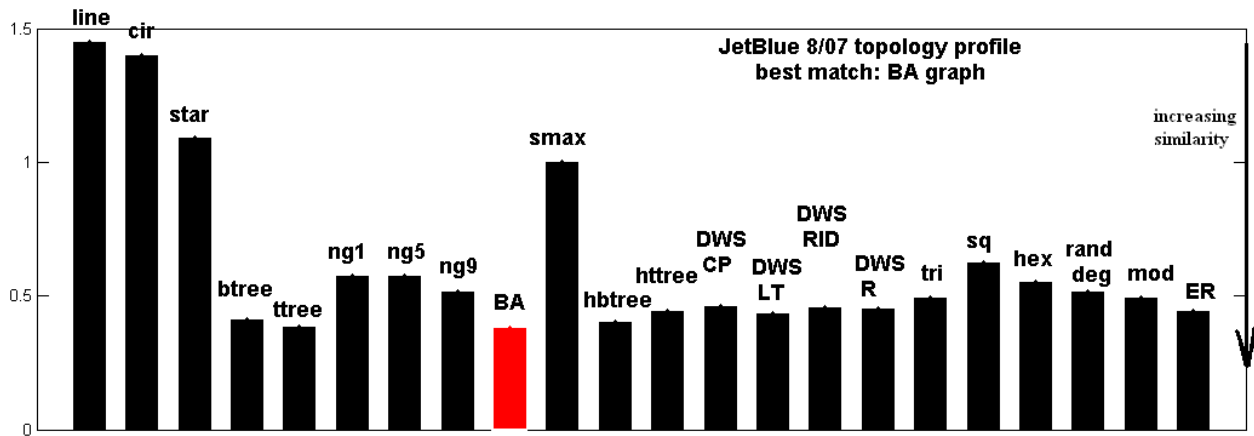


Figure 3-23: JetBlue 8/07 topology profile, best match is a BA graph.

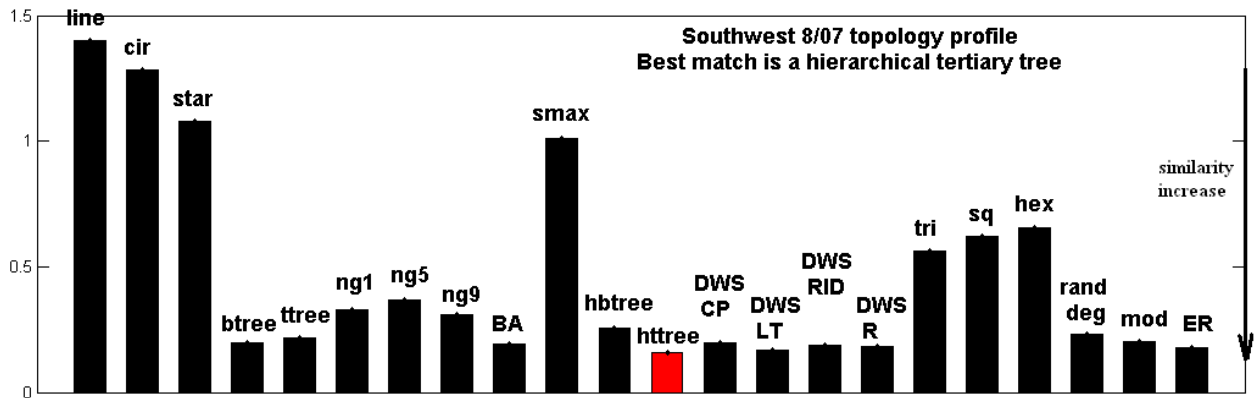


Figure 3-24: Southwest 8/07 topology profile, best match is a hierarchical tertiary tree graph.

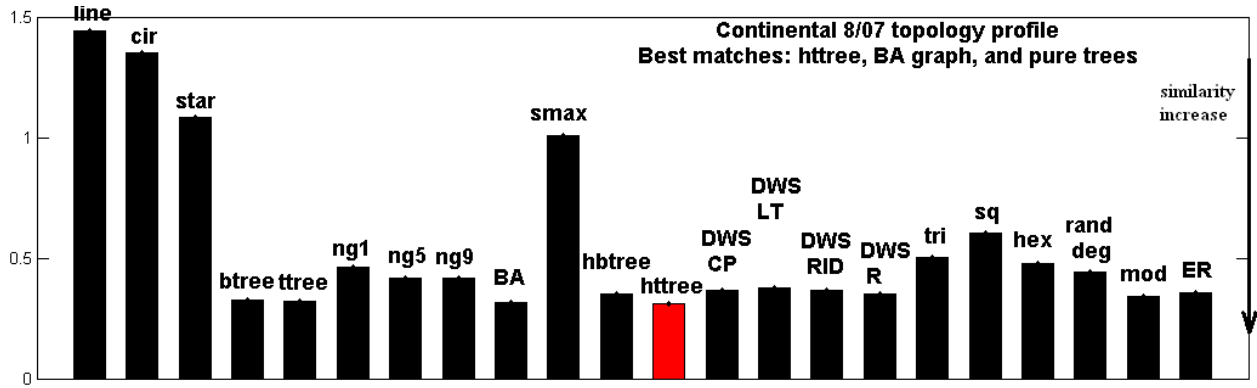


Figure 3-25: Continental 8/07 topology profile, best match is a hierarchical tertiary tree graph, next BA graph.

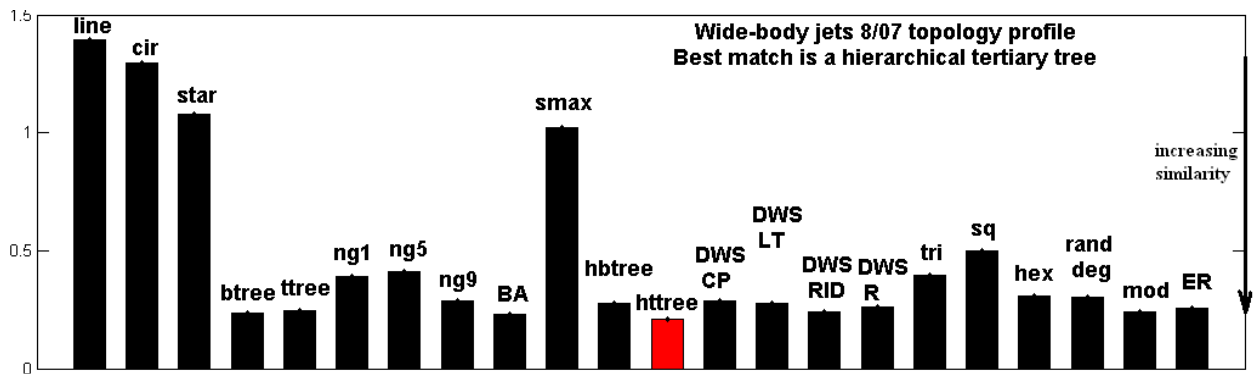


Figure 3-26: Wide-body jets 8/07 topology profile, best match is a hierarchical tertiary tree graph.

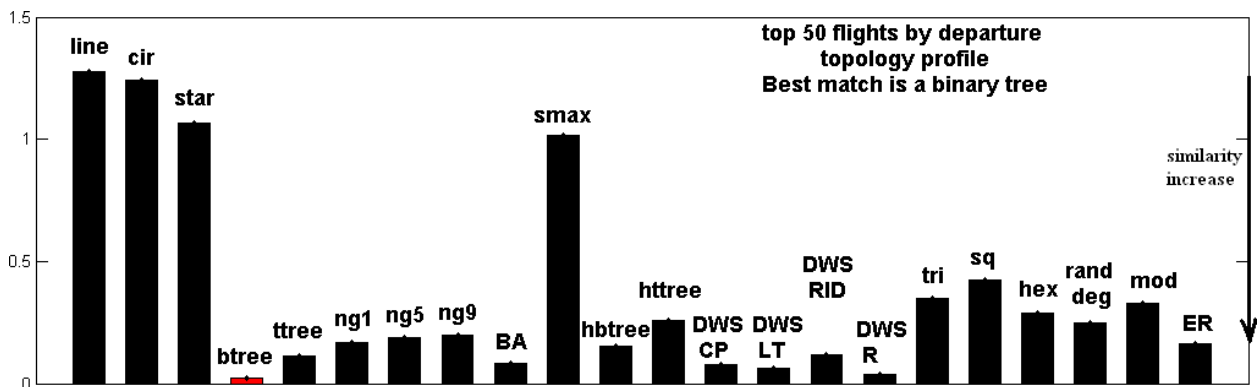


Figure 3-27: Top 50 flights by departure topology profile, best match is a binary tree.

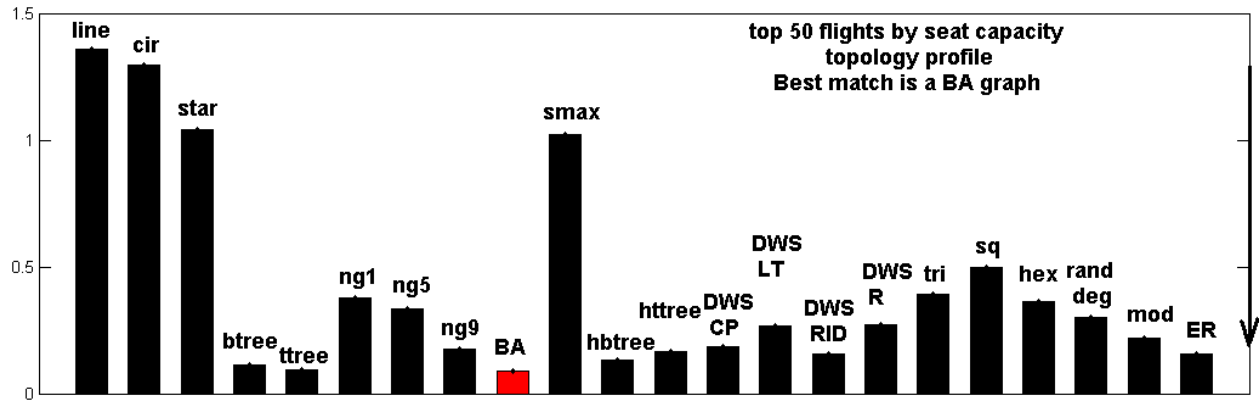


Figure 3-28: Top 50 flights by seat capacity topology profile, best match is a BA graph.

To verify the observations from Figures 3-23 through 3-28, we correlated the profile vectors of each pair of networks. The results are presented in Table 3.3.2. Clearly, the highest correlations are between JetBlue 8/07, Continental 8/07, the wide body jets 8/07 and the top 50 flights by seat capacity. Southwest 8/07 and the top 50 flights by departures stand out.

Table 3.4: Table of relative topology profile correlations. JetBlue, Continental, the wide body jets and the top 50 flights by seat capacity tend to be the most similar.

	JB 8/07	SW 8/07	CO 8/07	wide 8/07	dep top 50	seats top 50
JB 8/07	1	0.9572	0.9932	0.9876	0.9669	0.9847
SW 8/07		1	0.9746	0.9662	0.9621	0.9717
CO 8/07			1	0.9949	0.9769	0.9906
wide 8/07				1	0.9748	0.9897
dep top 50					1	0.9740
seats top 50						1

3.3.3 Degree distributions

Degree distributions are the distributions of nodal degree (total number of connections at a node). They give a visual and mathematical (via slopes) sense of how concentrated, or uniform the network is, based on the distribution. Higher degree for the airlines example means more connections at an airport. Figure 3-29 shows the weighted cumulative degree distribution of all US airlines in 8/07. "Weighted" means that instead of counting one leg (towards the degree of a node), all the departures along that leg are counted. The figure shows that there are three regimes of airports. Huge hubs, with the steepest slope (-2.7), mid-range large airports (with slope -0.7) and small airports, with really slow growth in departures (slope -0.15). The large airports corresponding to these points are: JFK New York, Atlanta, Chicago, Philadelphia, Denver, Charlotte, Los Angeles, San Francisco, Boston, Orlando, Salt Lake City, Seattle, Las Vegas, Phoenix, La Guardia, Detroit, Houston, Dallas, Minneapolis, and Covington, KY.

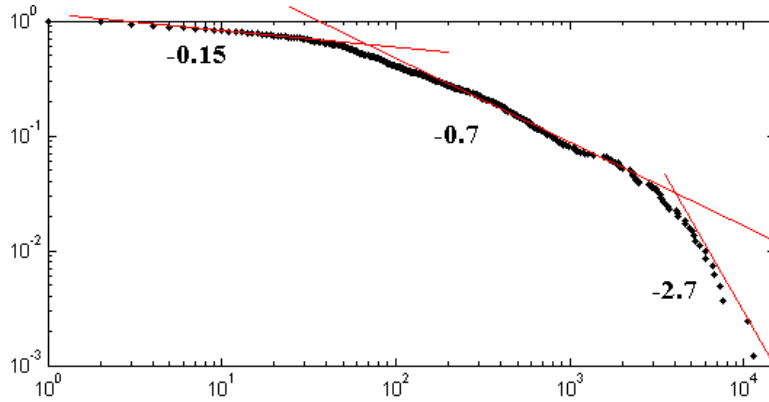


Figure 3-29: All airlines 8/07 weighted degree distribution. The weights are number of departures from a node. Top airports (steepest slope) are JFK New York, Atlanta, Chicago, Philadelphia, Denver, Charlotte, Los Angeles, San Francisco, Boston, Orlando, Salt Lake City, Seattle, Las Vegas, Phoenix, La Guardia, Detroit, Houston, Dallas, Minneapolis, and Covington, KY.

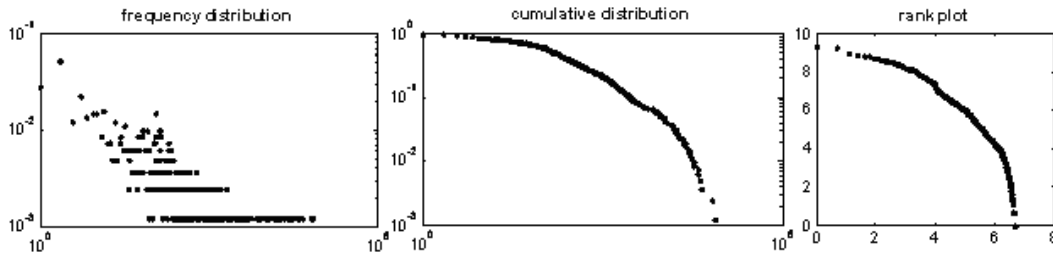


Figure 3-30: Unweighted (simple graph) degree distribution of the entire US airline system. Plots from left to right are: frequency distribution (or probability density plot), cumulative degree distribution (cumulative density plot) and rank plot ($k \sim \log d_k$).

The unweighted degree distribution, reflecting the simple graph behind the airline network, is clearly exponential (see Figure 3-30). This is confirmed both by the cdf and by the rank plot. The frequency distribution shows a very fat tail of the distribution, signifying few airports with high degree, but with great variation in degree. This means that US-wide, there are not 2-3 leading airports, but a set of 15-20 that dominate the scene, with large variation among them. A similar division in three regimes, as with the weighted distribution is seen here, especially in the rank plot. This means that the airports rank similarly in connectivity and in number of departures, overall.

This variation is seen in the plots of betweenness versus unweighted degree. Depending on the data slice, entire dataset (Figure 3-31) or long haul (Figure 3-32), different airports dominate. For long haul flight, Las Vegas (LAS) outshines Atlanta (ATL), both by number of connections, as well as in betweenness. This is interesting given that Las Vegas is never talked about as a mega-hub. When you add all routes, Atlanta becomes the (expected) winner in number of connections, but by far beaten by Anchorage (ANC) in betweenness. Anchorage has been found to be a top betweenness airport in previous studies [17]. Even more interesting, another Alaskan airport makes the cut - Fairbanks (FAI), and mid-west airport - Minneapolis St Paul (MSP). Clearly, Alaska is in the picture here, because of the many local regional, turbo-prop flights, as well as long haul flights from the mainland US to Anchorage. The NYC airports and Boston do not appear in this picture,

because they serve (domestically) coastal end-point markets, so their betweenness will be high in the international network, but not in the domestic.

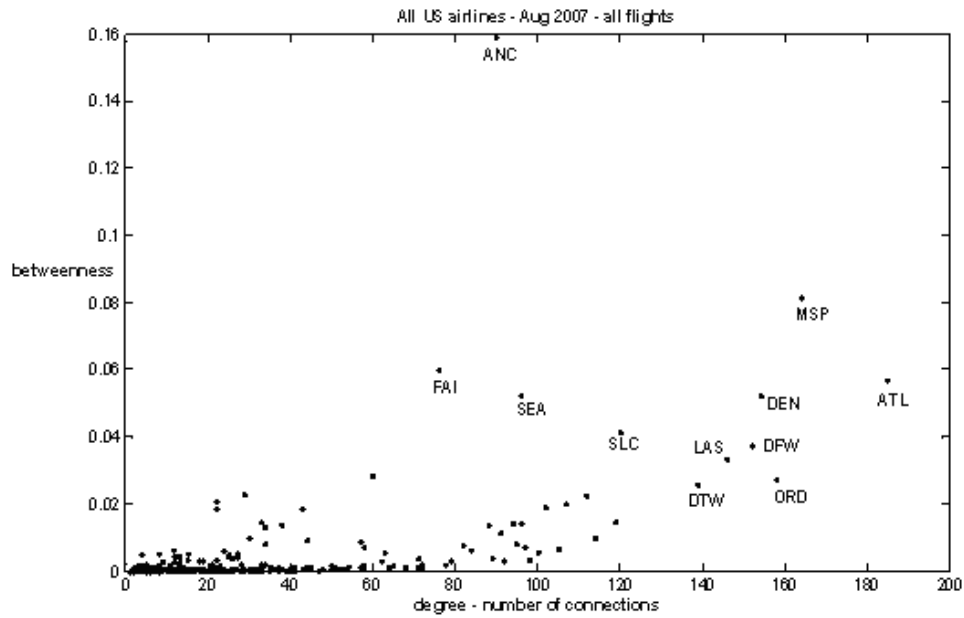


Figure 3-31: Betweenness versus number of connections (degree) for all flights, US airlines 8/07. Atlanta is the most connected airport, but Anchorage has the highest betweenness. The linear-like betweenness-degree relationship is broken here, which means that the network is more modular and balanced rather than centralized with 1 or 2 dominating hubs.

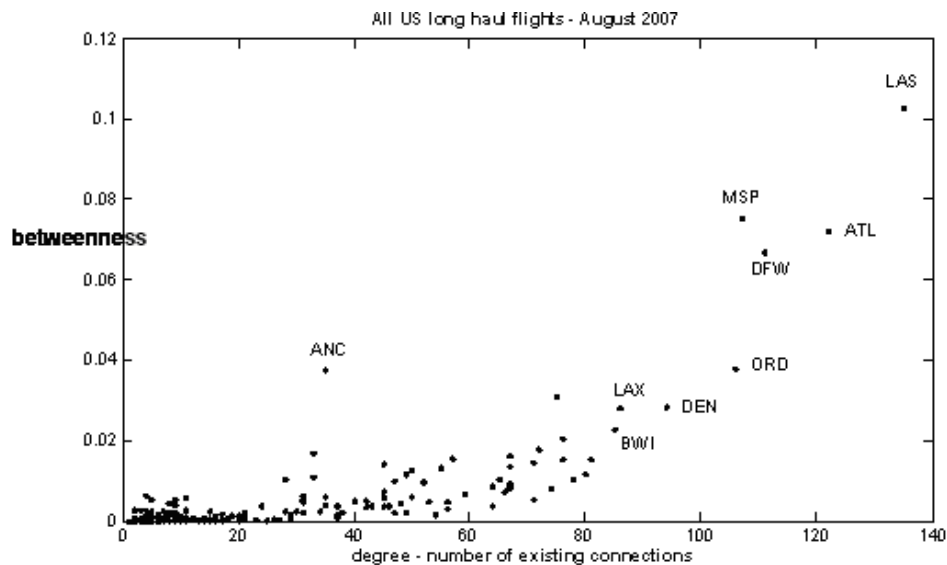


Figure 3-32: Betweenness versus number of connections (degree) for all long haul flights (> 500 mi), 8/07. In long-haul flights, the centralized pattern is more present. Las Vegas is the most connected and most “intermediate” airport. Anchorage still has high betweenness but not that many long-haul flights.

Plotting degree distributions for single airlines, confirms some of the topology profile observations from the previous section. JetBlue 8/07 (Figure 2-8) and the top 50 flights by seats (Figure 3-36) have a power-law like degree distribution - inline with their similarity to BA graphs. Every other network shows exponential degree distributions, with strange regime transitions for Southwest and Continental midway (Figure 3-33 and Figure 3-34). Just as seen in the entire airline dataset, smaller airports seem to operate in a different regime than larger airports. The wide body jets slice has a weak exponential distribution, close to a power-law, but not quite one.

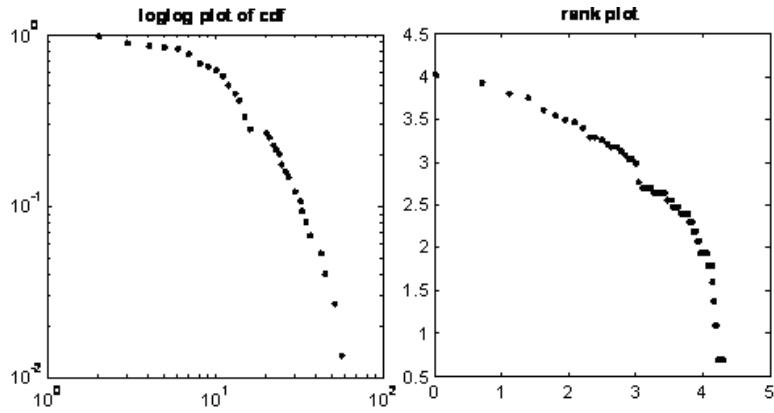


Figure 3-33: Southwest log-log plot of cumulative degree distribution and rank plot. Exponential with a steep power-law-like cutoff.

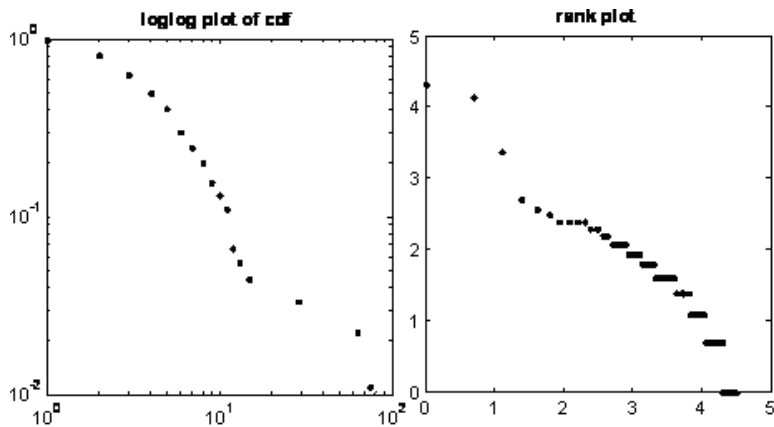


Figure 3-34: Continental log-log plot of cumulative degree distribution and rank plot. The three airports that stand out in the distribution are Newark, Houston and Cleveland. The kink in the distribution means that there are more flights out of these hubs than predicted by an exponential cut-off.

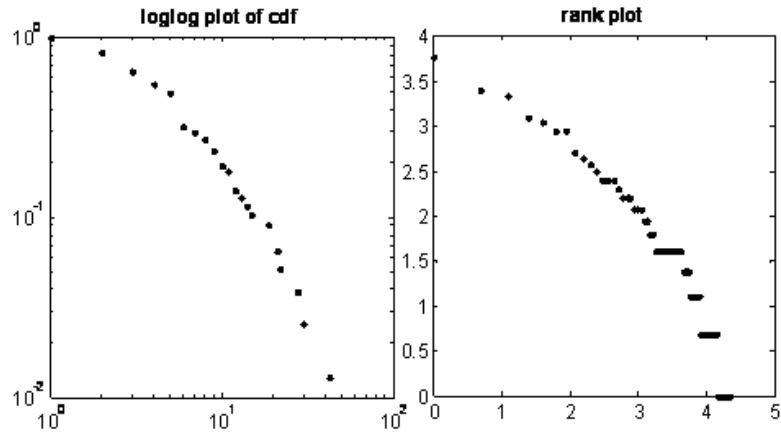


Figure 3-35: Wide-body jets log-log plot of cumulative degree distribution and rank plot. A clear exponential distribution.

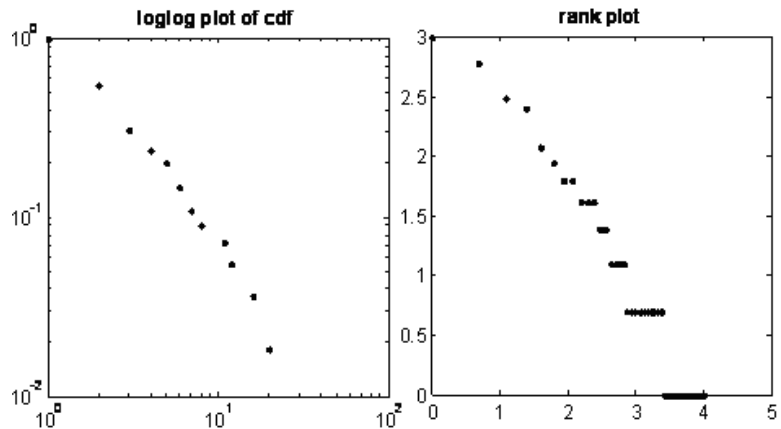


Figure 3-36: Top 50 seats flights log-log plot of cumulative degree distribution and rank plot. Power law distribution.

The degree distributions do not say much about topology. They do show that JetBlue, the wide-body jets and the top 50 flights by seat capacity have similarly skewed distributions, which agrees with the topology profile results from Figure 3-28. The motif analysis in the next section further supports this evidence.

3.3.4 Motif analysis

In this section we show motif analysis for the down-selected airline networks: Southwest Airlines, Continental Airlines (Table 3.3.4), the wide-body jet flights (Table 3.3.4), and top 50 flights by departures and by seat capacity (Figure 3-40 and Table 3.3.4). JetBlue Airlines was analyzed as an example in Chapter 2. The motif analysis reveals some expected patterns from previous discussion in this chapter, but also contains some surprises.

First of all, the similarities between JetBlue 8/07, Continental 8/07, the wide body jets 8/07 and the top 50 flights by seat capacity are confirmed by the underlying motifs. The patterns found in all these networks are mostly the same. For simplicity and to match terminology used in the airline industry, we'll term these topologies hub-spoke. This comes with the disclaimer, that they

are not scale-free, or necessarily preferential attachment graphs, but we'll define them by their predominant motifs. Figure 3-37 below shows the predominant motifs in all of these hub-spoke topologies.

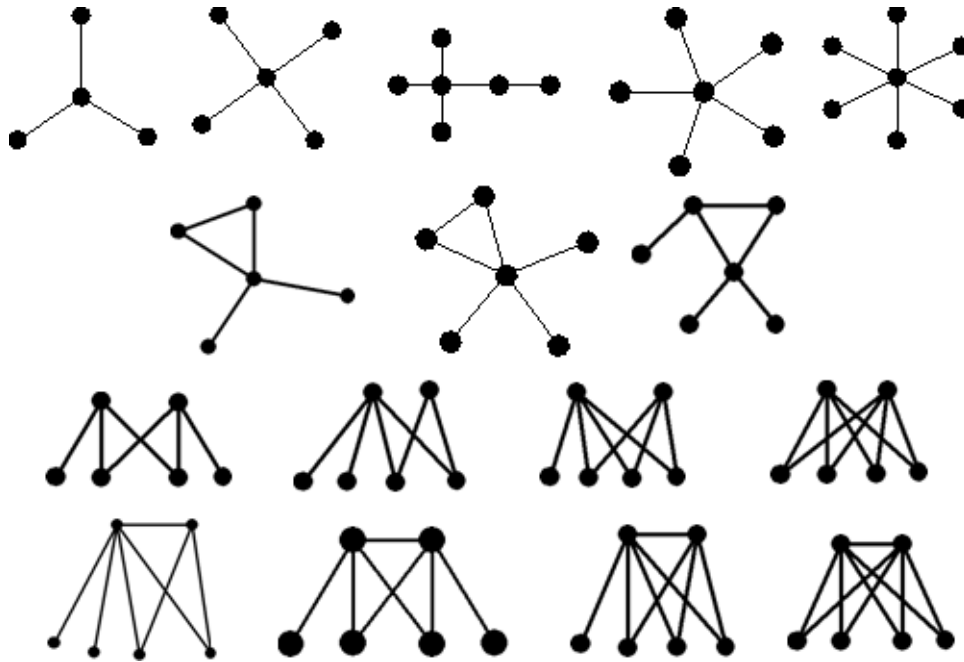


Figure 3-37: Predominant motifs in most airline (hub-spoke) topologies: stars, base-triangles, and bi-partite graphs.

These ubiquitous motifs are stars, of all variations, even with few nodes along an arm; base-triangle graphs, which have spokes off of a triangle from one or two nodes, and bi-partite graphs. Stars are natural occurring in airline networks because of the hub organization that the airlines impose on their routes. Base-triangles form from stars, after two spokes become popular enough that a direct route between them is worth it. If one of those new popular spoke does grow, it becomes a candidate for a new hub. We'll discuss this dynamic in more detail, as part of a growth model, in Chapter 4. Finally, bi-partite graphs are formations in which the airline services airports in parallel from two hubs. These are more advanced formations that occur after the initial stages of growth.

The following sections present the motif search results for Continental Airlines, the wide-body jets slice, the top 50 flights by seat capacity and Southwest Airlines. The JetBlue Airways results were presented in Chapter 2. Overall, the patterns expected from previous analysis in this chapter emerge: the motifs found in all data slices and “hub-spoke” airlines are the same, stars, base-triangles and bi-partite graphs and Southwest Airlines is an exception.

Continental Airlines

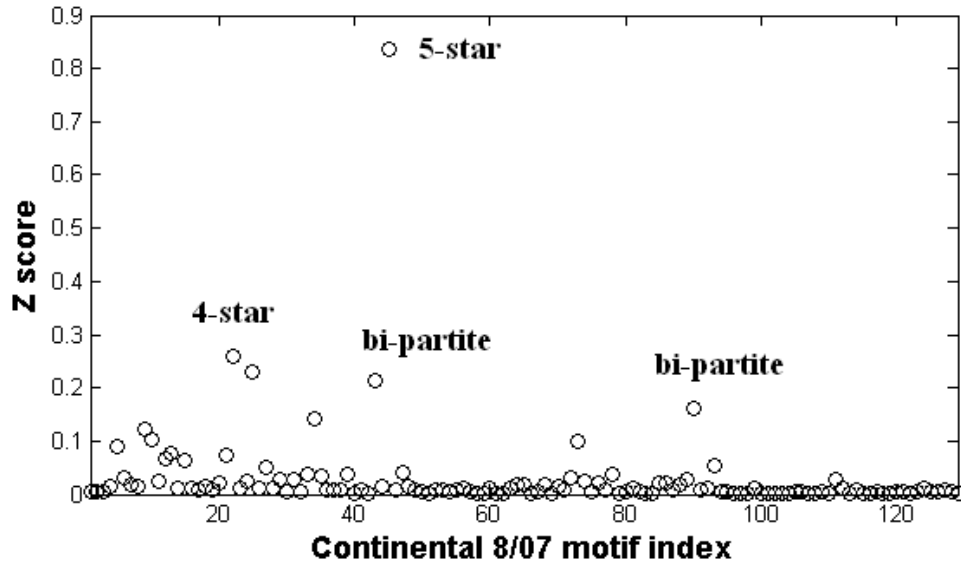

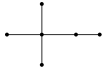


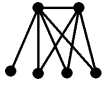

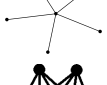



Figure 3-38: Continental Airlines Z-score profile. About 125 motifs are found, of which 6-7 have a Z-score that stands out. They fall in the families described in Figure 3-37.

Table 3.5: Continental Airlines (8/07) significant motifs statistics.

#	Motif	Z score	Count	Mean	Std
8		0.123611	9682228	296301.1	85092.75
9		0.101599	37563320	2341945	388499.4
21		0.257895	1839986	54270.74	7759.622
24		0.228926	16616937	345254.4	79654.4
33		0.142373	3043638	62587.88	23464.61
42		0.215049	244103	2595.12	1258.537
44		0.835139	24415026	153189.7	32556.43
89		0.162713	244734	4774.64	1652.669

Wide-body jets

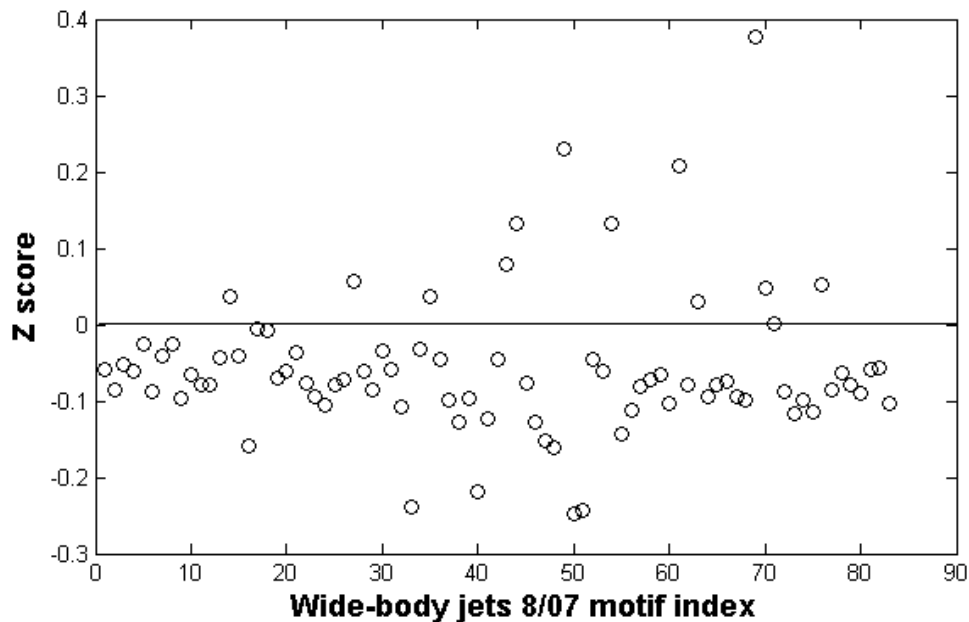
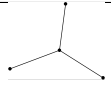
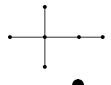





Figure 3-39: Wide-body jets data slice Z-score profile. From about 85 motifs found, 12 have a positive Z-score. All of them also fall in the families found in Figure 3-37.

Table 3.6: Wide-body jets (8/07) significant motifs statistics.

