Network Science **9** (3): 354–386, 2021
doi:10.1017/nws.2021.12

CAMBRIDGE UNIVERSITY PRESS

RESEARCH ARTICLE

Investigating scientific mobility in co-authorship networks using multilayer temporal motifs

Hanjo D. Boekhout[1,2,*], Vincent A. Traag[2] and Frank W. Takes[1]

[1]Department of Computer Science (LIACS), Leiden University, The Netherlands (e-mail: f.w.takes@liacs.leidenuniv.nl) and [2]Centre for Science & Technology Studies (CWTS), Leiden University, The Netherlands (e-mail: v.a.traag@cwts.leidenuniv.nl)
*Corresponding author. Email: h.d.boekhout@liacs.leidenuniv.nl

Action Editor: Ulrik Brandes

Abstract

This paper introduces a framework for understanding complex temporal interaction patterns in large-scale scientific collaboration networks. In particular, we investigate how two key concepts in science studies, scientific collaboration and scientific mobility, are related and possibly differ between fields. We do so by analyzing multilayer temporal motifs: small recurring configurations of nodes and edges.

Driven by the problem that many papers share the same publication year, we first provide a methodological contribution: an efficient counting algorithm for multilayer temporal motifs with concurrent edges. Next, we introduce a systematic categorization of the multilayer temporal motifs, such that each category reflects a pattern of behavior relevant to scientific collaboration and mobility. Here, a key question concerns the causal direction: does mobility lead to collaboration or vice versa? Applying this framework to scientific collaboration networks extracted from Web of Science (WoS) consisting of up to 7.7 million nodes (authors) and 94 million edges (collaborations), we find that international collaboration and international mobility reciprocally influence one another. Additionally, we find that Social sciences & Humanities (SSH) scholars co-author to a greater extent with authors at a distance, while Mathematics & Computer science (M&C) scholars tend to continue to collaborate within the established knowledge network and organization.

Keywords: co-authorship networks; scientific mobility; scientific collaboration; motif counting; multilayer temporal motifs; network motifs; concurrent edges

1. Introduction

Through technological advances and increasing digital communication, the world is becoming more and more connected. Small physical distances are no longer a necessity for interactions to occur. By modeling interactions between entities in complex systems as networks, the field of network science aims to understand these systems, their entities and interactions (Barabási, 2016). Network science approaches have provided new insights into a wide variety of complex systems. From social networks, identifying key persons within them (Das *et al.*, 2018), to protein networks, contributing to the understanding of protein structure, folding, stability, function, and dynamics (Chakrabarty & Parekh, 2016), to corporate networks, studying corporate governance practices through links of corporate ownership and shared directors (Takes *et al.*, 2018), and many more. In this paper, we focus on scientific collaboration networks, specifically *co-authorship networks* which capture interactions between authors who collaborated on scientific papers.

The study of co-authorship networks, and the study of networks more generally, often focuses on explaining macro-level properties of the network as a whole, using microlevel properties of

Figure 1. Example motifs implying mobility. Edge labels imply temporal order, and indicate either O(rganizational), N(ational), or I(nternational) collaborations. Collaboration type indicates the closest proximity between the known organizations of co-authors. For each motif, mobility can be inferred from the change of collaboration type on the parallel edges.

the nodes, such as node degrees (Bordons *et al.*, 2015; Molontay & Nagy, 2019). Some studies identified noteworthy patterns at the *meso-level* of co-authorship networks (Krumov *et al.*, 2011; Choobdar *et al.*, 2012), often conceptualized as so-called *network motifs*. A motif is a configuration of nodes and edges, usually only a few, that occurs at a high rate throughout the network (Milo *et al.*, 2002; Benson *et al.*, 2016). These studies focused on static motifs only, i.e., motifs that consist of edges on which no order is implied and which all model the same interaction type. However, co-authorship networks are inherently dynamic (Mali *et al.*, 2012), with new collaborations often resulting from their existing *knowledge network*, i.e., their past collaborations, either directly or indirectly. Our goal is to capture these dynamics with network motifs in an attempt to gain a better understanding of *scientific mobility*: scholars moving between organizations, a frequently studied concept in scientometrics (Mingers & Leydesdorff, 2015). Scientometrics, also known as quantitative science studies, is the field concerned with the study of quantitative features and characteristics of science and scientific research. We capture the dynamic evolution of collaborations based on the associated papers' publication years. Additionally, we represent physical distance between co-authors by distinguishing between collaborations at the organizational, local, national, and international level.

In recent years, methods have been proposed to deal with increasingly more complex motifs. Some recently introduced methods incorporated the evolution of networks over time in *temporal motifs*, to gain a greater understanding of the dynamic nature of *temporal networks*, also known as *dynamic networks* (Paranjape *et al.*, 2017; Holme & Saramäki, 2019). Other methods incorporated different types of interactions within motifs, i.e., *multilayer motifs* (Takes *et al.*, 2018). Recently, we proposed a method to efficiently count *multilayer temporal motifs* in large-scale networks (Boekhout *et al.*, 2019). Here, we build on that work and use multilayer temporal motifs to study the direct and indirect formation of scientific collaborations based on past collaborations. We believe this will contribute to a more fine-grained understanding of the evolution of such collaborations.

By making a distinction into different network layers based on collaborations at the organizational, local, national, and international level, we can infer scientific mobility from the configuration of some multilayer temporal motifs. For example, the motifs depicted in Figure 1 imply mobility events through a change in collaboration distance between two authors. In the motif on the left, two authors (top left and bottom nodes) that first collaborated at an organizational level (O) later collaborated on a national level (N), implying the two are no longer at the same organization. Similarly, in the the motif on the right, two authors that first collaborated at an organizational level (O) later collaborated on an international level (I), implying at least one of the two changed countries. This new perspective to study scientific mobility has several advantages: the methodology scales to large networks, considers the temporal order of co-authorships and directly places mobility events within the context of the relevant knowledge networks.

We extracted five large co-authorship networks, covering different fields. Each dataset consists of between 4 and 94 million collaborations on papers published in the period 2007–2016. Multilayer temporal networks are created by forming, for each pair of authors, one collaboration edge per paper based on their closest affiliations (organizational, local, national, or international layer), with the paper's publication year serving as timestamp. Our goal is to use these timestamps to impose a sequential order. However, using publication years as timestamps leads to many concurrent edges on which we do not want to infer an order. Existing algorithms (Boekhout *et al.*, 2019) cannot properly handle concurrent edges, which are prevalent in this dataset, as well as many other real-world systems. Furthermore, we wish to avoid counting motifs that are mostly

https://doi.org/10.1017/nws.2021.12 Published online by Cambridge University Press

the result of a collaboration on a single paper. For example, if a paper involved three authors, all three pairs of authors would be involved in a collaboration. Motifs that include multiple such collaborations tell us nothing about the evolution of collaborations over time. Therefore, they are uninteresting in the context of our study and we want to exclude them from the analysis. This too is not possible with existing algorithms. To overcome these two shortcomings, we extend existing motif counting algorithms to handle *concurrent edges*, i.e., allow for multiple edges to occur within a motif with the same timestamp, and to enforce a type of *edge attribute exclusivity*, so that in each counted motif every edge is formed from a different paper.

