

Cliques and a New Measure of Clustering: with Application to U.S. Domestic Airlines

Steve Lawford[†] and Yll Mehmeti

Data, Economics and Interactive Visualization (DEVI) group, ENAC (University of Toulouse),
7 avenue Edouard Belin, CS 54005, 31055, Toulouse, Cedex 4, France

[†]Corresponding author. Email: steve.lawford@enac.fr

Abstract

We propose a higher-order generalization of the well-known overall clustering coefficient for triples $C(3)$ to any number of nodes. We give analytic formulae for the special cases of three, four, and five nodes and show that they have very fast runtime performance for small graphs. We discuss some theoretical properties and limitations of the new measure, and use it to provide insight into dynamic changes in the structure of U.S. airline networks.

1 Introduction

Complex networks are widely used to describe important systems, with applications to biology, technology and infrastructure, and social and economic relationships [4, 5, 25, 59, 69, 83]. A network or “graph” involves a set of nodes or “vertices” that are linked by edges. For example, an airline company’s transportation of passengers can be thought of as a network of airports (nodes) joined by routes that have regular service (edges). The statistical physics and graph theory communities have focused in particular on the topology and dynamics of random and real-world networks, and have been successful in identifying robust structural features and organizational principles.¹ These include the small-world property, characterized by systems that are highly clustered but have short characteristic path lengths; and scale-free networks, which means that the number of neighbours of a node, or its “degree”, follows a power-law distribution whereby the topology of the system is dominated by a few high degree nodes [7, 22, 72].

One common property of networks is clustering or “transitivity”, which measures the relative frequency with which two neighbours of a given node are also neighbours of one another, forming a connected triangle of nodes. Many real-world networks display higher levels of clustering than would be expected if those networks were random, with nodes creating tightly connected groups [4, 58, 69, 72]. Clustering is especially important in economic and social networks, and there is strong evidence that it is related to cooperative social behaviour and beneficial information and reputation transfer [41, 42, 44, 59]. Other recent examples of the empirical application of graphs in economics include [6, 32, 42] (social networks) and [3, 23, 31, 40, 64] (financial networks).

In this paper, we focus our empirical application on air transportation. Recent work has considered air cargo networks [12, 53], the world-wide airport network [19, 37, 38, 51, 67, 71], and airline networks in the U.S. [2, 9, 35, 50, 65, 66, 73], Europe [63], and China [18, 30]; good surveys of research in this area appear in [52, 65, 79]. Typically, these papers report a selection of summary statistics to capture global or local aspects of the network, and provide insight into topology and dynamics that would not be available from other methods.

¹The field is continually expanding and is far too large to survey here. We thank one anonymous referee for pointing us towards recent work on edge prediction [80] and multiplex models [81, 82].

One widely used measure of clustering is the *overall clustering coefficient* or “transitivity” which is defined in [8, 59, 60, 61, 62] as, in our notation,

$$C(3) = \frac{3 \times \text{number of triangles in the network}}{\text{number of connected triples of vertices}}, \quad (1)$$

where a *connected triple* is a set of three distinct nodes u , v and w , such that at least two of the possible edges between them exist. In a social network this measures how often an individual’s “friends” are also friends with one another, on average, across the entire network. An alternative measure of clustering, the *average clustering coefficient*, takes a different approach to (1) and is computed locally for each node, and then averaged across all nodes.² We focus on overall clustering (transitivity) in this paper.

There is now substantial evidence that significant topological structures (known as “graphlets”, “motifs” or “subgraphs”), on more than three nodes, can be found in real-world networks, and that they may perform precise specialized functions [1, 10, 11, 14, 45, 55, 74]. Since the usual clustering coefficient $C(3)$ is based upon connected triples of nodes, it is natural to ask whether a similar measure can be derived for any number of nodes. A generalized clustering coefficient could potentially identify hidden higher-order clustering and enable a better understanding of the structure of real-world networks. The need to go beyond usual three-node *average* clustering has been addressed by [16] (cycles of length four), who find that “grid clustering” scales with node degree in a similar way to usual clustering, [34] (shortest paths of length greater than one between a node’s neighbours), who confirm the absence of clustering in Barabási-Albert preferential attachment (scale-free) models, and [46] (connectivity of more than one of a node’s nearest neighbours), who investigate scaling properties using data on urban street networks.

The work that is most closely related to the present paper is by Yin-Benson-Leskovec [76], hereafter referred to as YBL, who propose new overall and average clustering coefficients based on the relative frequency of cliques of order greater than three. Our work differs essentially from YBL’s *overall* coefficient in the way that we define the “relative frequency”, but also in the methods that are used for computation, and the motivation and empirical application. We draw careful comparisons between our method and YBL, for theoretical Erdős-Rényi random graphs and simulated small-world models, and argue that the two approaches are complementary.³

In this paper, we make the following specific contributions:

- We propose a new generalized clustering coefficient $C(b)$, based upon connected groups of b nodes, which nests the standard clustering coefficient $C(3)$. We develop a very fast analytic implementation for connected groups of three, four and five nodes, that we show to be up to 2,000 times faster than a naïve nested loop algorithm, for some small dense graphs.
- We examine the theoretical properties of $C(b)$ for Erdős-Rényi and small-world random graphs, and lollipop graphs, and draw comparisons with YBL. We show that it will become increasingly difficult to compute $C(b)$ efficiently as b becomes large, even using analytic formulae. Using dynamic data on U.S. airline networks, we also observe that $C(b)$ can be highly correlated across b , and with network density. When we control for lower-order clustering, we find low to moderate higher-order clustering in these networks. It is not known whether this finding holds generally for large classes of networks.

All of the analytic formulae that we use, and several proofs, are collected in Appendix A, and additional figures and tables are reported in Appendix B.

²Overall clustering (1) assigns the same weight to every triangle in the graph. Average clustering gives each node the same weight. Since high-degree nodes may be adjacent to more triangles than low-degree nodes, overall and average clustering can give different values.

³We became aware of the excellent [76] after the original version of our paper had been completed and submitted to the arXiv repository. The present paper contains substantial new material to address this omission. There is related work by YBL-Gleich [75] and by YBL [77].

2 Graph Theory and Clustering

We briefly review some relevant tools of graph theory. Important monographs include [29] (mathematics), [41] (economics of social networks) and [47] (algorithms). A *graph* is an ordered pair $G = (V, E)$ where V and E denote the sets of *nodes* and *edges* of G , respectively. We use $n = |V|$ and $m = |E|$ to represent the numbers of nodes and edges of G . A graph has an associated $n \times n$ *adjacency matrix* g , with representative element $(g)_{ij}$ that takes value one when an edge is present between nodes i and j , and zero otherwise. We also use $(i, j) \in E$ to denote an edge between nodes i and j , and say that they are *directly-connected*. A graph is *simple and unweighted* if $(g)_{ii} = 0$ (no self-links) and $(g)_{ij} \in \{0, 1\}$ (no pair of nodes is linked by more than one edge, or by an edge with a weight that is different from one). A graph is *undirected* if $(g)_{ij} = (g)_{ji}$. A *walk* between nodes i and j is a sequence of edges $\{(i_r, i_{r+1})\}_{r=1, \dots, R}$ such that $i_1 = i$ and $i_{R+1} = j$, and a *path* is a walk with distinct nodes. A graph is *connected* if there is at least one path between any pair of nodes i and j ; otherwise the graph is *disconnected*. A *bridge* is an edge the removal of which will disconnect the graph. In this paper, we consider simple, unweighted, undirected and connected graphs.

The *degree* $k_i = \sum_j (g)_{ij}$ is the number of nodes that are directly-connected to node i , and the (*1-degree*) *neighbourhood* of node i in G , denoted by $\Gamma_G(i) = \{j : (i, j) \in E\}$, is the set of all nodes that are directly-connected to i . The *density* $d(G) = 2m/n(n-1)$ is the number of edges in G relative to the maximum possible number of edges in a graph with n nodes. A graph $G' = (V', E')$ is a *subgraph* of G if $V' \subseteq V$ and $E' \subseteq E$ where $(i, j) \in E'$ implies that $i, j \in V'$. A *tree* is a connected graph with no cycles. A *spanning tree* on a connected G is a connected subgraph with nodes V and the minimum possible number of edges $m = n - 1$. A *complete graph* on n nodes, K_n , has all possible edges, and a complete subgraph on b nodes is called a *b-clique*. A *maximal clique* is a clique that cannot be made larger by the addition of another node in G with its associated edges, while preserving the complete-connectivity of the clique. A *maximum clique* is a (maximal) clique of the largest possible size in G , and the *clique number* $w(G)$ of the graph G is the number of nodes in a maximum clique in G .

Let $G(n, p)$ be an Erdős-Rényi random graph with nodes $V = \{1, \dots, n\}$ and edges that arise independently with a constant edge-formation probability p , giving a statistically homogeneous network that has, on average, $(n-1)p$ edges for a given node, and $\binom{n}{2}p$ randomly-distributed edges in total. We also use the lollipop graph $L(b, n-b)$, with n nodes and $\frac{1}{2}b(b-3) + n$ edges, and $3 \leq b \leq n$. The lollipop can be thought of as a b -clique K_b that is attached by a bridge to a path graph on $n-b$ nodes. Note that $L(n, 0)$ is the complete graph K_n .⁴ Using the notation of [1], we refer to particular topological subgraphs by $M_a^{(b)}$, where b is the number of nodes in the subgraph, and a is the decimal representation of the smallest binary number derived from a row-by-row reading of the upper triangles of each adjacency matrix g from the set of all topologically-identical subgraphs on the same b nodes (also see [48]).

2.1 Analytic formulae for a generalized clustering coefficient

The clustering coefficient $C(3)$ is bounded by $0 \leq C(3) \leq 1$, attaining the minimum when there are no triangles in the graph, and taking the maximum value for a complete graph K_n . Since each triangle contains three triples of nodes, a factor of three appears in the numerator of (1). A naïve algorithm that is based on nested loops, and considers every distinct triple of nodes in G , will run in $O(n^3)$ time. However, it is easy to write down an analytic version of $C(3)$, using the nested subgraph enumeration formulae in [1, equations (1) and (2)]:

$$C(3) = \frac{3|M_7^{(3)}|}{|M_3^{(3)}|} = \frac{\text{tr}(g^3)}{\sum_i k_i(k_i - 1)}, \quad (2)$$

and where $C(3)$ makes explicit the definition of clustering in terms of triples $M_3^{(3)}$ and triangles $M_7^{(3)}$.

⁴The lollipop was first introduced by [49, Example 2] in the one parameter case $b = n/2$, and was generalized to two parameters by [15]. It has applications in the fields of combinatorics (Ramsey theory) [33] and linear algebra (spectral theory) [13, 39].

If we instead interpret (2) as the average probability that any three connected nodes in a graph are also completely-connected, then a natural generalization follows to any number b of nodes, such that $3 \leq b \leq n$. In this paper, we define the *generalized clustering coefficient* as follows:

$$C(b) = \frac{a(b) \times \text{number of } b\text{-cliques in } G}{\text{number of } b\text{-spanning trees in } G}, \quad (3)$$

where *Cayley's formula* $a(b) = b^{b-2}$ gives the number of spanning trees in K_b , and ensures that $0 \leq C(b) \leq 1$. Clearly, $C(b)$ nests $C(3)$, and equals zero if and only if there are no b -cliques in the graph. It is natural that the generalized clustering coefficient should attain its maximum value for a complete graph, in the same way as $C(3)$, and we show this in:

Proposition 2.1. *Let G be a connected graph with at least b nodes ($b \geq 3$). Then $C(b) = 1$ if and only if G is complete.*

See Appendix A for a proof of Proposition 2.1.

A naïve algorithm for (3), based on nested loops, will run in $O(n^b)$ time. For example, the denominator of (3) can be calculated by considering every distinct set of b nodes in G , and counting the number of spanning trees on each subgraph. This will be excessively slow. If we instead think of $C(b)$ as a measure of the prevalence of b -cliques relative to all connected groups of b nodes, then it is clear that we can use analytic subgraph enumeration for counting the cliques and the spanning trees for the special cases $C(4)$ and $C(5)$, in the same way as for (2):

$$C(4) = \frac{16 |M_{63}^{(4)}|}{|M_{11}^{(4)}| + |M_{13}^{(4)}|} = \frac{4 \sum_i \text{tr}(g_{-i}^3)}{\sum_i k_i(k_i - 1)(k_i - 2) + 6 \sum_{(i,j) \in E} (k_i - 1)(k_j - 1) - 3 \text{tr}(g^3)}. \quad (4)$$

$$C(5) = \frac{125 |M_{1023}^{(5)}|}{|M_{75}^{(5)}| + |M_{77}^{(5)}| + |M_{86}^{(5)}|} = \frac{25 \sum_i \sum_{j \in \Gamma_G(i)} \text{tr}(((g_{-i})_{-j})^3)}{\sum_i k_i(k_i - 1) \{ (k_i - 2)(k_i - 15) - 24 \} + 12 \sum_{(i,j) \in E} (k_i - 1)(k_i + k_j - 8)(k_j - 1) - 48 \sum_i (g^3)_{ii}(k_i - 2) + 12 \sum_{i \neq j} (g^4)_{ij} - 12 \text{tr}(g^3)}. \quad (5)$$

The numerator terms $|M_{63}^{(4)}|$ and $|M_{1023}^{(5)}|$ are the number of 4-cliques and 5-cliques respectively. The denominator terms are the counts of the 4-star ($|M_{11}^{(4)}|$), the 4-path ($|M_{13}^{(4)}|$), the 5-star ($|M_{75}^{(5)}|$), the 5-arrow ($|M_{77}^{(5)}|$), and the 5-path ($|M_{86}^{(5)}|$), which are illustrated in Figures 1 and 2. Since there are sixteen possible spanning trees on any given four nodes in the graph, all of which will occur in K_4 , the factor $a(4)$ equals 16. Similarly, counting the distinct 5-spanning trees in K_5 gives $a(5)$ equal to 125. We do not recommend using the right-hand-sides of (4) and (5) for computation. We report the runtime performance of the analytic formulae in Appendix B.1 for small graphs.