#	Motif	Z score	Count	Mean	Std
48		0.12465596915095475	27112	20072	2207
60		0.11195313854135773	6017987	3803141	773009
68		0.20359337758136667	1799805	765291	198541
85		0.33847698098345691	197876	78120	13825
101		0.62246943048245051	1282658	251340	64737

Top 50 Seats Motif Analysis

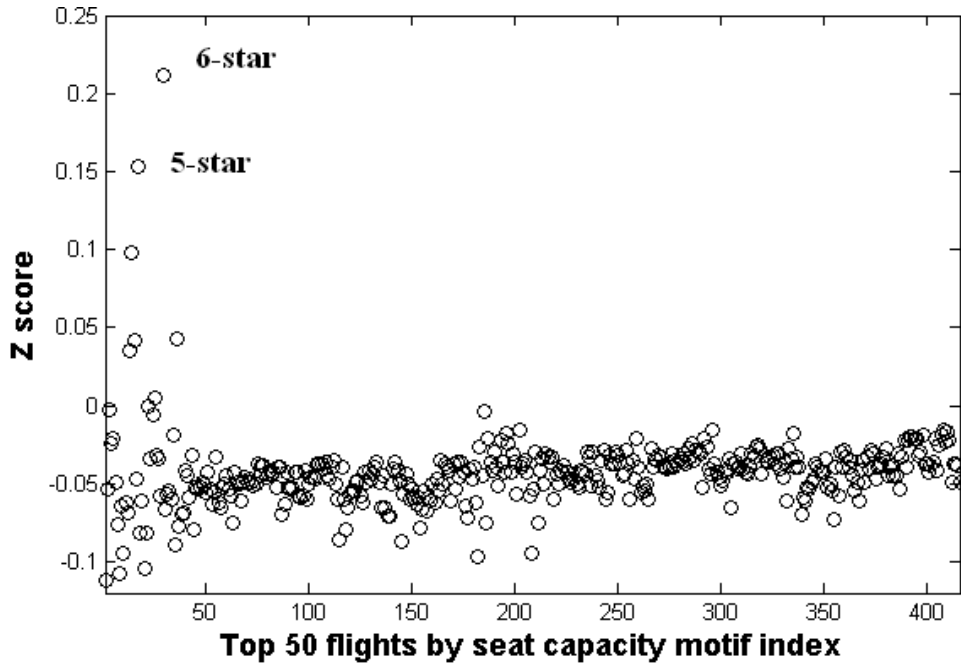


Figure 3-40: Z-scores for all motifs found in the top 50 by seat capacity flights networks. Most Z scores are below and close to zero. The only significant motifs correspond to stars.

Table 3.7: Top 50 by seat capacity significant motifs.

16		0.152974	21218	6677.08	2994.711
28		0.211938	48191	9236.28	5790.711

The surprises in the motif analysis are that both the top 50 flights by departures and the Southwest 8/97 network do not show any significant motifs. In the case of top 50 this is easier to explain. As discussed before, this is a really small network, almost a tree. This means that random graphs with the same distribution are highly constrained by the low-density and also end up being close to trees (to be connected) - hence the lack of difference in motifs found in the real network versus the random background. The bottom line here is that the top 50 flights by departure form a close-to-pure tree, and that is the topology.

Southwest Airlines

Southwest brings a surprise in motif finding. There are no significant motifs, compared to random graphs, though we tested a few snapshots of the airline's history (1/1990, 8/1997, 8/2007). Figure 3-41 shows the low positive Z-score scatter for all motifs found in the 8/07 network. Mathematically, this says that Southwest is no different from a random network. While the topology profiles also confirm that Southwest is close to its random-graph equivalents, it is hard to claim that an airline

operates at random. The logical conclusion here is that the topology analysis performed so far, that uncovers quite well patterns for other airlines, does not provide enough insight for Southwest.

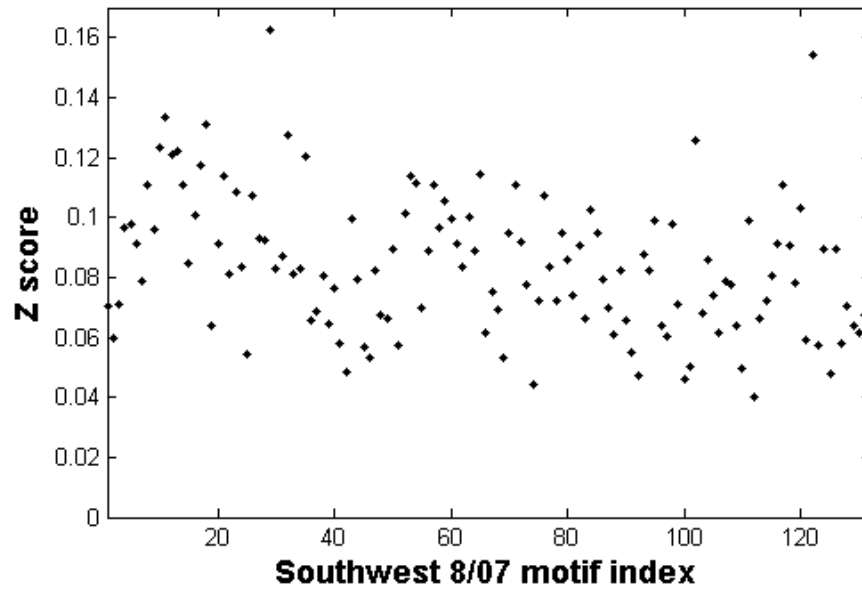


Figure 3-41: Southwest motifs z-score profile. No significant (high Z-score) motifs.

Other things need to be considered than simple where the airline flies. A simple extension is to look at departure frequencies. For Southwest departures equals scaled seats because there is only one type of aircraft. Consider an augmented graph for which an edge exists only if in the real data, that edge has weight above some threshold. For example, for Southwest in 8/07, the highest number of departures on any given leg is 361 (roughly 10 a day). In analogy of the top 50 slices, we extract the top 60 Southwest network, which consists of all legs that have weight within 60% of the top weight ($>0.4 \times$ maximum). This can be thought of as "filtering out the noise". If indeed the airline tries and leaves many temporary destinations somewhat randomly, this extraction filters out the flights that are being counted on for revenue. It turns out that this network has 2 significant motifs, a star and a bi-partite graph, shown below (Figure 3-42).

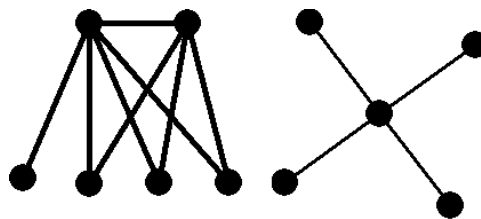


Figure 3-42: Southwest top 60 significant motifs.

These are the same motifs found in JetBlue, the wide-body jets or the more conventional topologies. We will use this idea in the evolution chapter to show that the hub-spoke motifs are only a recent phenomenon in Southwest. Maybe the airline is getting more centralized to deal with the common challenges of the rest of the industry. The weights of top departure flights in 8/07 are shown in Figure 3-43.

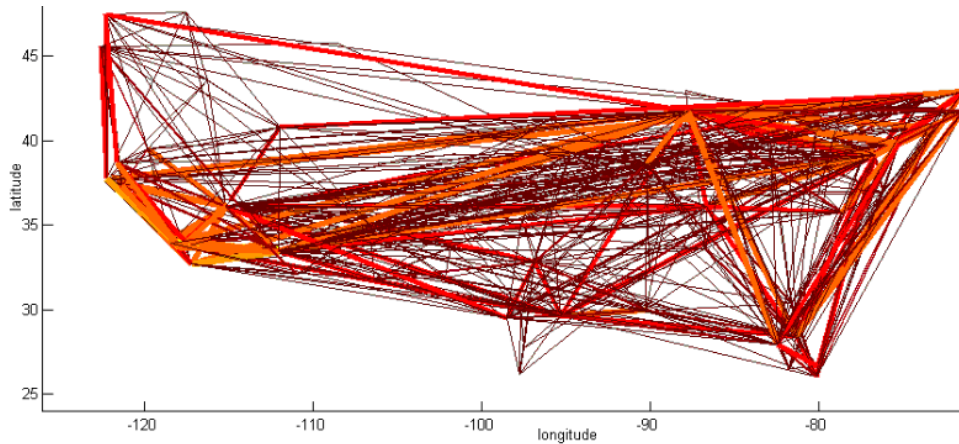


Figure 3-43: Southwest concentrates capacity in 8/2007. Line thickness is proportional to the number of seats offered on that leg.

3.4 Conclusion

In this chapter we reviewed the US airline industry from various points of view. First, we summarized current trends in traffic growth and aircraft patterns. Then we described the data and our approach for analyzing by slices. And most importantly, we presented an extensive analysis from statistical and graph-theoretical point of view of few carefully selected airline networks. We found that statistical, especially domain-specific, measures do not always correlate with graph-theoretical measures, but that most agree on splitting the airline networks in three general groups: low-cost, legacy airlines in 1990 and legacy in 2007. Southwest Airlines consistently showed to be an outlier. The in-depth analysis included JetBlue Airlines 8/07, Southwest Airlines 8/07, Continental Airlines 8/07, the wide-body jet flights 8/07, and the top 50 flights by departures and seat capacity over the entire 1990-2007 period.

We found that JetBlue, Continental, the wide-body jets and the top 50 seats flights have a hub-spoke topology and showed the underlying motifs that unite all of these networks. Using these patterns we discussed a growth model that will be investigated in Chapter 4.

Southwest Airlines appears to be a random network using the same network tools. With adding one layer of information, such as departure frequency, we discovered that there is more under the surface - some of the same motifs are present in the backbone of the network of its top frequency flights. This information will be used to analyze both the raw and the filtered Southwest network for its evolution in Chapter 4.

Chapter 4

Evolution of US Airline Route Networks

In this chapter we study how networks evolve over time. We begin examining the time-dependence of graph-theoretical statistics for the major US airlines, as well as interesting low-cost carriers which gives an overview of the industry over time. This is followed by a section of analysis of topology changes over time where the subject is the network itself. We compare snapshots of the graph topology in consecutive months, as well as to the final month, using the topology vector. Then, using the same representation, we compare the monthly snapshots of a network to canonical networks built with the same statistics and correlate the changes with topology with associated patterns. Using these findings, we propose two simple growth models, tailored to hub-spoke airlines (such as JetBlue) and a specific model for Southwest Airlines. We discuss the performance of these models on the background of canonical topologies, and conclude with remarks about topology evolution as a function of growth in the context of the airline industry.

4.1 Graph-theoretical statistics over time

In terms of number of destinations, all low-cost carriers experience steady growth since their start date except for ATA (launches before 1/1990 and sees decline in the 2000s) and USA3000, which is a young airline with unsteady presence. JetBlue, Southwest, Frontier and Airtran see steady, linear-like growth in number of destinations, with more oscillations in recent years for Southwest and Frontier. This is plotted in Figure 4-1. JetBlue is the fastest growing airline, followed by Airtran, Frontier and Southwest, though the average growth rate of Airtran is the highest (see Table 4.1). Table 4.1 also shows that ATA is declining and that Spirit Airlines has very slow growth. Of all low-cost carriers, Southwest flies to the most number of destinations by the end of 2007, and grows by 50% in the 1990-2007 period.

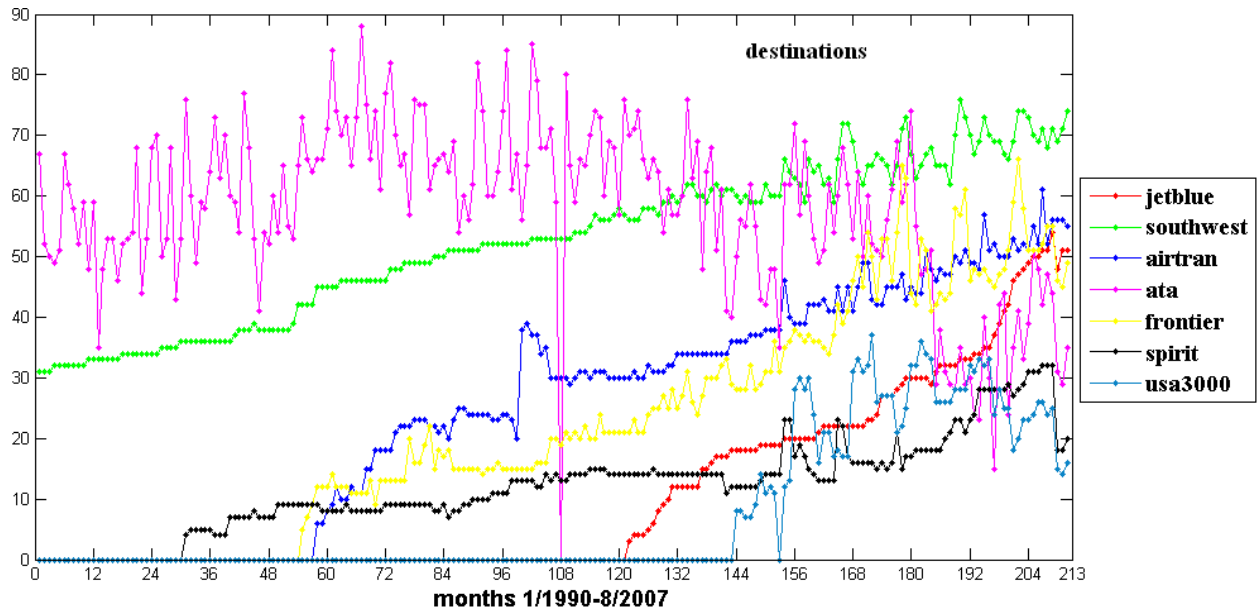


Figure 4-1: Low-cost airlines growth in number of airports, over the period 1/1990-8/2007. Airlines plotted are JetBlue (starts 2/2000), Southwest, Airtran (starts 10/1994), ATA, Frontier (starts 7/1994), Spirit (starts 7/1992) and USA3000 (starts 12/2001)

Table 4.1: Low-cost carriers average growth rate, in terms of number of new destinations per month, 1990-2007.

Airline	Average Growth Rate
JetBlue	0.2417
Southwest	0.2038
Airtran	0.2607
ATA	-0.1517
Frontier	0.2322
Spirit	0.0948
USA3000	0.0758

Figure 4-2 shows the same plot, but also including the big 8 legacy carriers: American Airlines, United, Northwest, Delta, Continental, US Airways and Alaska and America West. Legacy carriers in the past 17 years have not seen much growth in destinations, on the contrary they have steadily downsized, especially US Airways. Continental Airlines is the only one with a general upward trend (up until Aug 07). The plot shows a clear delineation between legacy and low-cost, with the first operating roughly above 100 airports, and the second roughly below 80. By that rule, Alaska and America West fall in the low-cost category. Interestingly, these are the only two carriers among the legacy, that have a positive average growth rate, as seen in Table 4.2. Southwest’s steady growth in number of destinations is most similar to America West’s. Also, the low-cost carriers have more steady operations from month to month in number of destinations. The legacy carriers show a higher seasonal and overall variation. Growth and variation are two factors that may affect topology. Later on, we discuss their effect on topology transitions, and show whether topology can also oscillate seasonally, and whether it goes through major phases as a function of growth.

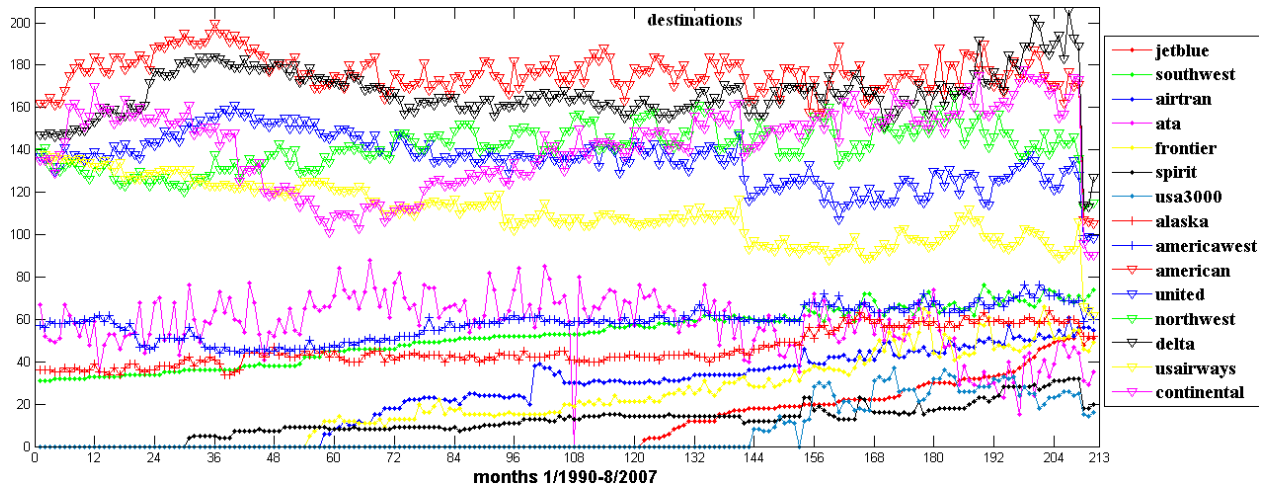


Figure 4-2: Top eight airlines (by passengers carried) and low-cost airlines growth in number of airports, over the period 1/1990-8/2007. Airlines plotted are JetBlue, Southwest, Airtran, ATA, Frontier, Spirit, USA3000, Alaska, America West, American Airlines, United, Northwest, Delta, US Airways and Continental Airlines.

Table 4.2: Legacy carriers average growth rate, in terms of number of new destinations per month, 1990-2007.

Airline	Average Growth Rate
Alaska	0.0758
America West	0.0142
American	-0.2701
United	-0.1848
Northwest	-0.1137
Delta	-0.0948
US Airways	-0.3460
Continental	-0.2133

The figures in number of seats offered by the airlines are closer to the economic outlook of the industry. As a result, September 2001 shows up as a prominent spike on Figure 4-3, and marks a faster downturn for legacy carriers, which are already on a downsizing curve prior to the event. Low-cost carriers (Figure 4-4) see a smaller spike, and recover their previous growth rate. Southwest Airlines is off the low-cost scale in terms of seats, and unprecedentedly sweeps the industry, to start out at the level of Alaska and America West in 1990 and offer more seats than anyone in 8/2007. So this is an airline that flies to as few destinations as a low-cost carrier but offers as many seats as legacy carrier. This is possible if the following is true: Southwest flies more frequently on any single destination than the average, and it flies a *denser* network. The higher density and departures frequency is already an indicator that Southwest Airlines will have a different topology, and will be an exception in the industry.

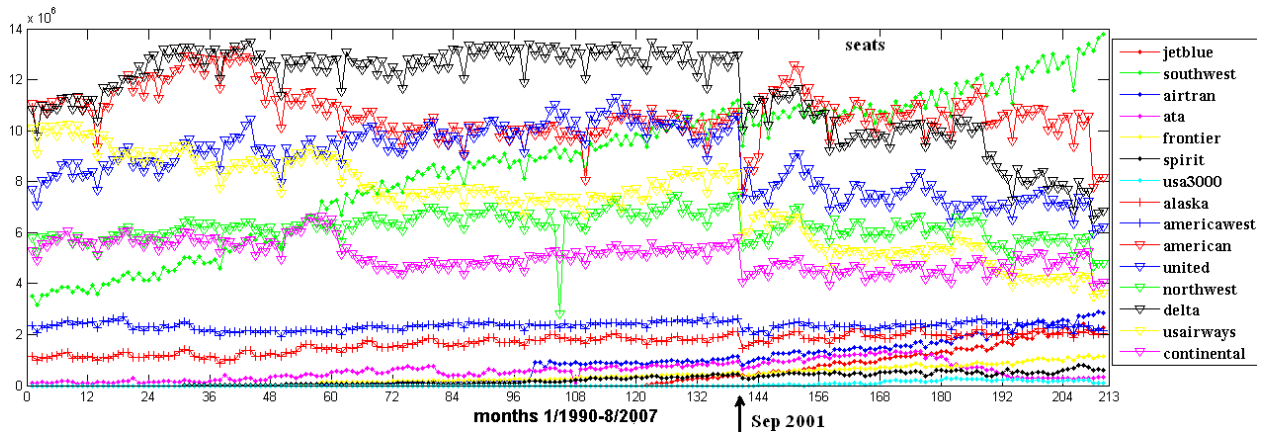


Figure 4-3: Top eight airlines and low-cost airlines growth in total number of seats offered monthly, over the period 1/1990-8/2007.

Notable events in low-cost carrier seat capacities are the Airtran merger with Valuate in 1998, which shows as a jump, followed by steady growth. JetBlue grows proportionally in capacity to growth in destinations, still fastest among low-cost carriers. ATA sees great decline starting in 2004, which is when the airline files its first bankruptcy (Oct 26, 2004). The second bankruptcy in April 2008 and end of operations of the airline, is beyond the timeline of this dataset. However, this is an interesting example of the demise of an airline, rather than its birth. A plot of the ATA topology evolution is included in the Appendix, Figure A-5.

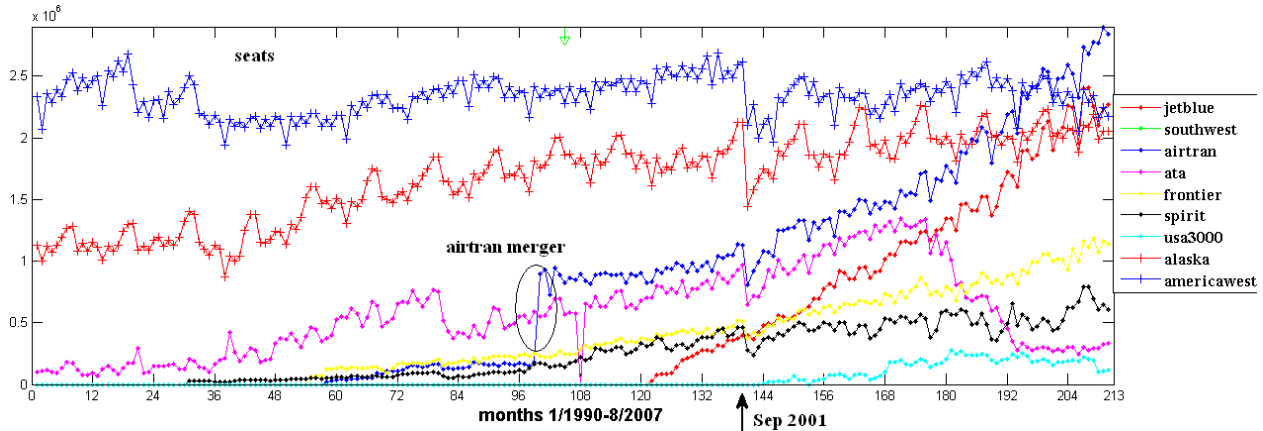


Figure 4-4: Low-cost airlines growth in total number of seats offered monthly, over the period 1/1990-8/2007. Notable events are the September 2001 spike and Airtran's merger with Valuate, beginning operations in September 1998.

The data of passengers carried monthly is a proof that this is a cyclical industry. The 12-month patterns are visible in both the legacy (Figure 4-5) as well as the low-cost (Figure 4-6) carrier plots. One cycle is marked on Figure 4-5 as an example. Collapsing the cycles by plotting only one month yearly, for example August, does smoothen the trends, but shows the same overall patterns. A plot of "smooth" number of passengers carried yearly (for 18 years) is available in the appendix (Figure A-4). In terms of size, Southwest becomes the largest carrier of passengers in 8/2007 and Airtran becomes the largest among the low-cost carriers, followed closely by JetBlue.

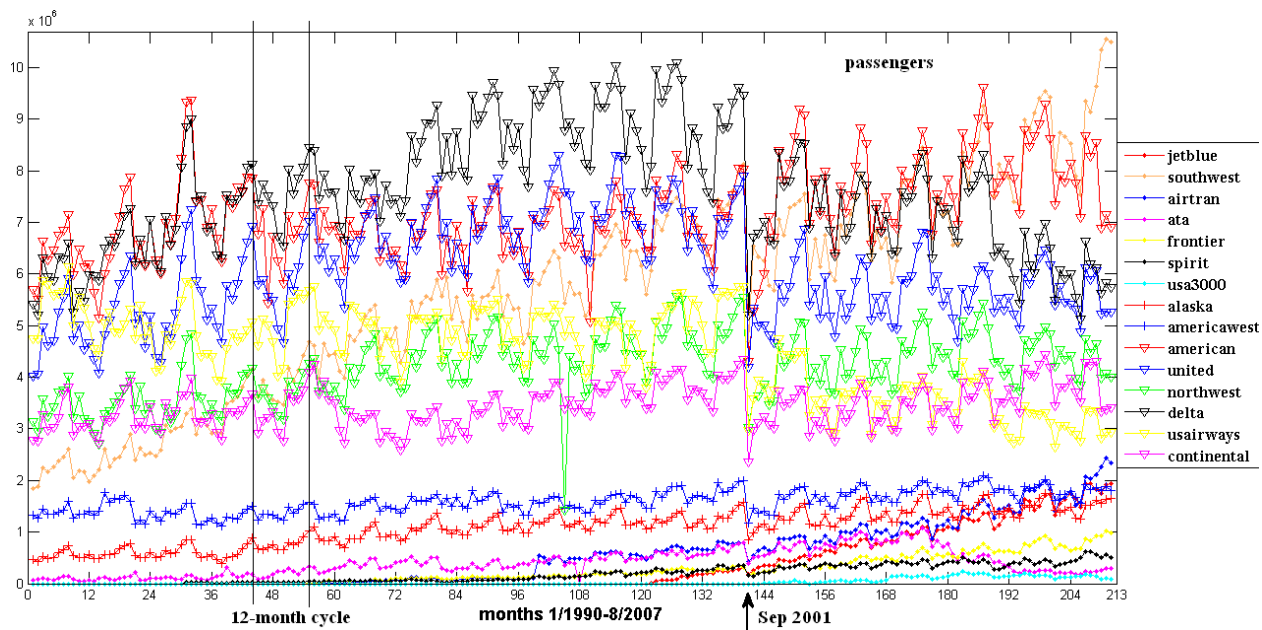


Figure 4-5: Top eight airlines growth in total number of passengers offered monthly, over the period 1/1990-8/2007.

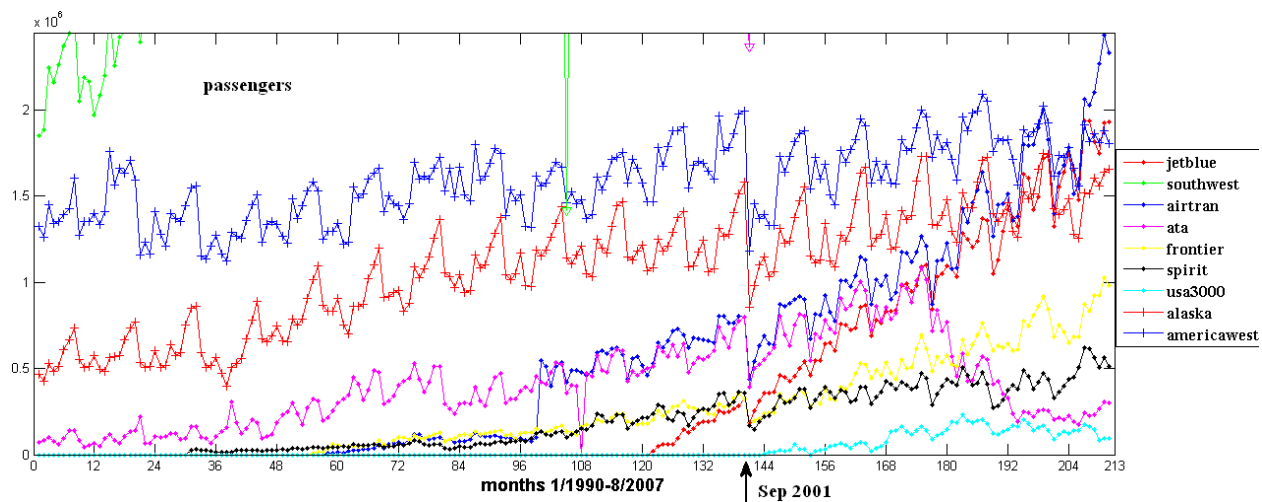


Figure 4-6: Low-cost airlines growth in total number of passengers offered monthly, over the period 1/1990-8/2007.

The minutes spent in the air are proportional to the number of departures and the distances flown. The history for all airlines plotted in Figure 4-7 shows the same patterns as other statistics, except for two observations concerning Southwest Airlines. First, they catch up in air time much later to the legacy carriers, than they do in number of passengers and seat capacity. That is very surprising, because they are supposed to fly more frequently, and have more efficient aircraft utilization. Two possible reasons for this are immediately obvious: they might be flying with higher load factors, and second, most of their flights are short haul. If American flies mostly long haul, and with lower load factors per flight, then they would be spending more time in the air but carrying

less passengers. All of this points to the short haul model as the distinguishing for Southwest. It is harder to claim that this model is the reason for their profitability - other factors such as fast turn-around time have to be considered. In terms of topology, we expect Southwest to have a lot of short and local connections, rather than cross-country flights. This description fits with the expectation of higher network density. Despite these considerations, it is still surprising that Southwest has less air time overall.

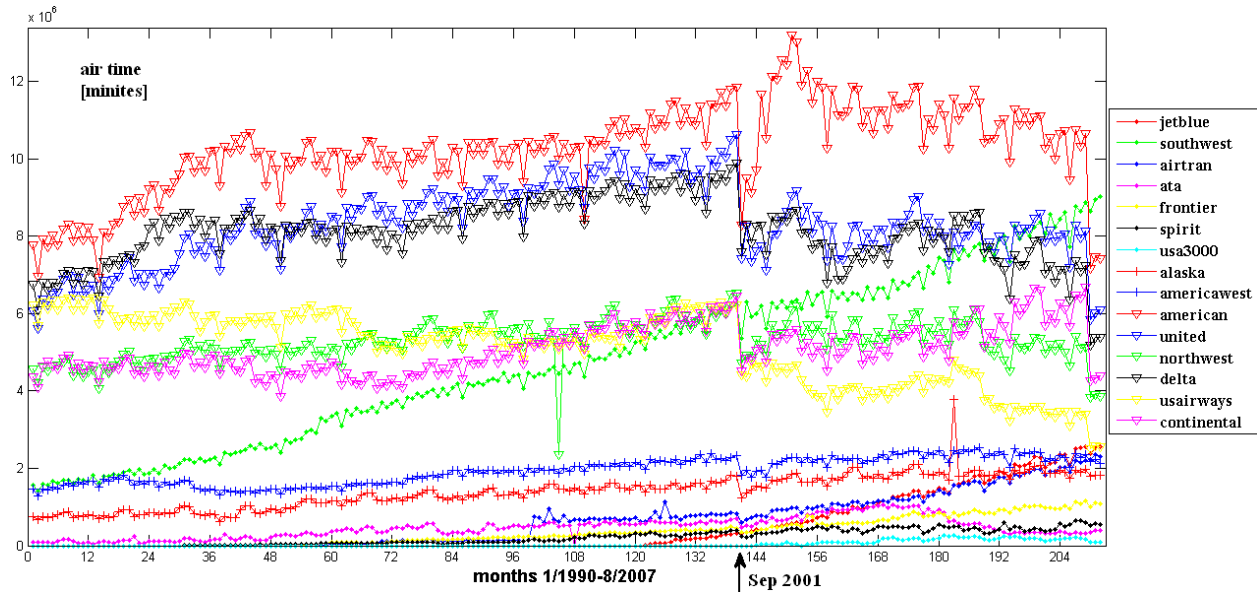


Figure 4-7: Top eight airlines growth in total number of minutes flown monthly, over the period 1/1990-8/2007.

The average path length for the entire industry is included in the Appendix in Figure A-3. All legacy carriers, almost without exception, fall in a bandwidth between 2 and 2.5 hops, as the average shortest number of legs (connections+1) to fly between any two airports. Here we plot the low-cost carriers only, in order to highlight some differences. Figure 4-8 shows America West, exactly at 2 for the entire period 1990-2007 and Alaska at 2.5. All carriers founded in that time period (JetBlue, Airtran, Frontier, Spirit), start out with average path lengths below 2, probably because of strictly single-hub initial operations, and quickly level out to 2. Two things are interesting to notice in this plot. Southwest does not stand out among carriers, for the first time. Despite the multiple local connections we expect to find out, and its allegedly point-to-point structure, clearly the airline must have cross-cutting routes (not local, intrastate), enabling easier travel. The other interesting observation that is ATA is the only airline in the industry flying with a number of average hops higher than 2.5. The reason could be that ATA also performed military and commercial charter flights across the world. Whether the business model is tied to the network structure and the reasons for bankruptcy and demise is not the point of this study, but interesting nevertheless.

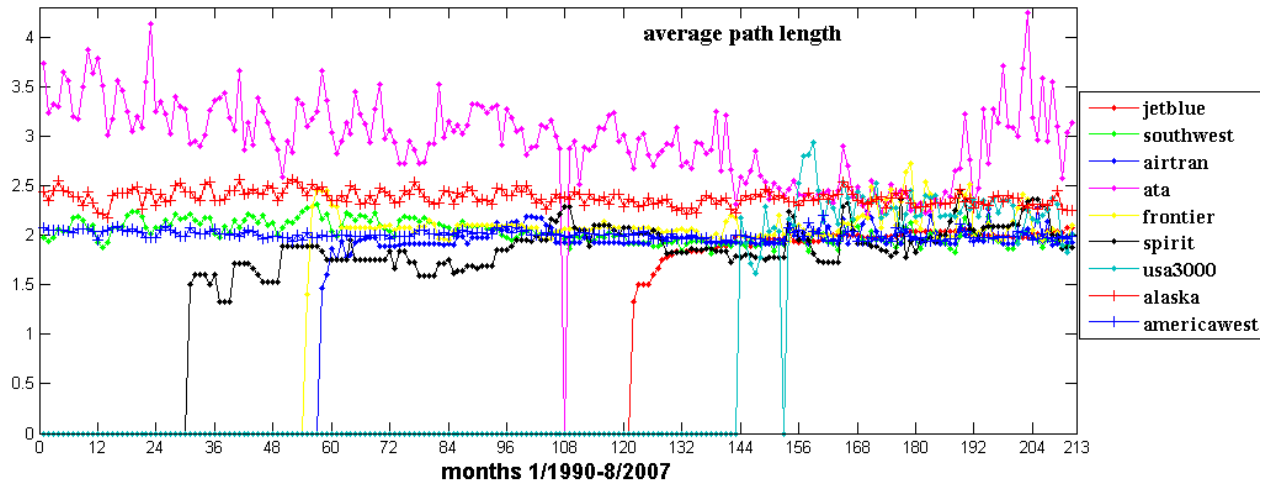


Figure 4-8: Average path length over time for all low-cost carriers: JetBlue, Southwest, Airtran, ATA, Frontier, Spirit, USA3000, Alaska, America West

In this section we reviewed the airline industry statistically for the period 1/1990-8/2007. We found that US carriers have largely downsized their networks, with faster decline since September 2001. Also, among the low-cost carriers JetBlue is the fastest growing, while Southwest grows the most, in size, and across all other scales. The statistics reflects well major events, including mergers and bankruptcies. In terms of structure, we expect Southwest to be different from all others, with many local connections, short flights and some cross-over long-haul flights. Its network is likely to be a lot denser, and expand in a different manner. The legacy carriers are like to have steady topologies over time, without many transitions, while for the low-cost airlines there are likely to be changes associated with growth.