Existing multilayer temporal motif counting algorithms (Boekhout *et al.*, 2019) have a time complexity of $O(m\lambda^2)$, with λ the number of layers and m the number of edges. We show that our extensions to handle concurrent edges and enforce edge attribute exclusivity can be accomplished through smart traversal of the edges, adding only a small constant factor to the complexity. Furthermore, the attribute exclusivity is applicable not only to co-authorship networks, but to any one-mode network projected from a two-mode network, where the one-mode edge attributes are based on node attributes of the projected mode.

The interpretation of motifs depends on the real-world complex system modeled by the network. Furthermore, the multitude of different multilayer temporal motifs, makes their interpretation exceedingly difficult. Therefore, we systematically assign each motif to categories that represent some real-world meaning that is relevant to the domain of scientific collaboration and mobility. By studying the prominence of motif categories in certain fields and countries and studying the interplay between the various categories, we are able to draw conclusions about typical behavior with respect to scientific collaboration and mobility, globally, as well as per country (see Supplementary Material B), for each scientific field.

One aspect of scientific mobility that we are especially interested in, is how collaborations lead to scientific mobility, and how scientific mobility fosters collaboration. Studies investigating causes of international mobility (Guth & Gill, 2008; Baruffaldi & Landoni, 2010) found that insertion in international knowledge networks play an important role in the motivations for international mobility. On the contrary, Kato & Ando (2017) concluded that the relationship between international mobility and collaboration goes in one direction only: mobility resulting in collaboration. The authors state that networks created through international collaboration are not a factor in international migration. Although we are not able to identify specific causes of individual mobility events, the results from our analysis suggest that the relationship between international mobility and collaboration exists in both directions.

To sum up, the contributions of this paper are as follows:

1. we extend existing motif counting algorithms to be able to handle concurrent edges;
2. we extend existing motif counting algorithms to enforce edge attribute exclusivity, such that no two edges in a counted motif can have the same attribute value;
3. we introduce a systematic categorization of the meaning of multilayer temporal motifs in the context of scientific collaboration and mobility;
4. we infer typical behavior with respect to scientific co-authorship and mobility in general and for specific scientific fields (and countries); and
5. we show that the relationship between international mobility and collaboration exists in both directions, shedding new light on the debate by Kato & Ando (2017).

The remainder of this paper is structured as follows. First, relevant related and previous work is presented in Section 2. Then, necessary background and definitions for motif counting are provided in Section 3. The motif counting algorithms from previous work and our new methodological extensions are discussed in Section 4. Next, Section 5 describes the network datasets and their extraction from Web of Science. Then in Section 6, we add meaning to each

https://doi.org/10.1017/nws.2021.12 Published online by Cambridge University Press

motif configuration through systematic categorization. Subsequently in Section 7, we perform experiments and interpret results with the use of these categories. Finally, we summarize our results and contributions and discuss future work in Section 8.

2. Related work

In this section we first discuss literature related to the motif counting problem, followed by literature investigating co-authorship networks and studies into scientific collaboration from a network context. Finally, we consider literature on scientific mobility.

2.1 Motif counting

Recently a comprehensive survey on subgraph counting methods, i.e., motif counting, was performed by Ribeiro *et al.* (2021). The authors provided a comprehensive review on exact, approximate, and parallel methods. However, this work focussed only on methods for simple static motifs and only briefly referenced methods for more complex motifs, such as motifs in multilayer networks. A different survey by Jazayeri & Yang (2020) also looked at methods dealing with temporal networks. One such method was introduced by Paranjape *et al.* (2017) to count a set of temporal motifs. The authors proposed algorithms that were able to efficiently count these motifs (in $O(m)$ time, with m the number of edges). Boekhout *et al.* (2019) extended these algorithms to count multilayer temporal motifs and handle partial timing. However, this methodology still implied an order on the "untimed" edges based on their order in the dataset. Here, we expand on this previous work by proposing a methodology that can adequately handle and efficiently produce counts for all temporal motifs, including those with concurrent edges.

2.2 Co-authorship networks and collaboration

Kumar (2015) provides an extensive review of the literature, up to 2015, on co-authorship networks. Research into co-authorship networks mostly follows three themes. First, there are papers that focus on specific fields or countries, which aim to understand them by using, for example, centrality measures to find the most prolific or influential scholars (Molontay & Nagy, 2019). These studies tend to analyze small static networks and focus on micro- and macro-level network properties. Second, there are papers that try to link node-specific social network measures to academic performance (Bordons *et al.*, 2015; Hu *et al.*, 2019). Our research falls in the third research theme, studies of collaboration itself, which we review in detail below.

As we study collaboration, we must realize that co-authorship and collaboration are not the same thing. Melin & Persson (1996) discussed to what extent co-authorship data reflects actual collaboration. The authors stated that there is hardly a tendency for collaboration to be underrepresented when studying co-authorships. However, we should acknowledge that this is field dependent as, for example, in social sciences much collaboration is not expressed through co-authorships, but through acknowledgements (Paul-Hus *et al.*, 2017). When it comes to the overall structure of collaboration networks, Barabási *et al.* (2002) found co-authorship networks to be scale-free and their evolution to be governed by preferential attachment, i.e., new collaborations were more likely to connect to scholars with a high degree of collaborations. Wagner & Leydesdorff (2005) found that the growth of international co-authorships overall could be attributed to preferential attachment (individual scientists collaborating in search of recognition and reward). Glänzel & Schubert (2005) found that co-authorship domesticity, the likelihood of collaborations to remain inside a country, was clearly influenced by country size and country "remoteness" (geographically, linguistically, politically, etc.). These studies tend to view collaboration edges as independent, whereas we attempt to find (meaningful) collaboration patterns through the analysis of motifs that incorporate more than one edge.

https://doi.org/10.1017/nws.2021.12 Published online by Cambridge University Press

In literature on co-authorship networks, the term "collaboration pattern" generally refers to a set of node and path measures that are characteristic for collaboration (Newman, 2004). However, we use it to refer to network patterns of collaboration edges in co-authorship networks, such as motifs. Krumov *et al.* (2011) analyzed the correlation of a small set of single-layered static motifs with citation frequencies. The authors showed that the impact of individual authors or publications depends unexpectedly strongly on the meso-level structure of co-authorship networks. Choobdar *et al.* (2012) used motif fingerprints of a set of single-layered static motifs to assess similarity across scientific fields. They found that some motifs were overrepresented in some fields, identifying characteristic collaboration behavior. Our approach differs from the previous work in this direction as follows: (1) we consider dynamic, not static, motifs; (2) we consider multiple types of collaboration (organizational, local, national, and international), i.e., multilayered motifs; and (3) we consider all motifs of a certain size, rather than a preselected set of motifs.

2.3 Scientific mobility

Early research into scientific mobility consisted of, often small-scale, qualitative research into "Brain drain", "Brain gain", and "Brain circulation" (Gaillard & Gaillard, 1997; Stark *et al.*, 1997). Laudel (2003) was the first to propose the use of bibliometric methods to investigate mobility, i.e., using the address field of publications to identify mobility patterns. The advent of author disambiguation methods for large bibliometric databases such as Scopus and Web of Science (Caron & van Eck, 2014), allowed researchers to track authors and their affiliations, as listed on their published papers, over time.