However, while this approach seems promising, it will rapidly become hard to derive analytic formulae for larger values of b , because the number of denominator terms will explode. Essentially, we would need to find a formula for *every* non-isomorphic tree on b nodes. For example, $C(6)$ would require evaluation of six denominator terms (Figure 3). Numerical values for the number of trees on n unlabelled nodes are given as series A000055 in the Online Encyclopedia of Integer Sequences (<http://oeis.org/A000055>). For example, $C(7)$ has 11 denominator terms, $C(8)$ has 23 denominator terms, and $C(36)$ has more than 6.2×10^{12} terms! This creates an intrinsic bound on the general applicability of analytic formulae for $C(b)$: we can reasonably expect to use them for $C(3)$, $C(4)$, $C(5)$ and perhaps $C(6)$ and $C(7)$, but not beyond. There has been considerable research on efficient numerical algorithms for generating all possible spanning trees of a simple undirected connected graph; see [17] for a review and comparison

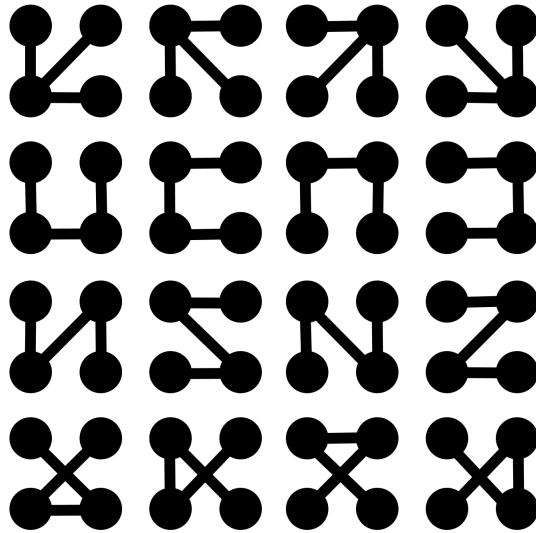


Figure 1: The sixteen spanning trees on four labelled nodes: four 4-stars $M_{11}^{(4)}$ and twelve 4-paths $M_{13}^{(4)}$. This illustrates Cayley's formula $a(b) = b^{b-2}$, which appears in the numerator of the generalized clustering coefficient, for $b = 4$.

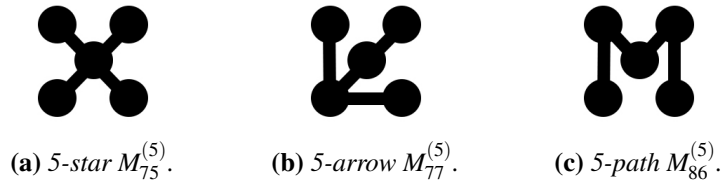


Figure 2: The three spanning trees on five unlabelled nodes: the 5-star $M_{75}^{(5)}$, the 5-arrow $M_{77}^{(5)}$ and the 5-path $M_{86}^{(5)}$. The total count of these subgraphs appears in the denominator of the generalized clustering coefficient $C(b)$, for $b = 5$.

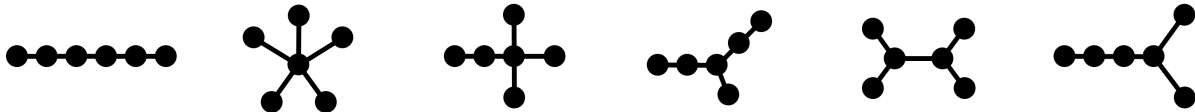


Figure 3: The six non-isomorphic spanning trees on six unlabelled nodes. The total count of these subgraphs appears in the denominator of the generalized clustering coefficient $C(b)$, for $b = 6$. The number of spanning trees that must be counted in order to compute $C(b)$ increases very rapidly with the number of nodes b in the subgraph.

of different methods. It is possible that numerical methods could be used to extend $C(b)$ to higher values of b , although computation of $C(b)$ requires consideration of *every* possible set of b connected nodes in the graph, and it is well known that the number of spanning trees of a graph increases exponentially in the number of nodes.⁵

2.2 The generalized clustering coefficient C_{b-1} of Yin-Benson-Leskovec

In recent work, Yin-Benson-Leskovec [76] develop an overall generalized clustering coefficient, based on clique expansion, that is closely related to $C(b)$.⁶ Their coefficient (Equation 4 in [76]) is

$$C_\ell = \frac{(\ell^2 + \ell) |K_{\ell+1}|}{|W_\ell|}; \quad \ell \geq 2,$$

where $K_{\ell+1}$ is the set of $(\ell + 1)$ -cliques and W_ℓ is the set of ℓ “wedges” (they define a wedge as an ℓ -clique with one additional node that is adjacent to any node in the clique). Despite the apparently different formulation, we can write YBL’s coefficient in the notation of our paper, with $\ell = b - 1$, as

$$C_{b-1} = \frac{(b^2 - b) |K_b|}{|L(b-1, 1)|}; \quad b \geq 4,$$

where $L(\cdot, \cdot)$ is the lollipop graph. Note that the lollipop $L(2, 1)$ is not typically defined, and so we need $b - 1 \geq 3$. YBL nest the usual clustering coefficient in their generalized framework by implicitly defining $L(2, 1)$ as a 3-path with directed edges (double-counting the undirected 3-path), which gives the coefficient 6 in the numerator rather than the usual 3. This is just a counting issue, and so we assume in the rest of the paper that $C(3) = C_2$.

We now compare the higher-order clustering coefficients on four and five nodes:

$$C(4) = \frac{16 |K_4|}{\text{number of 4-spanning trees in } G} = \frac{16 |M_{63}^{(4)}|}{|M_{11}^{(4)}| + |M_{13}^{(4)}|}; \quad C_3 = \frac{12 |K_4|}{|L(3, 1)|} = \frac{12 |M_{63}^{(4)}|}{|M_{15}^{(4)}|},$$

where $M_{15}^{(4)}$ is the tadpole subgraph; and

$$C(5) = \frac{125 |K_5|}{\text{number of 5-spanning trees in } G} = \frac{125 |M_{1023}^{(5)}|}{|M_{75}^{(5)}| + |M_{77}^{(5)}| + |M_{86}^{(5)}|}; \quad C_4 = \frac{20 |K_5|}{|L(4, 1)|} = \frac{20 |M_{1023}^{(5)}|}{|M_{127}^{(5)}|},$$

where $M_{127}^{(5)}$ is the kite subgraph (see [48] for the count formula). Hence YBL’s coefficient C_{b-1} fits naturally into the analytic framework of our paper. Likewise, our coefficient $C(b)$ can potentially use similar computational techniques to those in YBL. Intuitively, both methods will face computational difficulties as b becomes large, and in fact YBL do not go beyond $b = 5$. In practice, this is unlikely to be a serious issue, as shown by the empirical results of YBL, and by Section 3 in this paper. Small values of b appear to be sufficient to capture much of the higher-order clustering that is present in some real-world networks.

The essential difference between $C(b)$ and C_{b-1} is in the definition of the “relative frequency”. We compute the frequency of b -cliques relative to the number of minimally connected subgraphs on b nodes. On the other hand, YBL compute the frequency of b -cliques relative to the number of lollipops $L(b-1, 1)$ i.e. a b -clique where all but one of the edges adjacent to one node have been removed. This reflects the two possible interpretations of the 3-path

⁵Note that the analytical algorithm is not a property of the generalized clustering coefficient itself, but a means to compute it efficiently for small b , and for small to moderate n . As in other areas, when exact analytic methods become intractable, it often becomes necessary to use numerical techniques and approximations instead, such as approximate sampling algorithms and asymptotic results.

⁶“The novelty of our interpretation of the clustering coefficient is considering it as a form of clique expansion rather than as the closure of a length-2 path, which is key to our generalizations in the next section.” (Page 052306-2 in [76])

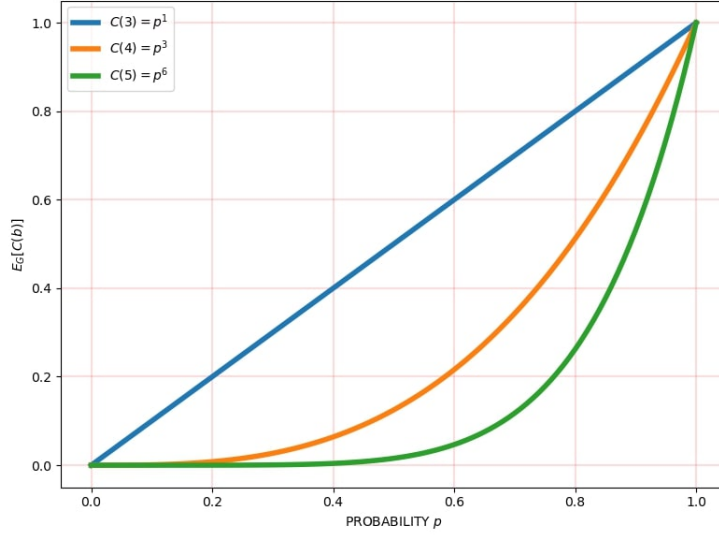


Figure 4: The theoretical expectation of the clustering coefficient $C(b)$ for the Erdős-Rényi random graph $G(n, p)$ is $\mathbb{E}_G[C(b)] = p^{(b-1)(b-2)/2}$ as n becomes large, with edge-formation probability $0 \leq p \leq 1$. We observe that clustering is monotonically increasing in probability p , as the network moves from a set of disconnected nodes to a complete graph. For a given p , the expected level of clustering is decreasing in b .

in the denominator of $C(3)$, either as a spanning tree on three nodes (our paper) or as a 2-clique with an additional adjacent node (YBL), and the two natural generalizations of $C(3)$ to higher-order clustering.

2.3 Analysis of generalized clustering $C(b)$ and C_{b-1} for the $G(n, p)$ model

In the special case of the Erdős-Rényi random graph $G = G(n, p)$, it follows from (3) that the expectation of $C(b)$ is given by $\mathbb{E}_G[C(b)] = p^{(b-1)(b-2)/2}$ as n becomes large, since there are $\binom{n}{b} p^{\binom{b}{2}}$ b -cliques and $b^{b-2} \binom{n}{b} p^{b-1}$ b -spanning trees in $G(n, p)$, on average. Also see Figure 4. We implicitly assume that both $C(b)$ and C_{b-1} are well-defined on G . Numerical values of $\binom{b-1}{2}$ are given in A161680 of the Online Encyclopedia of Integer Sequences (<http://oeis.org/A161680>). We can also show that $\mathbb{E}_G[C_{b-1}] = p^{b-2}$ as n becomes large, since there are $\binom{n}{b} p^{\binom{b}{2}}$ b -cliques and $(b^2 - b) \binom{n}{b} p^{\binom{b-1}{2}+1}$ lollipops in $G(n, p)$, on average (see also Proposition 2(1) in [76] for this result). It follows immediately that $\mathbb{E}_G[C(b)] \stackrel{\geq}{\leq} \mathbb{E}_G[C_{b-1}]$ as $(b-2)(b-3) \stackrel{\leq}{\geq} 0$, with $p \neq 0, 1$. Directly, $\mathbb{E}_G[C(b)] \leq \mathbb{E}_G[C_{b-1}]$, with equality when $p = 0, 1$ (for all b) or when $b = 3$ (for all p). Intuitively, there will be more b -spanning trees than there are lollipops $L(b-1, 1)$ in G . For example, there is one 4-star and two 4-paths in the tadpole; and there can also be 4-spanning trees that are not in a tadpole. In Figure 5 we examine the expected difference between $C(b)$ and C_{b-1} for $G = G(n, p)$, where $\mathbb{E}_G[C_{b-1}] - \mathbb{E}_G[C(b)] = p^{b-2} (1 - p^{(b-2)(b-3)/2}) \geq 0$, with equality at $p = 0, 1$. We see that the difference can be substantial, and that it increases in the edge-formation probability p (as the graph G becomes more dense) up to a certain point, and then falls to zero; but that it can increase or decrease in b , the order of the clustering, for a given p .