4.2 Detecting changes in topology

The first cornerstone of this thesis was to describe and define the term network topology, and to analyze it in the context of airline routes. We used a set of combined metrics to compare real networks to canonical topologies and to each other in order to understand underlying structural mechanisms. This same idea can be used to track the changes in a single topology over time. For example if a star network adds cross-links which break the star topology, the next snapshot will no longer have a zero clustering coefficient. Other metrics will differ too.

In Chapter 2, Section 2.2.2 we discussed the validity of the graph similarity score and found that it is useful in mapping continuous changes in the same topology. The graph similarity measure is computed using a graph similarity matrix which is the converged solution to the hub scores and authorities flow equations (see Section 2.2). The entries in the matrix $M_{i,j}$ show how similar nodes i and j are, where i and j are nodes in two distinct matrices. In this case, the two matrices compared are representative of two consecutive snapshots of the same network. Finally, from the graph similarity matrix M , the best matching sequence of nodes is extracted and the corresponding sequence of matching scores is taken as a sum of squares to compose the graph similarity measure.

The alternative measure for graph similarity, described in Section 2.1.2, is the Euclidean distance of topology vectors. This measure was used to calculate the topology profile of a graph. The two measures are compared in Figure 4-9 which shows the similarity of consecutive snapshots of a random graph. These snapshots are taken at every step of adding a new link in the process of

building an Erdős-Rényi graph. Over time, the snapshots become more similar to each other, which means that the topology 'stabilizes'. Early on, with very few nodes, it oscillates a lot, and eventually reaches a final state. The two measures operate on vastly different scales. Even the early oscillations are on the order of 0.001 for the graph similarity measure. As explained in its assessment (Section 2.2.2), this metric does not perform well on an absolute scale, but it shows changes relatively quite well. With that in mind, the two measures are positively correlated (0.6) and show similar patterns, which is a validation for both because their principle of measurement is different. The graph similarity measure uses flow based on the adjacency matrix, while the topology vector combines statistical metrics of different properties of the graph, such as clustering, reachability and degree correlation.

The conclusion about the random graph from Figure 4-9 is that the topology reaches a relative equilibrium. The first actual data example we chose is ATA Airlines (Figure 4-10), because as the history shows, this airline is in process of decline throughout the studied period, especially in the last three years. The topology consecutive snapshot comparison shows the reverse process than the case of the random graph. With higher number of oscillations because the graph is sparser, the oscillations increase with time as the network declines.

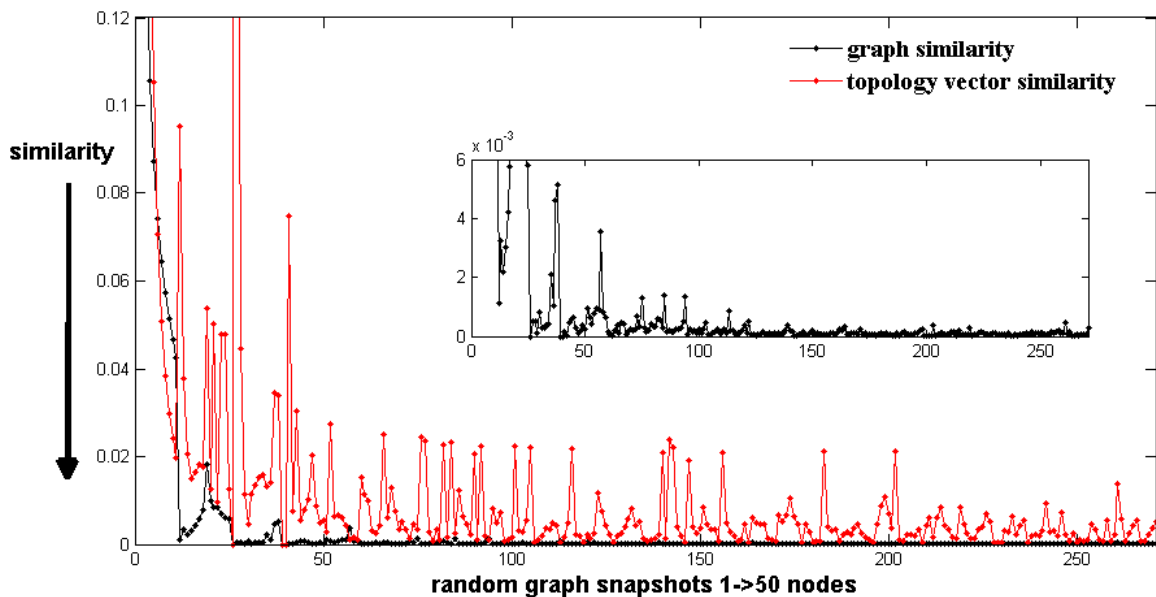


Figure 4-9: Example of tracking a random graph topology with the two similarity measures. The zoomed plot shows the graph similarity variation at a magnified scale.

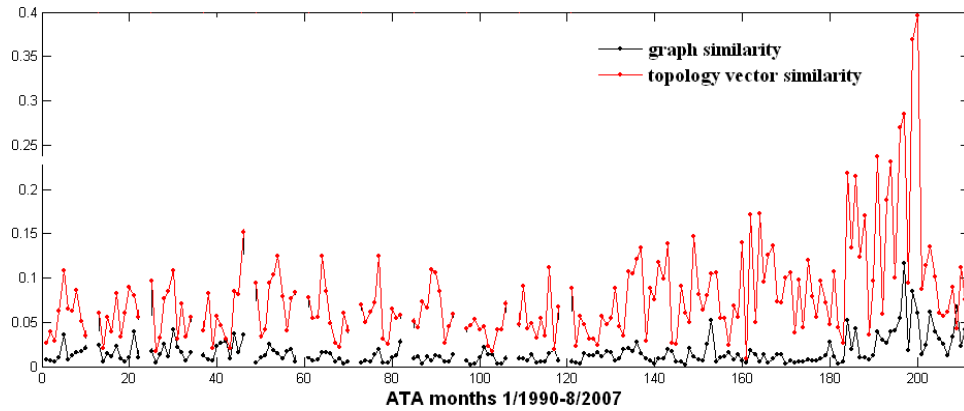


Figure 4-10: ATA topology changes month to month, from 1/1990 to 8/2007. Graph similarity and topology vector similarity measures.

JetBlue’s network stabilizes as the airline grows, especially after month 22 (Figure 4-11). There is a spike in that month, because the topology goes through a major transition from a pure star to a denser network with a second hub. The topology vector metric reflects that a lot better than the graph similarity measure.

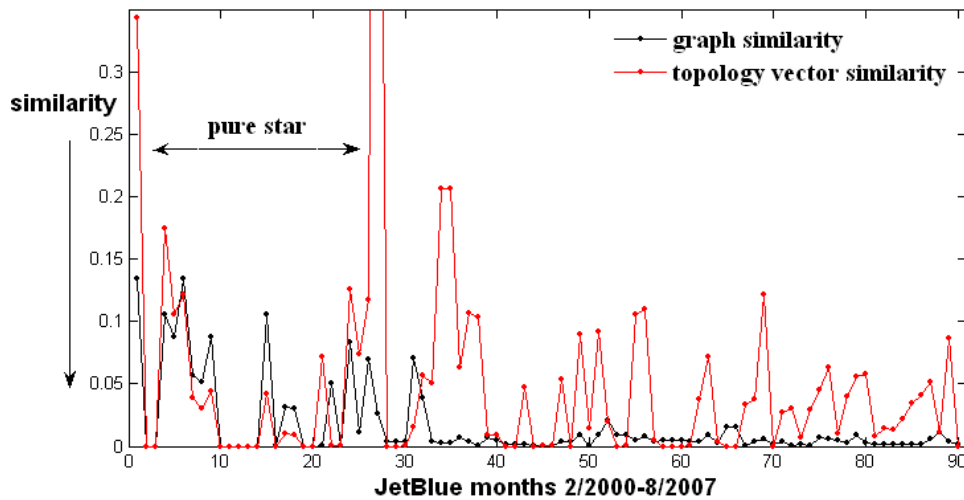


Figure 4-11: Plotting the two similarity measures, graph similarity (black) and topology vector (red) over time (JetBlue monthly index from 1 to 91).

Southwest is also growing in size in the period 1/1990-8/2007, even though it starts to operate in the 70s. It grows from 31 airports to 74 in 17 years, more than double, and super-linearly in number of routes. As seen in the statistical comparison with legacy carriers earlier, it grows from a low-cost rank airline to carrying more passengers than any network airline. Figure 4-12 shows that in this time the topology has been changing, with oscillations damping as the network gets bigger. With the same observations for ATA (in reverse) and JetBlue, it is clear that size is a factor in these metrics, but it is not the single reason for topology changes. The pure star transition for JetBlue’s 22nd month confirms that, as well as the steady topology patterns observed in the cases of the legacy carrier, Continental and the wide-body jet network. These topology changes are plotted in Figures 4-13 and 4-14 respectively.

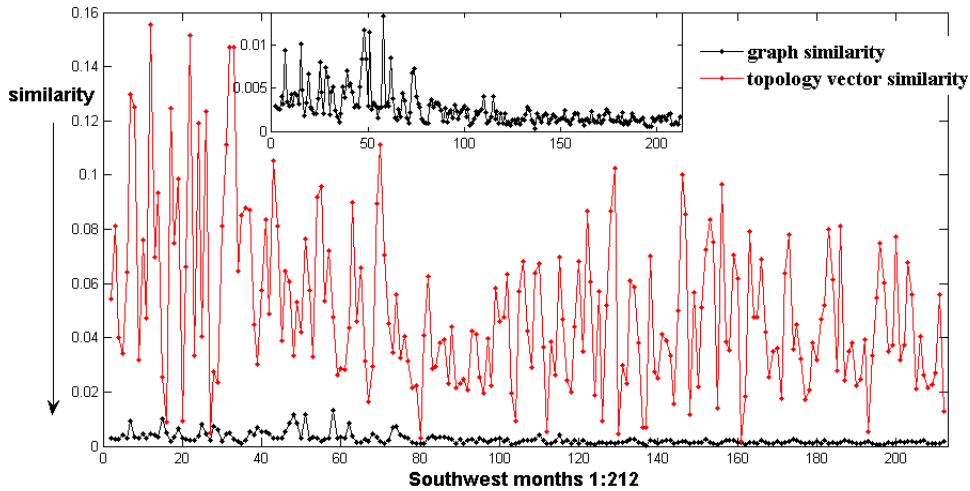


Figure 4-12: Southwest topology changes over 212 months from 1/1990 to 8/2007.

Continental Airlines shows the same oscillatory behavior except that the oscillations do not grow or decrease in size. A lot of the peaks are 12 months apart, signifying seasonal behavior. We know from Section 4.1 that Continental, as all legacy carriers, downsizes its network, so this is an example of topology being relatively constant with size. While the airline downsizes and concentrates, it keeps roughly the same operations.

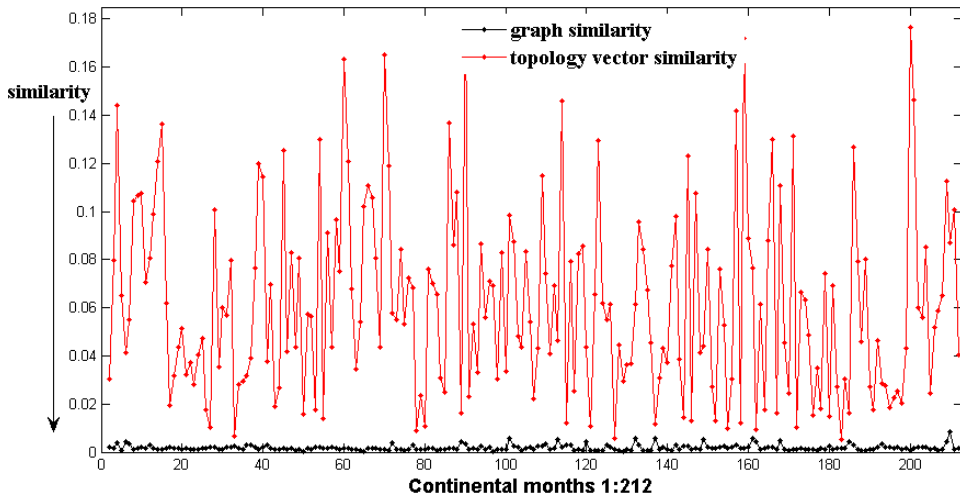


Figure 4-13: Continental topology changes over 212 months from 1/1990 to 8/2007.

The story is similar for the wide-body jet network - constant almost seasonal oscillations, but no major changes from month to month overall (Figure 4-14).

The above examples show that network topology varies with time cyclically for airlines, probably related to seasonal traffic patterns, such as summer and holiday peaks. For the large network carriers and the data slices (wide-body jets) these cycles are also cyclical and noisy without much significant variation over time. This means that topology transitions do happen when systems grow, not only in the early stages (ex: JetBlue), but also in later stages of growth (ex: Southwest), and when they experience major decline (ex: ATA).

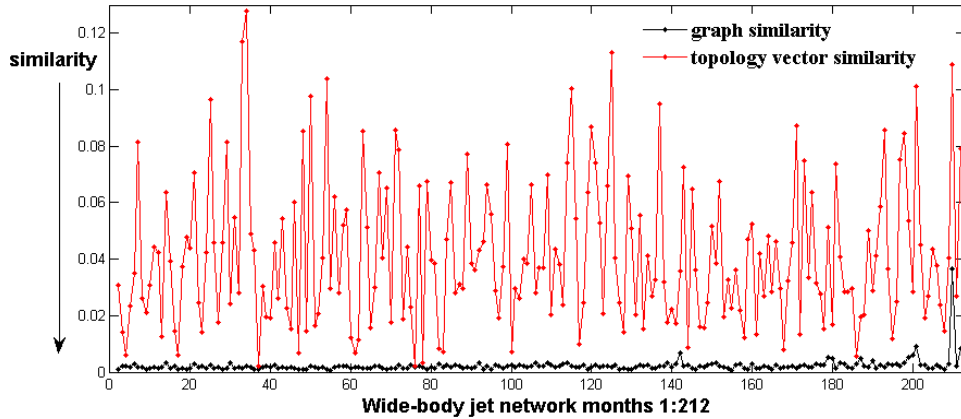


Figure 4-14: Wide-body jet network topology changes over 212 months from 1/1990 to 8/2007.

4.3 Comparing topology to canonical networks over time

This section tracks the topologies of the down-selected airline networks in terms of their similarity to canonical topologies. The canonical networks are the same set, reviewed in Chapter 3 (Section 3.3.2) for the topology profile discussion. For every time snapshot of the airline network (an instance of a graph), the corresponding set of canonical networks is created, with the same specifications (number of nodes, edges, or density, or degree distribution, or number of modules) depending on how the canonical network is generated. Then the real snapshot is compared to the canonical snapshot via a Euclidean distance of their topology vectors. So the comparison is point-wise for every month of the history of the airline. Since some of the generative algorithms are random, the topology history plots show stochastic averages of the resulting comparison, not single runs of these algorithms.

Figure 4-15 shows the topology comparison for JetBlue airlines for 91 months from 2/2000 to 8/2007. The plot shows two major topology transitions for JetBlue - one in the vicinity of month 22 (22-27) and one around month 70. The early transition is from a pure star network to a non-star (s-max jumps from 0 to 1). The second transition is away from a BA graph and towards hierarchies, which means that the topology has more interlinking or links between spokes, rather than pure hub-spoke edges. Figure 4-16 shows the network plotted geographically for months 23, 70 and 91. The snapshots show the transitions clearly: month 23 has just added a flight to Washington Dulles out of Fort Lauderdale, to break the pure star around JFK. In month 70, Boston Logan, Long Beach (LA) and Fort Lauderdale (FLL) have emerged as secondary hubs. The third 'phase' shows the topology moving away from hierarchical trees with no better match. This means that the change is not modeled well by any of the canonical networks. The snapshots plot shows (Figure 4-16) month 91 to be more interlinked than month 70, with all hubs experiencing growth, but unevenly. In particular, BOS has grown faster than Fort Lauderdale (FLL) and Long Beach (LGB). Also, the bi-partite pattern discussed in Chapter 2 has emerged.

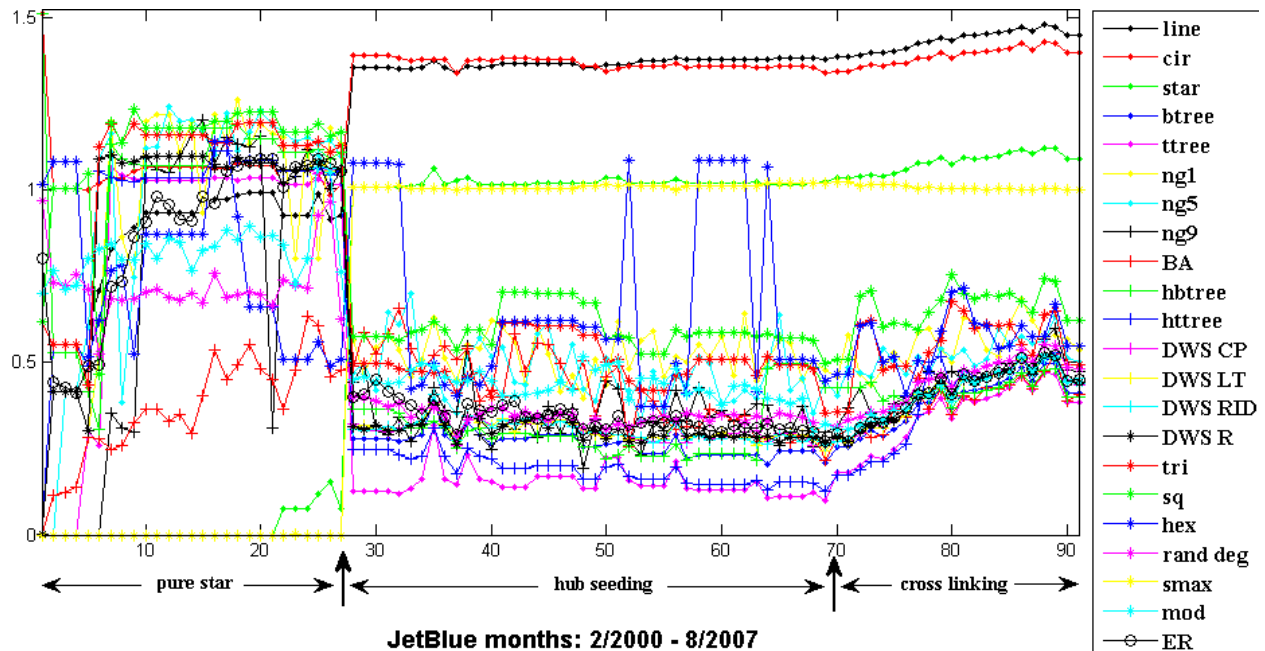


Figure 4-15: JetBlue topological vector distance to canonical networks. Time period of 91 months from 2/2000 to 8/2007. Two transitions around months 22 and 70.

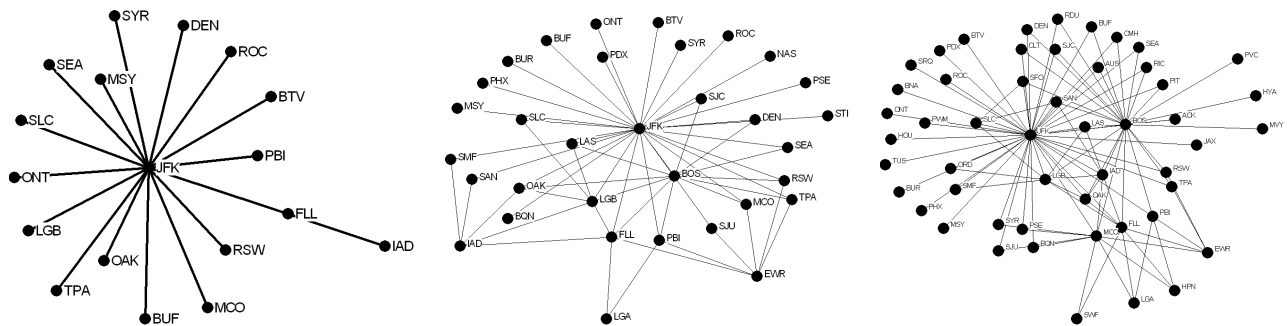


Figure 4-16: JetBlue months 23, 70 and 91

The patterns observed in the history of JetBlue airlines suggest a natural simple growth model for hub-spoke networks. There are three stages. The airline starts out as a star and remains so for awhile, then two (or more) spokes connect, probably in reality depending on market demand or airline strategy. The spoke connectivity strength can grow and single out one of the spokes as a potential new hub. If that happens this new hub starts to add connections at a higher rate than the average spoke¹, and these are connections of two types: to existing spokes of the older hub and to new (its own) spokes. The new-spoke formations become the star subgraphs, while the old-spoke connections are the reasons for the ubiquitous bipartite graphs in the network. We will call this model the "hub seeding growth model".

The case of Southwest is different. No major topology transitions can be seen, even though, as seen in the previous section, the airline is growing in number of total departures and seats offered.

¹The term *spoke* here is used more generally than a single-degree node to mean a non-hub.

Perhaps major transitions can be detected in the early 70s when the airline was founded. Southwest also carries more passengers by orders of magnitude since 1990. This is contrary to the observation that topology changes with growing size, and it will be interesting to find out whether one topology can handle growing traffic indefinitely, or what the limits are. The legacy carrier data suggests that the topology will stabilize, but the answer could be different for a network that can grow much larger than 200 nodes, and which does not have the constraints of transportation systems.

As far as closeness to canonical networks, Southwest is most similar to hierarchical binary trees, DWS graphs and all the variations of random graphs. The top matches are Dodds-Watts-Sabel type graphs or hierarchies with interlinking - among those random interdivisional interlinking (RID), core periphery (CP) and hierarchical tertiary tree. Figure 4-17 shows a scaled plot of only the best matches for the sake of clarity. Despite the lack of strong signatures, two weak transitions can be detected by looking at the similarity to hierarchical binary trees, which is weaker originally, stronger mid-way (late 90s) and then weak again in recent past. In the first phase, months 1-45 roughly, random graphs dominate the similarity; in the second phase, months 45-150, the network is closest to hierarchies with interlinking and then goes back to random graphs, but more weakly so. These patterns only suggest the inadequacy of the simple canonical topologies in comparison. Figure 4-18 shows the network plots of graphs in the three periods of time. Though they look random to the eye, the strong structure of local interlinking can be discerned, as the graph appears to get denser over time. It is clear that Southwest has a completely different growth dynamic than the patterns observed in the history of JetBlue.

In order to look at earlier Southwest history and test this statement, we added 5 data points of single months from the 1980s, found in visual format online: 7/1980, 2/1982, 2/1983, 1/1984 and 4/1988. Figure 4-17 and Figure 4-26 show these as the first 5 points in the timeline. The topology comparison indicates that these earlier stages of the airline do not look much different, i.e. there are no topology transitions. To further justify this statement, data from the 1970s is necessary.

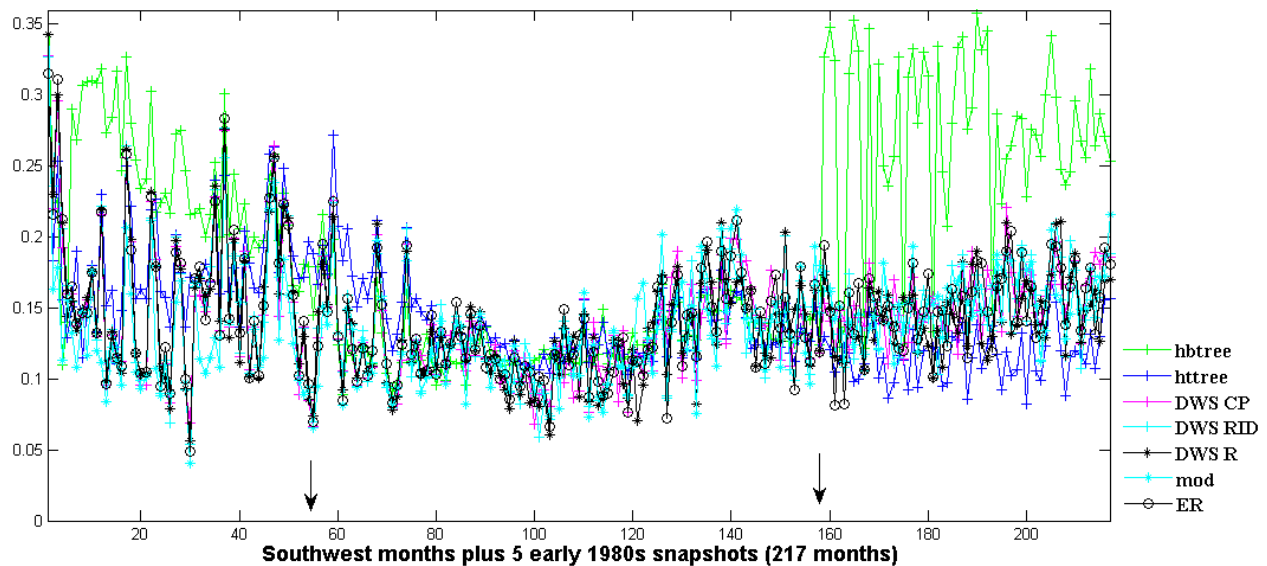


Figure 4-17: Southwest topological comparison to canonical network over the period 1/1990-8/2007. Weak topological transitions at month 45 and 150. 217 months including the 5 1980s snapshots: 7/1980, 2/1982, 2/1983, 1/1984 and 4/1988.

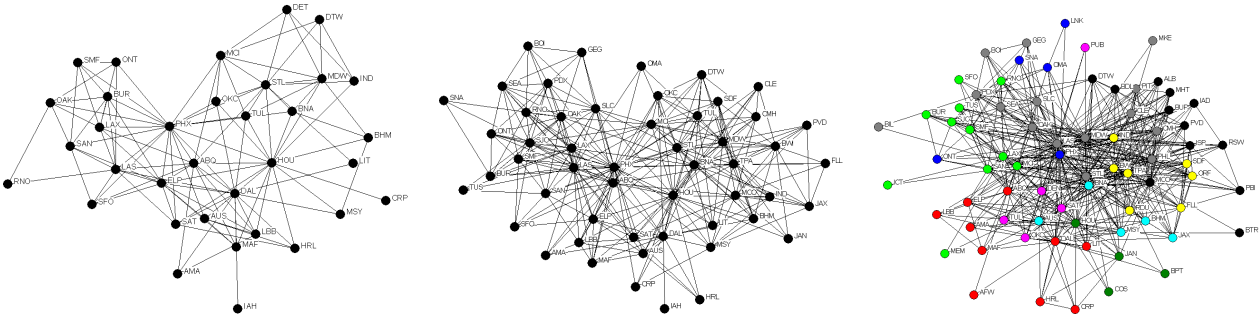


Figure 4-18: Southwest months 24, 96 and 212

The history of the wide-body jet network is quite steady. Across the studied time period, this is a preferential attachment-like (BA) network (Figure 4-19). The next best match is a tertiary tree. Even though the network does vary in size, its topology seems to remain constant over time. This finding confirms the conclusions from Section 4.2 where comparisons of consecutive snapshots of the graphs showed constant topology. The wide-body comparison plots are shown on Figure 4-19. The Continental story is almost identical, closest to preferential attachment graph and tertiary tree, consistently over time (Figure 4-24), with the exception of a hierarchical tertiary tree-like topology around month 60. This confirms earlier indications of similarity between Continental and the wide-body jets. While JetBlue was seen to have the same motif patterns and power-law-like degree distribution, its history looks rich in topology oscillations, compared to the consistency of Continental, the wide-body jets network and even Southwest. This could be related to topology transitions related to growth, which a legacy carrier will not experience in later stages of its history.

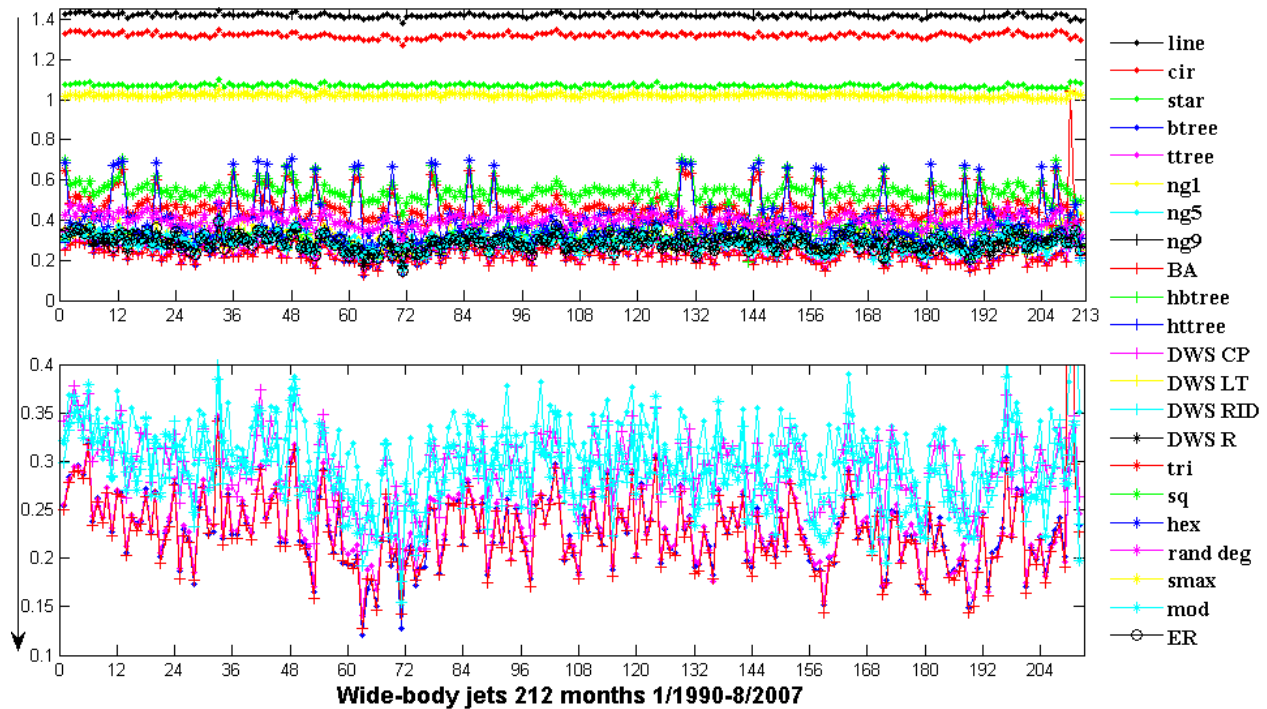


Figure 4-19: Wide-body jets topological vector similarity to canonical network. 212 months; (top) all matches, (bottom) closest matches only - BA graph (BA), tertiary tree (ttree), binary tree (btree), Newman-Gastner graph (ng5), hierarchy with random interlinking (DWS R).

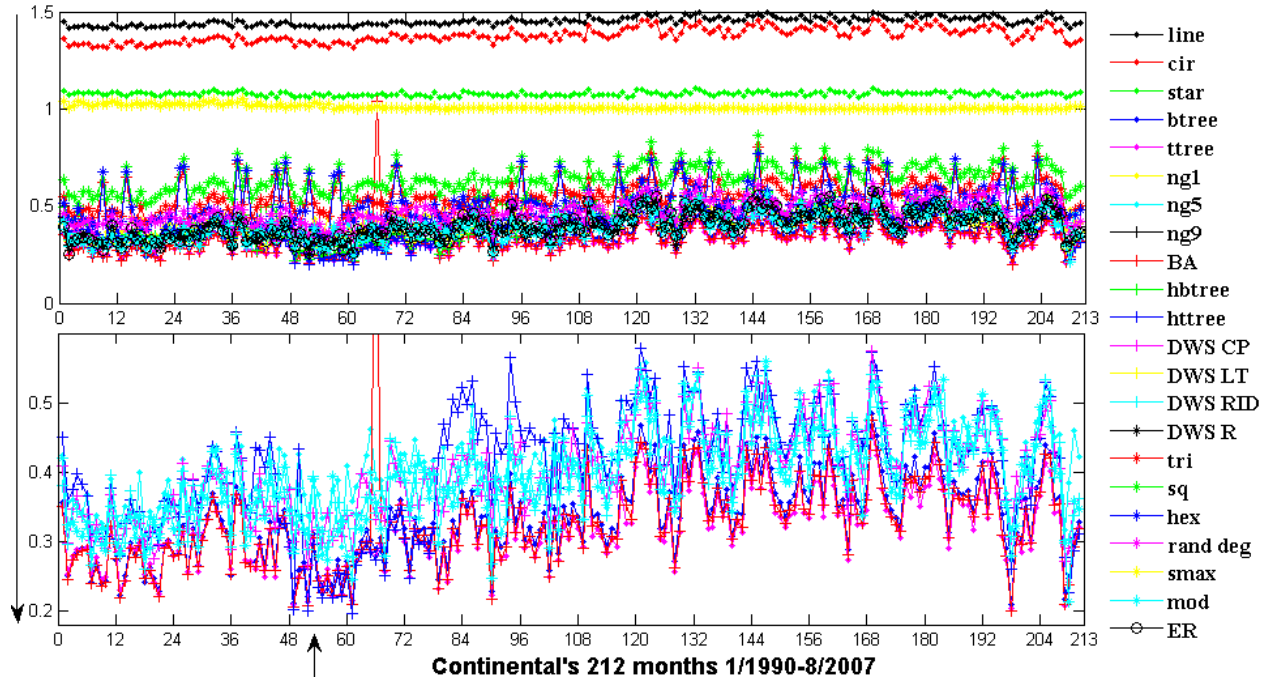


Figure 4-20: Continental Airlines topological vector similarity to canonical network. 212 months; (top) all matches, (bottom) closest matches only - tertiary tree (ttree), BA graph (BA), binary tree (btree), random modular graph (mod), Newman-Gastner graph (ng5).