Moed *et al.* (2013) concluded that a bibliometric study of scientific migration using Scopus was feasible and provided significant outcomes. This sparked various lines of research. Appelt *et al.* (2015) concluded that collaboration appeared to be a major factor associated with the mobility of scientists. Their analysis showed that the mobility of scientists particularly relied on flows of tertiary-level students in the opposite direction, from destination to origin country. Aman (2018) explored the relation between CV data and Scopus data in regard to tracking international mobility of scientists. Aman concluded that Scopus bibliometric data are suitable to identify a scientist's international mobility. Czaika & Orazbayev (2018) provided an empirical assessment of global scientific mobility over the past four decades. The authors found an increasing diversity of origin and destination countries, a shift of the center of gravity of scientific knowledge production eastwards, an increase in average migration distances and found that visa restrictions form a significant barrier to international mobility.

Similar as for Scopus, research using Web of Science was sparked. Notably, Chinchilla-Rodríguez *et al.* (2017) compared the networks of international collaboration and mobility. The authors showed that researchers collaborate internationally to a much higher degree than they become internationally mobile. Chinchilla-Rodríguez *et al.* (2018) compared the flow of mobile researchers and the number of publications in international collaboration. The authors found that there was a significant relationship between the flow of mobile researchers and the capacity for publishing with foreign partners in the more prolific countries, but found that mobility was always lower than collaboration. Furthermore, they found that the more resources available in a country (both scientific and economic) the greater the likelihood of attracting foreign partners and mobilizing human capital.

Unlike these related works, we do not directly obtain mobility information from affiliation data, but we use affiliation data to determine collaboration distances and imply mobility from changing collaboration distances over time. Every mobility event we capture is directly associated with collaborations, which tells us more about the structure of the mobile author's scientific knowledge network.

Other relevant work, related to scientific mobility, focusses on the motivations for mobility. Guth & Gill (2008) found that the actual moves themselves were often due to "chance" encounters

https://doi.org/10.1017/nws.2021.12 Published online by Cambridge University Press

or opportunities, but found contacts to also play an important role. Leyman (2009) demonstrated that researchers that are encouraged by their supervisor to go abroad show more interest in international mobility. Notably, Baruffaldi & Landoni (2010) found that insertion in international knowledge networks and the presence of links with the source country increased the probability of future mobility. On the contrary, Kato & Ando (2017) found that networks created through international collaboration are not a factor in international migration. The authors concluded that the relationship between international mobility and collaboration goes in one direction: from mobility to collaboration. Based on the collaboration motifs we find we try to shed new light on this debate.

3. Background, notation, and definitions

In this section, we provide definitions and introduce notation used to describe the algorithms discussed in this paper. We follow the notation and definitions introduced in Paranjape *et al.* (2017) and build upon the definitions in Boekhout *et al.* (2019).

3.1 Network notation and definitions

The two basic building blocks of any network are nodes and edges. An *edge* is a directed link between an ordered pair of nodes (u, v), which denotes u as the source node and v as the target node. Given a node set V of size $n = |V|$, a *multilayer temporal graph* $H = (V, E)$ is defined by a set E containing edges $e_i = (u_i, v_i, t_i, l_i)$, for $i = 1, 2, \ldots, m$, with $u_i, v_i \in V$, timestamp $t_i \in \mathbb{R}^+$ and layer $l_i \in \{1, 2, \ldots, \Lambda\}$, with Λ the number of layers. If $\Lambda > 1$, this is a *multilayer* network, if $\Lambda = 1$, this is a *single-layer* network. *Concurrent edges*, edges with the same timestamp, are allowed and parallel edges with the same direction and layer are also possible. The *underlying static graph* G of a multilayer temporal graph H is the graph formed by ignoring all timestamps and layers and removing any resulting duplicate edges. Although co-authorship networks are undirected, we assume edges to always be directed for the definitions and algorithms in this paper. This enables us to design algorithms that handle both directed and undirected networks, since the counts for undirected networks can be obtained through a simple post-processing step of the equivalent directed network, which we describe in Section 6.1.

3.2 Multilayer temporal motifs

In our previous work (Boekhout *et al.*, 2019), we gave the following definition for multilayer temporal motifs.

Definition 1. *A* r-node, s-edge, δ-temporal, λ-layer motif *is a sequence of* s *edges,* $M = ((u_1, v_1, t_1, l_1), (u_2, v_2, t_2, l_2), \ldots, (u_s, v_s, t_s, l_s))$ *that are time-ordered within a* δ *duration, i.e.,* $t_1 < t_2 < \ldots < t_s$ *and* $t_s - t_1 \leq \delta$*, and range over at most* λ *different layers, such that the underlying static graph, induced by* M*, is connected and has* r *nodes.*

The definition requires all edges in a motif to occur within δ time. This requirement gives us control over the period of time between interactions (edges) that we consider short enough to imply a relation between the interactions. For example, in a co-authorship network, co-authorships that are a year apart are very relevant to each other, while in a social network, such as Twitter, the relation between interactions that are a year apart is likely less meaningful. Furthermore, λ defines an upper limit on the number of layers involved. The definition allows for λ different layers in a motif M, but also allows fewer layers. This means that, for example, every 3-node, 3-edge, δ-temporal, 2-layer motif is also a 3-node, 3-edge, δ-temporal, 3-layer motif.

https://doi.org/10.1017/nws.2021.12 Published online by Cambridge University Press

Definition 1 induces a strict total order on the edges based on the timestamps. Because this ordering is strict, it does not allow for concurrent edges to occur within a motif. We redefine multilayer temporal motifs below, such that it encapsulates concurrent edges.

3.2.1 Concurrent edges

To facilitate concurrent edges, we would only have to change the strict total order ($<$) to a total order ($\leq$) in Definition 1. However, this change would introduce ambiguity as different orderings of concurrent edges could be considered different motifs. Instead, we first define a rank order as

Definition 2. *The rank order of element x_i in a set $(x_1, x_2, ..., x_m)$ is an integer $o_i \in \mathbb{N}^+$ (so $o_i \geq 1$) such that $o_i < o_j$ if and only if $x_i < x_j$, $o_i = o_j$ if and only if $x_i = x_j$ and $\min_j (o_j) - o_i = 1$ for $o_i < o_j$, with $\min_i (o_i) = 1$.*

Next we redefine a multilayer temporal motif allowing for concurrent edges as follows.

Definition 3. *A* r-node, s-edge, δ-temporal, λ-layer motif *is a sequence of s edges, $M = ((u_1, v_1, t_1, l_1), (u_2, v_2, t_2, l_2), \ldots, (u_s, v_s, t_s, l_s))$ with rank ordering $o = (o_1, o_2, \ldots, o_s)$, where o_i is the rank order of timestamp t_i and $o_1 \leq o_2 \leq \ldots \leq o_s$, such that $t_s - t_1 \leq \delta$, $(l_1, l_2, \ldots, l_s)$ range over at most λ different layers, and the underlying static graph, induced by M, is connected and has r nodes.*

This definition covers the full set of multilayer temporal motifs given some values for r, s and λ. To be able to count these motifs, we must distinguish between different configurations of the edges, their direction, temporal order, and layers. We define a multilayer temporal motif configuration as follows.