2.4 Invariance of C_{b-1} on graphs with vanishing density

The YBL coefficient C_{b-1} ($b \geq 4$) has a peculiar invariance property for a particular class of graphs, that is neither a feature of the usual clustering coefficient $C(3) = C_2$ nor of our generalized clustering $C(b)$. Note that $C(b)$ increases as the number of b -cliques increases or the number of b -spanning trees falls, while C_{b-1} increases as the number of b -cliques increases or the number of lollipops $L(b-1, 1)$ falls. It follows that C_{b-1} will give the *same* value for

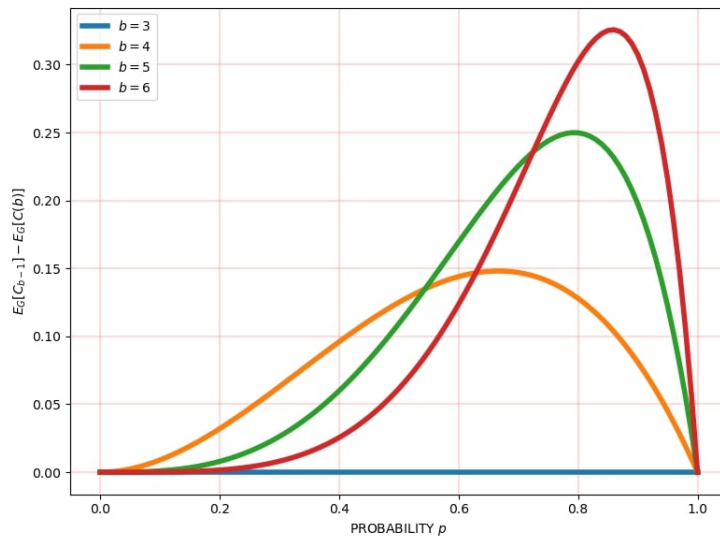


Figure 5: The theoretical difference in expectation between the clustering coefficient C_{b-1} of [76] and our coefficient $C(b)$ for the Erdős-Rényi random graph $G(n, p)$ is $\mathbb{E}_G[C_{b-1}] - \mathbb{E}_G[C(b)] = p^{b-2}(1 - p^{(b-2)(b-3)/2})$ as n becomes large, with edge-formation probability $0 \leq p \leq 1$. We note that $C(3) = C_2$. We observe that $C(b)$ and C_{b-1} are numerically identical when $p = 0$ (a set of disconnected nodes) and $p = 1$ (a complete graph). For a given b , the expected difference is positive and increases in p up to a certain point, and then falls to zero. It is easy to show that the maximum is attained at $p = (2/(b-1))^{2/(b-2)(b-3)}$, which gives $p = 2/3$ (for $b = 4$), $p = 2^{-1/3} \approx 0.7937$ (for $b = 5$) and $p = (2/5)^{1/6} \approx 0.8584$ (for $b = 6$). It is clear that $C_{b-1} > C(b)$ in expectation for all $p \neq 0, 1$. The numerical difference between the two statistics can be quite substantial, given that $0 \leq C_{b-1} \leq 1$ and $0 \leq C(b) \leq 1$. The maximum difference is $4/27 \approx 0.1481$ (for $b = 4$), $1/4$ (for $b = 5$) and $(108/3125)^{1/3} \approx 0.3257$ (for $b = 6$).

a lollipop graph $G = L(b, 1)$ as for *any* graph G_2 on more than $b + 1$ nodes that does not contain any additional lollipops $L(b - 1, 1)$ beyond those contained in G .⁷ We illustrate with two specific examples. First, take the lollipop $L(3, n - 3)$ with $n \geq 4$. It is easy to show that $C(3) = C_2 = 3/(n + 1)$, which is decreasing in n . This makes intuitive sense: as the lollipop becomes progressively less dense, there is less clustering, and in the limit $n \rightarrow \infty$ the lollipop resembles a path graph. Second, take the lollipop $L(4, n - 4)$ that has $C(4) = 16/(n + 22)$ for $n \geq 6$: our fourth order clustering decreases in n . The usual clustering $C(3) = 12/(n + 10)$ (for $n \geq 5$) also falls in n . Again, this is intuitively correct: the density of $L(4, n - 4)$, for $n \geq 5$, is $d(G) = 2(n + 2)/(n(n - 1)) \rightarrow 0$ as $n \rightarrow \infty$. However, $C_3 = 12|M_{63}^{(4)}|/|M_{15}^{(4)}| = 0.8$ for all $n \geq 5$ i.e. it is invariant as n increases.⁸

This is interesting for two reasons. First, C_3 will not change even as the graph becomes infinitely sparse in the limit. Second, C_3 does not behave in the same way as the usual clustering coefficient C_2 in this respect. The problem is not specific to the lollipop, and C_{b-1} will display this behaviour whenever a lollipop $L(b, 1)$ is extended in a non-trivial way so that no additional $L(b - 1, 1)$ structure is added to the graph. Consider $G = L(5, 1)$ with $n = 6$ nodes and $m = 11$ edges, so that $C_4 \approx 0.8333$. Then consider a connected graph G_2 with an arbitrarily large number of nodes, that includes one copy of G as a subgraph, and that has additional non-trivial topological structure outside of the subgraph G , that can include 17 of the possible 21 non-isomorphic subgraphs on five nodes [48], but with no additional $L(4, 1)$ structure. Nevertheless, C_4 will take the same value on G_2 as on G .

2.5 Analysis of generalized clustering $C(b)$ for the small-world model

We now investigate the properties of $C(b)$ for a simulated small-world model that can interpolate between regular and random behaviour. Many real-world networks exhibit small-world behaviour [4, 8, 43, 54, 57, 72, 78]. Following the original one-parameter model of [72], we start with a regular ring lattice on n nodes, where each node is adjacent to its k nearest neighbours in both the clockwise and anti-clockwise directions, giving a total nk edges. To have a sparse but connected network, we assume that $n \gg 2k \gg \ln(n) \gg 1$.⁹ We choose an edge that connects a node u to its nearest neighbour v in a clockwise direction, and rewire this edge, uniformly with probability p , to an edge (u, w) , avoiding self-loops and duplicated edges. With probability $1 - p$ the original edge is left in place. We continue in a clockwise direction around the ring until all nodes have been considered once. We then look at edges that connect each node to its second-nearest neighbour, and so on, continuing around the ring k times, until each edge in the original lattice has been examined once. The edge rewiring creates more randomness in the network: when $p = 0$, the ring lattice is unchanged and completely regular, and when $p = 1$ all edges are rewired randomly. It is well-known [72] that intermediate values of p give graphs with the small-world property, characterized by low average path length (as in a random graph) but high clustering (as in a regular graph). This is due to the presence of a small number of “short-cut” edges that connect nodes that would otherwise be far apart in a regular graph.

In Figure 6, we report the expected clustering coefficient $\mathbb{E}_G[C(b)]$ from 250 replications of a small-world graph G on $n = 50$ nodes, with $k = 7$. The small-world model is connected by construction, and so $C(b)$ will be well defined. We note that clustering falls in b (given p) but that it is not monotonic as p increases (given b): at some point, introducing more randomness actually increases clustering.¹⁰ Finding a closed-form formula for the expected value of $C(b)$ in a finite small-world model is an open problem. Some partial results are available [4, 8, 72]. When $p = 0$,

⁷If G_2 contains any positive contribution to higher-order clustering (in the sense of more b -cliques) then both $C(b)$ and C_{b-1} will detect it.

⁸We observe qualitatively the same result for the lollipop $L(5, n - 5)$ with $C(3) = 30/(n + 28)$ for $n \geq 6$ and $C(5) = 125/(n + 203)$ for $n \geq 8$ but $C_4 = 5/6 \approx 0.8333$ for $n \geq 6$ despite a vanishing density $d(G) = 2(n + 5)/(n(n - 1)) \rightarrow 0$ as $n \rightarrow \infty$.

⁹If $n = 2k + 1$ then the graph is complete, and $C(b) = 1$ from Proposition 2.1, for $3 \leq b \leq n$. If $n > 2k + 1$ then, trivially, we have that $C(b) = 0$ for $b > k + 1$ on the regular ring lattice with no rewiring.

¹⁰The observation that expected clustering falls in b for a given p is qualitatively the same as Figure 3 in [76], who also consider this small-world model, but for their *average* clustering coefficient, and with (apparently one replication of) a small-world model on $n = 20,000$ nodes, and $k = 5$. Figure 3 in [76] also suggests that their average clustering decreases monotonically in p (for a given b) for this large n . See Figure B.5 for simulation results on the expected overall clustering coefficient $\mathbb{E}_G[C_{b-1}]$ of Yin-Benson-Leskovec, and Figure B.6 for simulation results on the difference in expected clustering $\mathbb{E}_G[C_{b-1}] - \mathbb{E}_G[C(b)]$, for the small-world model of [72].

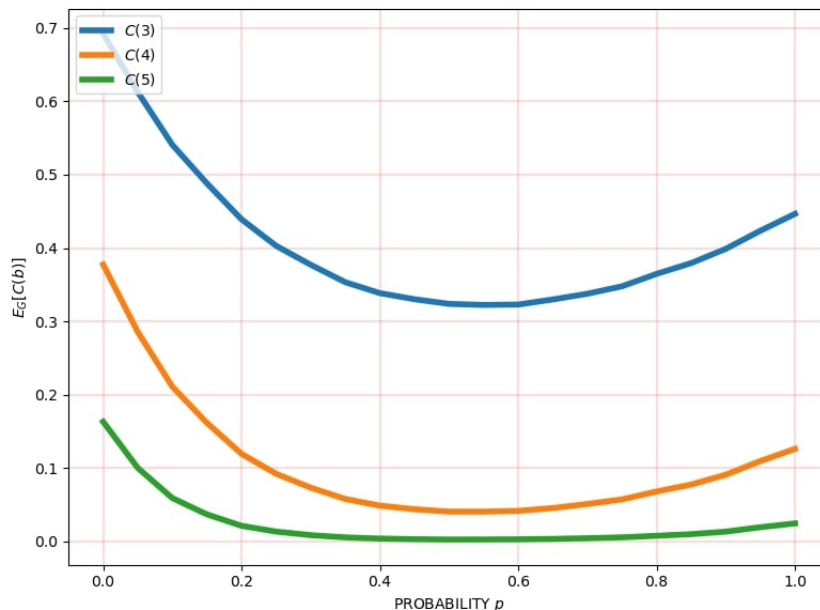


Figure 6: The simulated expected clustering coefficient $\mathbb{E}_G[C(b)]$ from 250 replications of a small-world graph with $n = 50$ nodes, each of which has degree $2k = 14$, and edge-rewiring probability $0 \leq p \leq 1$, which parameterizes the network from a regular graph to a random one. As in [72], we begin with a regular ring lattice, where each node is connected to its k nearest neighbours in both the clockwise and anti-clockwise directions. Moving around the lattice in a clockwise direction, each edge between a node u and its nearest neighbour v is randomly and uniformly rewired with probability p to another edge (u, w) , where self-loops and repeated edges are not allowed. We then continue with the second nearest neighbour, and so on. We observe that clustering falls in b but that it is not monotonic as p increases. As p increases from $p = 0$, the randomness that is added to the graph “breaks” the regular clusters and reduces $C(b)$. At some point, though, introducing more randomness actually increases clustering. Finding a closed-form formula for the expectation of $C(b)$ in a small-world model with $n < \infty$ remains an open problem, except for $b = 3$ and $p = 0$ whereupon $C(3) = 3/2 \times (k - 1)/(2k - 1)$ (see [4]), which equals $9/13 \approx 0.6923$ in this example.

we have $C(3) = 3/2 \times (k - 1)/(2k - 1)$ for all n large relative to k . There are $n \binom{k}{2}$ triangles as n is large relative to k , and $nk(2k - 1)$ connected triples. In the example of Figure 6, $C(3) = 9/13 \approx 0.6923$ for $p = 0$. Asymptotically ($n \rightarrow \infty$), it can be shown that $C(3) \approx 3/2 \times (k - 1)/(2k - 1) \times (1 - p)^3$ when $p \neq 0, 1$, and $C(3) \sim 2k/n$ when $p = 1$. So, $C(3)$ is asymptotically monotonically decreasing in p from $3/2 \times (k - 1)/(2k - 1)$ to zero.

3 Empirical Results on Air Transport Networks

To illustrate the behaviour of $C(b)$, we construct quarterly networks for eight airline carriers over the period 1999Q1 to 2013Q4, using publicly-available data from the U.S. Department of Transportation’s DB1B Airline Origin and Destination survey and T-100 Domestic Segment (All Carriers) census.¹¹ The DB1B is a 10% random sample of quarterly ticket-level itineraries, collected from reporting carriers. The T-100 is a monthly 100% census on domestic nonstop flight segments, including number of enplaned passengers and available capacity. Both datasets have been widely used in empirical work in economics e.g. [2, 20, 26]. We do not observe the actual date of flight or purchase, ticket restrictions, or the buyer’s characteristics.

We merge the DB1B and T-100, retaining all scheduled nonstop round-trip tickets, for domestic carriers, between

¹¹The carriers are American Airlines (AA), Alaska Airlines (AS), Delta Air Lines (DL), AirTran Airways (FL), Spirit Airlines (NK), United Airlines (UA), US Airways (US), and Southwest Airlines (WN).

Carrier	Nodes	Edges	Density	apl	apl _{rand} ^{conn}	C(3)	C(3) _{rand}	C(4)	C(4) _{rand}	C(5)	C(5) _{rand}	Connected %
AA	71	153	0.06	1.94	3.01	0.120	0.061	0.018	0.000	0.002	0.000	44.6
AS	34	49	0.09	2.00	3.12	0.037	0.085	0.000	0.001	0.000	0.000	18.2
DL	85	221	0.06	1.98	2.84	0.146	0.061	0.021	0.000	0.002	0.000	65.4
FL	38	78	0.11	1.94	2.63	0.154	0.108	0.008	0.001	0.000	0.000	61.7
NK	29	92	0.23	1.95	1.96	0.379	0.223	0.097	0.011	0.016	0.000	97.3
UA	48	158	0.14	2.03	2.23	0.346	0.138	0.122	0.003	0.034	0.000	96.0
US	58	113	0.07	2.09	3.03	0.115	0.067	0.011	0.000	0.000	0.000	35.3
WN	88	522	0.14	1.99	2.04	0.335	0.136	0.106	0.002	0.031	0.000	99.9

Table 1: Descriptive statistics for eight carrier networks in 2013Q4. The carriers are American Airlines (AA), Alaska Airlines (AS), Delta Air Lines (DL), AirTran Airways (FL), Spirit Airlines (NK), United Airlines (UA), US Airways (US), and Southwest Airlines (WN). The networks are generally small (nodes and edges) and sparse (density). The average path length (apl) is the number of edges in the shortest path between two nodes, averaged over all pairs of nodes: it is close to two for all networks. We compare the average path length and clustering coefficients $C(3)$, $C(4)$ and $C(5)$ to realizations from Erdős-Rényi random graphs $G(n, p)$ with n equal to the number of nodes in the observed network, and edge-formation probability p equal to its density. The random average path length (apl_{rand}) for $G(n, p)$ is averaged over 1000 replications. Some of the random graphs are not connected (and Connected % gives the percentage of connected realizations across all replications), and have an infinite average path length. For that reason, we compute the average path length only across connected realizations, with clustering coefficients averaged over all realizations of $G(n, p)$, both connected and disconnected, as $C(3)_{rand}$, $C(4)_{rand}$ and $C(5)_{rand}$. The percentage of connected realizations is positively related to the network density. There is some evidence (strongest for Southwest Airlines) that the airlines have a small-world property, with similar average path lengths to a random graph, but higher three-node clustering; although the average path lengths for connected random graphs are generally higher than those observed in the real networks. Alaska Airlines has some evidence of randomness, with particularly low clustering. We see that generalized clustering $C(4)$ and $C(5)$ are typically higher than random for all carriers.

airports in the continental U.S. We do not keep tickets that were sold under a codesharing agreement, that have unusually high or low fares, or that are considered unreliable by the data provider. Some carriers (e.g. JetBlue Airways and Southwest Airlines) report large numbers of business and first-class tickets. We only use coach class tickets, unless more than 75% of a carrier’s tickets are listed as business or first-class, in which case we keep all tickets for that carrier.¹² Individual tickets are then aggregated to non-directional route-carrier observations. We omit route-carriers with an especially low number of passengers, that do not have a constant number of passengers on each segment, or that are not present over the full sample period. In building the route networks, a node is an airport that was served as a route origin or destination, and an edge is present if some passengers travelled on a direct route between two nodes, for a given carrier-quarter. Our eight empirical networks are connected in every quarter of the sample. Further details of the data treatment are available from the authors.