4.4 Topology-derived growth models

In Chapter 1 we reviewed generative network models from the network theory literature. They fall into several categories: random graphs models (Erdős-Rényi [39]), node-degree based (BA etc [14]), edge-centric models (hierarchies, Dodds-Watts-Sabel [9]), spatial distribution algorithms (Newman-Gastner [42]), node and module copying (gene duplication and change propagation [13]) and econometric models (MITRE [58]). The growth algorithms proposed in these models were used to create a spectrum of canonical topologies with certain characteristics, such as preset number of nodes, edges, number of modules or degree sequence. The previous section showed how well these topologies match the airline networks studied over time, using the topology vector as a point of reference. The emerging patterns over time are used to devise two simple growth models, corresponding to the two distinct phenomena in the airline industry: the "hub seeding" and the Southwest phenomenon. One of the key findings was that there are three families of predominant motifs in airline networks - stars, base-triangles and bi-partite graphs. Section 3.3.4 discussed how these motifs suggest a hub-seeding dynamics. The Southwest model is derived from the lack of "hub seeding" motifs, i.e. higher randomness, and the finding that topologically, the airline is closest to a randomized hierarchy, with strong local linking.

In this section, we review these two topology-derived models and show that they outperform the canonical topology models. In many ways, this is expected, because these topology-derived models were tailored to the patterns found using canonical networks and motifs. What is exciting is that the two models, are simple, probabilistic, and despite the fairly rough probability estimates, outperform the other models consistently over time. So while they may not be "generative" (i.e. actually generating the real-world topologies), they match the evolutionary patterns and have ex-

planatory power. An explanation here means understanding the dynamics confirmed by various statistical measures, topology measures and underlying motifs. To fully understand an airline, one would have to look into strategy, market conditions, fleet and so on. This is beyond the scope of this thesis.

The hub seeding growth model

This model is intuitively derived from the patterns observed in the topology evolution of JetBlue Airlines, though it is likely to perform well for other “hub-spoke” airlines. It consists of three stages, of which the last two can be iterative. First, the network starts out with a pure star topology (one hub and many spokes). In the second stage two spokes connect to each other, and depending on popular demand one of these spokes has the potential to develop into a new hub. Notice that, a *spoke* here does not mean simply a single-degree node, but more generally, a non-hub. A spoke becomes a hub after it receives a certain threshold of links, which can be different for different networks. The third stage models the growth of the new hub - getting its own new spokes or connecting to spokes of the old hub. This is the entire model in summary, with the potential to continue hub development indefinitely as the network grows. The unknowns in this model are the probabilities (or frequencies) of addition of new spokes, p_1 , the probability of two spokes connecting to each other, p_2 , and the chance of a spoke to become a hub, p_3 , and the probability of a new hub of connecting to old spokes, p_4 . Overall, this accounts for four probability variables. Table 4.4 shows the outline of the model.

Table 4.3: Outline of the hub seeding growth model.

phase	events
(stage 1)	pure star topology; spokes added with probability p_1
(stage 2)	two (or more) spokes connect with probability p_2
(stage 3)	one of the (highly) connected spokes becomes a new hub with probability p_3
(stage 4)	new hub connects to old hub’s spokes with probability p_4

Deciding what the four probabilities in the model are can be done in various ways. Initially, we experimented with some common sense estimates such as $p_1=0.5$, $p_2=0.4$, $p_3=0.4$, $p_4=0.4$, which means roughly 40% of any event happening. A more informed way is to count the events occurring in the history of the airline, using some simple assumptions, and then computing the frequencies of 1/new links arriving, 2/two random spokes connecting, 3/ a potential hub developing into a new hub and 4/new hubs connecting to old spokes. A node becomes a hub after it receives a threshold number of links. For JetBlue, for the example, the threshold was set at 5, based on knowledge of the network in 8/2007. Obviously, a hub with five connections is a small secondary hub. In Southwest’s case, most nodes have more than 5 links, so a higher threshold makes sense. Table 4.4 shows the probabilities found by counting events in JetBlue, Airtran, Spirit, Southwest and Continental. As far as the frequencies found, the higher the threshold, the more likely the potential hub will develop into a new hub. Also new hubs connect to old spokes roughly half of the time. Random spokes connect with higher probability for larger carriers, because they can probably afford to jump into smaller markets and pull out more often. While these observations are interesting, the frequencies counted in Table 4.4 are averages over 212 months of (in some cases) different periods and environments of growth, so they are not ultimate indicators of airline strategy. That said, this type of estimating the probabilities and feeding them into the model, is common supervised learning approach and a good start at calibrating the hub seeding model.

Table 4.4: Counting frequencies of events according to the hub seeding model in various airlines.

Airline	hub threshold	p_1	p_2	p_3	p_4
JetBlue	5	0.473	0.138	0.554	0.462
Airtran	5	0.484	0.217	0.570	0.484
Spirit	5	0.328	0.125	0.352	0.323
Southwest	15	0.544	0.426	0.929	0.544
Continental	10	0.495	0.505	0.99	0.495

Figure 4-21 shows the pseudocode of the hub seeding model. This model was used to generate snapshots of graphs with the same number of nodes, as the months in airlines history and compared over time to the other canonical topologies.

```

initialize graph, hubs, spokes, probabilities
while total number of nodes < n

    • add new spoke with probability  $p_1$  preferentially to larger hubs, i.e.
      probability  $\frac{p_1}{h}$ , where h is the hub order by decreasing size; add node to
      spokes
    • select two spokes at random; connect with probability  $p_2$ 
    • find all nodes with threshold degree that are labeled as a spoke; of those
      select a random node and promote it to hub with probability  $p_3$ ; add new node
      to hubs, and remove from spokes
    • if there are no new hubs, continue; else: select random spoke (not of new
      hub); connect to new hub with probability  $p_4$ 

return graph

```

Figure 4-21: Pseudocode for the hub seeding model.

Figure 4-15 shows the result for JetBlue. Canonical topologies are compared to the real topology via the topology vector, as in Figure 4-15 including the hub seeding model snapshots. As mentioned before, the comparisons on this plot are statistical averages of running the comparison multiple times (in this case 50). The top of the plot shows the trends overall, as discussed before. The bottom plot concentrates on the best matches only - hub seeding, BA, tertiary trees pure and hierarchical, followed by hierarchies, (DWS graphs). The hub seeding model performs really well over time. The pure-star beginnings of JetBlue are a fact, and the probabilities are not adjusted to reflect that (since they are averages over the whole period), so the model does not do well (neither does the BA) in the first 22 months. It is encouraging that over time the match stays consistently better than other topologies, which means that the hub seeding mechanism with star and bi-partite graph formation may really be the right idea. Experimenting with other sets of probabilities (not strictly from the history of JetBlue) also shows better performance than the other topologies, signifying that the principle of growth is stronger than fitting the model precisely to the data.

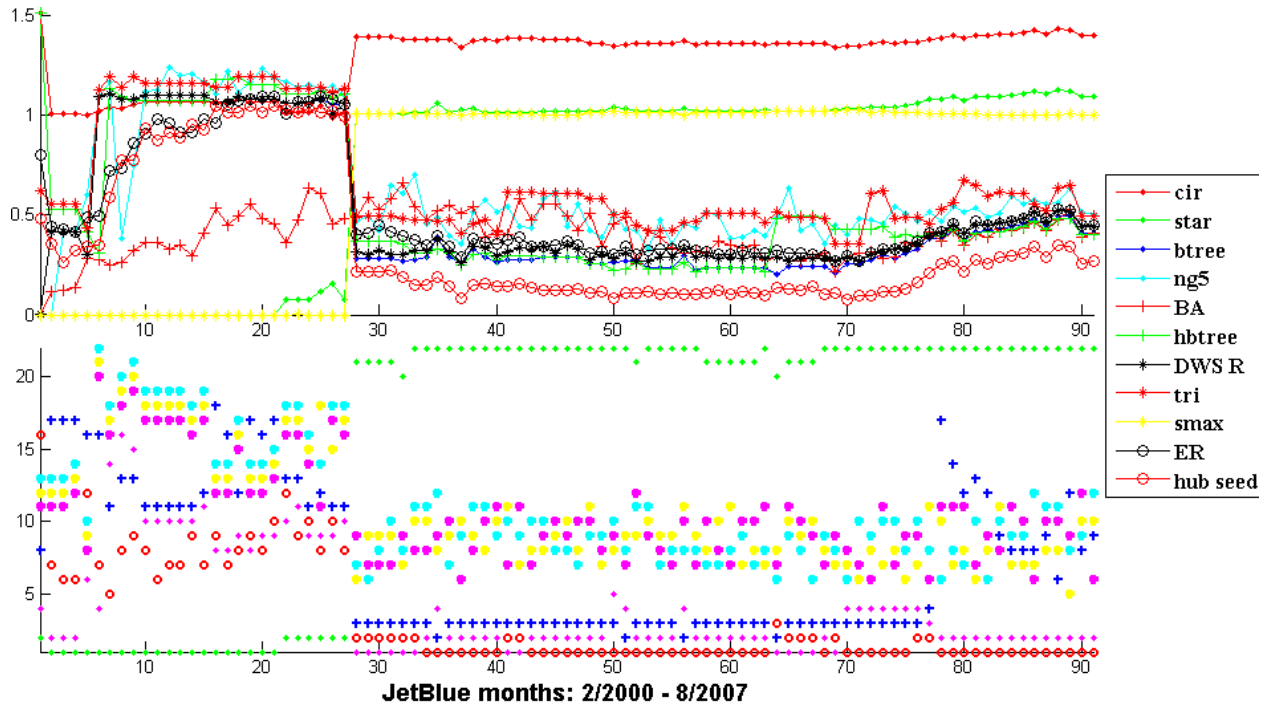


Figure 4-22: Topology over time comparison: JetBlue topology comparisons with the new hub seeding model (red circles). The hub seeding model outperforms the rest of the canonical topologies for most data points, and consistently for last 15 months. The top plot shows the topology vector distance to the real graph; the bottom plot shows the topology rank based on that distance.

Given that this model was derived from the patterns found in JetBlue’s structure, it is debatable whether it is universal. We use the probabilities shown in Table 4.4 to also apply it to Airtran’s and Continental’s histories. Airtran was chosen as another low-cost airline with known hub operations. Continental is chosen as a representative of the legacy carriers and one of the downselected airline networks for further study.

Airtran shows similar history to JetBlue’s - early transitions from pure-star-like topology and s-max graph with closest match to the hub seeding model after month 66 (Figure 4-23). Here we do not concentrate on the fine details of the topology of Airtran but just use the general trends in Figure 4-23 to show the performance of the hub seeding model.

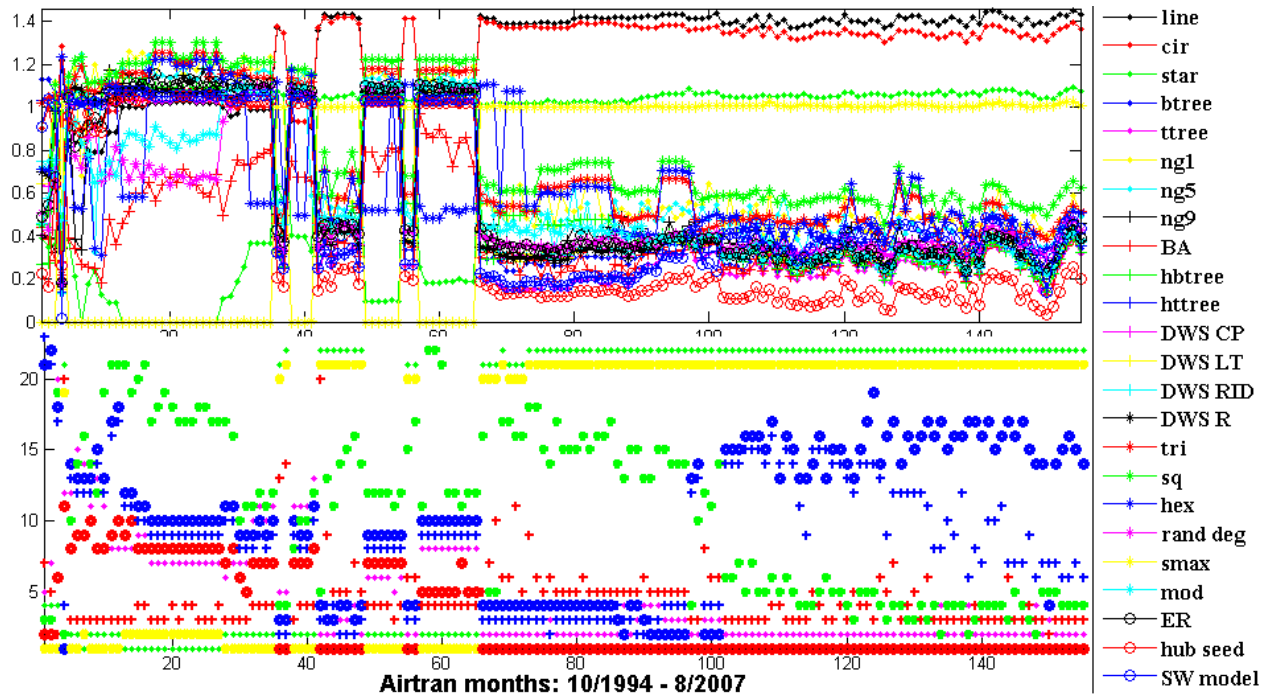


Figure 4-23: Topology over time comparison: Airtran topology comparisons with the new hub seeding model (red circles). The top plot shows the topology vector distance to the real graph; the bottom plot shows the topology rank based on that distance.

The hub seeding graph snapshots are the top match across the entire history of Continental. Figure 4-24 shows the overall pattern with the new model a notch better than other canonical topologies. The conclusion is that this model fits some real data, and certainly while not all airlines grow by this mechanism, it matches well patterns and statistics in the history of "hub-spoke" airlines such as JetBlue, Airtran and Continental.

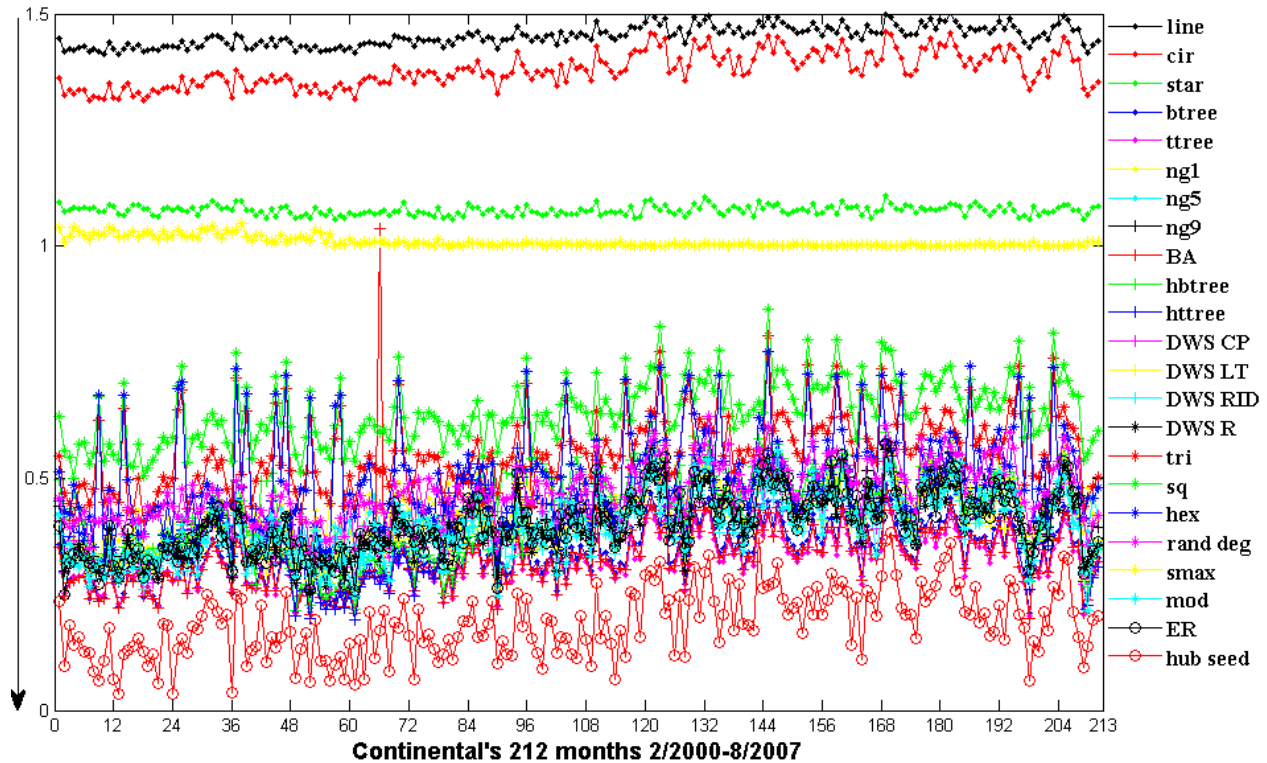


Figure 4-24: Topology over time comparison: Continental's topology best matches over time including the new hub seeding model (red circles).

The Southwest expansion model

(Growth hierarchical trees with strong local and weak random interlinking)

This growth model is called the Southwest model because it is unprecedented in airline growth, and because it is devised by observing the growth of the Southwest Airlines network. It is not as intuitive as the hub seeding model. In fact, it is hard to imagine why an airline would decide to operate as Southwest does, looking at pure topology only. Before we discuss the airline strategy, we will show that the model fits the topology evolution very well. In summary, the network expands as a tertiary hierarchical tree locally. As soon as a new "leaf" (airport) is added, this airport is connected to all airports within some threshold proximity (2-3 hops away, for example). This creates a natural *duplication* of patterns, similar to the ideas in the duplication-divergence model explained in Section 1.4.4. On top of that, at every step, random interlinks are added from various airports, with slight bias towards high-degree nodes (profitable markets). Table 4.4 shows an outline of the main steps.

Table 4.5: Outline of the Southwest expansion model.

phase	events
(stage 1)	create a hierarchical tree backbone structure with the desired number of nodes
(stage 2)	add multiple local connections from all spokes (or leaf nodes), connect the same way as “geographically similar nodes”, (for example connect directly to all nodes within 2-3 hops away (or the hops/distance can be a percentage of the number of airports) to reflect growth
(stage 3)	add random interlinks with bias towards high-degree nodes
(.....)	iterate the local expansion, with new local connections and random interlinks

Figure 4-25 presents the pseudocode for the Southwest model routine. There are two probabilities and one parameter associated with the model. In stage 2, nodes are linked locally within a certain distance threshold. We chose $0.05n$ because as the network grows, the notion of “local” should also change, but slowly. Five percent of 50 nodes is 2.5, so a small starting threshold. This is still a parameter, and can be varied in more detailed studies. The only real probability is the probability that the two local nodes will receive a direct flight. We chose that to be 0.6, higher than average yet not as high as certain, because otherwise, that would create an almost complete graph. The last probability is associated with random interlinking across the network. A random link between nodes i and j is added with probability $\frac{deg_i + deg_j}{2deg_{max}}$ which puts weight on popular routes.

```

initialize graph by building a hierarchical tertiary tree,  $n$  nodes
while total number of edges  $< m$ 

    • calculate all paths in the graph (with a Dijkstra routine)
    • for all nodes with a distance between them  $\leq 0.05n$  (5% of network size,
      since  $n - 1$  is the largest possible diameter) connect the two with some high
      probability ( $>0.5$ )
    • compute the degrees of all nodes; select two random nodes
    • connect the two nodes with probability proportional to their combined
      degree, i.e.  $\frac{deg_i + deg_j}{2deg_{max}}$ 
    • if number of edges is  $\geq m$ , return graph

return graph

```

Figure 4-25: Pseudocode for the Southwest expansion model.

Figure 4-26 shows that this model outperforms the other canonical topologies during the period 1/1990-8/2007 for Southwest. This makes sense because it combines hierarchy with randomness and matches the most recent concentrated traffic patterns at hubs like Chicago. The last point was illustrated in Chapter 3, Figure 3-43 where we plotted the density of high-seat capacity flights and showed that Southwest does concentrate capacity in hub-like operations, not unlike the other carriers.

The strategy behind this model is not obvious and not entirely based on topology. Part of the local linking has to do with the way Southwest schedules aircraft rotations. On many occasions a single aircraft performs a string of 3-6 flights in a chain-like pattern, between cities that are on average 1.5-2 hours away. Those may be numbered as the same flight, or as different flights. The local pattern has to do with the single aircraft Southwest operates - the 737 which does really well

short-haul. But that cannot be the only driving factor that determines that network structure.

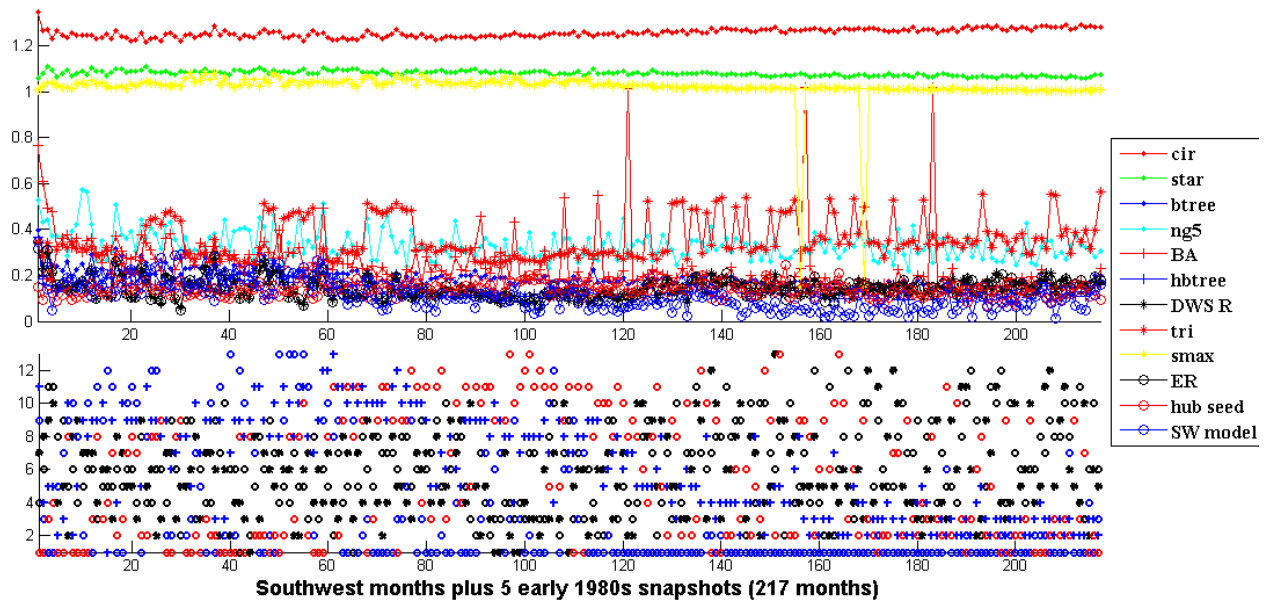


Figure 4-26: Topology over time comparison for Southwest Airlines. Comparing the "Southwest model" to canonical networks. The bottom plot shows the best topologies only. Includes the 5 months from the 1980s. No major topology transitions are seen between the 1980s and 1990s growth.

The logic behind this model is not obvious, because the topology we see is only a signature of more complex considerations. Part of the local linking has to do with the way Southwest schedules aircraft rotations.

The closest match to the Southwest topology are hierarchies with random interlinking. So there is an underlying grid, some random operations, with some general guidelines, and more recently heavier traffic on profitable routes. But that recent strategy has not been consistent. Since 1990 traffic share has been fairly randomly distributed across the network. One possible explanation is that the airline curbs cost and captures a great passenger segment so well, that it can afford to try and keep or leave different routes in a dynamic random-like way. The history certainly shows that this is how the airline operates. The slow tree-like geographic expansion is also a fact. The fact that the airline flies to secondary airports mostly (not exclusively) cannot be correlated with the random-like flight pattern, but most likely is a huge contributor to cost. In some cases, the airports are far enough (to be easily connected by public transportation) that they capture different market segments than the main airport itself (ex: Logan and Manchester), so that Southwest does not compete directly with major airlines.

Figure 4-27 shows the topological history plot, but for the Southwest top 80 flights every month. Top 80 means that a leg is in the network if it offers a seat capacity within 80% of the top seat capacity flight in the network. The bottom of the plot shows the best match topology only. As expected, the top 80 percent of the network looks a lot more centralized. For the last 50 months, the hub seeding topology dominates the history.

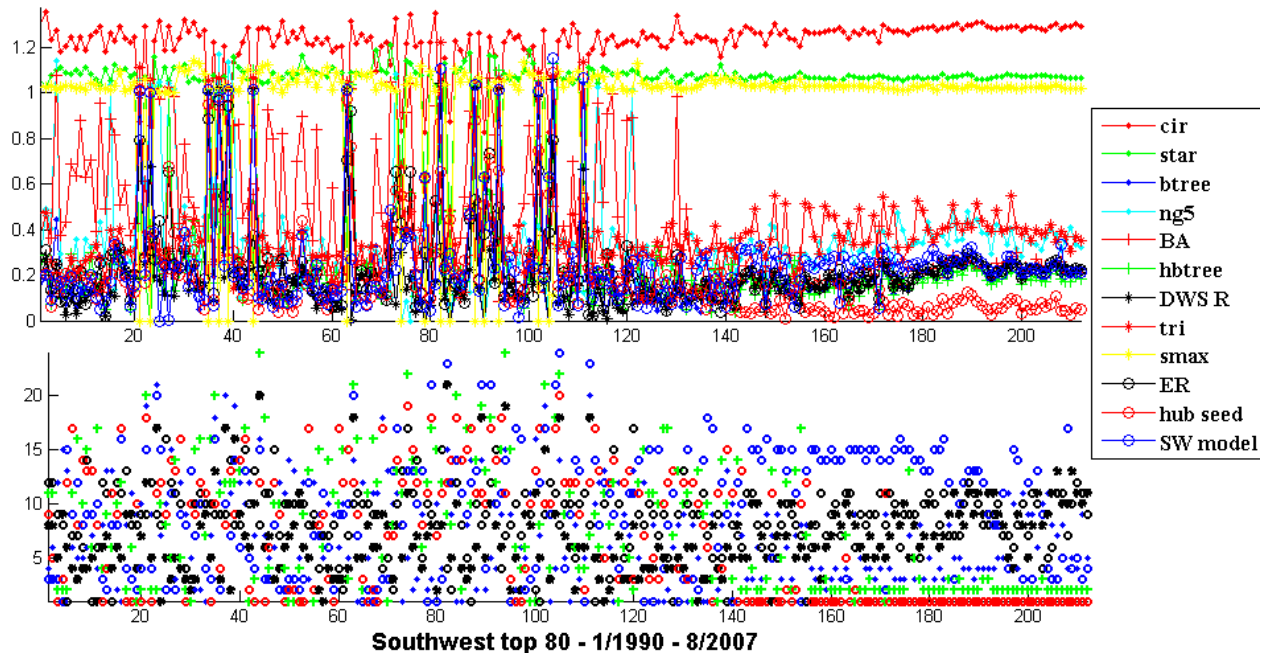


Figure 4-27: Southwest top 80 topology evolution: 1/1990-8/2007. Hub seeding model clearly outperforms the rest as expected, in the last 60 months. The bottom plot shows topology rank (rather than vector distance) - which canonical graph (or model) is number 1, 2,... 24.

This result is an indication that Southwest is getting more centralized, possibly due to dealing with cost and operations getting complex as they grow. The random-like flight patterns on top of this centralized network are possibly ways for the already profitable airline to probe new markets. Those are either maintained or abandoned depending on profitability. There are other cases, where the airport may not attract business or get permission for expansion (ex: Detroit).

4.5 Conclusion

In this chapter we closed the arguments of the thesis and addressed all the research goals. The major findings related to topology in general and airline networks are summarized below.

Topology transitions do happen. These are usually seen in early stages of growth, ex: stars to non-star formations, but not exclusively. They are often associated with major changes in network size, up or down. Steady topology over time is also observed, but not usually in the beginning of an airline. This means that the legacy carriers exhibit stable topologies. This is true also for large data slices of the airlines dataset, such as the wide-body jets. Topology also transitions at major events (ex Sep 2001), and cyclically with seasonal traffic patterns. But it is also seen to remain stable after major market fluctuations. For example, Continental sees stable topological history, but overall decline in the last 7-8 years with major downsizing. This means that there are other subsets of the system, other than its structure that can be a buffer for financial trouble or other changes.

Simple topology-derived probabilistic growth models outperform "canonical" models. Looking at the recurrent patterns in airline networks, we devised two simple models for growth, the hub seeding growth model and the "Southwest expansion model". With both being probabilistic, we

used statistics to learn the probabilities of event occurrence for the first model, and used estimates for the second. Though derived from JetBlue's history only, the hub seeding model is seen to perform well for Airtran and Continental Airlines, which are two examples chosen for benchmarking. The Southwest model is unique, as in, not applicable to any other airline, but it does capture the patterns in Southwest's history better than the other canonical models studied.

Southwest is an outlier by far, but simple in-depth analysis shows that the airline is similar in some way when the "noise" is filtered out. While operating various random "volatile" flights, its highest frequency flights resemble the hub-spoke operations of a traditional airline. Especially in later years. It could be that the airline is experiencing some of the cost-related pressure of other airlines and trying to concentrate logistics as well.

Graph-theoretical models can explain patterns in some airlines but not others (ex: Southwest) where more information is needed. For example, JetBlue Airways, Continental Airlines and Airtran have hub-spoke topology - which means that they have similar underlying patterns, (stars and bipartite formations) and favor the hub seeding model in their topological history. Southwest operates with a hierarchical backbone, a denser network overall, with many local links and random cross links that connect major markets.

Chapter 5

Evolution of Language Wikipedias

One of the goals of this thesis is to show that the techniques for analyzing systems represented as networks and their evolution are generalizable. This chapter addresses this goal by applying the same tools used to analyze airline networks in Chapter 3 and Chapter 4 to language Wikipedias. The Wikipedia is a web-based encyclopedia which is administered by a small number of people, and can be edited by anybody. The content is peer-reviewed. The Wikipedia project started in 2002 and has been very successful since, with Wikipedia becoming one of the prime sources of online information. A Wikipedia grows by the addition on new articles, i.e. pages with a unique title. Pages have content that is relevant to the title, and which contains references to other relevant articles. We consider the network of articles and references, in the form of hyperlinks, between them. Other than the representation, there is no similarity between this dataset and the airline networks. The goal is to show that using the same tools and some understanding of the content of Wikipedia pages, one can gain insight into the structure of Wikipedias and how they grow.

A language Wikipedia is a Wikipedia in only one language, Russian, Spanish, or English. Of course, different Wikipedias are connected to each other, but what makes them interesting separately is that they are developed by unique communities (overlap is possible) and develop slightly differently, as they cater to different cultures and ways of representing knowledge. Splitting Wikipedia by languages also allows the analysis of manageable datasets [61]. As of November 30th, 2008, the English Wikipedia is cited to have 2 640 206 articles and 7 times more individual pages, with more than 8 million registered users (you have to be registered to make edits). With the computational resources available, we could not consider the English Wikipedia, but managed to download histories and analyze some of the largest Wikipedias, such as French and Spanish, but for only approximately the first year of their history. Table 5 is summarized from [62] and lists the Wikipedias considered in this work, with some major statistics, such as number of articles and number of registered users.

Table 5.1: Table of Language Wikipedias analyzed in this thesis. The rank is by size among all Wikipedias. The code is used for the web address of the Wikipedia. The depth column is a rough indicator of how frequently the Wikipedia is updated.

rank	language	code	articles	users	depth ¹
3	French	fr	744954	517059	115
6	Italian	it	526002	337229	57
9	Spanish	es	429139	917533	93
10	Russian	ru	342956	204233	74
12	Chinese	zh	216431	560308	71
22	Esperanto	eo	108216	11830	13
39	Simple English	simple	44467	39270	21
106	Interlingua	ia	4381	2319	21

A Wikipedia, represented as a network, is a set of articles, which are connected to each other via hyperlinks. There are many more pages than articles in Wikipedia, so this difference was carefully accounted for. Other pages are about changes, summaries, stubs, authors' forums, profiles, history and admin pages. For the purposes of clarity and the desire to analyze these as "knowledge networks" we only consider article pages (they have a topic/name and address: `xx.wikipedia.org/Topic`, where `xx` is the language code). The data dumps for each Wikipedia were downloaded directly from [61]. The data mining code and parsed networks are courtesy of Dimitar Bounov [63].

5.1 The Language Wikipedias: Early Growth

Growth rate was a key factor in selecting Wikipedias for analysis. The early growth of these networks can be extremely slow, or highly discontinuous. The sudden addition of an entire topic by simply creating header pages (to be filled) can increase the size suddenly and also create a dominant connected component. There are examples of Wikipedias staying "dormant" for a year before many pages get written, as in the case of the Italian Wikipedia. This means that for many months there are very few articles (order of 10) and they stand alone, disconnected. Figure 5-1 shows the growth of the Wikipedias selected for analysis.

¹The depth is calculated as $(\text{Edits}/\text{Articles}) \times (\text{Non-Articles}/\text{Articles}) \times (\text{Stub-ratio})$ [62].

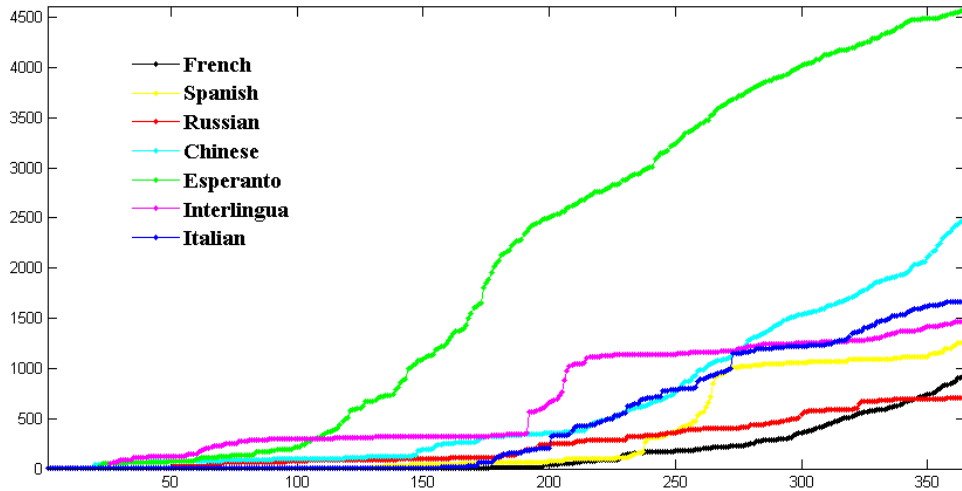


Figure 5-1: Number of articles versus days: first year growth in number of articles of the French, Italian, Spanish, Russian, Chinese, Esperanto and Interlingua Wikipedias. The Esperanto Wikipedia grows the fastest, though is currently outgrown by all the active language Wikipedias. All other Wikipedias see very slow growth in the first 200 days. The Spanish Wikipedia actually gets its first edge on day 285, and has few disconnected pages prior to that, but was plotted from 285 to 650 days to show comparative growth.