Definition 4. *A* multilayer temporal motif configuration, *$M_{a,b,c,d}$, of a r-node, s-edge, δ-temporal, λ-layer motif, is a combination of:*

a. *a structural configuration, i.e., an assignment of the s edges over the r nodes forming, for example, an (e) edge motif, (s) star motif, or (t) triangle motif;*
b. *a temporal configuration, i.e., an assignment of a rank order to each of the s edges defining a rank ordering $o = (o_1, o_2, \ldots, o_s)$;*
c. *a directional configuration, i.e., an assignment of a direction to each of the s edges, with 2^s possible configurations; and*
d. *a layer configuration, i.e., an assignment of a layer, from $\{1, \ldots, \lambda\}$, to each of the s edges, with λ^s possible layer configurations.*

The *static motif configuration* of a multilayer temporal motif configuration $M_{a,b,c,d}$ is given by $M_{a,c}$, i.e., the structural and directional configurations. The full set of 2-node and 3-node, 3-edge, δ-temporal motif configurations is depicted in Figure 2. Here, we only show single-layer motifs, because every δ-temporal λ-layer motif can be associated with a single δ-temporal motif (Boekhout *et al.*, 2019). For each of the 88 configurations shown in Figure 2, there exist λ^s layer configurations. Note that, for some motif configurations $M_{a,b,c}$, such as $M_{e,2,1}$, not every layer configuration is unique. After all, interchanging the layers of the concurrent edges of $M_{e,2,1}$ results in an identical motif. Furthermore, note that the same rank ordering with the rank orders assigned to different edges in the same static motif configuration can constitute different temporal configurations, for example, $M_{e,2,2}$ and $M_{e,5,2}$.

Each occurrence of a motif configuration in a multilayer temporal graph H is called an instance and is defined as follows.

https://doi.org/10.1017/nws.2021.12 Published online by Cambridge University Press

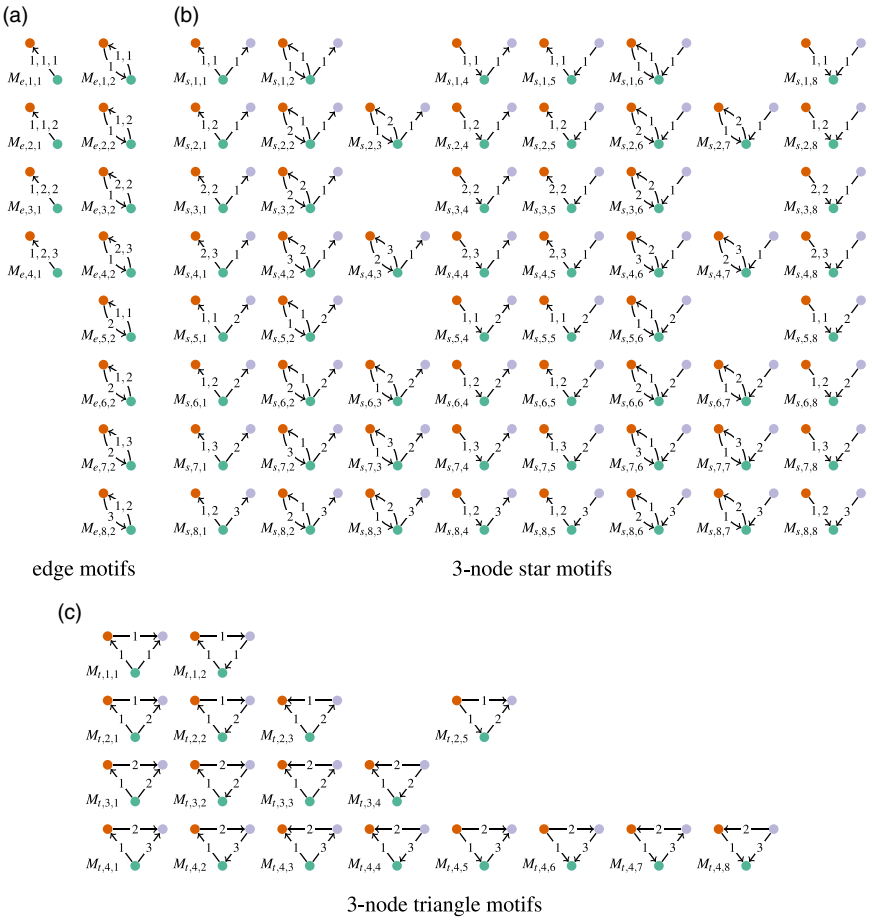

Figure 2. All 2-node and 3-node, 3-edge δ-temporal single-layer motif configurations allowing for concurrent edges. Edge numbers indicate their rank order. Rows have consistent temporal configurations and columns have consistent directional configurations.

Definition 5. *An* instance *of a multilayer temporal motif configuration $M_{a,b,c,d}$ in a multilayer temporal graph H, is a sequence $S = ((w_1, x_1, t'_1, l'_1), \ldots, (w_s, x_s, t'_s, l'_s))$ of s unique edges in H with rank ordering $o' = (o'_1, o'_2, \ldots, o'_s)$, where o'_i is the rank order of timestamp t'_i and $o'_1 \leq o'_2 \leq \ldots \leq o'_s$, such that*

1. *there exists a bijection f such that $f(w_j) = u_i$, $f(x_j) = v_i$, $l_i = l'_j$ and $o_i = o'_j$; and*
2. *the edges all occur within δ time, i.e., $t'_s - t'_1 \leq \delta$.*

Note that this definition requires the sequence S to have the same rank ordering as the motif configuration, but not the exact same edge ordering for concurrent edges. Therefore, we must be vigilant of equivalent concurrent edge orderings in our counting algorithms.

Because Paranjape *et al.* (2017) showed that a general algorithm quickly becomes inefficient as the motif size (r and s) increases, we focus on specific size (2-node and 3-node, 3-edge) motifs and define a separate algorithm for each structural configuration. The main problem, for which algorithms are proposed in Section 4, is as follows:

https://doi.org/10.1017/nws.2021.12 Published online by Cambridge University Press

Problem statement. *Given values for δ and λ and a multilayer temporal graph H, compute the number of instances of every 2-node and 3-node, 3-edge, δ-temporal, λ-layer motif.*

3.2.2 Edge attribute exclusivity

In addition to allowing concurrent edges, the second methodological contribution we make is *edge attribute exclusivity* within motifs. That is, we only count motifs that have no common attribute values on their edges. We define an additional edge attribute p_i for each edge ($i = 1, 2, \ldots, m$). Enforcing edge attribute exclusivity yields the following definition of a multilayer temporal edge-attribute-exclusive motif.

Definition 6. *A r-node, s-edge, δ-temporal, λ-layer edge-attribute-exclusive motif is a sequence of s edges, $M = ((u_1, v_1, t_1, l_1, p_1), (u_2, v_2, t_2, l_2, p_2), \ldots, (u_s, v_s, t_s, l_s, p_s))$ with rank ordering $o = (o_1, o_2, \ldots, o_s)$, where o_i is the rank order of timestamp t_i, such that $t_s - t_1 \leq \delta$, $(l_1, l_2, \ldots, l_s)$ range over at most λ different layers, the underlying static graph, induced by M, is connected and has r nodes and such that for all $i \neq j$ with $1 \leq i, j \leq s$ we have $p_i \neq p_j$.*

Our definition of a motif configuration remains unchanged for edge attribute exclusivity, but we do require a motif instance to adhere to the additional requirement that no two edges in the sequence S may have the same edge attribute value.

The algorithms we provide in Section 4 are able to enforce edge attribute exclusivity for a particular type of edge attributes. These attributes must be directly linked to the edge timestamp, i.e., if the attribute values are equal then the timestamps must be equal as well, but equal timestamps do not need to imply equal attribute values. This will always hold for one-mode networks that are projected from a two-mode network when the one-mode edge attribute uniquely identifies a node in the two-mode network. After all, the timestamp, and all other edge attributes, in the one-mode network originate from the same node in the two-mode network.