3.1 Descriptive statistics and small-world characteristics of airline networks

Table 1 reports descriptive statistics on the eight carrier networks in our dataset for 2013Q4. The networks are generally small (nodes and edges) and sparse (density 6–23%). We compute the average path length and clustering coefficient $C(b)$ for $b = 3, 4, 5$ for each real-world network, and compare these to values from simulated Erdős-Rényi random graphs $G(n, p)$, with n equal to the number of nodes in each observed network, and the edge-formation probability p set equal to its density.¹³ As in [72], there is some evidence that most U.S. airline networks are small-world, with average path lengths that are close to those from a random graph, but higher levels of clustering.

¹²The data treatment is quite standard in the empirical air transport literature e.g. [21] motivate our filtering of tickets by fare class.

¹³The clustering coefficients reported in Table 1 for $G(n, p)$ are based upon simulations and are not the theoretical asymptotic values. See Table B.3 for descriptive statistics on the Yin-Benson-Leskovec statistic C_{b-1} .

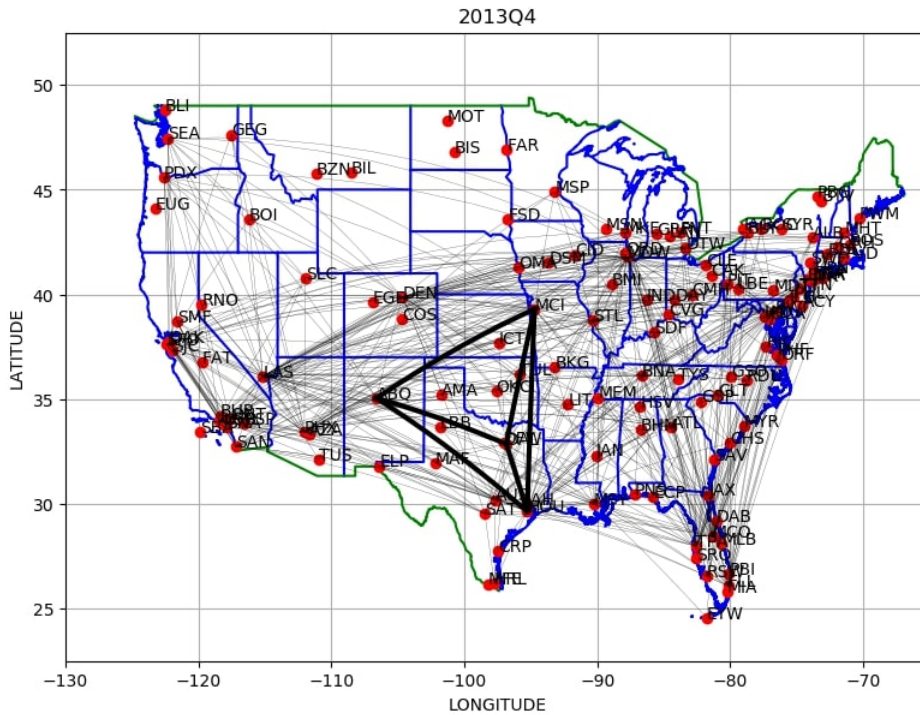


Figure 7: The Albuquerque–Dallas–Houston–Kansas City maximal 4-clique in Southwest’s network, found using the Bron-Kerbosch algorithm, is one example of a completely-connected set of four nodes. Southwest Airlines served every pairwise route among these four airports in 2013Q4. Since this is a maximal clique, no other airport in Southwest’s network can be added to the clique while preserving its complete connectivity (to create a 5-clique).

3.2 The distribution of maximal cliques

Since $C(b)$ is intimately related to the relative number of cliques, we use the Bron-Kerbosch algorithm to identify all cliques in a given network. Figure 7 displays the 2013Q4 network of Southwest Airlines, and we highlight one maximal 4-clique for illustration, between Albuquerque, Dallas, Houston, and Kansas City. We report the airport codes and co-ordinates in Table B.1 in Appendix B. It is interesting to see how many maximal cliques of any given size there are in a network, and whether this distribution is stable over time. For Southwest’s network, in Figure 8, we observe that the distribution is more spread out, and that more larger cliques appear, over time. There is a maximum clique size of eleven, which corresponds to 12.5% of all of the airports served by Southwest in 2013Q4. This might seem surprising, given that Southwest’s network is relatively sparse, with a density $d(G)$ close to 15% in that quarter. Since every airport in the maximum clique has at least 11 connections, we can think of it as a group of “important” airports, that are also very highly connected among themselves.¹⁴ An operational reason for developing such groups could be to enable the opening of a large number of new indirect routes between airport pairs, at relatively low cost, with the addition of a few well-chosen direct routes. It seems likely that Southwest, through its network expansion, has focused on increasing the size and connectivity of a moderate number of “central” airports while also creating links from non-central airports into this group.¹⁵

¹⁴We find evidence that nodes that belong to maximal cliques in Southwest’s network are more connected, on average, than nodes that are not in maximal cliques, and that the average degree of nodes in maximal cliques increases in the order of the clique (Figure B.2).

¹⁵Not all networks evolve in this way e.g. the distribution of maximal cliques for American (AA) is far more stable over time (Figure B.3).

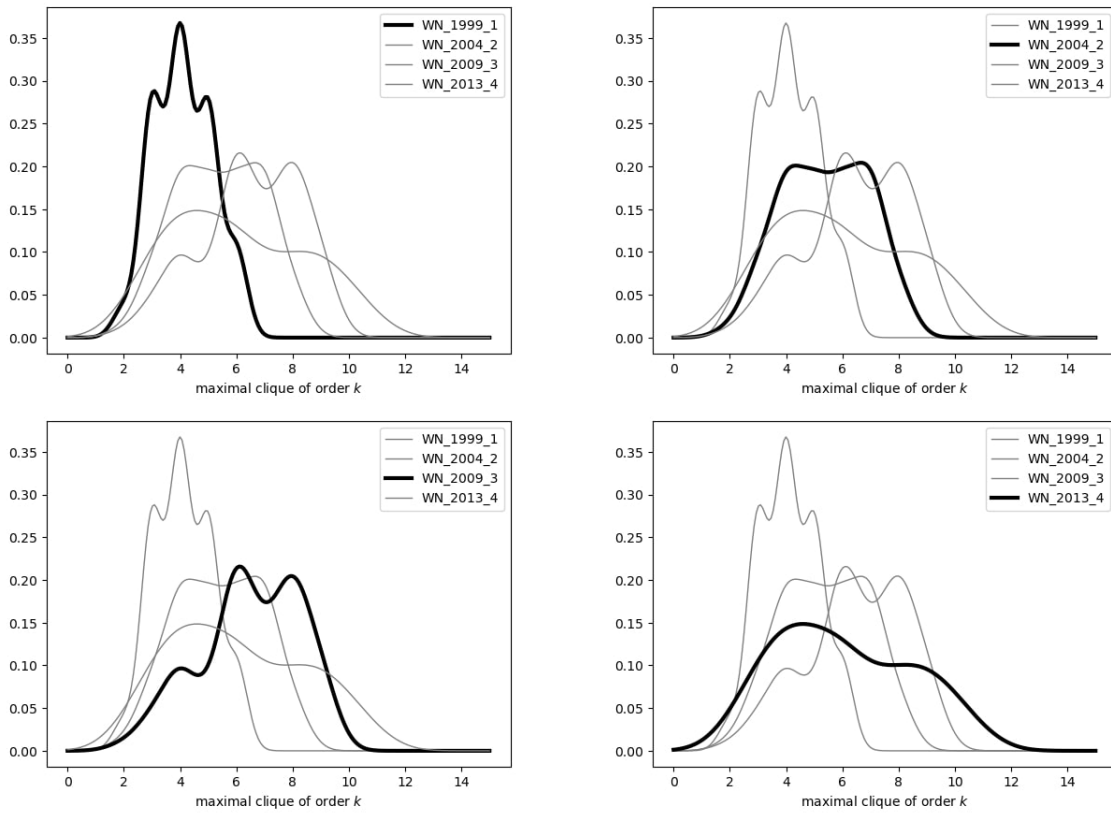


Figure 8: *The distribution of maximal k -cliques in Southwest's network, in 1999Q1, 2004Q2, 2009Q3 and 2013Q4. This shows that Southwest is creating increasingly large groups of very highly interconnected airports over time.*

3.3 Dynamic variation in higher-order clustering

Since there are many cliques with more than three nodes, we examine how $C(3)$, $C(4)$ and $C(5)$ vary across carriers, and over time (Figure 9). We make the following remarks:

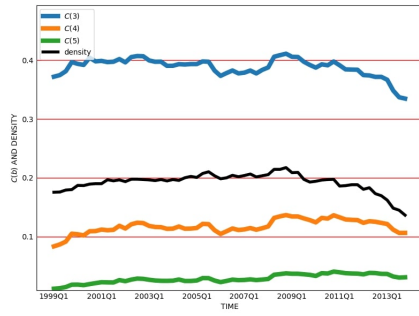
- There is considerable heterogeneity across carriers. For instance, Southwest has quite a stable $C(3)$ over 1999Q1 to 2013Q4, despite its significant expansion in terms of airports and routes. On the other hand, United has far more clustering (in triangles) from 2009 onwards, while Alaska has progressively less.
- Generalized clustering $C(b)$ is highly positively correlated across b , for some networks e.g. Delta, and is also positively correlated with network density (see Table B.2 in Appendix B). Some of this follows by construction e.g. for every newly formed 4-clique in a network, there will be between two and four new 3-cliques, while every newly formed 5-clique will create between two and five new 4-cliques and between three and ten new 3-cliques. High correlation reduces the information-content of $C(4)$ and $C(5)$, but it is unclear whether this result holds for other real-world networks. To control for this correlation, we consider the following regressions: $C(4) = \text{constant} + \beta C(3) + \text{error}$ and $C(5) = \text{constant} + \beta C(3) + \gamma C(4) + \text{error}$, where the residuals could be used rather than $C(4)$ and $C(5)$ themselves. We illustrate the $C(4)$ procedure in Figure B.1, for US Airways and Southwest (WN). Since $C(3)$ and $C(4)$ display evidence of a unit root (US) and a unit root and trend (WN), we first run regressions of the form $\Delta C(b)_t = \alpha + \delta t + u(b)_t$, for $b = 3, 4$. We then regress the difference and trend stationary $\hat{u}(4)$ on a constant and $\hat{u}(3)$, and find that 76% (US) and 89% (WN) of the variation in $C(4)$ is “explained” by $C(3)$. In this sense, $C(4)$ is moderately informative once $C(3)$ has been accounted for. It is unclear if other networks will give the same results, but similar behaviour is noticed on C_{b-1} by [76] for other real-world networks.
- Table 1 and Figure 9 provide evidence that $C(3) > C(4) > C(5)$, and we might think that this holds quite generally. However, Figure 5 shows that this is not necessarily the case for $G(n, p)$. We can also construct a series of counterexamples using the lollipop graph $L(b, n - b)$. As n increases given b , the number of complete subgraphs of order no more than b does not change (for instance, there are four triangles and one 4-complete subgraph in the lollipop graph $L(4, n - 4)$ of Figure 10). Furthermore, increasing n after a certain point will only add paths of length b to the denominator of $C(b)$, and no other spanning trees (e.g. b -stars). For $L(4, n - 4)$, we have already seen that

$$C(3) = \frac{12}{n+10}; n \geq 5; \quad C(4) = \frac{16}{n+22}; n \geq 6,$$

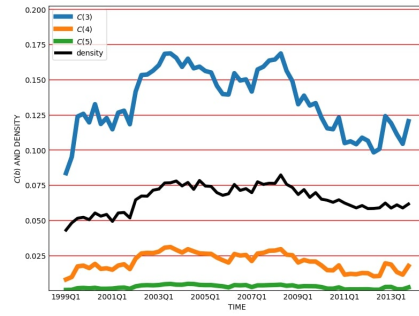
and it follows that $C(3) \geq C(4)$ as $n \leq 26$, with equality when $C(3) = C(4) = 1/3$. In the lollipop $L(5, n - 5)$ of Figure 11, the number of 4-complete and 5-complete subgraphs is constant as n increases. Beyond a certain point, only 4-paths and 5-paths are added to the denominators of $C(4)$ and $C(5)$ and no further 4-stars or 5-stars or 5-arrows are created. We can show that

$$C(4) = \frac{80}{n+95}; n \geq 7; \quad C(5) = \frac{125}{n+203}; n \geq 8,$$

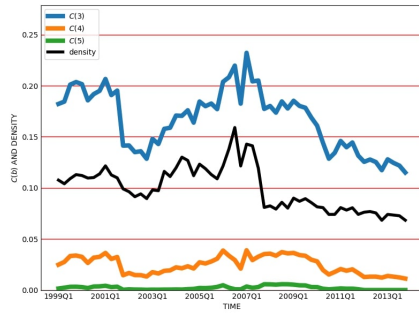
from which $C(4) \geq C(5)$ as $n \leq 97$. Equality occurs when $C(4) = C(5) = 5/12$. Incidentally, for this graph, $C(3) \geq C(4)$ as $n \leq 12.2$, i.e., $n < 13$. We could use this construction to show that $C(b) < C(b+1)$ for any $b < n$ and sufficiently large n .