Most of the Wikipedias on Figure 5-1 have slow initial growth and then they take off, with the exception of the Chinese and the Esperanto Wikipedias. The Esperanto grows fastest, with almost no plateaus, but this growth must slow down, since the Wikipedia remains 22nd, behind the rest (except for Interlingua). All other networks grow continuously, with some jump growth for the Spanish and the Interlingua wikis, which will be discussed in more detail in this chapter.

Link growth is shown in Figure 5-2 by plotting the average number of links per article (edge to node ratio). By that measure, all networks remain fairly sparse, with a maximum edge-to-node ratio for the Italian Wikipedia of 9. For the early stages of slow growth, the average links per node is very low, which means most Wikipedias start out as trees, or forests (collections of disconnected trees). The edge-to-node ratio below indicates many single nodes, i.e. articles without any connections. More unusual are the leaps of links per node, especially in the case of the Italian Wikipedia, but also the French. This occurs when a set of nodes arrives with a dense local structure, such as a clique, or a set of nodes, that are fully related, hence introduce a dense new cluster (or subgraph of the network, ex: months of the year). This statement is supported by the fact that the density decreases with time, as more nodes arrive, because their arrival with fewer links outweighs the local density. Yet, it is interesting to understand why some Wikipedias grow a lot denser than others. Given that many users are allowed to contribute, the cognitive limit of 7 ± 2 may not be a constraint in the number of associations.

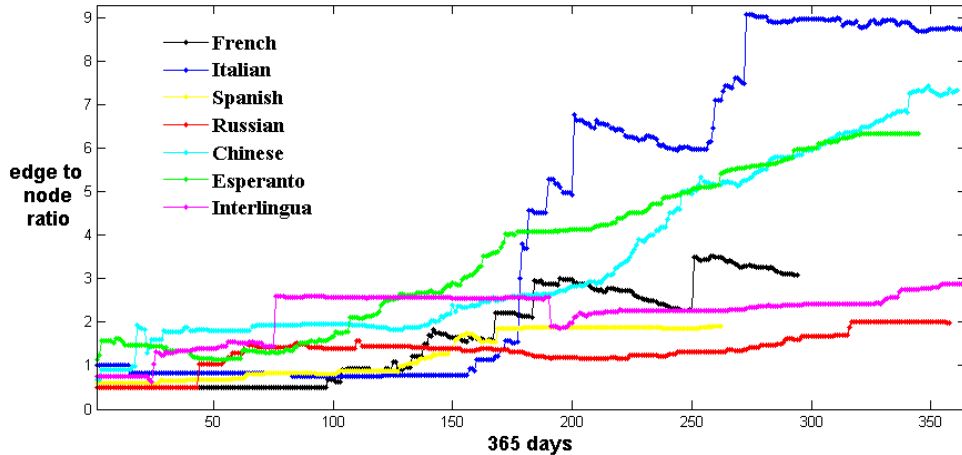


Figure 5-2: Average number of hyperlinks per article for the first year of the French, Italian, Spanish, Russian, Chinese, Esperanto and Interlingua Wikipedias. All Wikipedias are normalized to the same date (as if they started to grow together). For example, the Spanish Wikipedia does not see growth until day 285, so its timeline on this figure is day 285 to day 650.

To summarize, in this chapter, we study in detail the topologies and their histories of a set of large language Wikipedias, by concentrating on the first year of their active growth. In particular, these are the Interlingua, Esperanto, Spanish, Russian, and Chinese Wikipedias. Main takeaways for the Simple English, French and Italian are moved to the Appendix A.2 for conciseness.

The goal of this chapter is to test the tools developed for network analysis with a different example. The hope is to learn what the structure of Wikipedia networks is, and whether it is common among Wikipedias and different from airlines. More fundamentally, we expect that Wikipedia networks evolve differently than transportation systems. Studying growth of a cognitive network versus a technical (physical) network can potentially uncover novel principles that distinguish between these two types of systems.

5.2 The Interlingua Wikipedia

Interlingua is an international auxiliary language developed in 1930s-1950s and is probably the third most widely used international auxiliary language. The most popular such language, Esperanto, is analyzed in section 5.3. Interlingua is based on both Romance and Germanic languages, and is immediately understandable for at least the Romance language speakers. The Wikipedia's first pages appear in January 2003 and is claimed to have over 4000 articles as of December 2008. It is ranked 106th in the list of Wikipedias by number of pages [62], Table 5. Figure 5-3 shows the growth in terms of number of articles and number of hyperlinks daily for the first year of the Wikipedia's existence.

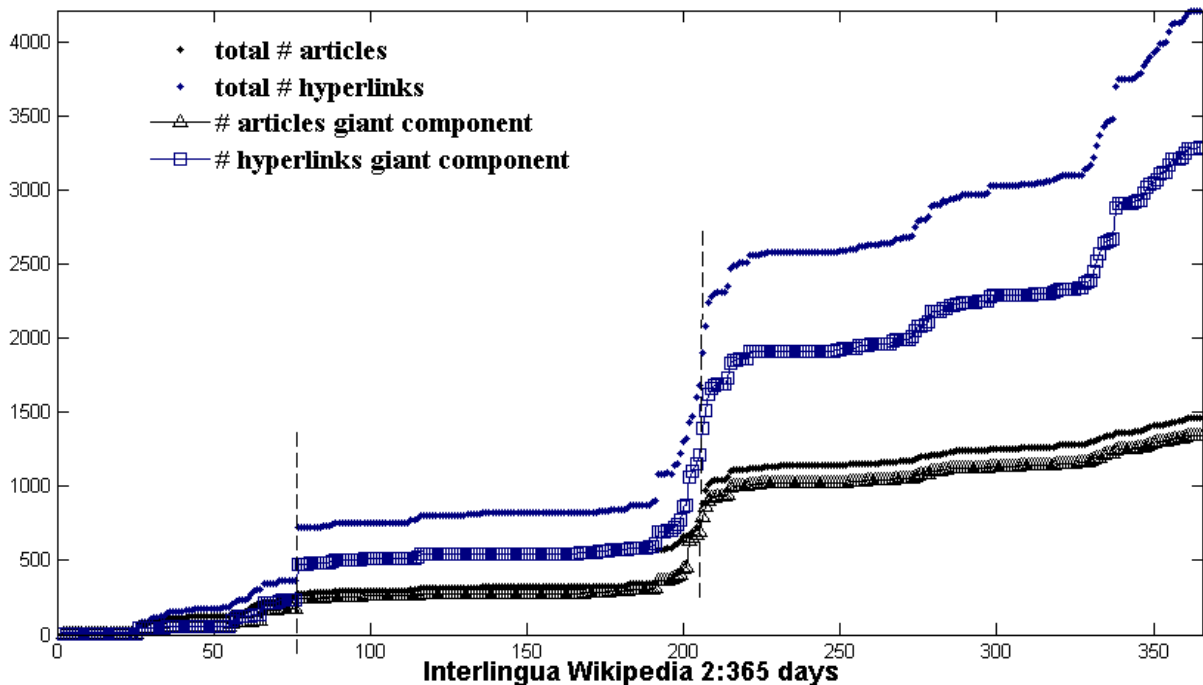


Figure 5-3: The first year of the Interlingua Wikipedia: total number of articles and hyperlinks and number of articles and hyperlinks in the giant connected component.

The data on Figure 5-3 shows that the Wikipedia is disconnected through the entire first year, but the size of the giant connected component is not far off the size of the entire network. There are two major transitions in growth, the first around day 75 and the second between days 200 and 210. Growth in the stages in-between is fairly steady. The first transition from day 65 to roughly 75-80 happens because of names of countries around the world are added (in addition to other "steady-growth" pages). The second transition (days 190-215) is related to the introduction of religion-related themes, from terms in Christianity (apostles, famous dates in the Bible) to articles about major creeds, affiliations (ex: asceticism, polytheism, esotericism). The theology-related wave is interspersed with medicine-related terms, actually all scientific terms with Greek roots.

From the yearly history, day 75 is chosen for analysis because it is of manageable size (255 articles, 179 in the giant component) and it marks a transition period in the history of the Interlingua Wikipedia. Figure 5-4 shows the actual network with all disconnected components. Two of the larger disconnected clusters are on topics Space/Astronomy with articles such as "sun", "satellite", "solar system" and Science, introducing what major sciences study. The numerous star formation in the giant cluster is centered around the word "language" and features major languages.

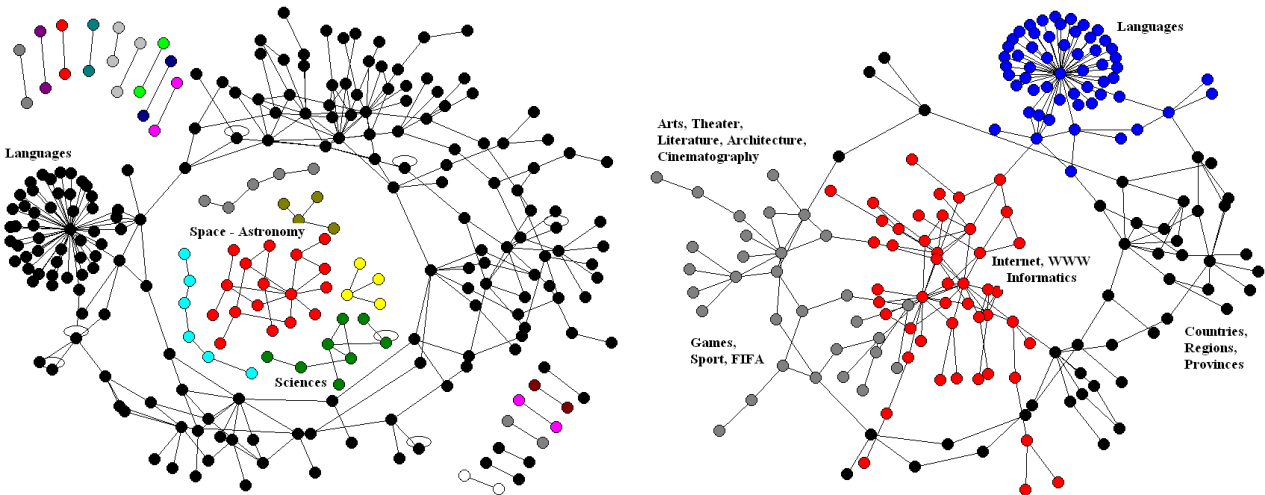


Figure 5-4: (left) Day 75 of the Interlingua Wikipedia. The star cluster is centered around the topic "languages". The two smaller components in the middle contain articles about space/astronomy and sciences. (right) Network modules of the giant component of day 75 (of the network on the left).

Interestingly, this Wikipedia is easily modularized, meaning that the Newman-Girvan algorithm, for example, gives high scoring (Q-metric) to the modules found. Even more interesting is that different modules make sense thematically. Figure 5-4 (right) shows a four-module split of the giant component for day 75. As point out earlier, the star cluster is centered around the word "language" and features different world languages. The middle component is about web, internet and informatics-related articles. The articles related to countries, and geographical regions (added in the first major transition) appear as a connector between informatics and languages. Another connector is a module about the arts, and games/sports. This picture supports the conjecture that topics develop in separate clusters and eventually form weak links to become a larger connected component. It is also interesting that "informatics" is at "the center of the world" in day 75.

The topology profile of day 75 confirms these conclusions (Figure 5-5) - the network is fairly sparse, close to binary trees and sparse hierarchies (with low interlinking). With this density, this class of topology is also close to preferential attachment graphs. Every topic cluster is centered around a few key terms, and then loosely connected to the rest of the clusters. The star around "language" reinforces this pattern.

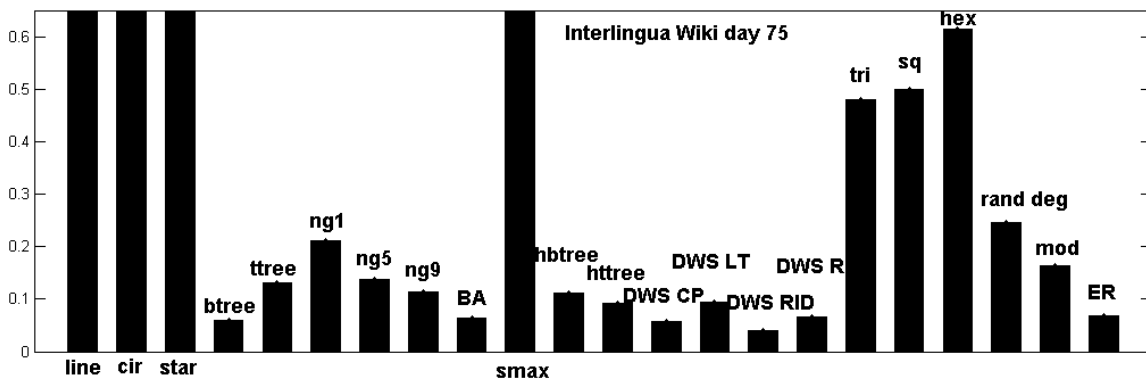


Figure 5-5: Topology profile for day 75 of the Interlingua Wikipedia.

There are only four significant motifs (with Z-score above zero) for day 75. These are the same popular motifs found in traditional airlines, stars and a base triangle. The 5-star has a very high Z-score of 0.9 (Note that stars with multiple spokes will be very significant for this graph due to the dominant star cluster). The motifs are shown in Figure 5-6.

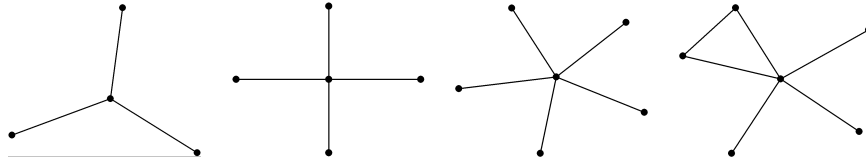


Figure 5-6: Significant motifs for day 75 of the Interlingua Wikipedia: stars and base-triangle.

The topology evolution shows four major phases which correlate well with the growth of the network. Figure 5-7 shows the nodes/edges growth in the first 75 days as well as the topology matching to canonical topologies and growth models. The first phase up until day 24 has very few nodes and edges and a large disconnected network of small "tree" components. In fact, in day 22 the Wikipedia is a simple three-spoke star (Figure 5-8). In the second phase, there is one large star (centered around "languages") and many other smaller disconnected components. Around day 33, the other components are outweighing the star, and the giant component is most similar to a tertiary tree with sparse cross-linking. Eventually the giant tree-like components and the star join (day 75) into one component, which is still sparse and tree-like, and most similar to random interdivisional hierarchies (DWS RID).

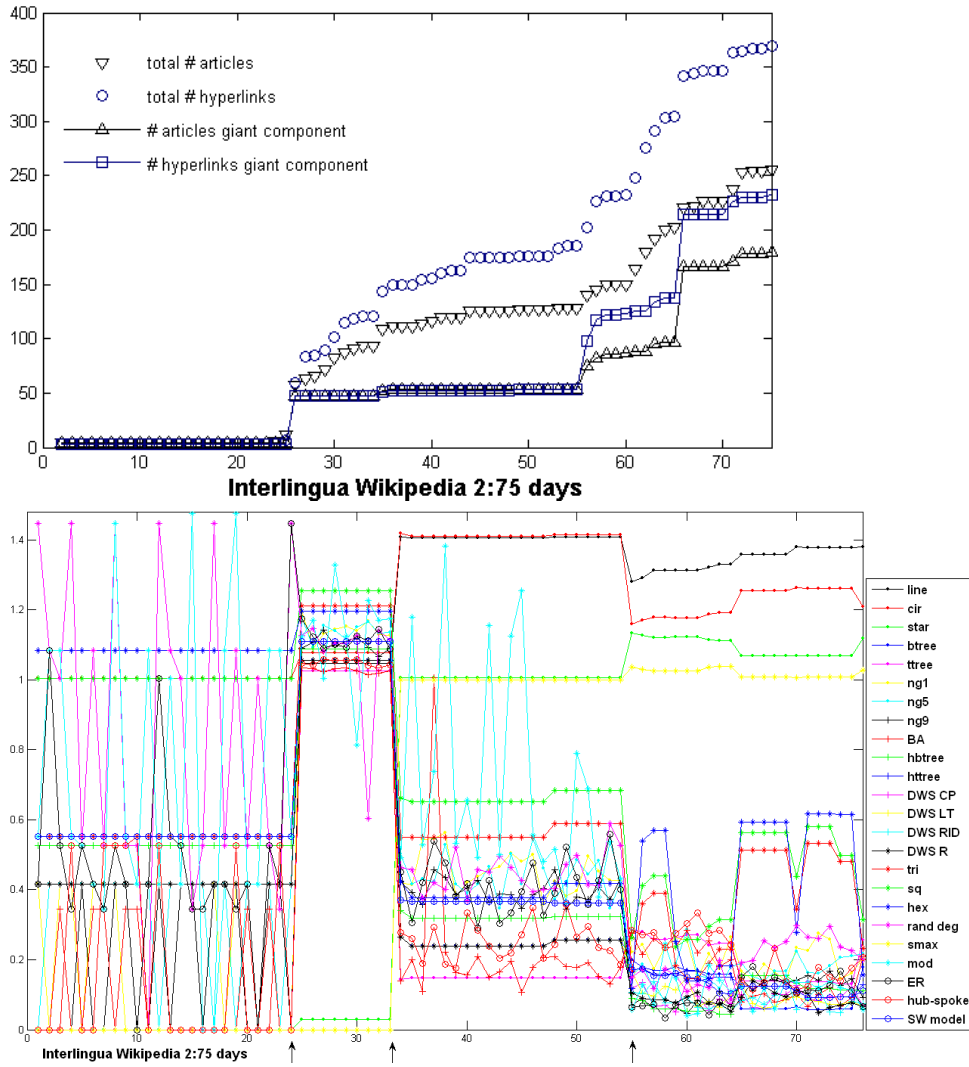


Figure 5-7: Topology evolution of the Interlingua Wikipedia. Days 2 to 75. Three major transitions

Figure 5-8 shows the network snapshots of days 20, 30, 45, and 75, each representing a "topology phase".

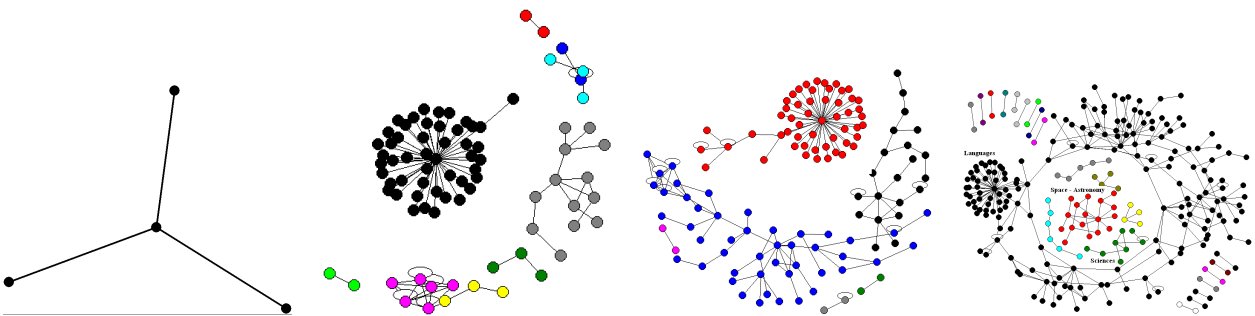


Figure 5-8: Days 20, 30, 45 and 75 of the Interlingua Wikipedia

To summarize, the Interlingua Wikipedia grows by accretion. Various topics emerge separately

and eventually coalesce via weak connector words. This is supported by the observed phases of growth, as well as by the module themes in Figure 5-4. This is a fundamentally different type of growth than the one observed in airline networks, as the two types of networks perform different functions and are subject to different constraints and operating environments. Further analysis (Figure A-6) showed that the Interlingua is a scale-free network, but not a small world. Path lengths increase over time, while the network stays close to its corresponding s-max graph. In fact, it is no surprise that the network is scale-free, since different topic-related modules grow similarly and then connect weakly which is a natural mechanism for repeating patterns at different scales. A repetitive mechanism for growth is a common feature with airlines, where airports get connected in a consistent manner, and probably common with other systems too - but the weak links are unique to Wikipedia. A similar phenomenon was seen in change propagation networks [13].

5.3 The Esperanto Wikipedia

The Esperanto language is the most widely spoken constructed international auxiliary language, first described in 1887. The vocabulary of the language also comes from Romance and Germanic languages. The language has been actively used for over more than a century and has 200-2000 native speakers, but it is not adopted officially in any country. The Esperanto Wikipedia has over 100000 articles and is ranked in the top 30, at 22 among all language Wikipedias.

The first 200 days of the Esperanto Wikipedia show very steady growth as seen in Figure 5-9. Originally the giant connected component contains a small portion of all nodes in the network (<50%) and eventually grows up to 80-90%. This is an indication of the network growing out originally as many disconnected clusters that eventually coalesce in a mostly connected graph.

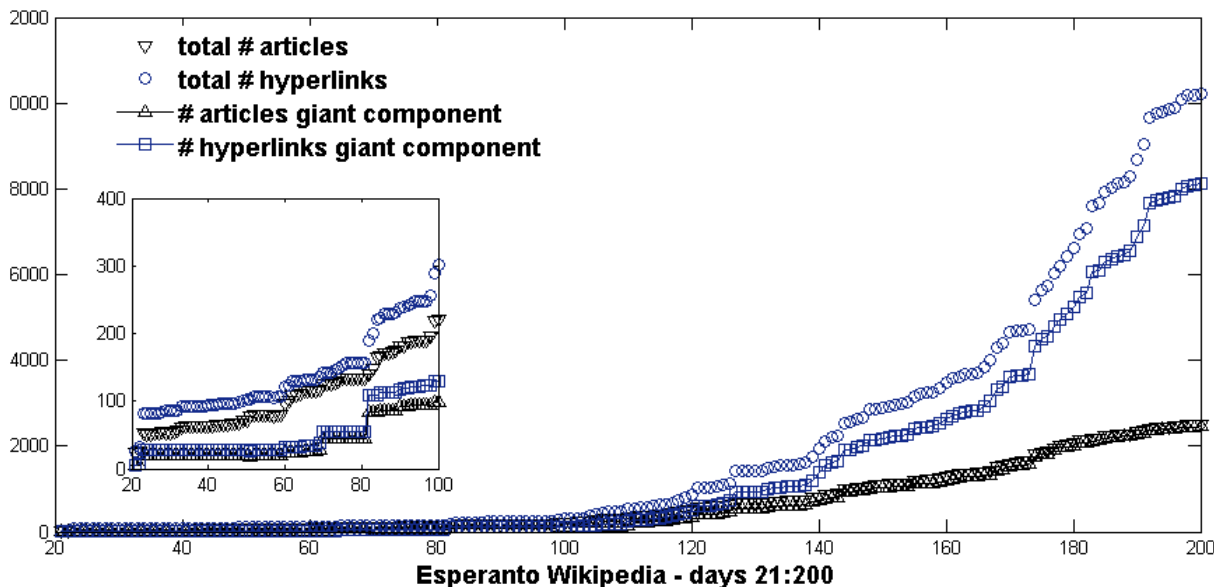


Figure 5-9: The 200 days of the Esperanto Wikipedia: total number of articles and hyperlinks and number of articles and hyperlinks in the giant connected component. The giant component is small compared to the overall size in the first 100 days and then catches up and becomes 80-90% of all nodes.

Day 100 of the Wikipedia is plotted on Figure 5-10 with the giant component only on the right. The other disconnected components other than the giant cluster around topics such as

geographic regions (including many Chinese provinces), Internet terminology, and Beijing parks, cultural centers and places of interest. The main component is more eclectic in themes and is centered around the work "encyclopedia". It is interesting to see the Chinese influence in these early days, probably due to participation of a larger portion of Chinese users. Esperanto was brought in China by Russian merchants in the late 1800s, and strongly supported by the government in early 1900s, with the community of speakers centered in Shanghai.

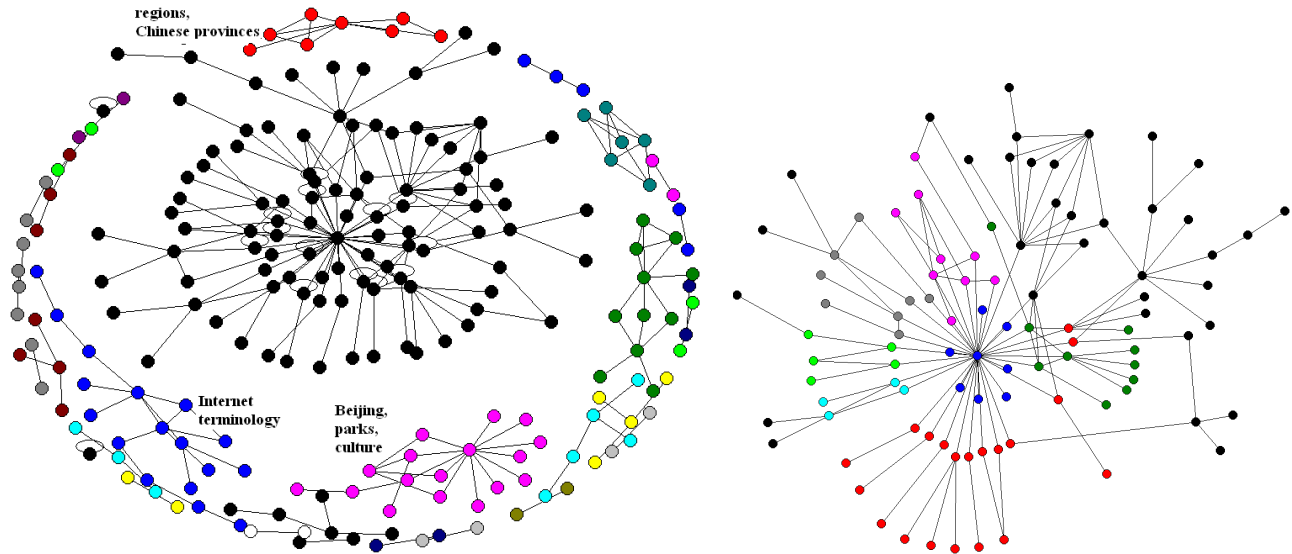


Figure 5-10: Day 100 of the Esperanto Wikipedia: largely disconnected, with the main component centered around the word "encyclopedia" and featuring various topics. Other connected components include topics such as "Internet terminology" (ex: server, unicode), "Beijing places of interest and culture", names of geographic regions, lots of Chinese provinces. (right) Day 100 modules by the Newman eigenvector method.

The right plot on Figure 5-10 shows network modules of the day 100 network found by the Newman eigenvector method. The modules found connect in an unusual pattern compared to the usual loosely connected compact modules. Instead modules are connected centrally (to the word "encyclopedia") and develop thematically in depth growing out of the center. This is different from the chain-like arrangement of modules in the case of the Interlingua Wikipedia.

It is no surprise then that the topological profile shows the Esperanto Wikipedia to be closest to a hierarchy with local interlinking (DWS LT). The local links are the interlinks that define the modules streaming out of the central node. Other close topologies are other hierarchies, hierarchical tertiary tree and a pure random graph (ER) with the same density.

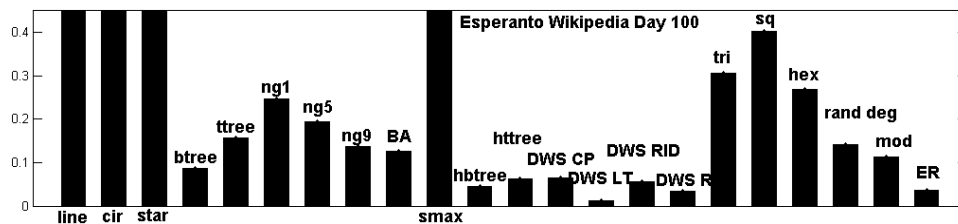


Figure 5-11: Topology profile for day 100 of the Esperanto Wikipedia.

The significant motifs are in the same family, stars and base-triangles, also bi-partite graphs with weaker significance, but present. This is similar to the hub-seeding phenomenon in airlines, though the same growth principles are absent here. Hubbing is natural because knowledge is organized in categories, and sub-categories, but while the emergence of a second super-category (second hub) is entirely possible, it does not have to be a general phenomenon. For example, “biology”, “physics” and “chemistry” can all be filed under “science”, and also “field” or “study” and “discipline”, but the latter are more general than “science”, so the underlying hierarchy will still be there. This is supported by the low Z-score of the bi-partite motif.

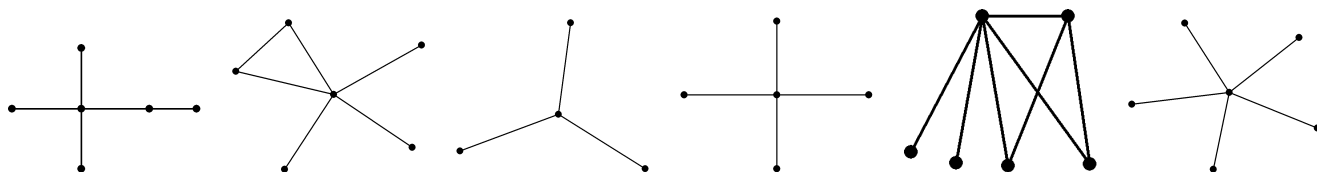


Figure 5-12: Significant motifs for day 100 of the Esperanto Wikipedia: stars and base-triangle, with Z-scores 0.121, 0.102, 0.195, 0.447, 0.01, 0.817 respectively

The topology of the Esperanto Wikipedia does not go through major transitions. There are three visible stages in the evolution plot (Figure 5-13), but they are not drastic. The giant component which eventually becomes the majority of the network grows steadily into a hierarchy with local interlinking and away from pure topologies such as pure trees, stars and lattices. This is also evident in Figure 5-14 showing three snapshots of the network in days 40, 70 and 100 respectively.

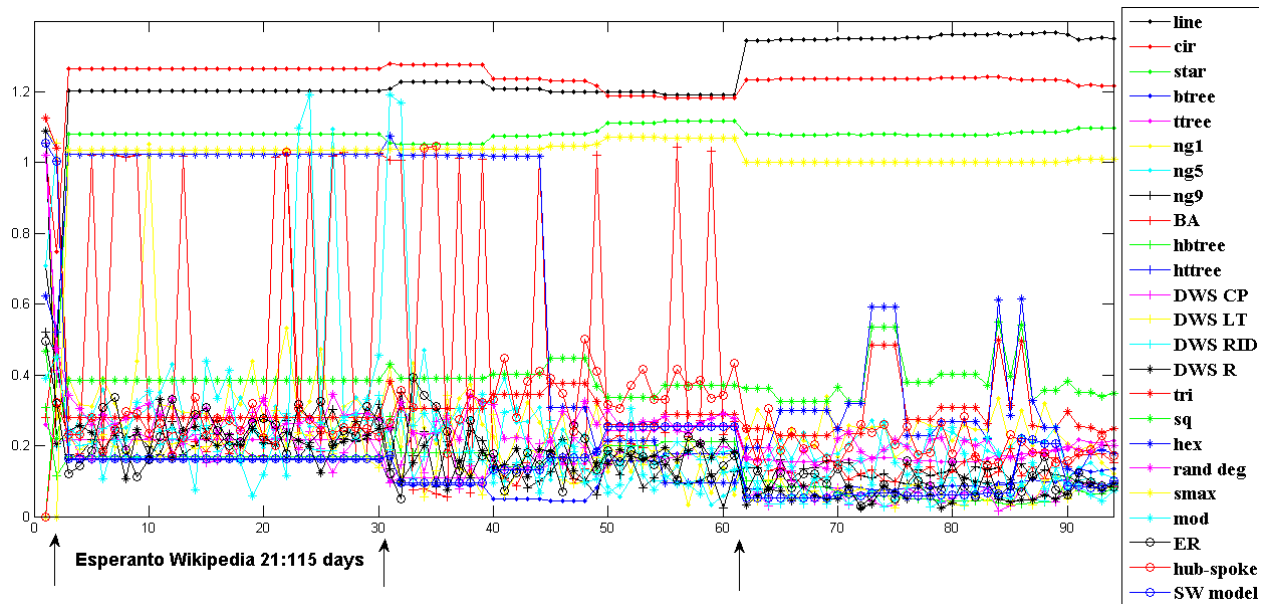


Figure 5-13: Topology evolution of the Esperanto Wikipedia. Days 21 to 115. Three transitions.

The same phenomenon of many separate clusters growing into a whole is evident here, however they emerge into a centralized backbone, rather than a linear series as with the Interlingua Wikipedia.

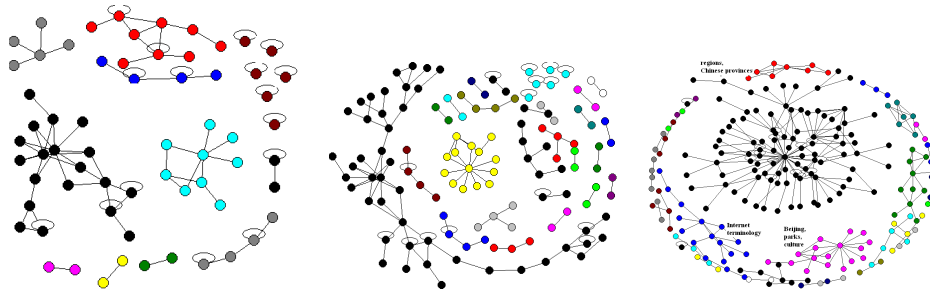


Figure 5-14: Days 40, 70, and 100 of the Esperanto Wikipedia

In summary, the Esperanto Wikipedia grows without major topology transitions, steadily in both size and structure. The same accretion growth has been found, as with the Interlingua network, but the emergent backbone is different. Instead of the scale-free linear backbone, the Esperanto Wikipedia is closest to a centralized hierarchy with local interlinking, at least up until the hundredth day.

5.4 The Spanish Wikipedia

Though Spanish is the second or third most spoken language in the world, it trails behind in its Wikipedia size, compared to languages with much smaller speaking population such as Japanese, Portuguese, Polish and Dutch. Probably other geo-political factors such as country GDP and population income level affect the representation of these people online. Transliteration has been cited as a problem, since web addresses are exclusively written in English, this hinders development in other languages, for difficulties such as spelling with accents in web addresses, or other characters. Interestingly, languages with much greater challenges in that respect, such as Chinese and Japanese, are more present.

Figure 5-15 shows the slow growth of the Spanish Wikipedia in its first 200 days, with a steep stage of growth between 240 and 270 days and then another steady period till the end of the first year. As with the other Wikipedias analyzed so far, the giant connected component catches up in size to the rest of the network, which starts out very disconnected originally.