4. Motif counting algorithms

In this section, we present methodological extensions to existing motif counting algorithms (Paranjape *et al.*, 2017; Boekhout *et al.*, 2019) that allow for concurrent edges and enforce edge attribute exclusivity, the two requirements for successfully applying motif counting to evolving scientific collaboration networks. We discuss the basic concepts and functionality of the existing algorithms, which we consider vital knowledge for understanding our extensions, their efficiency, and correctness, in Section 4.1. In Section 4.2, 4.3 and 4.4, we describe our reformulation and extensions of the existing algorithms for counting, respectively, edge, star, and triangle motifs (see Figure 2). A detailed discussion is provided for all nontrivial changes required to achieve the extensions, and for why these changes add only small constant factors to the time complexities of the algorithms. Algorithmic details, such as counter definitions and pseudocode, are available in Supplementary Material A.

4.1 Existing algorithms

Paranjape *et al.* (2017) introduced three algorithms to count temporal motifs with a strict temporal order, a general algorithm and two specialized algorithms for two specific 3-node, 3-edge structural configurations. These algorithms were extended to count multilayer temporal motifs by Boekhout *et al.* (2019). The approach for each of the algorithms is to count all motif instances in an input sequence (S) in a single pass, thereby achieving a minimal number of considerations of each edge. The formation of the input sequences and functionality of the counting algorithm differs between the three algorithms. Below, we first concisely compare the format of the input sequences for the three algorithms in Section 4.1.1 and then focus on their functionality in Section 4.1.2.

https://doi.org/10.1017/nws.2021.12 Published online by Cambridge University Press

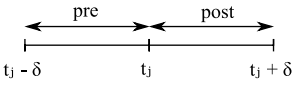

Figure 3. δ-timeframe

4.1.1 Input sequences

The general algorithm, which focuses on a single undirected static motif configuration, i.e., structural configuration (M_a), at a time, determines a separate input sequence for every instance of M_a. This is efficient for *edge motifs*, motifs that consist only of edges between two nodes, because each edge will only belong to one instance of M_a and will therefore be added to only one input sequence. On the contrary, Paranjape *et al.* (2017) showed that an edge may appear in a great number of instances of M_a for motifs that cover more than two nodes ($r > 2$). The authors concluded that their general algorithm is only efficient ($O(m)$ time) for edge motifs. The extension of the general algorithm to multilayer temporal motifs by Boekhout *et al.* (2019), increased the time complexity to $O(m\lambda^2)$. However, for a small number of layers, λ^2 is negligible with respect to the time complexity.

The two specialized algorithms reduce the number of input sequences an edge can appear in. For *star motifs*, motifs that consist of a center node u and edges to $r-1$ neighbors, this is achieved by grouping together all star motifs with the same center node. Only one input sequence is gathered for each center node $u \in V$ by gathering all edges connected to u. Thus, every edge (u, v) is only added to the two input sequences with, respectively, u and v as center nodes and a time complexity of $O(m\lambda^2)$ is achieved.

For *triangle motifs*, motifs whose edges form a triangle, the number of input sequences an edge appears in is reduced by assigning each static triangle to a pair of its nodes, i.e., one of its edges. Specifically, each static triangle is assigned to the node pair u, v that is connected by the greatest number of multilayer temporal edges. An input sequence is gathered for each node pair $u, v \in V$, to which at least one static triangle is assigned, by collecting the edges connecting u and v and the edges connecting them to their common neighbors as determined by the assigned static triangles. Paranjape *et al.* (2017) proved that this reduces the time complexity of counting triangle motifs from $O(m\tau)$ for the general algorithm to $O(m\sqrt{\tau})$, with τ the number of static triangles. Therefore, the extended algorithm to multilayer temporal motifs in Boekhout *et al.* (2019) has a time complexity of $O(m\sqrt{\tau}\lambda^2)$.

4.1.2 The delta-timeframe

As mentioned above, all three algorithms count motif instances in a single pass over the input sequence S. To accomplish this, the sequence is first preprocessed to produce sequence S' such that S' is time-ordered and all layers that are not of interest to a specific study are filtered out. For now, the sequence S' can still be considered strictly time-ordered because the existing algorithms do not yet consider concurrent edges. As such, when we iterate over the edges in S', we also move sequentially through time. As we iterate over the edges in the sequence S', at time t_j we consider e_j the *current edge* and we know that all motifs that include e_j consist of edges in the time window $[t_j - \delta, t_j + \delta]$, which we call the *$\delta$-timeframe*, depicted in Figure 3. Because all motif instances that include e_j occur in its δ-timeframe, each edge in the sequence S' has to be processed at most three times: (1) when it enters the δ-timeframe; (2) when the edge is the current edge; and (3) when it leaves the δ-timeframe.

The general algorithm only uses the "pre" segment of the δ-timeframe. Because the algorithm considers a single instance of a structural configuration (M_a) at a time, we have knowledge of all nodes in the input sequence and we can generate all possible combinations of edges up to length s, the number of edges in the target motifs. The algorithm maintains a counter for all such combinations that form subsequences of motif configurations ($M_{a,b,c,d}$) under investigation. For example, given motif configuration $M_{s,4,2}$ on nodes a, b and c, as

https://doi.org/10.1017/nws.2021.12 Published online by Cambridge University Press

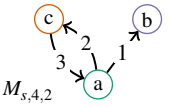

Figure 4. Example configuration instance

Figure 5. All temporal configurations of star motifs provided no concurrent edges

depicted in Figure 4, the set of edge combinations for which a counter is maintained is $\{((a,b)),((a,c)),((c,a)),((a,b),(a,c)),((c,a),(a,c)),((a,b),(a,c),(c,a))\}$. The counters are maintained such that at time t_j they indicate how often each of the edge combinations occur within time window $[t_j - \delta, t_j]$, i.e., how often they occur in the "pre" segment. Thus, an edge e_x is considered the last edge in the temporal configuration at time $t_x = t_j$, after which it fulfills the role of earlier edges until $t_x < t_j - \delta$ and it is removed from the counters. So, the general algorithm has to consider each edge in the input sequence S' only twice.

Because the number of edge combinations, and thus the number of counters, explodes as the number of nodes under consideration increases, the specialized 3-node, 3-edge star, and triangle algorithms are not able to use this same counting method. Instead they use the full δ-timeframe and consider not the last edge in the temporal configuration at time t_j, but consider a specific, strategically chosen, edge in the structural configuration as the *pivotal edge*. Counters are then maintained for all edge combinations of the remaining two edges in the configuration within the full δ-timeframe, such that at time t_j all motif instances with the current edge e_j as the pivotal edge in the configuration can be counted. However, unlike the general algorithm, which defines a counter for every specific edge combination, knowledge of the exact edges is discarded for the two edge combinations. The counters simply specify specific temporal, directional and layer configurations of the edges. We discuss why this is possible below.

For star motifs, the single edge, e.g., edge (a, b) in Figure 4, is chosen as the pivotal edge. Assuming a strict temporal order, Figure 5 shows the various temporal configurations. Excluding the pivotal edge, the remainder of the structural configuration consists of two parallel edges connecting the center node to the same neighbor. Now, if we discard knowledge of the neighbor to which the parallel edges connect, every combination of the parallel edges with a third edge from the center node forms either an edge motif or a star motif. Because we can count edge motifs in $O(m\lambda^2)$ time using the general algorithm, we can compensate for the edge motifs that we incorrectly counted as star motif by deducting the edge motif counts for center node u to all its neighbors. This is preferable as it reduces the number of counters required, as well as the time complexity, by a factor n.