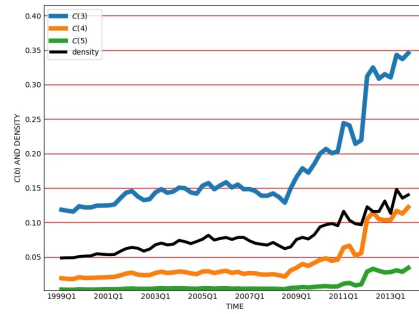
(a) Southwest Airlines.



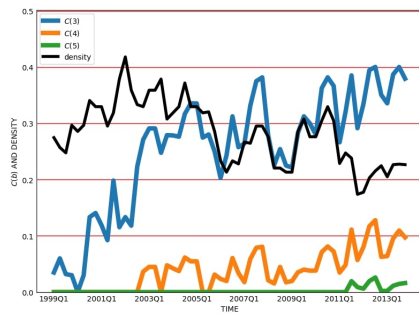
(b) American Airlines.



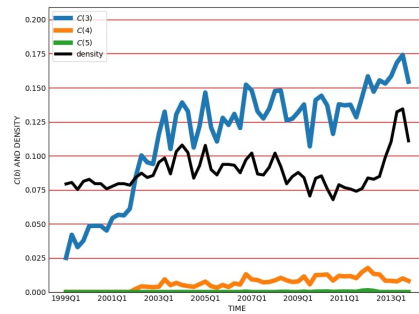
(c) US Airways.



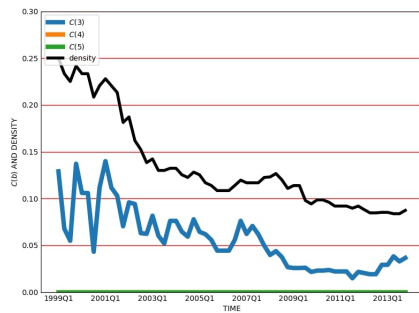
(d) United Airlines.



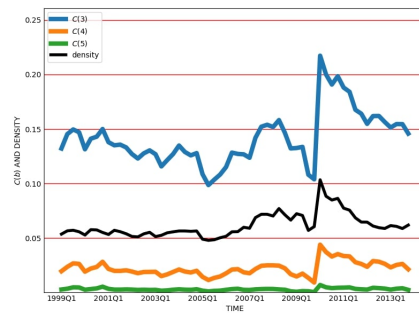
(e) Spirit Airlines.



(f) AirTran Airways.



(g) Alaska Airlines.



(h) Delta Air Lines.

Figure 9: The dynamic behaviour of $C(3)$, $C(4)$, $C(5)$ and density from 1999Q1 to 2013Q4 for eight airline carriers. We observe substantial heterogeneity across carriers, a high degree of correlation between generalized clustering $C(b)$ for different b , and evidence that $C(b)$ is decreasing in b (in the text, we give theoretical counterexamples to the latter observation).

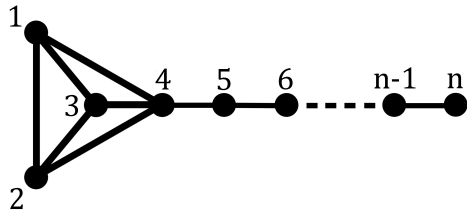


Figure 10: The lollipop graph $L(4, n-4)$ is a counterexample to $C(3) \geq C(4)$, which does not hold for $n \geq 27$.

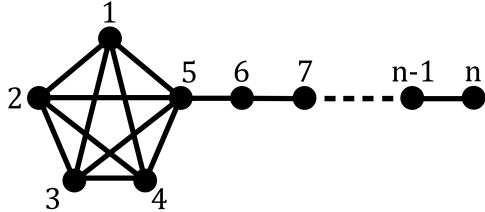


Figure 11: The lollipop graph $L(5, n-5)$ is a counterexample to $C(4) \geq C(5)$, which does not hold for $n \geq 98$.

4 Conclusions

We have proposed a generalized clustering coefficient $C(b)$ of order greater than or equal to three, that nests the usual overall clustering, or transitivity, $C(3)$. We investigate the properties of $C(b)$ on random and small-world graphs, and propose an algorithm for implementation that is based on analytical subgraph counts, and that is practical for $b = 3, 4, 5$ when the graph is not too large. Our work complements the recent paper by Yin-Benson-Leskovec [76], who generalize overall clustering in a different way, and we draw careful comparisons with their findings. We illustrate the performance of our measure using data on U.S. airline route networks, and provide new insight into the strategic behaviour that leads carriers to develop small groups of highly connected airports. Extending these ideas to generalize the notion of a “hub” node to multi-node hubs, in both transportation and other real-world networks, is a promising avenue for future work. We expect analytic formulae for subgraph enumeration to have application in other areas of applied graph theory (e.g. motif detection [1]), although it currently seems too difficult to derive analytic formulae for all subgraphs on more than five nodes (for complete results on five node subgraphs, see [48]). Future work linking graphs and econometrics should also lead to a better understanding of the economic, strategic and spatial factors that drive dynamic clustering in real-world networks e.g. [27, 28, 36] suggest possible applications in economics to game-theoretic network formation models and production networks and, in social networks, to the coordination of clustered individuals on collective actions.

A Proofs

Derivations for the 3-star $M_3^{(3)}$, the triangle $M_7^{(3)}$, the 4-star $M_{11}^{(4)}$, the 4-path $M_{13}^{(4)}$, the tadpole $M_{15}^{(4)}$, and the 4-complete $M_{63}^{(4)}$ are given in [1, Proposition 2.1]. For completeness, we repeat the results here, without proof, in Proposition A.1. We also include the three spanning trees on five nodes: the 5-star $M_{75}^{(5)}$, the 5-arrow $M_{77}^{(5)}$ and the 5-path $M_{86}^{(5)}$, as well as the 5-complete $M_{1023}^{(5)}$, all with their corresponding proofs.

Proposition A.1 (Analytic formulae for nested subgraph enumeration).

$$|M_3^{(3)}| = \sum_i \binom{k_i}{2} = \frac{1}{2} \sum_i k_i(k_i - 1).$$

$$|M_7^{(3)}| = \frac{1}{6} \text{tr}(g^3). \quad (6)$$

$$|M_{11}^{(4)}| = \sum_i \binom{k_i}{3} = \frac{1}{6} \sum_i k_i(k_i - 1)(k_i - 2).$$

$$|M_{13}^{(4)}| = \sum_{(i,j) \in E} (k_i - 1)(k_j - 1) - 3|M_7^{(3)}|.$$

$$|M_{15}^{(4)}| = \frac{1}{2} \sum_{k_i > 2} (g^3)_{ii} (k_i - 2).$$

$$|M_{63}^{(4)}| = \frac{1}{24} \sum_i \text{tr}(g^3_{-i}). \quad (7)$$

$$|M_{75}^{(5)}| = \sum_i \binom{k_i}{4} = \frac{1}{24} \sum_i k_i(k_i - 1)(k_i - 2)(k_i - 3). \quad (8)$$

$$|M_{77}^{(5)}| = \sum_{(i,j)^* \in E} \binom{k_i - 1}{2} (k_j - 1) - 2|M_{15}^{(4)}|. \quad (9)$$

$$|M_{86}^{(5)}| = \frac{1}{2} \sum_{i \neq j} (g^4)_{ij} - 2|M_3^{(3)}| - 9|M_7^{(3)}| - 3|M_{11}^{(4)}| - 2|M_{13}^{(4)}| - 2|M_{15}^{(4)}|. \quad (10)$$

$$|M_{1023}^{(5)}| = \frac{1}{5} \sum_i |M_{63}^{(4)}(g_{-i})| = \frac{1}{120} \sum_i \sum_{j \in \Gamma_G(i)} \text{tr}(((g_{-i})_{-j})^3). \quad (11)$$

Remark A.1. In (7), g_{-i} is the adjacency matrix corresponding to the subgraph induced by the neighbourhood $\Gamma_G(i)$ of i , which we denote by $G_{-i} = (V(\Gamma_G(i)), E(\Gamma_G(i)))$, and we use (6) to count the number of triangles.

Remark A.2. In (9), $\sum_{(i,j)^* \in E}$ denotes summation over all edges in E , in *both* directions (i, j) and (j, i) .

Remark A.3. In (11), $(g_{-i})_{-j}$ is the adjacency matrix corresponding to the subgraph induced by the neighbourhood $\Gamma_{G_{-i}}(j)$ of j , which we denote by $G_{-i-j} = (V(\Gamma_{G_{-i}}(j)), E(\Gamma_{G_{-i}}(j)))$, and we use (7) to count the number of 4-cliques.

Proof of Proposition A.1. We treat each subgraph separately, and only report proofs that are not presented in [1, Proposition 2.1].

- (a) $|M_{75}^{(5)}|$: Node i has edges to k_i neighbours, and any four of those edges will form a 5-star, centered on i . The result (8) follows immediately.

- (b) $|M_{77}^{(5)}|$: The method of proof is similar to that used for the count of the nested 4-path $|M_{13}^{(4)}|$ in [1]. Consider any edge $(i, j) \in E$, as the central edge in a 5-arrow. Let i and j have degrees three and two respectively, and let node i be directly-connected to nodes x and z , and let node j be directly-connected to node y . Node i has $k_i - 1$ possible neighbours (for nodes x and z) and node j has $k_j - 1$ possible neighbours (for node y). There are $\frac{1}{2}(k_i - 1)(k_i - 2)(k_j - 1)$ ways in which two neighbours of i can be paired with a neighbour of j , which gives a total of $\sum_{(i,j)^* \in E} \binom{k_i - 1}{2} (k_j - 1)$ across all possible central edges, in both directions (we use $(i, j)^*$ to denote “ (i, j) and (j, i) ”). This sum includes the unwanted cases $x = y$ and $y = z$, both of which form a tadpole. Since two of the four edges of the tadpole can be a candidate central edge (i, j) of a 5-arrow, we subtract $2|M_{15}^{(4)}|$ to give result (9).
- (c) $|M_{86}^{(5)}|$: A very similar but less transparent proof can be found in [56]. A 5-path is a walk of length 4 with no repeated nodes. Note that $\frac{1}{2} \sum_{i \neq j} (g^4)_{ij}$ gives the number of walks of length 4 from i to j , which does not only include 5-paths. There are five subgraphs in which we can find walks of length 4 that are not 5-paths:

	Subgraph				
	3-path $M_3^{(3)}$	triangle $M_7^{(3)}$	4-star $M_{11}^{(4)}$	4-path $M_{13}^{(4)}$	tadpole $M_{15}^{(4)}$
Number of other walks of length 4	2	9	3	2	2

So, by removing them from the sum, we have (10) as required.

- (d) $|M_{1023}^{(5)}|$: Consider a 4-complete subgraph $M_{63}^{(4)}$ comprised of nodes j, k, ℓ and m . Let each node be in the neighbourhood $\Gamma_G(i)$ of some node i such that $i \neq j \neq k \neq \ell \neq m$. Hence, the five nodes i, j, k, ℓ and m , and the edges between them, form a 5-complete subgraph $M_{1023}^{(5)}$. The quantity $|M_{63}^{(4)}(g_{-i})|$ gives the number of 5-complete subgraphs that contain node i , where g_{-i} is the adjacency matrix corresponding to the subgraph induced by $\Gamma_G(i)$. By symmetry, summing across all nodes i will give five times the total count of 5-complete subgraphs in the graph, and so we divide the sum by five to give result (11), which can be simplified further by using (7) to count 4-complete subgraphs in each subgraph G_{-i} .

□

Proof of Proposition 2.1. We consider the “if” and “only if” parts separately:

- **(if)** Let G be complete. Hence, each set of b nodes of G forms a b -clique and G contains exactly $\binom{n}{b}$ b -cliques. The number of b -spanning trees of G is equal to the number of b -spanning trees enclosed in any b -clique which is, using Cayley’s formula:

$$b^{b-2} \times \binom{n}{b},$$

from which (3) gives $C(b) = 1$.

- **(only if)** We prove this part by contrapositive. Suppose that G is not complete. Since G has at least b nodes, we can find a connected subgraph G' of G with b nodes such that G' is not a b -clique, and we can extract a b -spanning tree from G' by removing any cycles. Hence, there is at least one b -spanning tree in G which is not enclosed in a b -clique. It follows that:

$$\begin{aligned} \text{number of } b\text{-spanning trees in } G &\geq \text{number of } b\text{-spanning trees enclosed in a } b\text{-clique} + 1 \\ &> \text{number of } b\text{-spanning trees enclosed in a } b\text{-clique} \\ &= b^{b-2} \times \text{number of } b\text{-cliques in } G, \end{aligned}$$

and so $C(b) < 1$ from (3), which proves the proposition.