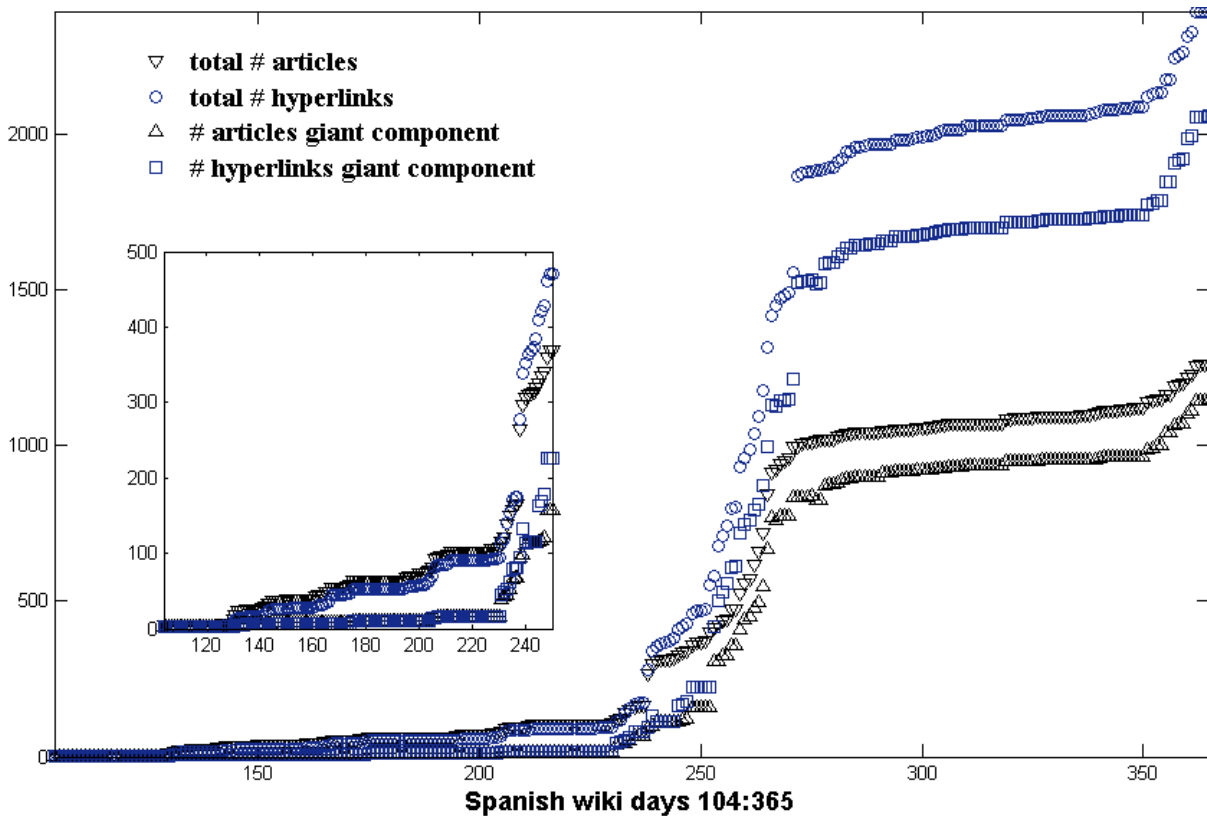


Figure 5-15: The first year of the Spanish Wikipedia: total number of articles and hyperlinks and number of articles and hyperlinks in the giant connected component. The giant component is small compared to the overall size in the first 200 days and then catches up and becomes 80-90% of all nodes. This transition happens during days 250-260.

Day 235 was chosen for further analysis as representative of the transition period. Different disconnected components are centered around topics such as software, electro-technics, continents, and Earth/planetary systems. The giant component in day 235 is a growing tree with low interlinking. This explains why the topological profile shows lower similarity to trees, and higher similarity to core-periphery hierarchies (Figure 5-17). As a reminder, the CP hierarchy has higher level of interlinking closer to the core, rather than deeper in the leaves. The similarity to a random graph is also interesting. Due to the low density and the high-level interlinking, no nodes stand out, and the degree distribution is closer to uniform, both in the core of the hierarchy, as well as in the periphery. There are no significant motifs in the network of day 235.

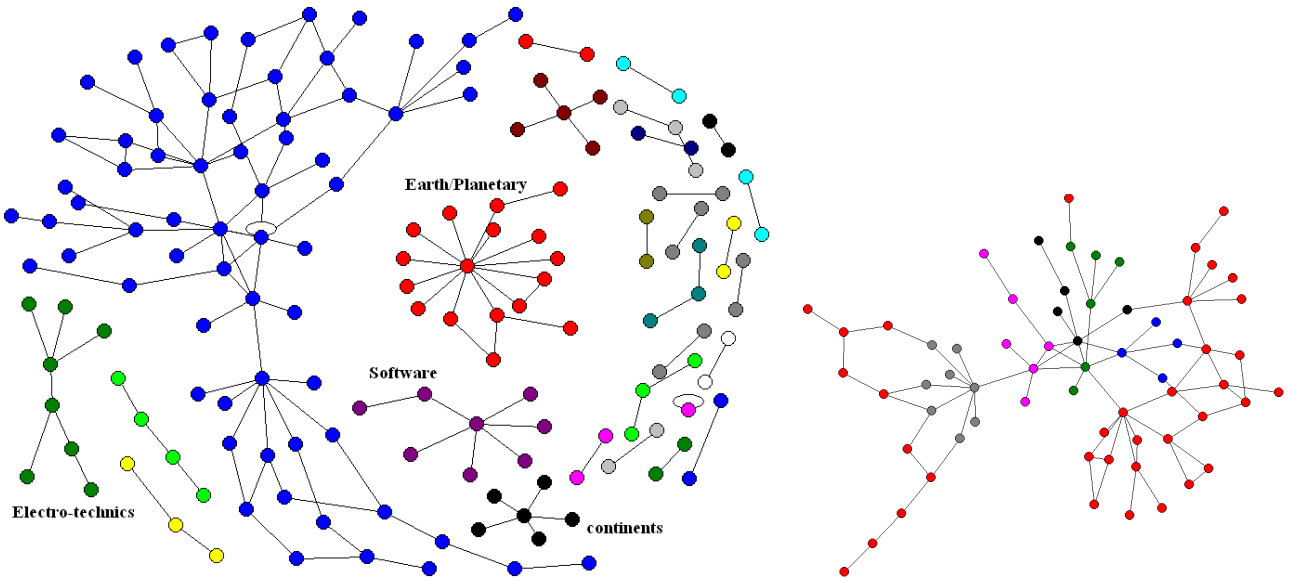


Figure 5-16: (left) Day 235 of the Spanish Wikipedia. Different components include topics such as electro-technics, software, continents and Earth/planetary terms; (right) Modularization of the giant component of day 235 of the Spanish Wikipedia.

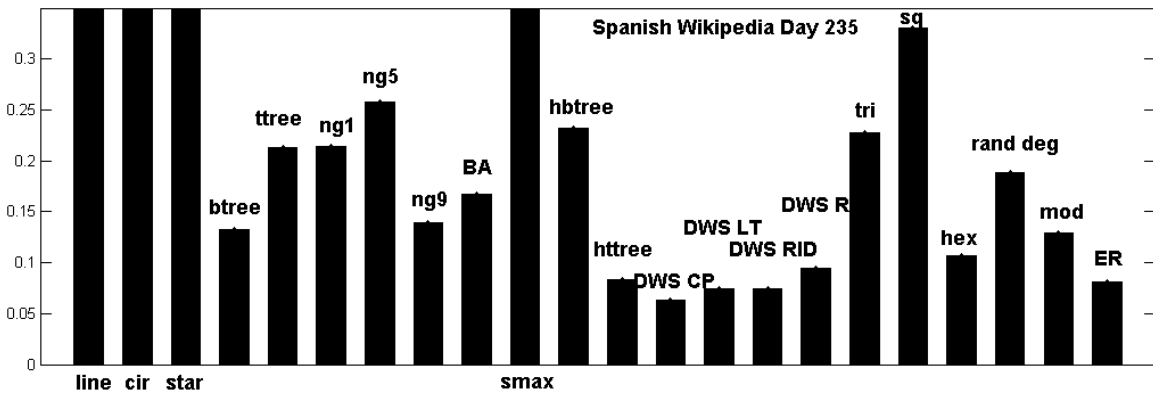


Figure 5-17: Topology profile of day 235 of the Spanish Wikipedia

As with the other Wikipedias, the Spanish topology transitions are related to jumps in growth (Figure 5-18). The network goes through being a collection of small disconnected stars, to bigger stars, that grow in collections of trees. This is still the case up until the 200th day as seen in Figure 5-19. On the topology similarity plot, it is evident that in the first 200 days the network is closest to its s-max graph equivalent, evidently because of the trees. The third phase on Figure 5-18 is only different because the linear trees become more dominant than the stars, so the similarity to BA graphs and the hub-seeding model disappear, and the similarity to lines, and Newman-Gastner type trees re-appears. The fourth phase becomes fundamentally different because the smaller connected clusters start to connect. They are still very tree-like, hence the similarity to hub-seeded graphs. Notice that though this model shows good match here, the mechanism behind the growth of the Wikipedia graph is very different. Hubs (or highly connected article pages) did not gradually become popular and gained links. They developed topically in separate clusters and joined through weak links later [46].

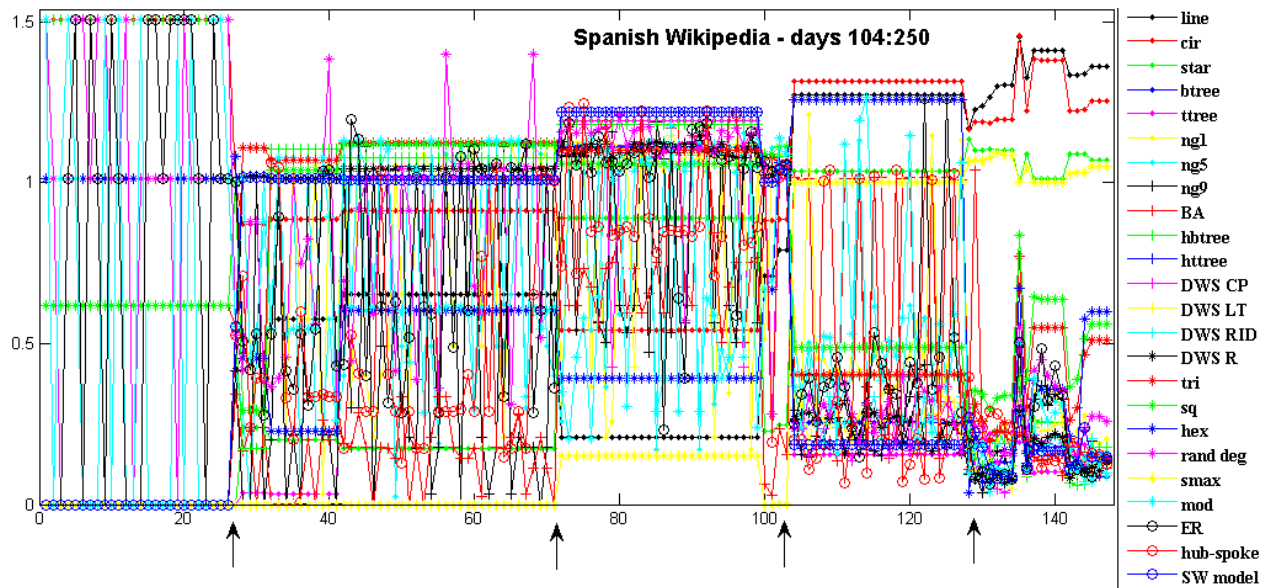


Figure 5-18: Topology evolution of the Spanish Wikipedia, days 104:250

Finally, the last stage contains all the thematic clusters, stars, trees and fully grown hierarchies into a weakly connected whole. The best example of this can be seen in day 250 of the Spanish Wikipedia, plotted in Figure 5-19. This history is almost exactly like the history of the Interlingua - slow growth of disconnected BA-like components into a weakly connected decentralized hierarchy. The Esperanto Wikipedia, in contrast, grew mostly out of one core component, without major phase transitions, and turned into a centralized hierarchy - still very modular in topics (and graph-wise), but based out of a single page (the word "encyclopedia").

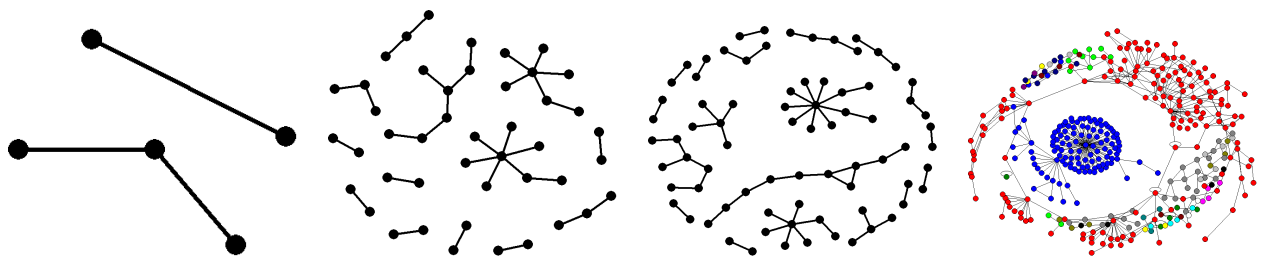


Figure 5-19: Days 124, 164, 204 and 250 of the Spanish Wikipedia

To summarize, the Spanish Wikipedia shows distinct phases of growth, definitely exhibiting growth by accretion. Over the course of 200 days the network grows from small disconnected linear trees and stars, into a weakly connected sparse hierarchy.

5.5 The Russian Wikipedia

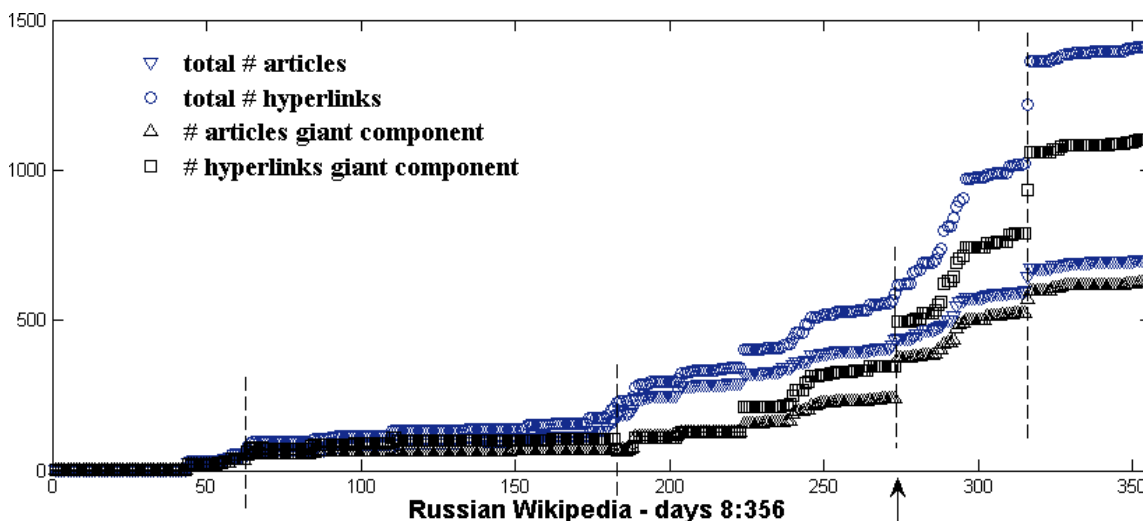


Figure 5-20: The first year of the Russian Wikipedia (days 8 to 356): total number of articles and hyperlinks and number of articles and hyperlinks in the giant connected component. The giant component becomes the majority of the network around day 275. About four major jumps in growth can be detected.

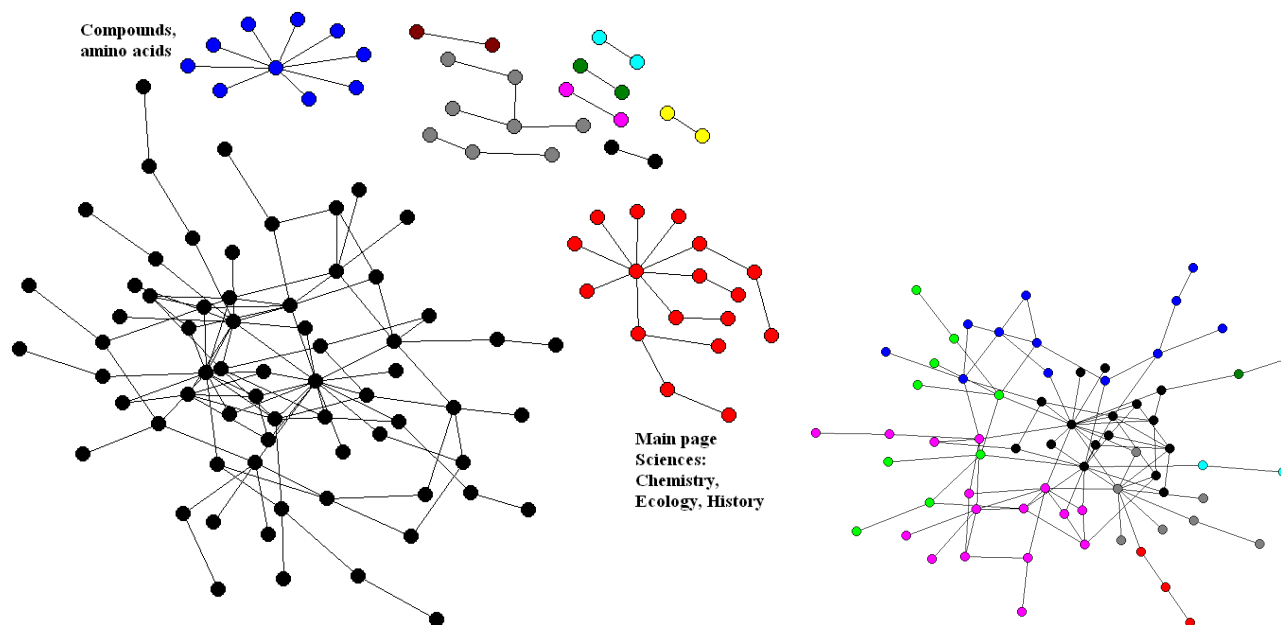


Figure 5-21: Day 180 of the Russian Wikipedia and its giant component modularized; The topics of separate components are 1/compounds, amino-acids and other molecules (this is the component that will outgrow the giant component on day 190) and 2/around the main page sciences and disciplines of study, such as chemistry, history, and ecology. It is interesting that the wiki main page is not in the giant component at that time. The most central article in the giant component is "Russia".

The topological profile of day 180 of the Russian Wikipedia (more precisely of its giant component) holds no surprises - just like other Wikipedias, and more so, the best match is to a random graph and to hierarchies. There are no significant motifs. Though in the case of this network, the single snapshot tells little of the whole story. Day 180's topology is quickly replaced as another component becomes the giant component - a star.

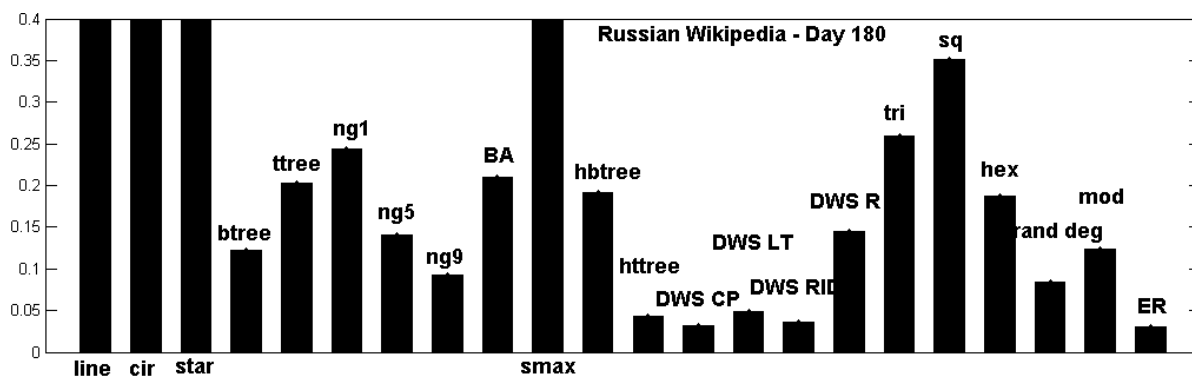


Figure 5-22: Topology profile of day 180 of the Russian Wikipedia. Best match are hierarchies, core-periphery and random-interdivisional and random graphs.

Figure 5-23 shows a distilled version of the topology matches, only concentrating on the closest canonical curves. Up until day 190 (140 on the figure), the network evolves as a hierarchy with local interlinking (best matches are DWS, ER, and SW model graphs). Then on day 190, the star centered around chemical compounds and molecules outgrows the giant component, and prevails the geometry until about day 210 (160 on the figure). The topology goes through two more transitions - outward branching of the giant star and then weak connection to the other large components in the network (including the original giant component). This history is evident on Figure 5-24, which shows snapshots of days 180, 200, 220 and 250 respectively.

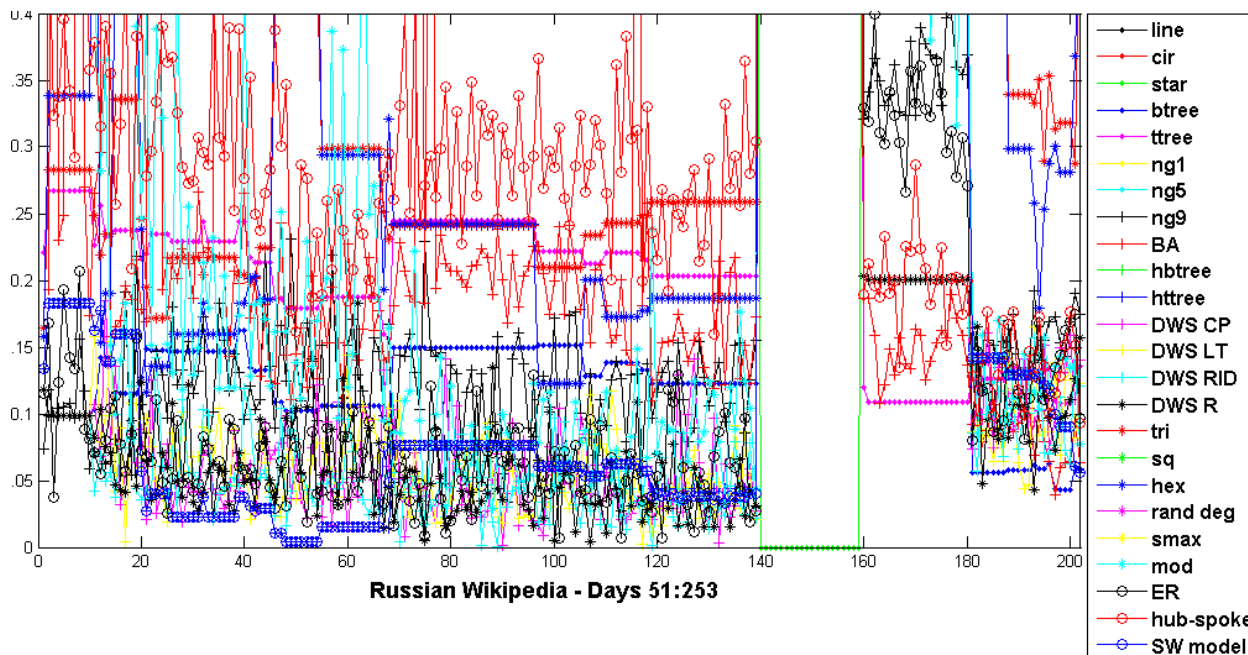


Figure 5-23: Topology evolution of the Russian Wikipedia, days 51 to 253. The first 50 days were too small to analyze (essentially two edges). A single component dominates the network, until day 190, when a star component centered around the topic "compounds/molecules" takes over.

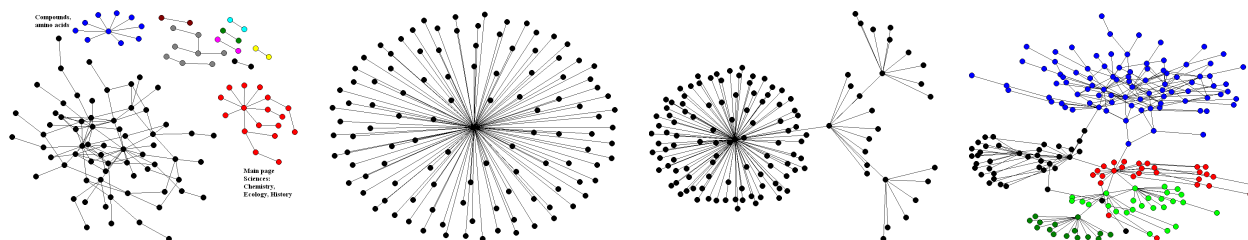


Figure 5-24: Snapshots of the Russian Wikipedia history: days 180, 200, 220 and 250

Again, the principle of growth here is growth by accretion. Modules arise separately, though some are core throughout the history of the network. Then spontaneous growth occurs, as some topics get more attention or more users, until different topics connect weakly into a whole.

5.6 The Chinese Wikipedia

The Chinese Wikipedia was started in October 2002, and has over 200000 articles as of today, ranked 12th by size among all Wikipedias. Given the size of the potential user pool, this Wikipedia will outgrow many of the other (especially Eastern European and Japanese) language Wikipedias. The Wikipedia site has been blocked by the Chinese government several times in mainland China, in addition to content filtering. Access to the English Wikipedia was restored prior to the Olympic Games and access to the Chinese Wikipedia was restored soon thereafter. These regulations probably affected the growth of the Wikipedia, as its largest potential user pool did not have access to it, for most of its lifetime.

The growth of the Chinese Wikipedia during its first year is plotted in Figure 5-25. The plot shows very steady growth, without jumps in size, high density, i.e. many links per node, and surprisingly, high connectivity. For the entire first year, most of the network is connected: as the figure shows, the giant component contains close to 99% of all nodes. This is unusual compared to other Wikipedias, and suggests that this network may be growing as a whole from the start, without the coalescence of various disconnected topics. This could be due to central management or a philosophy of contribution by extending the knowledge that is already there.

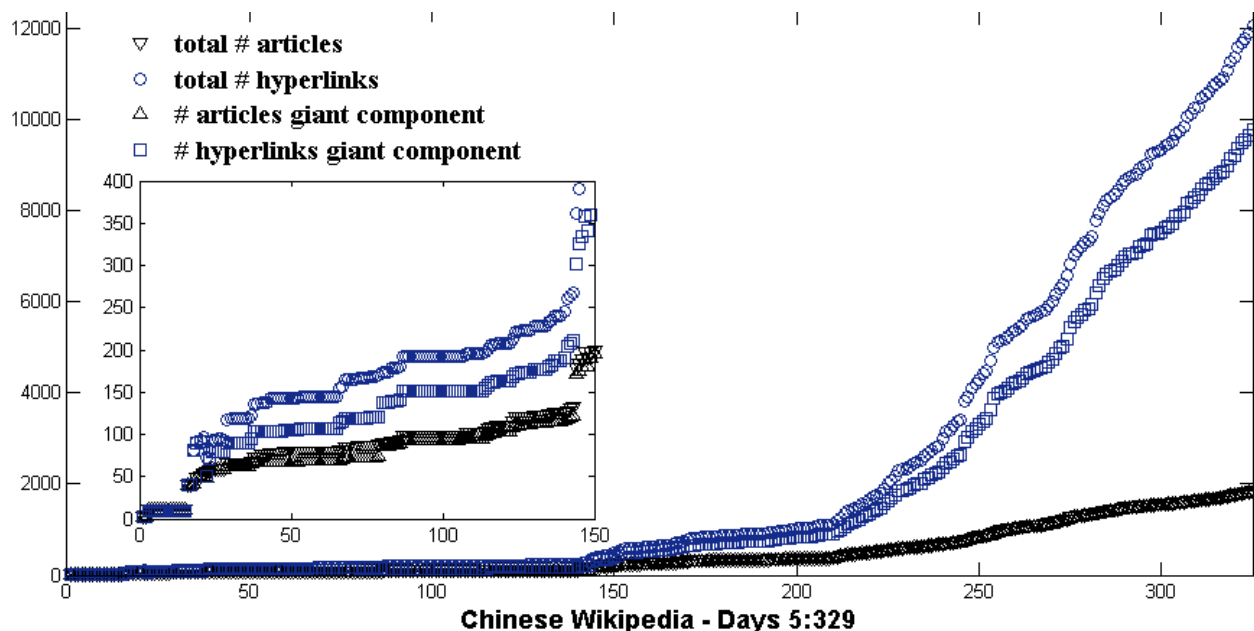


Figure 5-25: The first year of the Chinese Wikipedia (days 5 to 325): total number of articles and hyperlinks and number of articles and hyperlinks in the giant connected component. Unlike all other Wikipedias, this one stays largely connected: the giant component contains almost all nodes. Another novelty is the large density, i.e. many more links per node on average.

Day 100 of the Chinese Wikipedia does not look unconventional compared to other Wikipedias - several modules (possibly separate thematically) connect weakly to form the network. The significant motifs are only stars (Figure 5-27), likely due to the large star subgraphs seen in Figure 5-26. The larger stars are more statistically significant, as expected - the 5-node star has a twice as high Z-score than the 4-node star.

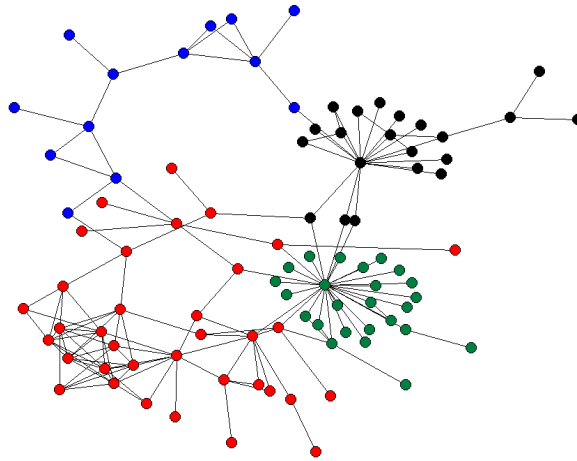


Figure 5-26: Day 100 of the Chinese Wikipedia (giant component), modularized using the Newman-Girvan algorithm.

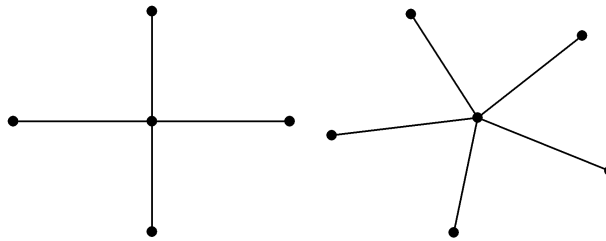


Figure 5-27: Significant motifs for day 100 of the Chinese Wikipedia: stars with Z-scores 0.166 and 0.338 respectively.

The topology evolution of the Chinese Wikipedia is shown in Figure 5-28. As with the growth, the topology is very steady over time, without major transitions. There is only one transition, from day 21 to day 22, the topology goes from a pure star to a multiple star with three hubs connected to all spokes (Figure 5-29 shows snapshots of these two days). This transition is really unusual, and seems to be related to connecting all pages to an older version of a main page. Except for that anomaly, the rest of the history holds no surprises. The network is consistently favoring the hub-seeding airline model. Other close matches are hierarchical tertiary trees, binary trees and BA graphs. The fact that the Wikipedia grows as whole, connected and that it is close to the hub-seeding model, is probably not a coincidence. The alternative growth by accretion cannot produce this pattern because smaller components with local hierarchies join weakly to form other hierarchies, with no adherence to few hubs. It is likely that authors do not spend time to link their material to everything it is relevant to, or even to the highest (most general) page in the same field. So in a way, the organization emerges, rather than being dictated from the top. That said, there is top-level organization in Wikipedia, because one can search major categories and their subcategories, but maybe that organization was not the driver in the case of other Wikipedias. Probably due to the nature of the Chinese Wikipedia, access restriction, and connected whole growth, central hierarchy played the key role in driving the growth, which resulted in the network looking more like a hub-spoke-like airline, rather than a randomized hierarchy.

The question of why the Chinese Wikipedia grows connected, unlike other Wikipedias, is difficult to answer. The content was only open to mainland China just before the Olympics of 2008, and prior

to that the user and author population came from outside of China. There is no apparent reason why the early growth would not be disconnected, because there is no known external regulation on the Wikipedia growth.

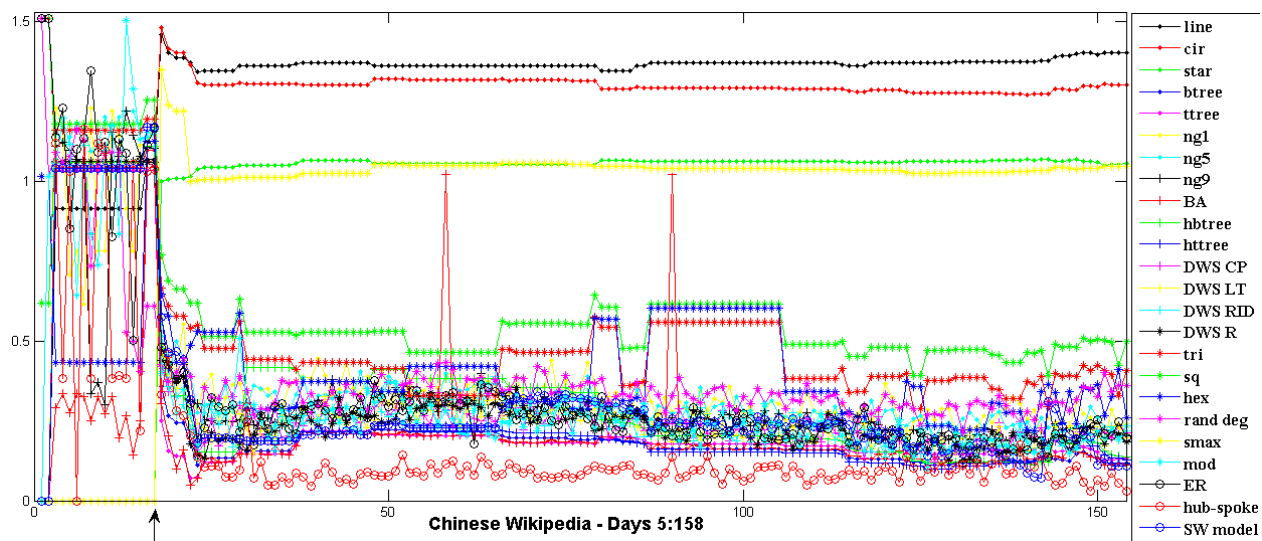


Figure 5-28: Topology evolution of the Chinese Wikipedia, days 5:158. One major transition from a pure-star to a multiple-star topology, at day 22.