For triangle motifs, where the input sequence is based on a node pair, the edge connecting this node pair is chosen as the pivotal edge. Now, the two remaining edges connect the node pair to a common neighbor and the pivotal edge connecting the node pair requires no knowledge of this neighbor at all. Thus, knowledge of the exact common neighbor can be discarded without issue for counters representing two edge combinations.

4.2 Edge motifs

The general algorithm introduced by Paranjape *et al.* (2017) was shown to be efficient only for *edge motifs*. Recall that edge motifs are motifs that consist of edges connecting only two nodes. Here, we reformulate the general algorithm to focus on counting edge motifs specifically and extend it to handle concurrent edges and enforce edge attribute exclusivity. With three (concurrent) edges we get four temporal configurations, which we label as shown in Figure 6. Adding edge directionality leads to the set depicted in Figure 2a.

https://doi.org/10.1017/nws.2021.12 Published online by Cambridge University Press

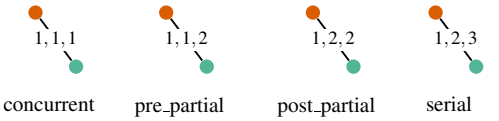

Figure 6. Edge motif temporal configurations

As discussed in Section 4.1.1 an input sequence (S) is gathered for every instance of a static edge, i.e., an input sequence is gathered for every connected node pair $u, v \in V$. Where existing algorithms only time-ordered the input sequence, we now preprocess it into a time-ordered sequence of sets of concurrent edges S', such that the rank order of every edge in a concurrent set of edges is equal. The sequence is further preprocessed to account for edge attribute exclusivity. Note that, for our purposes in this paper, these attributes are directly linked to the paper from which a co-authorship edge originates. This means that every edge that has the same attribute value (in our case, shares the same origin paper) also shares the same timestamp. It is exactly this observation that allows us to achieve edge attribute exclusivity within the current approach by grouping the edges with the same attribute value, within the sets of concurrent edges, together. This results in a sequence S'' of sets of sets of concurrent equal attribute value edges.

Like the general algorithm, only the "pre" segment of the δ-timeframe is used by our reformulation of the algorithm. However, we shift from counters for exact edge combinations to counters capturing the various temporal, directional and layer configurations, as used by the specialized algorithms. The definitions of these counters and the reformulated and extended edge motif counting algorithm pseudocode are provided in Supplementary Material A.1.

The approach of the extended algorithm remains unchanged from that described in Section 4.1.2. Where the existing general algorithm iterates over the input sequence one edge at a time, we now iterate over one set of concurrent edges, in input sequence S'', at a time. In fact, this is exactly the same behavior except that now more than one edge can share the same timestamp. Furthermore, as we iterate over the input sequence the counters are updated at the same points in time as well. The counters are updated when a concurrent set of edges becomes the current set ($coll_j$) and when a concurrent set of edges leaves the δ-timeframe. Thus, the main change to the algorithm comes from how and which counters are updated at those times.

Because every counter represents a specific combination of a structural and temporal configuration and because no two separate concurrent sets can have edges between them with the same edge attribute value, the extensions to accommodate concurrent edges and enforce edge attribute exclusivity are achieved through the simple addition of the appropriate counters and their update logic. As such, the counters and update logic used for the temporal configuration with no concurrent edges ("serial") remains unchanged from existing algorithms other than requiring to iterate over the edges in the concurrent set. We discuss how the extended algorithm handles concurrent edges in Section 4.2.1 and how edge attribute exclusivity is enforced in Section 4.2.2.

4.2.1 Concurrent edges

Because concurrent edges occur at the same time, we process all edges in a concurrent set at the same time. We use a single forward pass through the set of concurrent edges to form each combination of concurrent edges exactly once. By constructing all combinations of concurrent edges in a single forward pass, we prevent counting the same edge as two concurrent edges.

Because we use a single forward pass, the ordering of the edges within a concurrent set now determines the ordering of the directional and layer configurations counted. However, the ordering of the directional and layer configurations among concurrent edges has no meaning for edge motifs. After all, all concurrent edges connect the same two nodes and therefore their ordering is interchangeable, i.e., concurrent edge combinations $((u, v, A), (v, u, B))$ and $((v, u, B), (u, v, A))$, with layers A and B, are no different. Note that we are talking about changing

https://doi.org/10.1017/nws.2021.12 Published online by Cambridge University Press

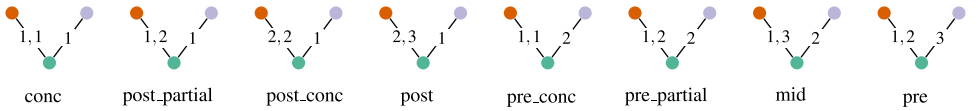

Figure 7. Star motif temporal configurations

the order of both the direction and layer at the same time. Because the ordering of the directional and layer configurations is interchangeable, we would want to count an occurrence of either as an occurrence of both. This is achieved in a post-processing step by adding the counted total of all equivalent directional and layer configuration permutations together and giving their sum as the result for each of them.

Note that for concurrent edges in an undirected network, the equivalence of layer configurations is no longer dependent on the directionality. For example, $((u, v, A), (v, u, B))$ and $((u, v, B), (v, u, A))$ are not equivalent in a directed network, but are equivalent in an undirected network. The resolution of these equivalences as an additional post-processing step, allow us to go from directed to undirected motif count results.

4.2.2 Edge attribute exclusivity

Earlier we stated that to realize edge attribute exclusivity with co-authorship networks, where the attribute uniquely identifies the source paper, we only need consider equal edge attribute edges when dealing with concurrent edges. We achieve this with the addition of various temporary counters. We update the temporary counters during passes over equal attribute value edge sets as stand-ins for the real counters and subsequently use them to update the real counters at the end of such a pass, thereby updating these counters for the entire set of equal attribute value edges at a time. As a result, larger combinations of concurrent edges formed using these real counters, never include two edges from the same set of equal attribute value edges, i.e., we never count combinations of concurrent edges with the same edge attribute value.

Thus, we are able to enforce edge attribute exclusivity without having to store information regarding the attribute as part of any counter nor the input sequence. We only require the addition of a small set of temporary counters and a minimal set of operations.

In short, we are able to deal with concurrent edges and realize edge attribute exclusivity through the simple addition of a few counters and a more systematic loop over the edges in the input sequence S''. Thus adding only a small constant number of operations per edge, based on the number of additional counters, and maintaining time complexity $O(m\lambda^2)$.

4.3 Star motifs

Star motifs are motifs that consist of a center node u and edges to $r-1$ neighbors (Paranjape *et al.*, 2017; Boekhout *et al.*, 2019). Given three nodes and three edges and allowing for concurrent edges, there are eight temporal configurations, which we label as shown in Figure 7. The full set of directed star motif configurations is depicted in Figure 2b.