□

B Additional Figures and Tables

code	x	y	code	x	y	code	x	y
ABQ	-106.61	35.04	ACY	-74.58	39.47	ALB	-73.80	42.73
AMA	-101.71	35.23	ATL	-84.43	33.64	AUS	-97.67	30.19
AZA	-111.66	33.31	BDL	-72.68	41.94	BHM	-86.75	33.57
BIL	-108.53	45.80	BIS	-100.75	46.78	BKG	-93.20	36.53
BLI	-122.53	48.80	BMI	-88.92	40.48	BNA	-86.68	36.12
BOI	-116.22	43.56	BOS	-71.00	42.36	BTW	-73.15	44.47
BUF	-78.73	42.94	BUR	-118.35	34.20	BWI	-76.67	39.18
BZN	-111.15	45.78	CAK	-81.44	40.92	CHS	-80.04	32.90
CID	-91.71	41.88	CLE	-81.85	41.42	CLT	-80.93	35.22
CMH	-82.88	40.00	COS	-104.72	38.82	CRP	-97.50	27.77
CVG	-84.67	39.05	DAB	-81.05	29.18	DAL	-96.85	32.85
DAY	-84.18	39.75	DCA	-77.04	38.85	DEN	-104.67	39.86
DFW	-97.04	32.90	DSM	-93.66	41.53	DTW	-83.35	42.21
ECP	-85.80	30.36	EGE	-106.92	39.63	ELP	-106.38	31.80
EUG	-123.22	44.12	EWR	-74.17	40.69	EYW	-81.77	24.55
FAR	-96.82	46.92	FAT	-119.72	36.77	FLL	-80.15	26.07
FNT	-83.74	42.97	FSD	-96.74	43.58	GEG	-117.53	47.62
GRR	-85.53	42.88	GSO	-79.94	36.10	GSP	-82.22	34.90
HOU	-95.28	29.65	HPN	-73.70	41.07	HRL	-97.75	26.20
HSV	-86.78	34.64	IAD	-77.46	38.94	IAG	-79.03	43.10
IAH	-95.34	29.98	ICT	-97.43	37.65	ILG	-75.60	39.68
IND	-86.29	39.72	ISP	-73.10	40.80	JAN	-90.08	32.31
JAX	-81.63	30.42	JFK	-73.78	40.63	LAN	-84.58	42.78
LAS	-115.17	36.08	LAX	-118.41	33.94	LBB	-101.83	33.67
LBE	-79.40	40.28	LGA	-73.87	40.77	LGB	-118.15	33.82
LIT	-92.22	34.73	MAF	-102.20	31.94	MCI	-94.73	39.29
MCO	-81.31	28.43	MDT	-76.76	40.19	MDW	-87.75	41.78
MEM	-89.97	35.07	MFE	-98.24	26.18	MHT	-71.44	42.93
MIA	-80.27	25.78	MKE	-87.90	42.95	MLB	-80.63	28.10
MOT	-101.28	48.27	MSN	-89.34	43.14	MSP	-93.22	44.88
MSY	-90.26	29.99	MYR	-78.97	33.70	OAK	-122.22	37.72
OKC	-97.60	35.39	OMA	-95.90	41.30	ONT	-117.60	34.06
ORD	-87.90	41.98	ORF	-76.20	36.90	ORH	-71.88	42.27
PBG	-73.47	44.65	PBI	-80.10	26.68	PDX	-122.60	45.59
PHF	-76.50	37.13	PHL	-75.24	39.87	PHX	-112.03	33.43
PIT	-80.23	40.49	PNS	-87.18	30.47	PSP	-116.50	33.83
PVD	-71.43	41.73	PWM	-70.30	43.65	RDU	-78.79	35.88
RIC	-77.32	37.50	RNO	-119.77	39.50	ROC	-77.67	43.12
RSW	-81.76	26.54	SAN	-117.18	32.73	SAT	-98.47	29.53
SAV	-81.20	32.13	SBA	-119.84	33.43	SDF	-85.74	38.17
SEA	-122.31	47.45	SFO	-122.38	37.62	SJC	-121.92	37.35
SLC	-111.97	40.79	SMF	-121.62	38.70	SNA	-117.87	33.67
SRQ	-82.55	27.40	STL	-90.37	38.75	SWF	-74.02	41.50
SYR	-76.12	43.12	TPA	-82.53	27.98	TTN	-74.81	40.28
TUL	-95.89	36.20	TUS	-110.94	32.12	TYS	-83.92	35.95

Table B.1: List of airports by IATA code in 2013Q4, all carriers, with their longitude (x) and latitude (y). The identity of the airport corresponding to each IATA code can be found at <http://www.iata.org/en/publications/directories/code-search/>.

Variable	$C(3)$	$C(4)$	$C(5)$	$density$
$C(3)$	1.000	0.394	-0.012	0.790
p -value	0.000	0.002	0.927	0.000
$C(4)$	-	1.000	0.910	0.365
p -value	-	0.000	0.000	0.004
$C(5)$	-	-	1.000	0.039
p -value	-	-	0.000	0.766
$density$	-	-	-	1.000
p -value	-	-	-	0.000

(a) Southwest Airlines.

Variable	$C(3)$	$C(4)$	$C(5)$	$density$
$C(3)$	1.000	0.992	0.963	0.864
p -value	0.000	0.000	0.000	0.000
$C(4)$	-	1.000	0.984	0.852
p -value	-	0.000	0.000	0.000
$C(5)$	-	-	1.000	0.861
p -value	-	-	0.000	0.000
$density$	-	-	-	1.000
p -value	-	-	-	0.000

(b) American Airlines.

Variable	$C(3)$	$C(4)$	$C(5)$	$density$
$C(3)$	1.000	0.897	0.659	0.781
p -value	0.000	0.000	0.000	0.000
$C(4)$	-	1.000	0.916	0.459
p -value	-	0.000	0.000	0.000
$C(5)$	-	-	1.000	0.102
p -value	-	-	0.000	0.438
$density$	-	-	-	1.000
p -value	-	-	-	0.000

(c) US Airways.

Variable	$C(3)$	$C(4)$	$C(5)$	$density$
$C(3)$	1.000	0.994	0.968	0.965
p -value	0.000	0.000	0.000	0.000
$C(4)$	-	1.000	0.989	0.936
p -value	-	0.000	0.000	0.000
$C(5)$	-	-	1.000	0.887
p -value	-	-	0.000	0.000
$density$	-	-	-	1.000
p -value	-	-	-	0.000

(d) United Airlines.

Variable	$C(3)$	$C(4)$	$C(5)$	$density$
$C(3)$	1.000	0.844	0.439	-0.256
p -value	0.000	0.000	0.000	0.049
$C(4)$	-	1.000	0.709	-0.414
p -value	-	0.000	0.000	0.001
$C(5)$	-	-	1.000	-0.450
p -value	-	-	0.000	0.000
$density$	-	-	-	1.000
p -value	-	-	-	0.000

(e) Spirit Airlines.

Variable	$C(3)$	$C(4)$	$C(5)$	$density$
$C(3)$	1.000	0.854	0.333	0.539
p -value	0.000	0.000	0.009	0.000
$C(4)$	-	1.000	0.682	0.156
p -value	-	0.000	0.000	0.235
$C(5)$	-	-	1.000	-0.255
p -value	-	-	0.000	0.049
$density$	-	-	-	1.000
p -value	-	-	-	0.000

(f) AirTran Airways.

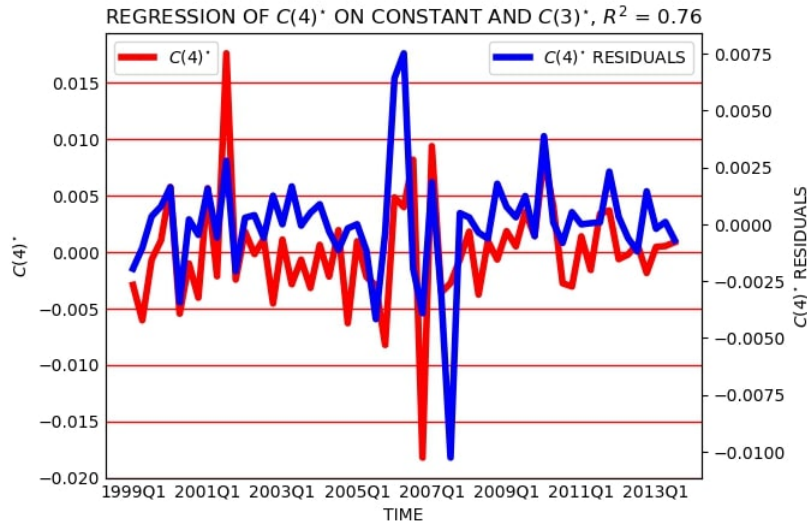
Variable	$C(3)$	$C(4)$	$C(5)$	$density$
$C(3)$	1.000	NA	NA	0.840
p -value	0.000	NA	NA	0.000
$C(4)$	-	1.000	NA	NA
p -value	-	0.000	NA	NA
$C(5)$	-	-	1.000	NA
p -value	-	-	0.000	NA
$density$	-	-	-	1.000
p -value	-	-	-	0.000

(g) Alaska Airlines.

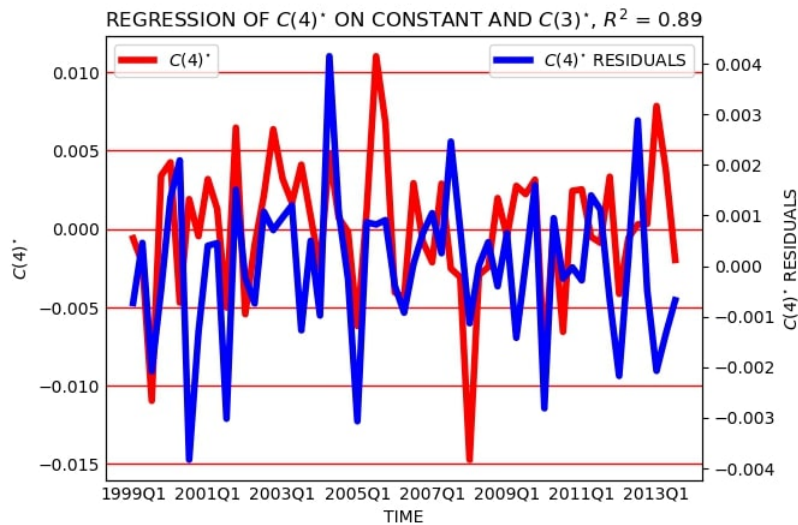
Variable	$C(3)$	$C(4)$	$C(5)$	$density$
$C(3)$	1.000	0.970	0.807	0.838
p -value	0.000	0.000	0.000	0.000
$C(4)$	-	1.000	0.919	0.743
p -value	-	0.000	0.000	0.000
$C(5)$	-	-	1.000	0.488
p -value	-	-	0.000	0.000
$density$	-	-	-	1.000
p -value	-	-	-	0.000

(h) Delta Air Lines.

Table B.2: Pearson's correlation test for $C(3)$, $C(4)$, $C(5)$ and $density$, for different networks. The usual clustering coefficient $C(3)$ is often highly correlated with the network density, but this does not always carry over to generalized clustering $C(4)$ and $C(5)$ (Southwest Airlines, US Airways, Spirit Airlines, AirTran Airways). Furthermore, $C(b)$ is often highly correlated with $C(b-1)$ for $b=4,5$, although there is less observed correlation between $C(3)$ and $C(5)$.



(a) *US Airways.*



(b) *Southwest Airlines.*

Figure B.1: Results of regressions of $C(4)^*$ on a constant and $C(3)^*$, for US Airways and Southwest Airlines, where the star notation indicates that both clustering coefficients have been corrected so that they are difference and trend stationary, before performing the regressions (see Section 3.3): this is common practice in applied econometrics and avoids potential spurious results from regressing one nonstationary series on another. The figures suggest that there is little clustering of order four once we have corrected for the presence of usual clustering of order three.

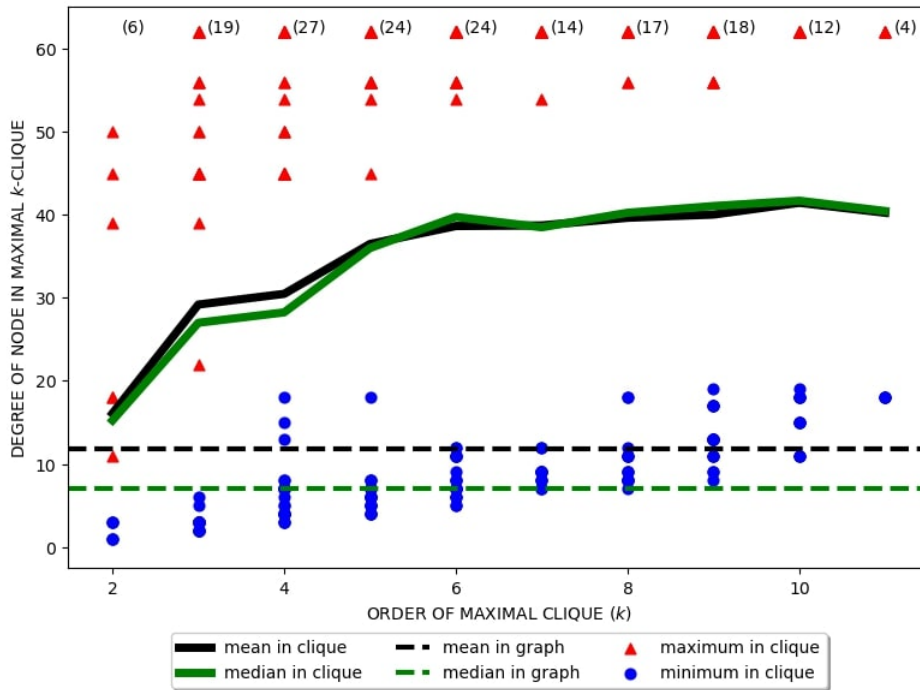


Figure B.2: The mean (black line), median (green line), minimum (blue circle) and maximum (red triangle) degree of nodes that belong to maximal cliques of order k , in Southwest's 2013Q4 network. For comparison, we plot the mean (black dashed line) and median (green dashed line) degree of all nodes in the network. Values in parentheses are the total number of maximal cliques of order k in the network. We see that airports that belong to maximal cliques are more connected, on average, than those that are not in a maximal clique, and that their degree increases in the number of airports in the clique.