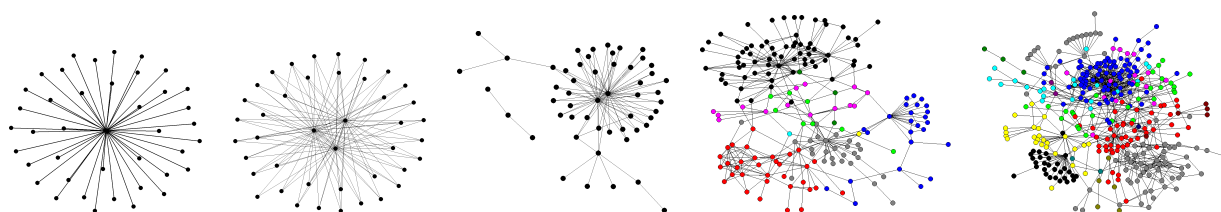


Figure 5-29: Days 21, 22, 25 and 150 and 200 of the Chinese Wikipedia.

5.7 Topology Comparison of Language Wikipedias. Conclusion

In this chapter, we presented detailed analysis of the evolution of topology in the early history of some prominent language Wikipedias, in particular the Interlingua, Esperanto, Spanish, Russian and Chinese Wikipedias. Overall, all networks proved to be very modular, with the modularity in a network theory sense (ex: Newman algorithms) corresponding to *modularity* in topics/themes. For example, often pages related to geography, names of countries, continents and provinces appear clustered together. This has been shown for all Wikipedias, except the Chinese, due to the language barrier. Other than modularity, Wikipedias exhibit a strong tendency for *hierarchical organization*. This arises from two directions: from the top, as often articles are organized in categories, and sub-categories and those are linked one or two levels up and down, and from the bottom, as pages in one categories naturally arise disconnected and eventually get connected hierarchically. The third prevalent characteristic of Wikipedias is *growth by accretion*. Since connectivity of flow is not a requirement, as in the case of airlines, Wikipedias do not have to be connected. This is enforced by the virtue of searchability. An article can be found without being connected to a giant component or the whole Wikipedia thematically. All articles are connected to the main Wikipedia page to

be searchable (as explained earlier, that connectivity does not appear in our model). This means that connectivity does not affect the visibility of information. With airline routes, if an airport is disconnected, that means it cannot be reached, hence that airport's market is not being served. As a result, Wikipedias grow as disconnected thematic clusters, which connect weakly by expansion, and eventually coalesce into a giant component which dominates the network. In combination with the hierarchical organization, they become weakly connected hierarchies, with varying levels of interlinking. This is the same growth principle seen in change propagation networks [13].

There is one exception to the above patterns - the Chinese Wikipedia. Here we summarize the findings about topology and growth of all studied Wikipedias.

The Interlingua and the Spanish Wikipedias have very similar topologies. Both grow by the weak connection of disconnected thematic clusters, with many topology transitions as different components dominate the network. In our analysis only the topology of the giant component is considered in the network history. This signifies a diverse beginning for both Wikipedias. Both exhibit a topology closest to a hierarchy with random interlinking, except that the Interlingua has more linking across levels (DWS RID), and the Spanish Wikipedia is more sparse, with more interlinking at the core of the hierarchy (DWS CP).

The Esperanto Wikipedia shows the same growth patterns, except that rather than many components oscillating to dominate the network, one prevails throughout the history. This is why the hierarchy looks more centrally based (centered around the word *encyclopedia*). Because of the single-component dominance, there are no major topology transitions, and the topology is a centralized hierarchy with local interlinking.

The Russian Wikipedia is the one closest to a random graph. With the same growth model, many topology transitions and diverse giant components, every component is more uniform in this network. The hierarchies tend to be sparse and not very centralized which makes the graph look random.

And finally, the outlier in the Wikipedias studied in this thesis is the Chinese Wikipedia. Throughout its first year, this network grows steadily, without jumps and major transitions and surprisingly, grows in a connected manner. The same giant component contains 99% of all nodes throughout the history of the Wikipedia, which starts out as a growing star. At day 22, multiple hubs appear which connect to the old spokes, much like in the second step of the hub-seeding algorithm. For the rest of the history of 150 days, the Wikipedia shows consistent closeness to the hub-seeding model. As discussed earlier, this difference from other Wikipedias and similarity to the airline growth model has a lot to do with connectivity and hierarchy as a driver in growth. The Chinese Wikipedia is strongly driven by hierarchy, i.e. mapping relevant topics that already exist in the giant component. This is the model of a top-managed system, in which the hubs naturally emerge as the super-categories. The growth by accretion cannot produce these patterns because organization there emerges rather than drives the growth.

In this chapter we have learned that there are underlying mechanisms in the early growth of web-based free encyclopedias, and that connectivity and management of hierarchy and driving factors in the evolution of topology.

Chapter 6

Conclusion

The subject of this thesis is the analysis of network topology of engineering systems and its evolution, with examples in US airline routes and language Wikipedias. We developed concepts and tools for network analysis, and especially for tracking topology patterns over time. Though challenged by the immensity of the data, we extracted interesting slices for analysis. The airlines analysis included the wide-body jet routes, the top 50 percentile flights in number of departures and seat capacity, and among individual airlines, JetBlue, Southwest and Continental Airlines, represented as networks. We learned that the underlying motifs of hub-spoke topologies fall in three families (stars, base-triangles and bi-partite subgraphs), and proposed growth models that fit their formation. Southwest was analyzed as an outlier in the industry and was found to be too complex to study with simple network metrics. There were no significant motifs found in the simple graph representation of Southwest. We proposed a probabilistic growth model for Southwest and a way to reduce “noise” in the data. This eventually showed that over time Southwest is becoming more centralized, quite like other more traditional airlines.

Finally, we applied the tools developed in the thesis to study the early growth of language Wikipedias and found that their underlying patterns are very similar to each other. Their growth dynamics is very different from the airlines - they grow by coalescence, i.e. growing separate connected modules eventually merge into a giant component, rather than centralized growth from a single initial node.

In summary, we have identified two trends in patterns of evolution in complex systems. Physical systems grow mainly through aggregation, adding nodes and links, by some design principle. Such principles are preferential attachment, hub seeding and hierarchy interlinking. These systems are governed by conservation laws (mass, energy) and are subject to constraints and performance objectives, such as efficiency. Social and knowledge systems tend to grow by coalescence, in which disconnected clusters gradually form giant components via weak links. These systems are governed by human behavior and cognitive limits. An example of a physical system (Lufthansa Airlines world routes) and a cognitive system (the Interlingua Wikipedia) are shown in Figure 6-1.

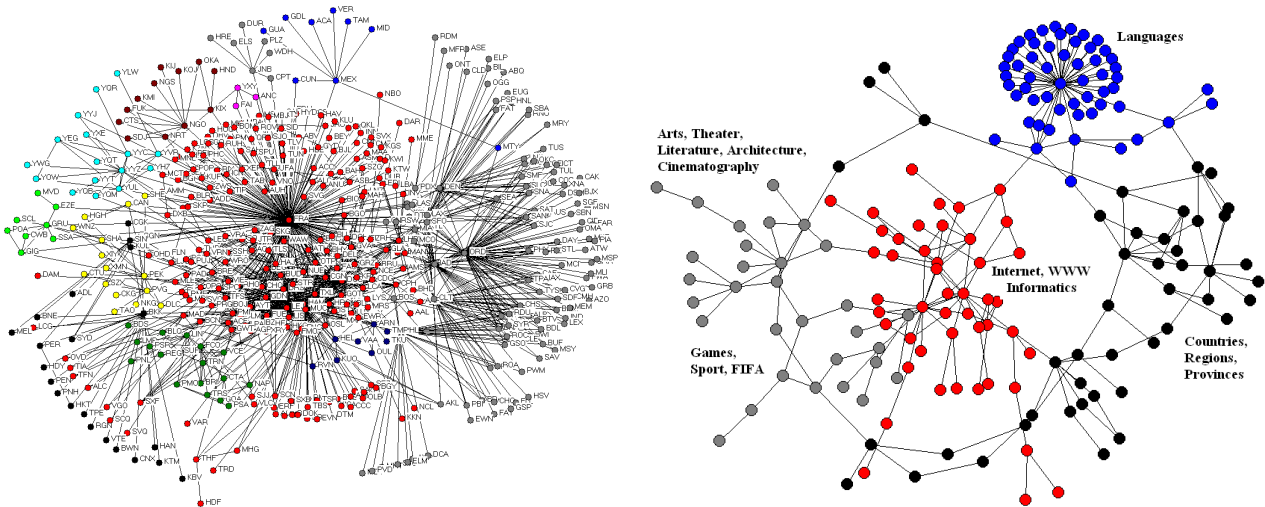


Figure 6-1: (Left) Lufthansa Airlines world routes 2006, modularized using the Newman-Girvan algorithm and (right) giant component of day 75 of the Interlingua Wikipedia, modularized using the Newman eigenvector algorithm.

Next we expand on these results, organized in the three areas we set out to address: network topology and its evolution, airline networks and tool and concept generalizability.

6.1 Contributions

6.1.1 Analyzing network topology

We combined graph theoretical metrics from the literature in a novel way to analyze network topology comprehensively. First, we surveyed a comprehensive list of "canonical" networks to create a "profile" of topologies aligned by increasing density and complexity. We used two measures to compare real systems relatively to this topology profile. We developed a best matching node sequence algorithm and validated a graph similarity measure based on the graph similarity ideas by Blondel et al [49]. We discussed the relationship between degree and betweenness as a distinguishing factor in network topology. In terms of pattern finding, we adapted a state of the art algorithm for single-motif search [5] with the idea of topologically generalized motifs by Kashtan et al [6] to create a comprehensive heuristic motif search uncapped by motif size, limited only by computer memory.

In summary, the contributions in network topology analysis are as follows:

- Best matching node sequence based on graph similarity; Euclidean distance of best similarity matrix diagonals as a measure of similarity.
- Developed the topology profile and the non-dimensional topology vector concepts.
- Adapted and augmented algorithms for unlimited size pattern finding in real networks. Combined a state of the art algorithm for single-motif search with the idea of topologically generalized motifs, to create a comprehensive heuristic motif finding algorithm for any size motif and graph.

6.1.2 Studying statistically system topology over time

To examine system topology evolution, we studied data series for various large networks, from 212 months of airline data, of single airlines and industry segments, to 365 days of early Wikipedia history. We detected clear topology transitions in the early history of these systems and identified the underlying topology patterns. We found similarities among the majority of airlines, and majority of Wikipedia networks. Most interesting were the exceptions: we showed mathematically that Southwest's topology grows in a completely different manner than the rest of airlines, and similarly, that the Chinese Wikipedia grows differently than the rest of Wikipedias. We proposed two probabilistic growth models for hub-spoke airlines, with hub-seeding and for Southwest, with strong local interlinking.

- Detected "topology transitions" and modes of growth, across systems in the same domain (Southwest vs JetBlue), as well as across domains (airlines vs Wikipedia).
- Proposed two models for airline network growth: the hub-spoke growth model and "the Southwest model" and showed that they match the topology history better than canonical topologies.

In summary, our contributions in the tools for analyzing network topology and its evolution culminate in a process for studying system structure over time. The pattern synthesis part of this process can be used to devise system-specific growth models. The network analysis process is described in Figure 6-2. The process starts with collecting snapshots of data (or system history) that can be represented as graphs. The data is then manipulated into format conducive to network analysis. The actual analysis contains three major steps. First is the statistical part, which involves studying the data historically using graph-theoretical and domain-specific metrics. The second part is the non-dimensional topology analysis summarized in Chapter 2, Section 2.1.2 and Section 6.1.1. The topology analysis contains topology profile analysis, motif finding and coarse-graining if necessary. The last analysis step involves synthesizing the patterns found and devising custom growth models based on the results. In the last stage, the custom models are matched to the real data and revised if necessary. The goal of this process is to gain an understanding of how system topology evolves and lay the ground work for domain analysis of factors driving its evolution.

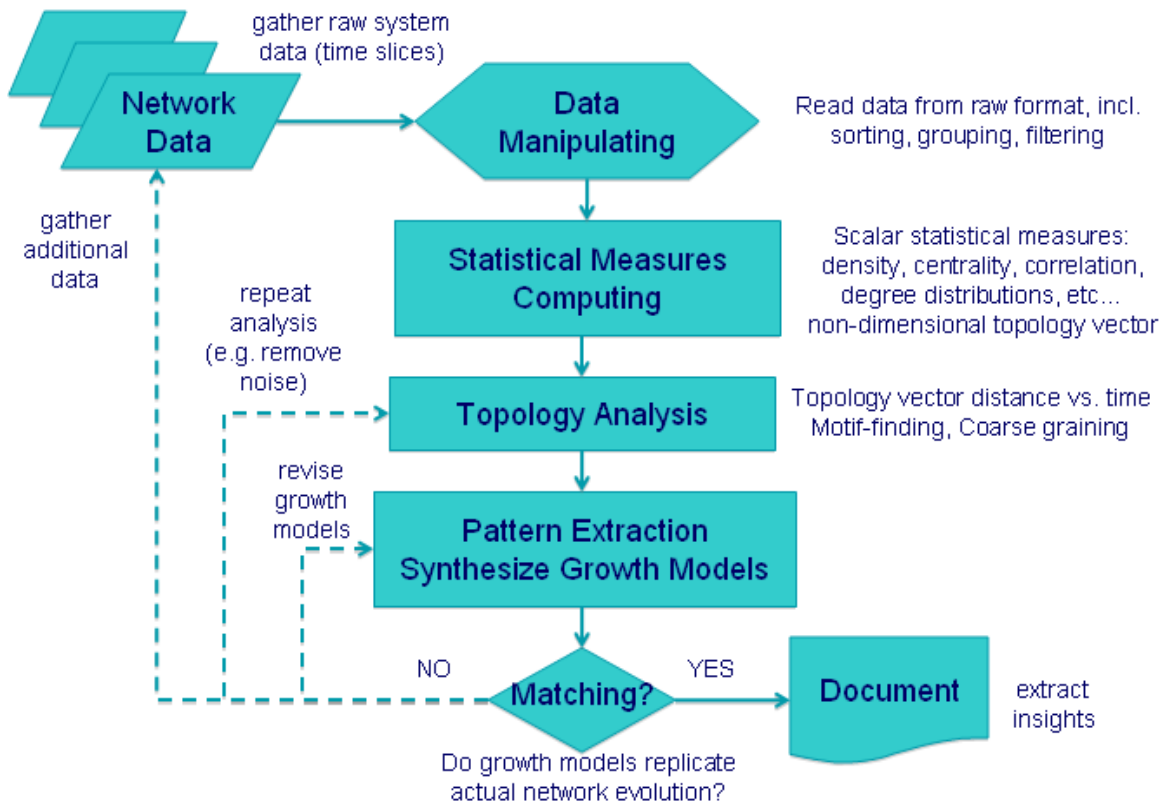


Figure 6-2: Network Evolution Analysis Process

6.1.3 Airline Networks

We analyzed airline routes comprehensively from a network theory point of view, by plotting statistics for various slices of the industry and individual airlines. We compared graph-theoretic with industry hubbing indices and found no strong correlations with small exceptions. Among airlines and data slices, we found that JetBlue, Continental and the route network of wide-body jets, as well as the top 50 slices by seat capacity have similar topologies - close to trees and BA graphs. We showed that Southwest is distinctly different, and that graph-theoretic metrics are not enough to understand the underlying patterns in the airline's structure. We also showed that over time it is likely that Southwest will become more centralized (more hubs, less point-to-point). The following is a summary of the above points:

- Compared graph-theoretical topology analysis with industry-relevant metrics to understand the state of the industry from a high-level
- Compared topologies of different airlines based on topology profile plots and underlying motifs.
- Found that JetBlue, Continental and all industry slices have similar topologies, as opposed to Southwest.
- Detect patterns of growth in airline networks: hub seeding, geographic expansion, and local interlinking.

The major takeaway from the airline analysis is that most airlines have a prevalent “hub-spoke” topology that can be matched with a hub-seeding generation model. The second major takeaway

is that Southwest Airlines was proved to have a different structure, using various size-related and non-dimensional statistics, as well as underlying motifs.

6.1.4 Wikipedia

We found that Wikipedia networks grow very differently from transportation systems. Local sub-structure can strongly influence the early days of a language Wikipedias. Also we found that most Wikipedias converge to a random hierarchical structure, strongly modular with weak links. And this happens by coalescence. Small clusters of knowledge first grow separately, in small local hierarchies (often simple trees) and eventually connect via weak thematic links. This model is pervasive, except in the case of the Chinese Wikipedia, which grows 99% connected from its first day, and exhibits a more centralized structure. The summary of Wikipedia-related contributions is below:

- First comprehensive analysis of the network topology evolution of articles in Wikipedia.
- Quantified patterns of evolution in knowledge networks: verified growth by coalescence.
- Proposed general static topology for language Wikipedias - random hierarchy of weakly-linked modules.
- Understood the early stages of growth of various Wikipedias, including the Esperanto, Spanish, Russian, Chinese, Interlingua, French and Italian Wikipedias. Similar patterns of growth by coalescence were observed in change propagation networks [13].

6.1.5 Computational contributions

Computation has been a challenge in this research because graph algorithms are not easily parallelizable, and the datasets of interest have 100s and 1000s of nodes. We have discussed computational complexity where relevant in the thesis. For example, the complexity of the components of the topology vector is included in Table 2.1.2. More on computational challenges is discussed in the following section 6.2.

- Created an extensive Python toolbox for modeling and analyzing data that can be represented as a network.
- Adapted single-motif finding and topologically generalized motifs to create a comprehensive motif finding algorithm with complexity $O(k! + \binom{n}{k})$, where n is the graph size and k is the motif size (Chapter 2, Section 2.4.).
- Proposed two simple probabilistic growth algorithms for airlines, and showed that stochastically they perform better than canonical growth models from the literature (Chapter 4, Section 4.4).

6.2 Limitations

6.2.1 Data/Computation

A major challenge in this research was computation time. Both the airline and the Wikipedia data in their entirety were too large to analyze as a full set. That is why slices of the airline data were carefully selected and only the first 100 or 365 days of Wikipedias in general were analyzed. The most intensive part of the computation, by far, is the motif finding. Full factorial motif finding is exponentially hard to do, but even probabilistically, with growing network or motif size the problem becomes very challenging. We found that for motif sizes above 9, the memory, not the time is the

limit.

All computation was performed across various platforms, which also presented some software version challenges. The machines used are:

- 3.6 GHz/2 GB RAM Dell desktop computer
- 2 GHz/0.99 GB RAM Dell laptop
- Dedicated Linux node (courtesy of the Aero/Astro department)
- Athena (Linux) station IBM dual processor
- Massachusetts General Hospital supercluster (courtesy of Dr. Anna Custo)

As mentioned, graph algorithms are not naturally parallelizable, which makes computing on various platforms, not only difficult due to Python versions and library differences, but also in putting together the results. Given better computational resources, the datasets we studied can be analyzed for longer parts of their history (since they mostly grow in size). This could answer questions related to whether topology transitions are only typical in the early stages of growth. And whether the beginnings of different airlines are mostly the same, and they diverge (ex: Southwest) or whether they have different histories from the start. Another advantage of better computational resources would be the ability to search for larger motif patterns. This should confirm some of the findings related to topologically-generalized motifs, by finding larger motifs in the same families, but it may also uncover new structures.

The potential for analyzing larger datasets such as the network of Internet routers, some biological datasets, power grids would be possible with better computational resources.

6.2.2 Approach limitations

Probabilistic motif finding (NP problem)

The full factorial motif finding was hard to implement given the computational resources described above, the size of the data and the limitations of the Python language. As described in Chapter 2, the motif finding is based on a randomly-seeded breadth-first-search across the graph. The rest of the motif search is deterministic. We have not compared our algorithm to other motif-finding software, neither in approach, nor in computational time. Such assessment might augment the results or speed up the search.

Topology vector/profile

The topology vector is an intuitive idea, but obviously cannot claim to be a comprehensive description of topology. It proved to be a fairly good relative measure, but cannot be used universally to classify topologies. The topology profile idea validates well when tested with canonical topologies, however the question of a topology continuum or topology space is still open. It could be that graph topology space is not metrizable, in the way we have approached the problem, so the vector distance is a *notional*, but not a real metric. Our research adds a small step in the direction of understanding topology space and whether it is metrizable.

Canonical topology comparison

The canonical topology comparison is done point-by-point rather than continuously. This means that at each step a new canonical graph is created for comparison with the real network, not grown alongside it. This is a forced comparison since every snapshot of real data is based on the previous one. The reason why the canonical networks were not grown along is that with the "wrong" initial

conditions, some of the random versions can stray too far from the real graph, while on average they might not be that different.

Southwest growth model probabilities not refined (Section 4.4)

The hub seeding and Southwest growth models are probabilistic, based on frequency of certain events in the real data. These frequencies / probabilities were not studied from the real datasets, for the Southwest growth model, but adjusted to match the topology vector better. The difficulty in calculating probabilities is that in the case of Southwest, hubs are harder to define, based on connectivity because airports are very well connected, and local connectivity is hard to define, because it is based on geography, and not necessarily path lengths. Certainly these probabilities could be studied further from the data.

6.2.3 Airline analysis

The airline analysis leaves many open trails. Aircraft rotations were not taken as a crucial component of route structure, which they certainly are. Neither was competition or congestion at airports considered. Southwest's flying a parallel network (same geographical area, but secondary airports) was not discussed, while it has huge implications for the ability of the airline to fly with such high density and yet perform so well. Also, in general we did not tie performance and cost with any of the metrics we discussed. The drivers behind topology transitions and airline history were not discussed explicitly. There are angles of analysis that we did not address, for example, node-role based approaches [16]. We also did not discuss the meaning of the patterns of evolution we found for airline operations and management.

6.3 Future Work

There is a lot of potential for further work, as this research has only touched the surface on the topics of network topology, topology evolution and pattern finding in real systems. The following list discusses various directions of future research, both in the theoretical direction, as well as the applications side.

Topology Analysis

- Develop further the idea of the topology vector:
The topology vector we developed features 5 metrics that easily non-dimensionalized and tend to scale well between 0 and 1. The question of scaling and dominance of the various metrics could still be explored further. For example, the s-max measure for all real systems tends to fall roughly between 0.5 and 0.9 (for airlines). It is an open question what the true range is, though 1 is clearly achievable by the s-max graph. If such scaling behavior is fine-tuned, the sensitivity of the vector distance could improve. Another open question is what the measures in that vector should be. We chose 5 metrics that measure very different properties of the graph, but they are certainly not the only cited in the literature. Some additional metrics are the rich club metric, average path length (vs diameter), minimum and maximum degree correlations with the same degree distributions, degree distribution exponents (depending on the distribution fit). These are some possible additions to the topology vector. Further work can analyze these together and search the literature for others, to understand whether adding metrics is actually beneficial. Preliminary analysis shows that reducing the size of the topology vector results in non-uniqueness, i.e. different graphs appearing to the have

the same topologies. To sum up, the question of which metrics and how many is open, and it is only on the surface of the question of topological space continuity and metrizable. A true distance may not be definable in the space of graph topologies. Verifying that this distance is a measure because it is hard to verify even the first condition: uniqueness. Simply, if the distance between two topology vectors is 0, does that mean that the two graphs are isomorphic?

- Weighted and directed graphs:

The non-dimensional topology analysis developed in this thesis is concerned only with undirected, unweighted graphs, though we have discussed implications of edge weights. In reality, airlines networks, for example, are neither undirected, not unweighted. Developing the modeling to address this issue would help look into questions such the effect on network structure of aircraft rotations (if data is available) or the effect of concentrating capacity in parts of the network, and how that affects growth.

- Tensor analysis or multi-dimensional graphs:

The engineering systems analyzed in this thesis are not simple graphs, but exist at the interface of many layers of technical, social, informational and other networks. These can be modeled separately, as we have done, or combined. This would involve looking multi-graph type of analysis and understand how the various layers interact. In the example of Wikipedia, it may be possible to map the network of users/authors to the articles they write and look for common patterns or communities. Understanding the socio-technical layers of an engineering system using network modeling would certify the usefulness and applicability of such methods.

- Similarity visualization:

Representing similarity, or how close two graphs are to a set of canonical topologies is challenging because the inherent distance between canonical graphs and how they should be arranged with respect to each other is an open problem too. The topology profile we present in Chapter 2 orders canonical topologies roughly by increasing density. With increasing density, “complexity” or interconnectedness also increases, though this is not a linear scale either. For example, the most dense graph, a complete graph is straightforward to understand. Experimenting with other visualization techniques than a linear scale can be helpful in interpreting the results better. An example is Multidimensional Scaling (MDS) which can be used to detect meaningful underlying dimensions between studies objects (various topologies). Preliminary experiments show that this method might indeed provide useful generalizations, though some fine-tuning and scaling is needed.

- Motif Search:

The motif search was an integral part of studying system topology. Motif size is related to the question of building block size and the what are the largest detectable patterns, if there are such. One could argue that it is not interesting to search for motifs larger than half of the size of the graph, but then a loop of any size might be interesting to find, depending on what the graph represents. One possible research direction is to improve computation (resources and algorithms) to do larger motif search. Another direction is to look for novel techniques to do combinatorial search or better sampling of the graph, than breadth-first-search from every node.

Airline Networks

Further work in airline networks can be done in both improving the models and in working beyond the scope of this thesis. In terms of models,

- The two proposed growth models, hub-seeding and the Southwest model are probabilistic and only the hub-seeding model probabilities were learned from data. This can be done for the

Southwest model to test whether the model matches the data better stochastically.

- Another extension to the modeling work would be to seed the models with discrete events, such as mergers or major disruptions (cancellations of departures) and simulate various “probability” scenarios based on these events. This work could be beginning of simulation and prediction of actual evolution scenarios.
- Given better computational facilities, we also like to study statistically the entire airline dataset as an evolving network. This might give significant insights into the evolution of the air transportation system as a whole.

In terms of domain research in airline routes evolution, there are a number of factors that can be incorporate in models or investigated as drivers. An important question is whether there is a correlation between airline performance and route structure. Southwest is the most (and often only) profitable airline in the US in the past 10 years or so - and its route structure is clearly an outlier. Would other airlines be profitable if they had the topology of Southwest, or are other economic factors prevalent? A summary of ideas for further exploration is below:

- Look for correlation between airline performance and route structure.
- Account for airline aircraft rotation schedules - see how these correlate with hub structure, and with the Southwest local highly-interlinked substructures.
- Match the airports where different airlines fly; account for Southwest’s ”parallel” network.

Wikipedia Networks

- In this research, we have explored the early growth of 5 Wikipedia datasets. Studying other languages could add knowledge to the base of similarities and differences among Wikipedias. Such an exploration would also reveal a wider set of interesting local structures.
- Part of the challenge of analyzing larger datasets would be to attempt to download the early history of the English Wikipedia and see if the largest and most popular Wikipedia has similar growth patterns.
- A question of cognitive interest is whether there is a limit to the number of links an article can have. Since these articles are not written or edited by one person, the answer can be unexpected. This could be found by computing the edge-to-node ratio (or average degree) of Wikipedia articles for longer part of their history (4-5 years). If the edge-to-node ratio converges or is similar for various Wikipedias, then there is some cognitive limit in how people form associations in knowledge.
- As discussed above, in terms of analyzing all layers of context of a system, future work can be done to understand the network of authors/editors of Wikipedia articles, in relation to the network of articles. It is likely to find that some of the thematic clusters correspond to clusters of authors with domain expertise in the same areas. General knowledge content will probably correspond to more diverse communities of authors.

Predictive models

An important consequence of this research can be prediction of the evolution of systems. We synthesized growth algorithms based on underlying patterns, but did not look at overall factors and testing these algorithms under different scenarios and applying them to different systems. A number of steps that can be taken in this direction are summarized below:

- Implement and test alternative growth models. These can be a range of the same models with different variations on probability or details. There can also be completely different models that can be derived from the underlying patterns. They can be tested on all of the airlines

in the data, or the entire industry as a whole to extract insights. The same models can also be tested with different initial conditions. Some research on social networks indicates that the early beginnings of many trends and popular products are randomly-seeded, but after the take-off the system dynamics is more predictable. Various scenarios, for example with seeding different hubs for growth can be played out to see if the system will evolve in a similar way.

- Look at the phenomenon of decay and shrinking versus growth in systems. There are many examples of systems that are down-sizing and ceasing operations. In our data, ATA is one of the airlines that go bankrupt and stop operations 6 months after the dataset ends, and though that was not studied, the decay in the last months could be detected from the topology patterns. Other systems that could be studied in this context are rail networks in the US in the 1950s, other airlines that ceased operations, and any networked systems that went obsolete.
- Algorithmic prediction is achieved by training algorithms on real data. The only aspect of this we have addressed is some supervised learning in extracting the probabilities for the hub seeding algorithm from the airline data. Since, for some of our examples, such as Wikipedia, there are many data snapshots (up to 1500 in days for example), the algorithms can be trained on various parts of the timeline and used to predict the immediate future. One can also insert noise in the data to account for unforeseen events, and further train the algorithms in a unsupervised way.

With these suggestions for future directions, we conclude with the hope that this thesis has added some knowledge and insight on topics of topology and evolution in the context of real systems.

Appendix A

Additional Materials

A.1 Airlines

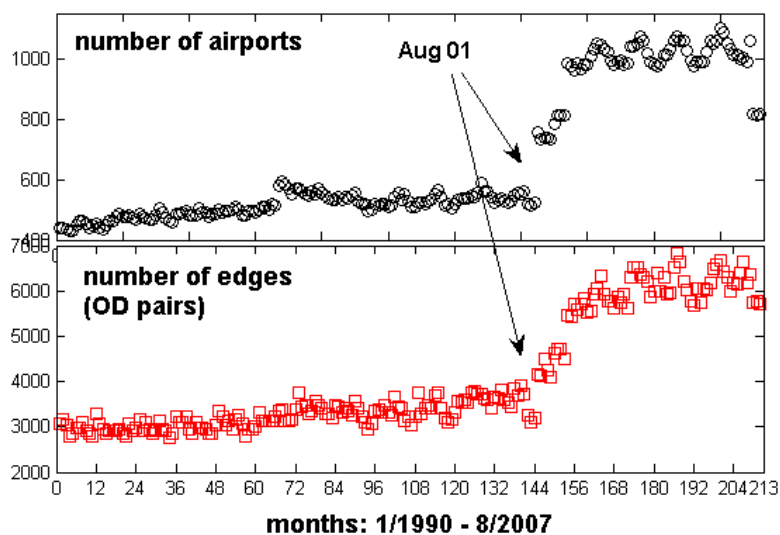


Figure A-1: Number of airports and number of OD pairs monthly for all US airlines reporting to the BTS [1]. In August 2001, the DOT proposes the addition of military, cargo and charter flights to the reports, hence the jump in the data (source: http://www.bts.gov/publications/federal_register/2001/html/bts_20010828.html)

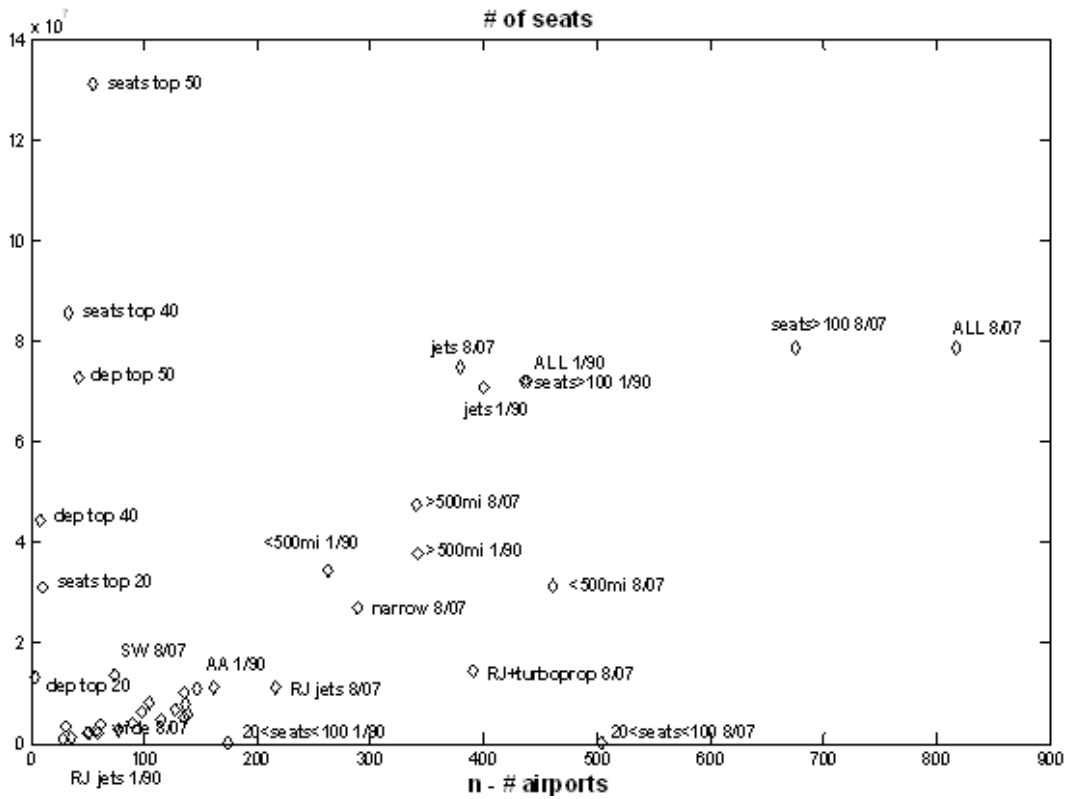


Figure A-2: Total number of seats offered versus number of destinations. All data slices. 1990 and 2007.