As discussed in Section 4.1.1 an input sequence (S) is gathered for every (center) node $u \in V$. These input sequences are preprocessed into sequences S'' consisting of sets of sets of concurrent equal attribute value edges, in the same manner as for edge motifs. As discussed in Section 4.1.2 the single, nonparallel, edge in the star motif configuration is chosen as the pivotal edge and counters are formed for all edge combinations for the remaining two parallel edges over the full δ-timeframe. The definitions of these counters and the extended star motif counting algorithm pseudocode are provided in Supplementary Material A.3.

https://doi.org/10.1017/nws.2021.12 Published online by Cambridge University Press

The approach of the extended algorithm remains unchanged from that described in Section 4.1.2. We process sets of concurrent edges when they enter the δ-timeframe, when they become the current set and when they leave the δ-timeframe. Similar to the edge motif algorithm, we simply move from processing one edge to processing a set of concurrent edges at a time.

Again, the extensions to accommodate concurrent edges and enforce edge attribute exclusivity are achieved through the addition of the appropriate counters and their update logic. Note that edge attribute exclusivity is achieved in exactly the same way as was the case for the edge motifs, through the addition of temporary counters. Therefore, we will not discuss this further for star motifs (see Supplementary Material A.3 for the exact counter definitions and update logic). We explore new complications that arise for concurrent edges in star motifs and explain the extensions made to solve these below.

4.3.1 Concurrent edges

For edge motifs we had the convenience that every edge in the motif connected the same pair of nodes. This meant that a pair of concurrent edges and its reverse order, for example $((u, v, A), (v, u, B))$ and $((v, u, B), (u, v, A))$, could be counted using the same counters and update logic. Therefore, all possible orderings of concurrent edges could be counted in a single forward pass and their true total counts could be obtained by resolving for equivalences in post-processing.

Unfortunately, three of the temporal configurations of star motifs ("conc", "post_partial", and "pre_partial") have concurrent edges connecting the center node (u) to different neighbors (v, w), where we consider the parallel edges to connect to neighbor v. To be able to count these temporal configurations in a single forward pass, we cannot assign a specific edge in the configuration as the pivotal edge, as we have done for star motifs up to this point, because this might not be the last edge considered in a traversal of a set of concurrent edges. For example, the concurrent edge combination $((u, v, A), (u, w, B))$ remains equivalent to its reverse order $((u, w, B), (u, v, A))$. However, counting star motifs given the latter order in a single forward pass, means that edge (u, w, B) must be processed before the parallel edge (u, v, A). This requires us to consider one of the parallel edges to v as the pivotal edge instead of the single edge to w. This presents a significant algorithmic problem. After all, if one of the parallel edges (u, v) is the pivotal edge, then we require a counter that represents a combination of the two remaining edges in the configuration, which connect to v and w, respectively, i.e., a counter that represents a connection of two edges to two different neighbors. As we previously discussed in Boekhout *et al.* (2019), this would inevitably lead to *neighbor loops*, which would, with a worst case of $n - 1$ neighbors, increase both the time (and space) complexity by a factor n. Therefore, the simpler and more efficient solution is to traverse each set of concurrent edges both forward and backward, such that each of the loops covers one of the two possible orders of two concurrent edges. This is far more efficient, because the backward loop adds just a small number of additional operations per edge and requires no additional counters. As such, it only adds a small constant factor to the time complexity, instead of factor n, in worst case.

After introducing the backward loop, there remains one problematic case involving the "conc" configuration. This temporal configuration consists of three concurrent edges. If a set of concurrent edges is ordered such that we have $((u, v), (u, w), (u, v))$, neither the forward nor the backward loop on its own can prevent one of the parallel edges from being considered the pivotal edge. To allow the middle edge of three concurrent edges to be considered the pivotal edge, we approach the problem in a similar way as the existing algorithms did for the "mid" configuration. First, during the forward loop, we count all one edge directional and layer configurations and store this in a new counter "conc_pre_nodes". During the following backwards loop, this counter keeps track of the number of one edge directional and layer configurations that may be considered the last edge in the order $((u, v), (u, w), (u, v))$. A second additional counter called "conc_mid_sum" is added, which keeps track of the number of pairs of parallel edges $((u, v), (u, v))$ of which one edge

https://doi.org/10.1017/nws.2021.12 Published online by Cambridge University Press

Figure 8. Triangle motif temporal configurations

occurs before and the other after the current edge under consideration in the backward traversal. As we traverse the edges in the backward loop, for each set of equal attribute edges we perform the following three actions:

1. as the edges in this set become the current edges, they can no longer be considered possible last edges in the order $((u, v), (u, w), (u, v))$ and we reduce counters "conc_pre_nodes" and "conc_mid_sum" accordingly;
2. we consider the edges in this set the current edges, i.e., pivotal edges, and update the counter for the "conc" configuration accordingly; and
3. as the edges have been fully processed as current edges, they now become preceding edges, i.e., the first edges in the order $((u, v), (u, w), (u, v))$, and we update the counter "conc_mid_sum" accordingly.

Like for edge motifs, counting star motifs also requires a post-processing step to account for equivalent directional and layer configurations. Because we directly count every possible ordering of concurrent edges that connect to different neighbors, these equivalences only occur for the three temporal configuration that have concurrent parallel edges ("conc", 'post_conc", and "pre_conc").

Similar to the existing star motif counting algorithms (Paranjape *et al.*, 2017; Boekhout *et al.*, 2019), we drastically reduce the number of counters by discarding the knowledge of the neighbor to which the two parallel edges are connected for counters that represent two edge combinations. As such the algorithm cannot ensure, for the pivotal edge, that neighbor $v \neq w$. The number of additional star motifs counted when $v = w$ are exactly the sum of the number of 3-edge edge motifs for u and each of its neighbors. Therefore, as a second post-processing step, we subtract the matching edge motif counts, based on matching temporal, directional and layer configurations, for every neighbor of u from the star motif counts. Note that all directional and layer configuration equivalences should be resolved for both the edge and star motifs before the subtraction is performed.

Although the extensions to accommodate concurrent edges and enforce edge attribute exclusivity have made the star motif counting algorithm more complex compared to existing algorithms, its time and space complexity have remained virtually unchanged. Both complexities increase by only a small constant factor and thus the time complexity remains $O(m\lambda^2)$.

4.4 Triangle motifs

The last structural configuration for which we extend the motif counting algorithm is that of *triangle* motifs. Triangle motifs are motifs whose edges form a triangle (Paranjape *et al.*, 2017; Boekhout *et al.*, 2019). Given three nodes and three edges and allowing for concurrent edges, there are four temporal configurations, which we label as shown in Figure 8. The full set of directed triangle motif configurations is depicted in Figure 2c.

As discussed in Section 4.1.1 an input sequence (S) is gathered for every node pair $u, v \in V$ to which a static triangle has been assigned. These input sequences are preprocessed into sequences S'' consisting of sets of sets of concurrent equal attribute value edges, in the same manner as for edge and star motifs. As discussed in Section 4.1.2 the edge connecting node pair u, v in the triangle motif configuration is chosen as the pivotal edge and counters are formed for all edge

https://doi.org/10.1017/nws.2021.12 Published online by Cambridge University Press

combinations for the remaining two edges connecting to the common neighbor over the full δ-timeframe. The definitions of the counters and the extended triangle motif counting algorithm pseudocode are provided in Supplementary Material A.4.

Similar extensions were made to the triangle motif counting algorithm as discussed for edge and star motifs, including directional and layer configuration equivalence post-processing. Note that in Figure 2c only motif configuration $M_{t,1,2}$, a concurrent circle, lends itself to layer configuration equivalence within the same directional configuration.