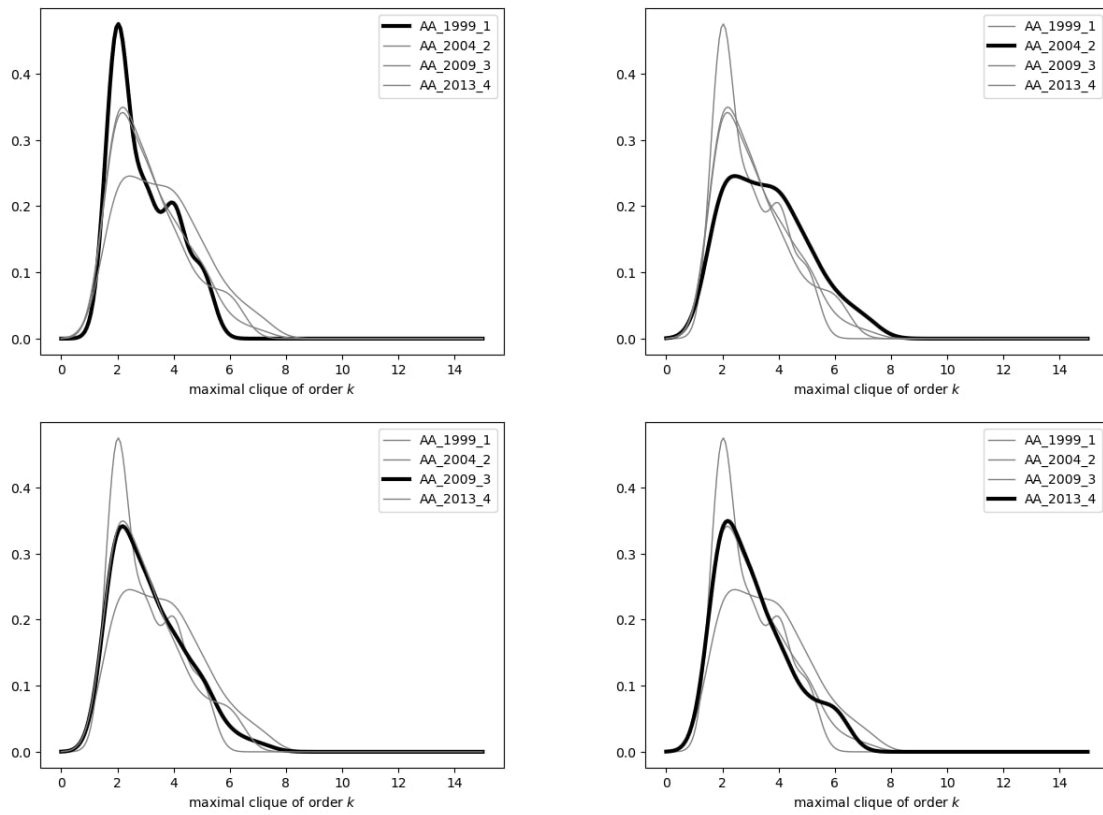
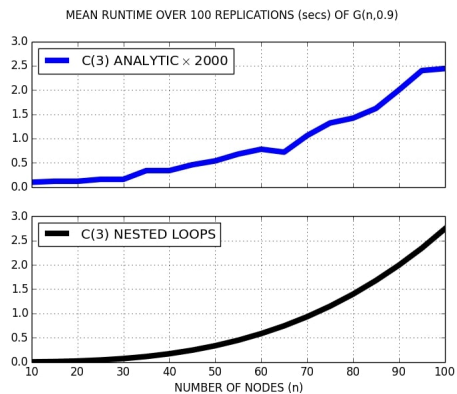


Figure B.3: *The distribution of maximal k -cliques in American's network, in 1999Q1, 2004Q2, 2009Q3 and 2013Q4. This shows that there is little variation over time in the number of groups of connected airports (of any size) in American's network. This is very different to what we observe for Southwest Airlines (Figure 8).*

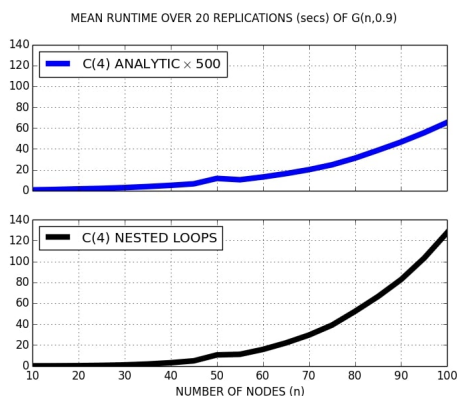
B.1 Computational performance of analytic formulae for $C(b)$

We simulated the actual runtimes of the analytic formulae for $C(b)$ for $b = 3, 4, 5$, on dense Erdős-Rényi graphs $G(n, 0.9)$, and compared these with the runtimes of a simple nested loop implementation. We are able to show that the theoretical asymptotic runtime of each of the analytic clustering formulae is lower than that of the nested loops.¹⁶ However, the small-sample runtime is much lower when analytics are used (Figure B.4): the analytic algorithm is roughly 2,000 times faster for $C(3)$ and more than 500 times faster for $C(4)$ and $C(5)$ for the dense $G(n, p)$. While analytic runtime gains are lower for sparse $G(n, p)$, they remain very substantial, and this contributes to making these generalized clustering coefficients a practical tool for small graphs. We expect that numerical algorithms similar to those used in [76] will be more appropriate as the size of the graph increases.

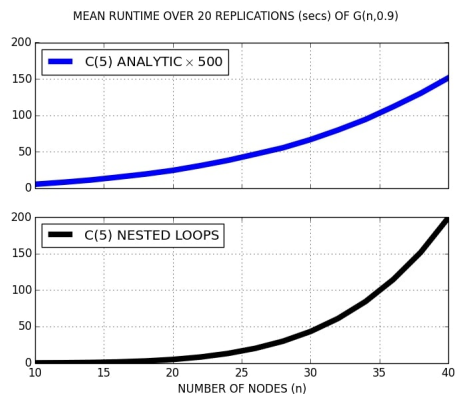
¹⁶The worst-case theoretical runtime of a nested loop implementation of $C(b)$ is $O(n^b)$, since there are b nested loops. In a very sparse graph, the actual runtime of nested loops can be much faster, and coding shortcuts can take advantage of the fact that not every b -tuple needs to be considered. Directly from (2), (4) and (5), we can see that the numerator will dominate the asymptotic runtime of the analytic formulae. We find that $C(3)$ is $O(n^\omega)$, $C(4)$ is $O(n^{\omega+1})$, and $C(5)$ is $O(n^{\omega+2})$, where ω is the exponent of matrix multiplication, for which current implementations give $2.38 \leq \omega \leq 3$. The very fast matrix multiplication algorithms due to [24] and [70] both have $\omega \approx 2.38$, the well-known algorithm due to [68] has $\omega \approx 2.81$, and a naïve algorithm has $\omega = 3$.



(a) $C(3)$.



(b) $C(4)$.



(c) $C(5)$.

Figure B.4: Simulated runtimes, in seconds, of $C(3)$, $C(4)$ and $C(5)$ analytic and nested loop algorithms, computed over 100 (or 20 for $C(4)$ and $C(5)$) replications of dense Erdős-Rényi graphs $G(n,0.9)$, where we only retain connected graphs. The analytic algorithm is roughly 2,000 times faster for usual clustering $C(3)$ and more than 500 times faster for generalized clustering $C(4)$ and $C(5)$.

B.2 Supplementary results for the Yin-Benson-Leskovec statistic

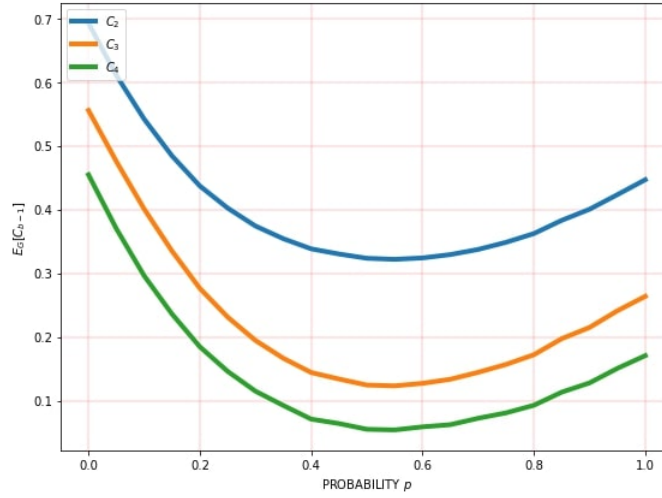


Figure B.5: The simulated expected Yin-Benson-Leskovec clustering coefficient $\mathbb{E}_G[C_{b-1}]$ from 250 replications of a small-world graph with $n = 50$ nodes, each of which has degree $2k = 14$, and edge-rewiring probability $0 \leq p \leq 1$. See Figure 6 for more details on the construction of the graph. As for $\mathbb{E}_G[C(b)]$ in Figure 6, we observe that expected clustering falls in b but that it is not monotonic as p increases.

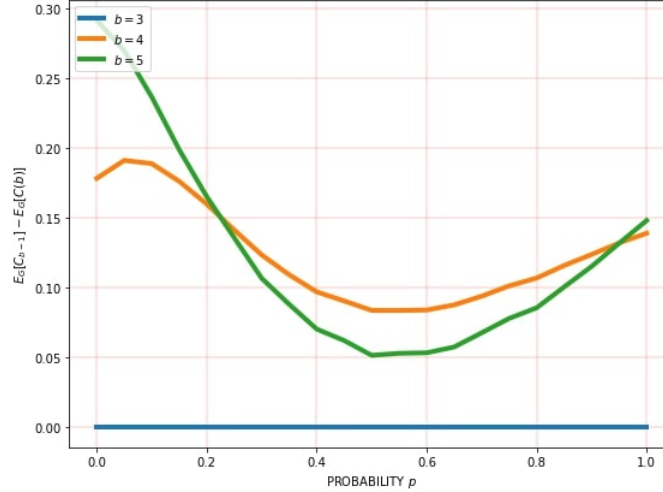


Figure B.6: The simulated difference in expectation $\mathbb{E}_G[C_{b-1}] - \mathbb{E}_G[C(b)]$ between the Yin-Benson-Leskovec clustering coefficient C_{b-1} and our coefficient $C(b)$ from 250 replications of the small-world graph of [72] with $n = 50$ nodes, each of which has degree $2k = 14$, and edge-rewiring probability $0 \leq p \leq 1$. See Figure 6 for more details on the construction of the graph. We note that $C(3) = C_2$. We observe that the expected difference is positive and can be quite substantial, as for the Erdős-Rényi random graph $G(n, p)$ in Figure 5. Similarly to the simulated small-world expected clustering in Figure 6 and Figure B.5, the expected difference is not monotonic as p increases.

Carrier	$C(3)$	$C(3)_{\text{rand}}$	$C(4)$	$C(4)_{\text{rand}}$	$C(5)$	$C(5)_{\text{rand}}$	Connected %	C_3	$C_{3,\text{rand}}$	C_4	$C_{4,\text{rand}}$	C_3 %	C_4 %
AA	0.120	0.060	0.018	0.000	0.002	0.000	45.3	0.101	0.003	0.075	0.000	100.0	5.7
AS	0.037	0.083	0.000	0.000	0.000	0.000	19.3	0.000	0.003	NA	0.000	95.4	1.7
DL	0.146	0.061	0.021	0.000	0.002	0.000	68.7	0.106	0.003	0.066	0.000	100.0	9.5
FL	0.154	0.108	0.008	0.001	0.000	0.000	62.7	0.038	0.007	0.000	0.000	100.0	10.4
NK	0.379	0.222	0.097	0.010	0.016	0.000	97.8	0.218	0.043	0.130	0.005	100.0	83.0
UA	0.346	0.138	0.122	0.003	0.034	0.000	95.6	0.259	0.017	0.189	0.001	100.0	65.6
US	0.115	0.067	0.011	0.000	0.000	0.000	34.0	0.054	0.003	0.000	0.000	100.0	4.7
WN	0.335	0.136	0.106	0.002	0.031	0.000	100.0	0.242	0.018	0.199	0.002	100.0	100.0

Table B.3: Descriptive statistics for eight carrier networks in 2013Q4. The carriers are American Airlines (AA), Alaska Airlines (AS), Delta Air Lines (DL), AirTran Airways (FL), Spirit Airlines (NK), United Airlines (UA), US Airways (US), and Southwest Airlines (WN). We compare the clustering coefficients $C(3)$, $C(4)$ and $C(5)$ to realizations from Erdős-Rényi random graphs $G(n, p)$ with n equal to the number of nodes in the observed network, and edge-formation probability p equal to its density. Some of the random graphs are not connected (and Connected % gives the percentage of connected realizations across all replications). Clustering coefficients are averaged over all 1000 realizations of $G(n, p)$, both connected and disconnected, as $C(3)_{\text{rand}}$, $C(4)_{\text{rand}}$ and $C(5)_{\text{rand}}$. These columns are given in Table 1. We also report the Yin-Benson-Leskovec [76] clustering coefficients C_3 and C_4 (note that $C_2 = C(3)$ and this is not included separately). When $C_{b-1} = 0$, so that there are no b -cliques in the network, then there cannot be any $L(b, 1)$ lollipops and so the statistic C_b will be undefined. In the empirical data, C_4 is not defined on the AS network. Conversely, $C(b)$ will always be defined on a connected graph. Both C_3 and C_4 detect higher-order clustering when it is present, and $C_3 > C_4$ for each network. There is also evidence that $C_3 \geq C(4)$ and $C_4 \geq C(5)$ whenever the Yin-Benson-Leskovec statistics are defined. The clustering coefficients C_3 and C_4 are averaged over all realizations of $G(n, p)$ for which the statistics are defined, and are reported as $C_{3,\text{rand}}$ and $C_{4,\text{rand}}$ (and C_3 % and C_4 % give the percentage of realizations for which each statistic is properly defined). We see that Yin-Benson-Leskovec clustering coefficients C_3 and C_4 are typically higher than random for all carriers, and that C_4 cannot be computed on many realizations of the sparse random graphs.

Acknowledgements

We are grateful to the editor Youjin Deng and two referees, whose comments greatly improved the paper, and to Chantal Roucolle and Tatiana Seregina. The usual caveat applies. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Keywords: Airline network, Clique, Higher-order clustering, Graph theory, Subgraph.