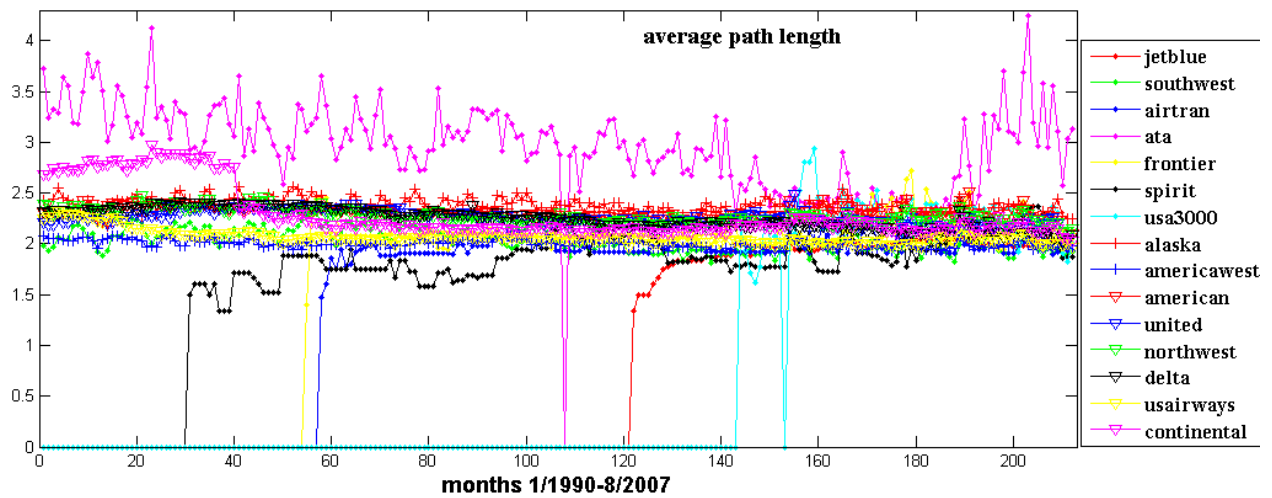


Figure A-3: Average path lengths for the eight major airlines and all low-cost airlines discussed in Chapter 4, 1/1990-8/2007

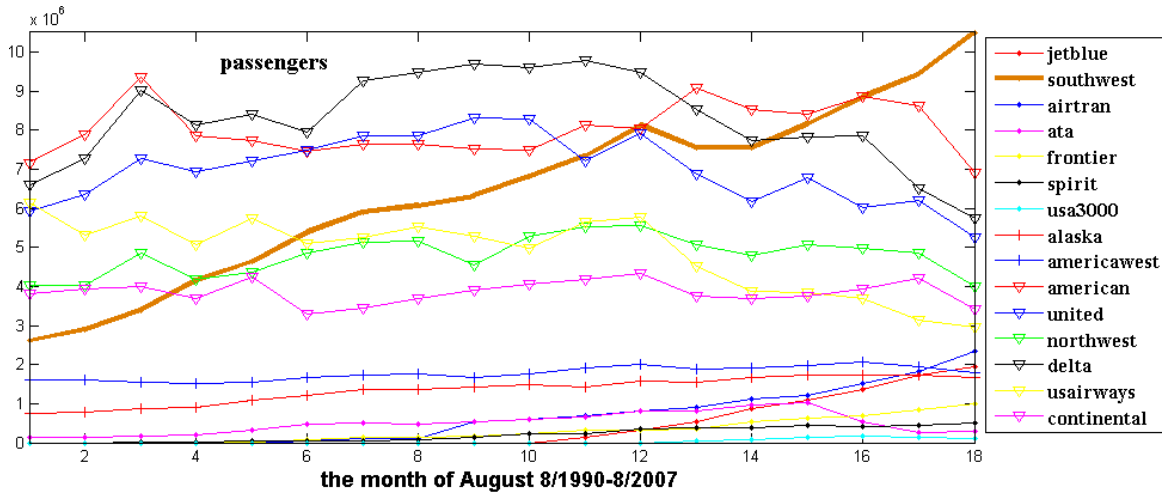


Figure A-4: Passengers carried for the eight major airlines and all low-cost airlines discussed in Chapter 4, plotted for the month of August only, yearly from 8/1990 to 8/2007.

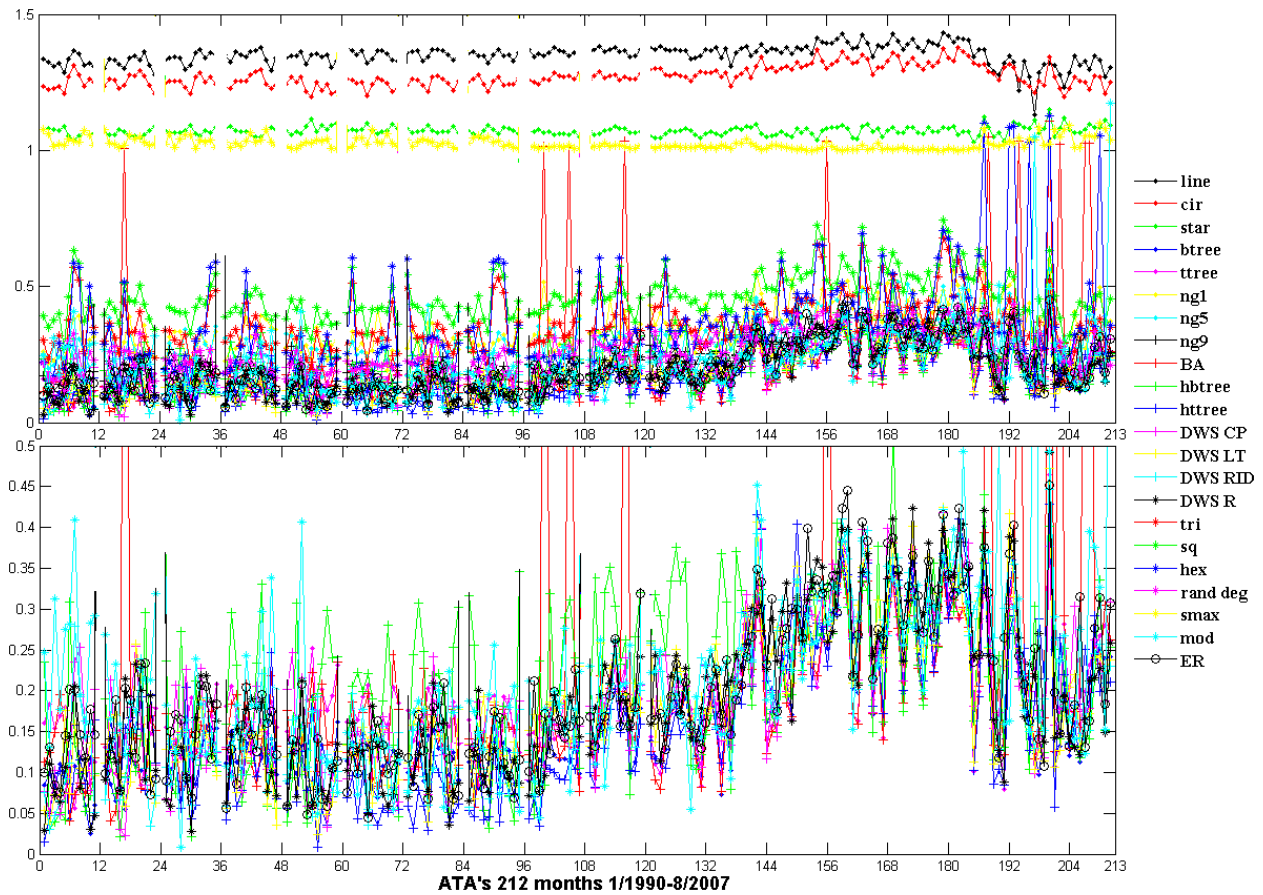


Figure A-5: ATA Airlines topological similarity to canonical topologies over time, 212 months, 1/1990-8/2007

A.2 Language Wikipedias

A.2.1 Interlingua Wikipedia

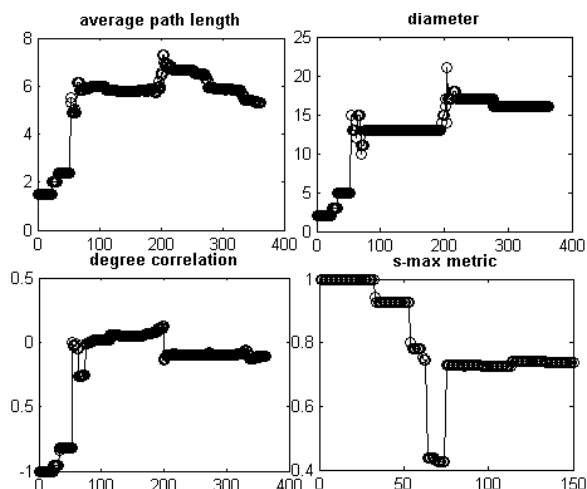


Figure A-6: Graph-theoretical metrics for the Interlingua Wikipedia: degree correlation, average path length, diameter and s-max measure.

A.2.2 The Simple English Wikipedia

The Simple English Wikipedia is a Wikipedia which only uses simple vocabulary and grammar. Users are encouraged to use simple and short sentences, useful and concise pages, so that the Wikipedia is accessible to children and people learning English. With currently more than 44000 articles, the Simple English Wikipedia is ranked 39 by size.

Due to the slow growth of the Simple English Wikipedia we analyze the period between 17 and 24 months or roughly 550 and 730 days. While much slower in growth, this Wikipedia is different from the others also in topology. Figure A-7 shows the article and hyperlink growth for that period.

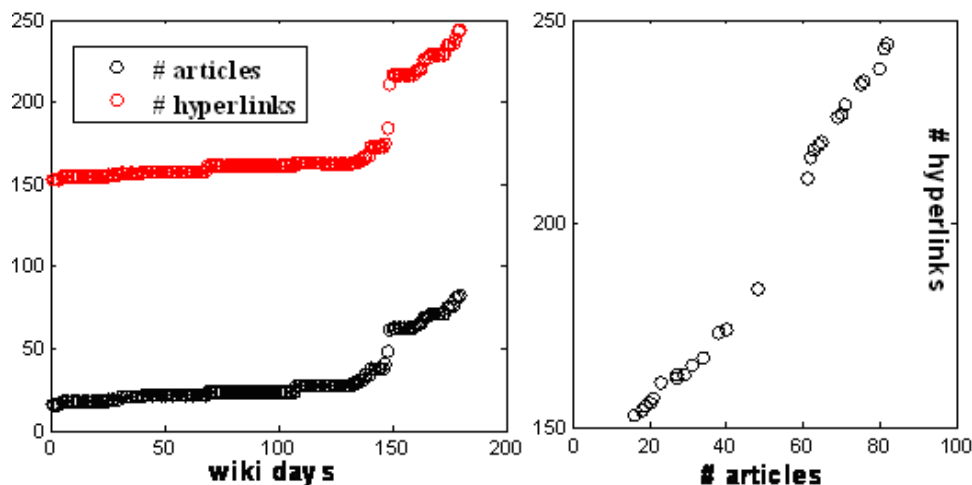


Figure A-7: Nodes and edges in the history of the Simple English Wikipedia from day 550 to day 730.

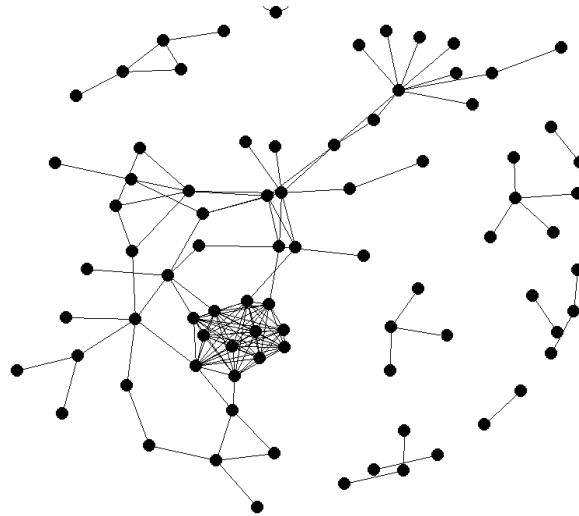


Figure A-8: Day 730 of the Simple English Wikipedia. The fully connected cluster in the middle consists of the months of the year (i.e. the pages for January, February, etc.).

The topology profile for day 730 shows an unusual similarity to lattice, triangular and square, more than any other canonical topologies (Figure A-9).

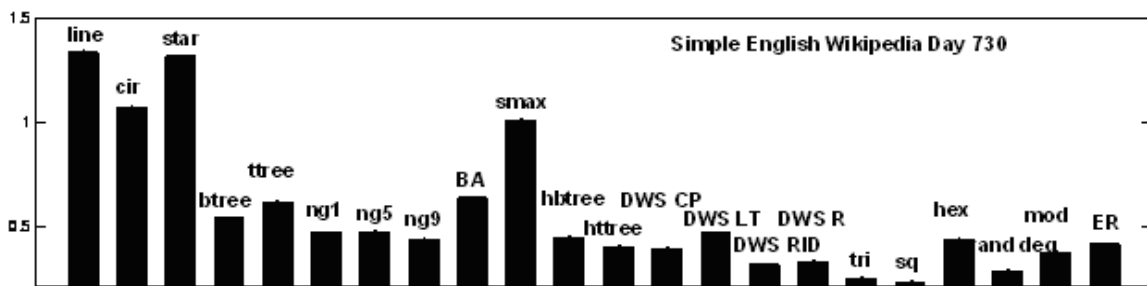


Figure A-9: Topology profile for day 730 of the Simple English Wikipedia.

The fully connected subgraphs (cliques) are very unusual in these sparse networks, so it is no surprise that the significant motifs for day 730 are fully connected 5 and 6-motifs and variations on the theme (Figure A-10).



Figure A-10: Simple English Wikipedia significant motifs for day 730 (2 years)

The canonical topologies matching (Figure A-11) confirms the tendency to look like a lattice over time. Given the size of the wiki this is probably large due to the fully connected subgraph of months of the year. This case of a clique is clearly an exception and it would be interesting if later in the history of this Wikipedia there are more cliques.

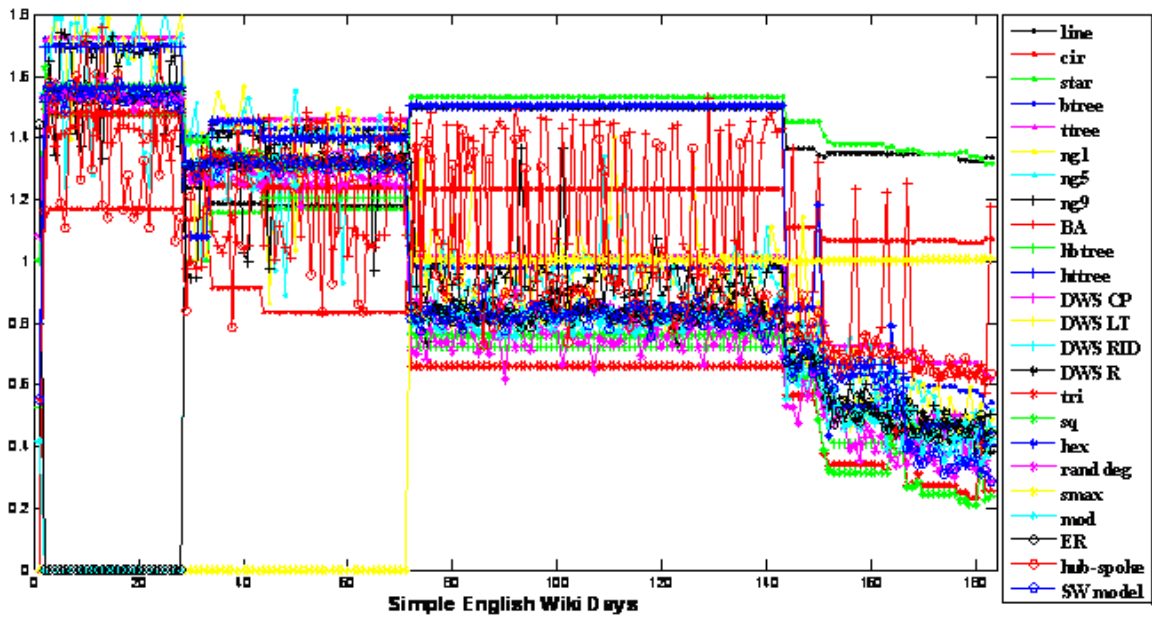


Figure A-11: Canonical networks comparison for the Simple English Wikipedia, days 550 - 730

A.2.3 French and Italian Wikipedias

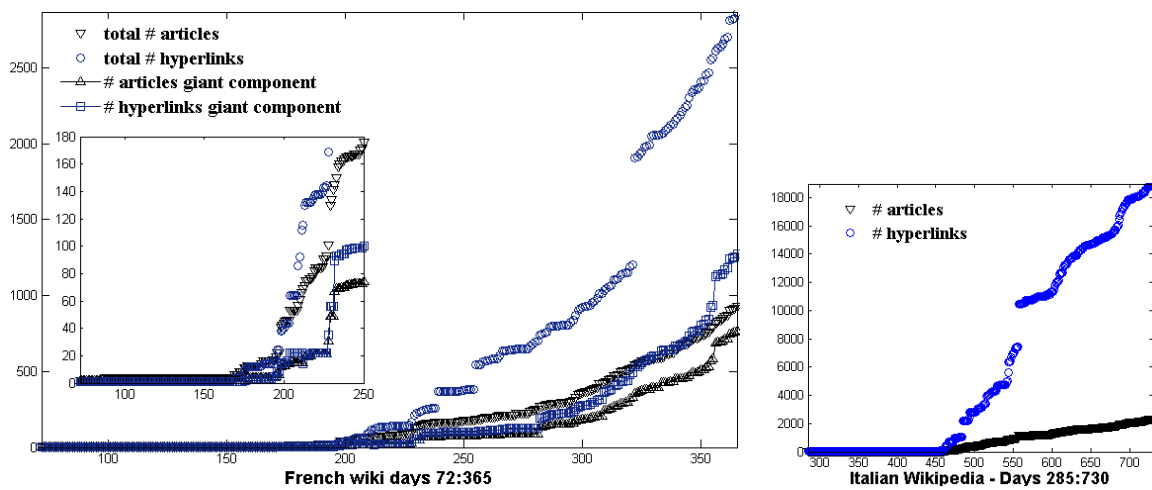


Figure A-12: Growth of the French and the Italian Wikipedias

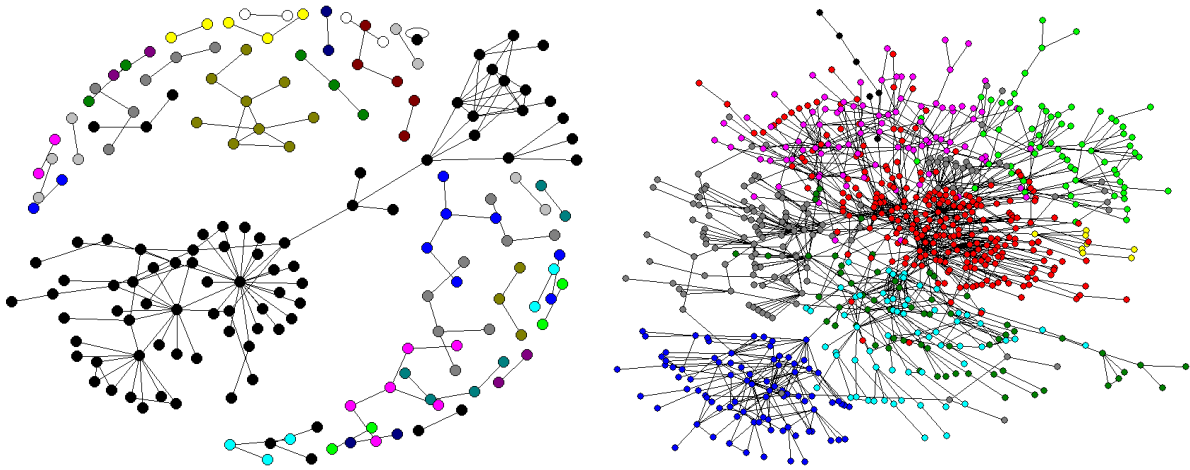


Figure A-13: Days 250 and 365 of the French Wikipedia

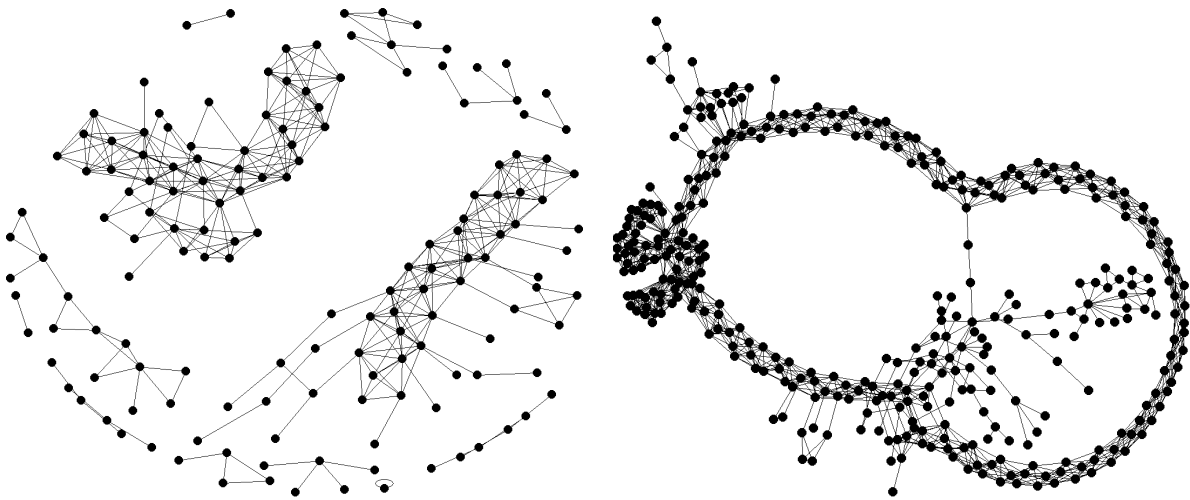


Figure A-14: Days 470 and 500 of the Italian Wikipedia

A.3 Airlines Data Sample

Table A.1: One line from the US BTS [1] data. A JetBlue (B6) flight from Boston to Austin in Jan 2007, 30 departures.

DEPARTURES SCHEDULED	30
DEPARTURES PERFORMED	30
PAYLOAD	786000
SEATS	3000
PASSENGERS	1426
FREIGHT	0
MAIL	29
DISTANCE	1698
RAMP TO RAMP	8643
AIRTIME	7890
UNIQUE CARRIER	B6
ORIGIN	BOS
DEST	AUS
AIRCRAFT TYPE	678
YEAR	2007
MONTH	1

A.4 JetBlue Airways 8/2007 airport codes

ORD - Chicago, IL
IAD - Washington Dulles, DC
DEN - Denver, CO
CLT - Charlotte, NC
SFO - San Francisco, CA
AUS - Austin, TX
PDX - Portland, OR
PWM - Portland, ME
ACK - Nantucket, MA
BOS - Boston, MA
HYA - Hyannis, MA
MVY - Martha's Vineyard, MA
PVC - Provincetown, MA
BUF - Buffalo, NY
BTV - Burlington, VT
CMH - Columbus, OH
FLL - Fort Lauderdale, FL
RSW - Fort Myers, FL
TPA - Tampa, FL
ROC - Rochester, NY
PBI - West Palm Beach, FL
MCO - Orlando, FL
SYR - Syracuse, NY
LGB - Long Beach, CA
SLC - Salt Lake City, UT
OAK - Oakland, CA

SEA - Seattle, WA
SMF - Sacramento, CA
SJC - San Jose, CA
BUR - Burbank, CA
SAN - San Diego, CA
LAS - Las Vegas, CA
TUS - Tucson, AZ
PHX - Phoenix, AZ
MSY - New Orleans, LA
PSE - Ponce, Puerto Rico
BQN - Aguadilla, Puerto Rico
SJU - San Juan, Puerto Rico
EWR - Newark, NJ
SWF - Newburgh, CT
SRQ - Sarasota, FL
JAX - Jacksonville, FL
RDU - Raleigh/Durham, NC
BNA - Nashville, TN
RIC - Richmond, VA
PIT - Pittsburgh, PA
ONT - Ontario, CA
HOU - Houston, TX
LGA - New York, NY
HPN - Westchester County, NY
JFK - New York, NY

A.5 Airline industry data slices

Table A.2: All data slices of the US airlines dataset.

SW 1/90	Southwest Airlines, January 1990
SW 8/07	Southwest Airlines, August 2007
B6 8/07	JetBlue Airways, August 2007
AS 1/90	Alaska Airlines, January 1990
AS 8/07	Alaska Airlines, August 2007
HP 1/90	America West Airlines, January 1990
HP 8/07	America West Airlines, August 2007
AA 1/90	American Airlines, January 1990
AA 8/07	American Airlines, August 2007
UA 1/90	United Airlines, January 1990
UA 8/07	United Airlines, August 2007
DL 1/90	Delta Airlines, January 1990
DL 8/07	Delta Airlines, August 1990
NW 1/90	Northwest Airlines, January 1990
NW 8/07	Northwest Airlines, August 2007
CO 1/90	Continental Airlines, January 1990
CO 8/07	Continental Airlines, August 2007
US 1/90	US Airways, January 1990
US 8/07	US Airways, August 2007
20<seats<100 1/90	all US flights offering between 20 and 100 seats, January 1990
20<seats<100 8/07	all US flights offering between 20 and 100 seats, August 2007
seats>100 1/90	all US flights offering more than 100 seats, January 1990
seats>100 8/07	all US flights offering more than 100 seats, August 2007
seats top 50	top 50% flights by seat capacity over the period 1/1990-8/2007
seats top 40	top 40% flights by seat capacity over the period 1/1990-8/2007
seats top 20	top 20% flights by seat capacity over the period 1/1990-8/2007
jets only 1/90	all US flights served by jet-powered aircraft, January 1990
jets only 8/07	all US flights served by jet-powered aircraft, August 2007
RJ jets 1/90	all US flights served by regional jets, January 1990
RJ jets 8/07	all US flights served by regional jets, August 2007
RJ+turboprop 8/07	all US flights served by regional jet and turboprop aircraft, August 2007
narrow 8/07	all US flights served by narrow-body jets, August 2007
wide 8/07	all US flights served by wide-body jets, August 2007
<500 mi 1/90	all US flights under 500 miles, January 1990
<500 mi 8/07	all US flights under 500 miles, August 2007
>500 mi 1/90	all US flights longer than 500 miles, January 1990
>500 mi 8/07	all US flights longer than 500 miles, August 2007
dep top 50	top 50% flights by number of departures over the period 1/1990-8/2007
dep top 40	top 40% flights by number of departures over the period 1/1990-8/2007
dep top 20	top 20% flights by number of departures over the period 1/1990-8/2007
ALL 1/90	all US flights, January 1990
ALL 8/07	all US flights, August 2007

Bibliography

- [1] Bureau of Transportation Statistics, source: <http://www.bts.gov>.
- [2] M.E.J. Newman. The structure and function of complex networks. SIAM Review, 45(2):167–256, 2003.
- [3] T. Kamada and S. Kawai. An algorithm for drawing general undirected graphs. Information Processing Letters, 31, 1989.
- [4] L. Li, D. Alderson, W. Willinger, and J. Doyle. A first-principles approach to understanding the internet’s router-level topology. In SIGCOMM ’04 Proceedings, 2004.
- [5] J.A. Grochow and M. Kellis. Network Motif Discovery Using Subgraph Enumeration and Symmetry-Breaking, Research in Computational Molecular Biology Annual Conference, 2007.
- [6] N. Kashtan, S. Itzkovitz, R. Milo, and U. Alon. Topological generalizations of network motifs. Physical Review E, 70, 2004.
- [7] R. Milo, S. Itzkovitz, N. Kashtan, R. Levitt, S. Shen-Orr, I. Ayzenshtat, M. Sheffer, and U. Alon. Superfamilies of evolved and designed networks. Science, 303, 2004.
- [8] Gergana Bounova. Graph-theoretical considerations in design of large telescope arrays for robustness and scalability. Master’s thesis, Massachusetts Institute of Technology, 2005.
- [9] P.S. Dodds, D.J. Watts, and C.F. Sabel. Information exchange and the robustness of organizational networks. PNAS, 100(21), 2003.
- [10] J. Hansman. Overview of recent forces & trends in the airline industry. Technical report, International Center for Air Transportation.
- [11] Sgouris Sgouridis. Symbiotic Strategies in Enterprise Ecology: Modeling Commercial Aviation as an Enterprise of Enterprises. PhD dissertation, Massachusetts Institute of Technology, Engineering Systems Division, August 2007.
- [12] Philippe Bonnefoy. Scalability of the Air Transportation System and Development of Multi-Airport Systems: A Worldwide Perspective. PhD dissertation, Massachusetts Institute of Technology, Engineering Systems Division, June 2008.
- [13] M. Giffin, O.L. de Weck, G. Bounova, R. Keller, C. Eckert, and J. Clarkson. Change propagation analysis in complex technical systems. Journal of Mechanical Design, 2008.
- [14] A-L Barabasi and R. Albert. Emergence of scaling in random networks. Science, 286(5439):509–512, 1999.

- [15] M.E.J Newman. Detecting community structure in networks. European Physical Journal B, 38, 2004.
- [16] R. Guimera, Sales-Pardo M., and L.A.N. Amaral. Classes of complex networks defined by role-to-role connectivity profiles. Nature Physics, 3, 2007.
- [17] R. Guimera, S. Mossa, A. Turttschi, and L.A.N Amaral. The worldwide air transportation network: Anomalous centrality, community structure, and cities' global roles. PNAS, 102(22):7794–7799, 2005.
- [18] Oliver W. Wojahn. Airline Networks. Peter Lang, 2001.
- [19] S. Wasserman and K. Faust. Social network analysis. Cambridge University Press, 1994.
- [20] S.N. Dorogovtsev and J.F.F. Mendes. Evolution of Networks: From Biological Nets to the Internet and WWW.
- [21] H. Sachs, M. Stiebitz, and R.J. Wilson. A historical note: Euler's königsberg letters. Journal of Graph Theory, 12, 1988.
- [22] M. Lima. Visual Complexity, source: <http://www.visualcomplexity.com>.
- [23] Database of Interacting Proteins, source: <http://dip.doe-mbi.ucla.edu/dip/Stat.cgi>.
- [24] The Official Google Blog. We knew the web was big..., July 2008. source: <http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html>.
- [25] M.E.J Newman. Modularity and community structure in networks. PNAS, 103(23):8577–8582, June 2006.
- [26] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network motifs: Simple building blocks of complex networks. Science, 298(5594):824–827, October 2002.
- [27] S. Itzkovitz, R. Levitt, N. Kashtan, R. Milo, M. Itzkovitz, and U. Alon. Coarse-graining and self-dissimilarity of complex networks. Physical Review E, 71:1539–3755, October 2005.
- [28] The Cooperative Association for Internet Data Analysis. source: <http://www.caida.org/home>.
- [29] M.E.J. Newman and M. Girvan. Finding and evaluating community structure in networks. Physical Review E, 69, 2004.
- [30] H.S. Osborne. A General Switching Plan for Telephone Toll Service, Bell system Technical Journal, July 1930, pp 429-447.
- [31] L. Li, D. Alderson, J. Doyle, and Willinger W. Towards a theory of scale-free graphs: Definition, properties, and implications. Internet Mathematics, 2(4):431–523, 2005.
- [32] D. Whitney and D. Alderson. Are technological and social networks really different? Interjournal, 2006.
- [33] J. Travers and S. Milgram. An experimental study of the small world problem. Sociometry, 32(4):425–443, 1969.
- [34] D. Watts and S. Strogatz. Collective dynamics of 'small-world' networks. Nature, 393, 1998.

- [35] M. Granovetter. The strength of weak ties: A network theory revisited. Sociological Theory, 1, 1983.
- [36] M. Fiedler. Algebraic connectivity of graphs. Czechoslovak Mathematical Journal, 23, 1973.
- [37] Mo-Han Hsieh. Standards as Interdependent Artifacts: the Case of the Internet. PhD dissertation, Massachusetts Institute of Technology, Engineering Systems Division, August 2007.
- [38] S. Kirkpatrick, C.D. Gelatt Jr., and M.P. Vecchi. Optimization by simulated annealing. Science, 220:671–680, May 1983.
- [39] P. Erdős and A. Rényi. On Random Graphs, Publ. Math. Debrecen 1959, 6, p.290-297.
- [40] D.J. de S. Price. Networks of scientific papers. Science, 149, 1965.
- [41] A. Fabrikant, E. Koutsoupias, and C.H. Papadimitriou. Heuristically optimized trade-offs: A new paradigm for power laws in the internet. International Colloquium on Automata, Languages and Programming. Springer-Verlag LNCS, 2002.
- [42] M.T. Gastner and M.E.J Newman. Shape and efficiency in spatial distribution networks. Journal of Statistical Mechanics: Theory and Experiment, 2006.
- [43] B. Cohanin, J. Hewitt, and O.L. de Weck. The design of radio telescope array configurations using multiobjective optimization: Imaging performance versus cable length. The Astrophysical Journal Supplement Series, 154, 2004.
- [44] A. Vazquez, A. Flammini, A. Maritan, and A. Vespignani. Modeling of protein interaction networks. ComPlexUs, 2002.
- [45] R. Sole, R. Pastor-Satorras, E. Smith, and T. Kepler. A model of large-scale proteome evolution. Advances in Complex Systems, 5(43), 2002.
- [46] Duncan J. Watts. Six Degrees: The Science of a Connected Age. W. W. Norton, 2004.
- [47] D. Whitney, 2006. Degree Correlation Introduction, Lecture 10, ESD.342 2006, MIT OpenCourseWare.
- [48] P. Bonnefoy and J. Hansman. Emergence of secondary airports and dynamics of regional airport systems in the united states. Technical Report ICAT-2005-02, International Center for Air Transportation, May 2005.
- [49] V.D. Blondel, A. Gajardo, M. Heymans, P. Senellart, and P. Van Dooren. A measure of similarity between graph vertices. with applications to synonym extraction and web searching. SIAM Review, 46(4), 2004.
- [50] J. Kleinberg. Authoritative sources in a hyperlinked environment. Journal of ACM, 46, 1999.
- [51] K. Hendricks, M. Piccione, and G. Tan. The economics of hubs: The case of monopoly. Review of Economic Studies, 62, 1995.
- [52] The Next Generation Air Transportation System: Status and Issues, Hearing Charter, US House of Representatives Committee on Science and Technology, Sep 2008.

- [53] The Next Generation Air Transportation System: Preliminary Analysis of Progress and Challenges Associated with the Transformation of the National Airspace System, GAO-06-915T, July 2006.
- [54] G. Bounova, H. Yan, J. Silvis, Q. Li, and L. Jianghai. Analysis and optimization of airline networks: A case study of china. Technical report.
- [55] M. Boguna, D. Krioukov, and K.C. Claffy. Navigability of complex networks. Nature Physics, 2008.
- [56] Y. Lee and R Kornfeld. Examination of the hub-and-spoke network: A case example using overnight package delivery. Number AIAA 2003-1334. AIAA 41st Aerospace Sciences Meeting and Exhibit, 6-9 January 2003, Reno NV.
- [57] G. Burghouwt, J. Hakfoort, and J. Ritsema van Eck. The spatial configuration of airline networks in europe. Journal of Air Transport Management, 9, 2003.
- [58] D. Bhadra and B. Hogan, 2005. A Preliminary Analysis of the Evolution of US Air Transportation Network.
- [59] B. Bollobas and O. Riordan. The diameter of a scale-free random graph. Combinatorica, 24(1):5–34, 2004.
- [60] B. Bollobas. The diameter of random graphs. Transactions of the American Mathematical Society, 267(1):41–52, 1981.
- [61] Wikimedia Database Dumps, source: <http://download.wikimedia.org/backup-index.html>.
- [62] List of Wikipedias, source: http://meta.wikimedia.org/wiki/List_of_Wikipedias.
- [63] Dimitar Bounov, 2008. Wikipedia Data Mining Project, source: <http://web.mit.edu/gerganaa/www/wikipedia.html>.