PACS numbers: 02.10.Ox (Combinatorics; graph theory), 89.40.Dd (Air transportation), 89.65.Gh (Economics; econophysics; financial markets; business and management), 89.75.-k (Complex systems).

JEL classification: L14 (Transactional Relationships; Contracts and Reputation; Networks), L22 (Firm Organization and Market Structure), L93 (Air Transportation), C65 (Miscellaneous Mathematical Tools).

References

- [1] M. Agasse-Duval and S. Lawford. Subgraphs and motifs in a dynamic airline network. Technical Report arXiv:1807.02585, 2018.
- [2] V. Aguirregabiria and C.-Y. Ho. A dynamic oligopoly game of the US airline industry: Estimation and policy experiments. *Journal of Econometrics*, 168:156–173, 2012.
- [3] F. Akbas, F. Meschke, and M.B. Wintoki. Director networks and informed traders. *Journal of Accounting and Economics*, 62:1–23, 2016.
- [4] R. Albert and A.-L. Barabási. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74:47–97, 2002.
- [5] L.A.N. Amaral and J.M. Ottino. Complex networks. *European Physical Journal B*, 38:147–162, 2004.
- [6] A. Banerjee, A.G. Chandrasekhar, E. Duflo, and M.O. Jackson. The diffusion of microfinance. *Science*, 341:1236498, 2013.
- [7] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286:509–512, 1999.
- [8] A. Barrat and M. Weigt. On the properties of small-world network models. *European Physical Journal B*, 13:547–560, 2000.
- [9] P. Baumgarten, R. Malina, and A. Lange. The impact of hubbing concentration on flight delays within airline networks: An empirical analysis of the US domestic market. *Transportation Research E*, 66:103–114, 2014.
- [10] A.R. Benson. *Tools for higher-order network analysis*. PhD thesis, Stanford University, 2017. Available from <http://arxiv.org/pdf/1802.06820.pdf>.
- [11] A.R. Benson, D.F. Gleich, and J. Leskovec. Higher-order organization of complex networks. *Science*, 353:163–166, 2016.
- [12] A. Bombelli, B.F. Santos, and L. Tavasszy. Analysis of the air cargo transport network using a complex network theory perspective. *Transportation Research Part E*, 138:101959, 2020.
- [13] R. Boulet and B. Jouve. The lollipop graph is determined by its spectrum. *Electronic Journal of Combinatorics*, 15, 2008.
- [14] G.A. Bounova. *Topological evolution of networks: Case studies in the US airlines and language Wikipedias*. PhD thesis, Massachusetts Institute of Technology, 2009.
- [15] G. Brightwell and P. Winkler. Maximum hitting time for random walks on graphs. *Random Structures & Algorithms*, 1:263–276, 1990.
- [16] G. Caldarelli, R. Pastor-Satorras, and A. Vespignani. Structure of cycles and local ordering in complex networks. *European Physical Journal B*, 38:183–186, 2004.
- [17] M. Chakraborty, S. Chowdhury, J. Chakraborty, R. Mehera, and R.K. Pal. Algorithms for generating all possible spanning trees of a simple undirected connected graph: An extensive review. *Complex & Intelligent Systems*, 5:265–281, 2019.

- [18] Y. Chen, J. Wang, and F. Jin. Robustness of China’s air transport network from 1975 to 2017. *Physica A*, 539: 122876, 2020.
- [19] T.K.Y. Cheung, C.W.H. Wong, and A. Zhang. The evolution of aviation network: Global airport connectivity index 2006–2016. *Transportation Research Part E*, 133:101826, 2020.
- [20] F. Ciliberto and E. Tamer. Market structure and multiple equilibria in airline markets. *Econometrica*, 77: 1791–1828, 2009.
- [21] F. Ciliberto and J.W. Williams. Limited access to airport facilities and market power in the airline industry. *Journal of Law and Economics*, 53:467–495, 2010.
- [22] G. Cimini, T. Squartini, F. Saracco, D. Garlaschelli, A. Gabrielli, and G. Caldarelli. The statistical physics of real-world networks. *Nature Reviews Physics*, 1:58–71, 2019.
- [23] E. Cohen-Cole, A. Kirilenko, and E. Patacchini. Trading networks and liquidity provision. *Journal of Financial Economics*, 113:235–251, 2014.
- [24] D. Coppersmith and S. Winograd. Matrix multiplication via arithmetic progressions. *Journal of Symbolic Computation*, 9:251–280, 1990.
- [25] L. da F. Costa, F.A. Rodrigues, G. Travieso, and P.R. Villas Boas. Characterization of complex networks: A survey of measurements. *Advances in Physics*, 56:167–242, 2007.
- [26] M. Dai, Q. Liu, and K. Serfes. Is the effect of competition on price dispersion non-monotonic? Evidence from the U.S. airline industry. *Review of Economics and Statistics*, 96:161–170, 2014.
- [27] Á. de Paula. Econometrics of network models. In B. Honoré, A. Pakes, M. Piazzasi, and L. Samuelson, editors, *Advances in Economics and Econometrics, (Proceedings of the 11th World Congress of the Econometric Society)*, pages 268–323. Cambridge University Press, 2017.
- [28] Á. de Paula. Econometric models of network formation. *Annual Review of Economics*, 12, 2020.
- [29] R. Diestel. *Graph Theory*. Springer, 5th edition, 2017.
- [30] W.-B. Du, X.-L. Zhou, O. Lordan, Z. Wang, C. Zhao, and Y.-B. Zhu. Analysis of the Chinese Airline Network as multi-layer networks. *Transportation Research Part E*, 89:108–116, 2016.
- [31] R. El-Khatib, K. Fogel, and T. Jandik. CEO network centrality and merger performance. *Journal of Financial Economics*, 116:349–382, 2015.
- [32] R. Faris and D. Felmlee. Status struggles: Network centrality and gender segregation in same- and cross-gender aggression. *American Sociological Review*, 76:48–73, 2011.
- [33] J. Fox. There exist graphs with super-exponential Ramsey multiplicity constant. *Journal of Graph Theory*, 57: 89–98, 2008.
- [34] A. Fronczak, J.A. Hołyst, M. Jedynek, and J. Sienkiewicz. Higher order clustering coefficients in Barabási–Albert networks. *Physica A*, 316:688–694, 2002.
- [35] A. Gautreau, A. Barrat, and M. Barthélemy. Microdynamics in stationary complex networks. *PNAS*, 106: 8847–8852, 2009.

- [36] B. Graham and Á. de Paula, editors. *The Econometric Analysis of Network Data*. Academic Press, 2020.
- [37] R. Guimerà and L.A.N. Amaral. Modeling the world-wide airport network. *European Physical Journal B*, 38: 381–385, 2004.
- [38] R. Guimerà, S. Mossa, A. Turttschi, and L.A.N. Amaral. The worldwide air transportation network: Anomalous centrality, community structure, and cities’ global roles. *PNAS*, 102:7794–7799, 2005.
- [39] W.H. Haemers, X. Liu, and Y. Zhang. Spectral characterizations of lollipop graphs. *Linear Algebra and its Applications*, 428:2415–2423, 2008.
- [40] Y.V. Hochberg, A. Ljungqvist, and Y. Lu. Whom you know matters: Venture capital networks and investment performance. *Journal of Finance*, 62:251–301, 2007.
- [41] M.O. Jackson. *Social and Economic Networks*. Princeton University Press, 2008.
- [42] M.O. Jackson. Networks in the understanding of economic behaviors. *Journal of Economic Perspectives*, 28: 3–22, 2014.
- [43] M.O. Jackson and B.W. Rogers. The economics of small worlds. *Journal of the European Economic Association*, 3:617–627, 2005.
- [44] M.O. Jackson, B.W. Rogers, and Y. Zenou. The economic consequences of social-network structure. *Journal of Economic Literature*, 55:49–95, 2017.
- [45] H. Jeong, B. Tombor, R. Albert, Z.N. Oltvai, and A.-L. Barabási. The large-scale organization of metabolic networks. *Nature*, 407:651–654, 2000.
- [46] B. Jiang and C. Claramunt. Topological analysis of urban street networks. *Environment and Planning B: Planning and Design*, 31:151–162, 2004.
- [47] D. Jungnickel. *Graphs, Networks and Algorithms*. Springer, 3rd edition, 2008.
- [48] S. Lawford. Counting five node subgraphs. DEVI/ENAC unpublished report, 2020.
- [49] G.F. Lawler. Expected hitting times for a random walk on a connected graph. *Discrete Mathematics*, 61:85–92, 1986.
- [50] J. Lin and Y. Ban. The evolving network structure of US airline system during 1990–2010. *Physica A*, 410: 302–312, 2014.
- [51] O. Lordan and J.M. Sallan. Core and critical cities of global region airport networks. *Physica A*, 513:724–733, 2019.
- [52] O. Lordan, J.M. Sallan, and P. Simo. Study of the topology and robustness of airline route networks from the complex network approach: a survey and research agenda. *Journal of Transport Geography*, 37:112–120, 2014.
- [53] P. Malighetti, G. Martini, R. Redondi, and D. Scotti. Air transport networks of global integrators in the more liberalized Asian air cargo industry. *Transport Policy*, 80:12–23, 2019.
- [54] S.A. Marvel, T. Martin, C.R. Doering, D. Lusseau, and M.E.J. Newman. The small-world effect is a modern phenomenon. Technical Report arXiv:1310.2636, 2013.

- [55] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network motifs: Simple building blocks of complex networks. *Science*, 298:824–827, 2002.
- [56] N. Movarraei and M.M. Shikare. On the number of paths of lengths 3 and 4 in a graph. *International Journal of Applied Mathematical Research*, 3:178–189, 2014.
- [57] M.E.J. Newman. Models of the small world. *Journal of Statistical Physics*, 101:819–841, 2000.
- [58] M.E.J. Newman. Clustering and preferential attachment in growing networks. *Physical Review E*, 64:025102, 2001.
- [59] M.E.J. Newman. The structure and function of complex networks. *SIAM Review*, 45:167–256, 2003.
- [60] M.E.J. Newman. Properties of highly clustered networks. *Physical Review E*, 68:026121, 2003.
- [61] M.E.J. Newman. Random graphs with clustering. *Physical Review Letters*, 103:058701, 2009.
- [62] M.E.J. Newman, S.H. Strogatz, and D.J. Watts. Random graphs with arbitrary degree distributions and their applications. *Physical Review E*, 64:026118, 2001.
- [63] A. Reggiani, P. Nijkamp, and A. Cento. Connectivity and concentration in airline networks: A complexity analysis of Lufthansa’s network. *European Journal of Information Systems*, 119:449–461, 2010.
- [64] D.T. Robinson and T.E. Stuart. Network effects in the governance of strategic alliances. *Journal of Law, Economics, & Organization*, 23:242–273, 2004.
- [65] C. Roucolle, T. Seregina, and M. Urdanoz. Measuring the development of airline networks: Comprehensive indicators. *Transportation Research Part A*, 133:303–324, 2020.
- [66] C. Roucolle, T. Seregina, and M. Urdanoz. Network development and excess travel time. *Transport Policy*, 94:139–152, 2020.
- [67] T. Ryczkowski, A. Fronczak, and P. Fronczak. How transfer flights shape the structure of the airline network. *Scientific Reports*, 7:5630, 2017.
- [68] V. Strassen. Gaussian elimination is not optimal. *Numerische Mathematik*, 14:354–356, 1969.
- [69] S.H. Strogatz. Exploring complex networks. *Nature*, 410:268–276, 2001.
- [70] V. Vassilevska Williams. Multiplying matrices in $O(n^{2.373})$ time. Mimeo (available at: <http://people.csail.mit.edu/virgi/matrixmult-f.pdf>), 2014.
- [71] T. Verma, N.A.M. Araújo, and H.J. Herrmann. Revealing the structure of the world airline network. *Scientific Reports*, 4:5638, 2014.
- [72] D.J. Watts and S.H. Strogatz. Collective dynamics of ‘small-world’ networks. *Nature*, 393:440–442, 1998.
- [73] D.R. Wuellner, S. Roy, and R.M. D’Souza. Resilience and rewiring of the passenger airline networks in the United States. *Physical Review E*, 82:056101, 2010.
- [74] O.N. Yaveroğlu, N. Malod-Dognin, D. Davis, Z. Levnajic, V. Janjic, R. Karapandza, A. Stojmirovic, and N. Pržulj. Revealing the hidden language of complex networks. *Scientific Reports*, 4:4547, 2014.

- [75] H. Yin, A.R. Benson, J. Leskovec, and D.F. Gleich. Local higher-order graph clustering. In *KDD 17 (Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining)*, pages 555–564, 2017.
- [76] H. Yin, A.R. Benson, and J. Leskovec. Higher-order clustering in networks. *Physical Review E*, 97:052306, 2018.
- [77] H. Yin, A.R. Benson, and J. Leskovec. The local closure coefficient: A new perspective on network clustering. In *WSDM 19 (Proceedings of the 12th ACM International Conference on Web Search and Data Mining)*, pages 303–311, 2019.
- [78] F. Zaidi. Small world networks and clustered small world networks with random connectivity. *Social Network Analysis and Mining*, 3:51–63, 2012.
- [79] M. Zanin and F. Lillo. Modelling the air transport with complex networks: A short review. *European Physical Journal Special Topics*, 215:5–21, 2013.
- [80] X. Zhu, H. Tian, and S. Cai. Predicting missing links via effective paths. *Physica A*, 413:515–522, 2014.
- [81] X. Zhu, H. Tian, X. Chen, W. Wang, and S. Cai. Heterogeneous behavioral adoption in multiplex networks. *New Journal of Physics*, 20:125002, 2018.
- [82] X. Zhu, J. Ma, X. Su, H. Tian, W. Wang, and S. Cai. Information spreading on weighted multiplex social network. *Complexity*, 2019:5920187, 2019.
- [83] Y. Zou, R.V. Donner, N. Marwan, J.F. Donges, and J. Kurths. Complex network approaches to nonlinear time series analysis. *Physics Reports*, 787:1–97, 2019.