As the same extensions are made to the triangle motif counting algorithm as was done for the star motif counting algorithm, here too we have only a small constant increase of our time and space complexity. Thus, the time complexity remains $O(m\sqrt{\tau}\lambda^2)$.

5. Data

In this section we discuss the co-authorship datasets used in this work. We discuss how the datasets were obtained from Web of Science (WoS) and define the various network layers.

We extracted our five global datasets from the in-house version of WoS at the Centre for Science and Technology Studies (CWTS). The CWTS version of WoS has been enriched with in-house author identifiers based on an improved author disambiguation algorithm (Caron & van Eck, 2014). We use these in-house author identifiers to associate authors to their respective oeuvres. Furthermore, this version has enriched organization information and more consistent and accurate assignment of papers to universities and organizations (Waltman *et al.*, 2012). Each extracted co-authorship network covers one main field and includes papers published in the period 2007–2016. A 10-year period was chosen so that there is an increased likelihood of mobility events to have occurred for each active author. Papers, and by extension co-authorships, are assigned to the fields on the journal level and can be associated with multiple fields. Papers with more than 25 authors are excluded to prevent papers with large author lists from skewing our results. For example, in the field of High Energy Physics publications with hundreds or thousands of authors are not uncommon. Given a mostly similar group of authors, just three such publications would generate such a large number of motifs that the balance of motifs found in the overarching field would be skewed toward motifs representing co-authorship in High Energy Physics. Additionally, for such publications the meaning of authorship with respect to individual contributions and collaboration is different compared to other fields (Birnholtz, 2006).

Co-authorship links are formed for every pair of authors on a paper, provided organization affiliation information for that paper was present in WoS for both authors. For each scientific field under study between 15% and 26% of organization affiliations is missing (see Table 1). Organization affiliations can be missing when, for (some) authors, it is not properly indicated on the published paper which authors were affiliated with which of the listed organizations. Each co-authorship link is assigned to a specific layer based on the proximity of the organizations to which the respective authors were affiliated. We define the following layers.

O. *Organizational* co-authorship, both authors were associated with the same organization.
L. *Local* co-authorship, the authors were associated with organizations based in the same city.
N. *National* co-authorship, the authors were associated with organizations based in the same country.
I. *International* co-authorship, the authors were associated with organizations based in different countries.

Because authors can be affiliated with multiple organizations at a time, multiple co-authorship links in different layers between two authors for the same paper are possible. When this occurs, only the link with the closest proximity ($O < L < N < I$) is included.

https://doi.org/10.1017/nws.2021.12 Published online by Cambridge University Press

Table 1. Descriptive global network dataset statistics

Field	SSH	B&H	P&E	L&E	M&C
Nodes (in millions)	1.0	7.7	4.6	3.1	1.2
Edges (in millions)	4.6	94.0	35.2	22.6	4.6
Static edges (in millions)	3.2	53.4	21.4	15.3	3.2
O(rganizational) edges (%)	55.8	67.6	65.6	61.8	63.4
L(ocal) edges (%)	3.3	3.8	2.6	2.8	2.8
N(ational) edges (%)	23.9	15.9	13.3	16.8	13.7
I(nternational) edges (%)	17.0	12.7	18.5	18.6	20.1
Papers (in thousands)	826	5,612	3,412	2,092	950
Interdisciplinary papers (%)	33.4	16.1	19.2	36.1	34.8
Missing org-affiliation information (%)	22.2	26.0	15.5	20.2	17.3

The publication year of a paper is used as the timestamp of co-authorship links associated with that paper. We use the publication year because the listed publication months in WoS are not always accurate, possibly leading to inaccurate timing of co-authorship links.

The five extracted global datasets cover, respectively: Social sciences & Humanities (SSH); Biomedical & Health sciences (B&H); Physical sciences & Engineering (P&E); Life & Earth sciences (L&E); and Mathematics & Computer science (M&C). Descriptive statistics are provided in Table 1, listing the number of nodes and edges, the number of static edges in the underlying static network, the percentage of edges in each of the layers, the number of papers from which the co-authorship edges are formed, the percentage of those papers that are interdisciplinary, i.e., that are associated with at least one other field as well, and the percentage of missing organization affiliations.

From the global datasets, country-specific datasets can be extracted and analyzed as well. This process is documented in Supplementary Material B.1.

6. Systematic interpretation of motifs in co-authorship networks

Although we have now defined our approach (Section 4) and datasets (Section 5), there is one final step to take: systematically assigning meaning to the various motif configurations by mapping them to categories relevant to the domain of co-authorship and scientific mobility. First we discuss how we retrieve undirected motif counts from directed results in Section 6.1. Then, we introduce three relevant categorization schemes:

- collaboration categories that capture the structural configuration (Section 6.2);
- international categories that deal with international collaboration and international mobility (Section 6.3); and
- mobility categories that describe different types of scientific mobility (Section 6.4).

A summary of the various categories is then given in Table 2. How the categories are mapped onto the full set of motifs is shown in Figure 10.

6.1 Directed to undirected results

Recall that the algorithms discussed in Section 4 count directed motifs. Co-authorships links, however, are an undirected relation between two authors. The set of undirected motifs is depicted in

https://doi.org/10.1017/nws.2021.12 Published online by Cambridge University Press

Table 2. An overview of all collaboration and mobility motif categories. In the examples, edge labels indicate the rank order and layer of the edges

(a) Collaboration categories		
Category	**Example**	**Description**
CC		continued collaboration between two authors
MC		central author with multiple co-authors
MEC		central author with multiple equidistant co-authors
MPEC		central author with multiple possibly equidistant co-authors
TC		team collaboration
ETC		equidistant team collaboration
EP		equidistant partner
EPC		equidistant partner likely caused by collaboration
EPE		equidistant partner likely cause of collaboration
OEP		organizational equidistant partner
OEPC		org. equidistant partner likely caused by collaboration
OEPE		org. equidistant partner likely cause of collaboration
(b) International categories		
Category	**Example**	**Description**
I		international co-authorship
IM		international mobility, unknown direction
IMI		incoming international mobility
IMO		outgoing international mobility
(c) Mobility categories		
Category	**Example**	**Description**
M		mobility event implied
CM		certain mobility event implied by an edge or star motif
MP		mobility event implied accompanied by a preceding edge
MS		mobility event implied accompanied by a succeeding edge
PM		possible mobility event implied by a triangle motif
MTC		possible (incoming) mobility event leading to collaboration
MSC		collaboration despite possible (outgoing) mobility event
M2		two mobility events implied
RFM		return or follow mobility
VM		visit mobility

https://doi.org/10.1017/nws.2021.12 Published online by Cambridge University Press

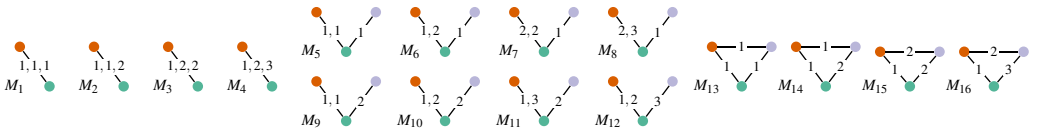

Figure 9. The set of undirected motifs, w.r.t. the set of directed motifs in Figure 2.

Figure 9. The numbering of the undirected motif configurations used throughout the remainder of the paper follows the one shown in this figure: edge motifs M_1–M_4, star motifs M_5–M_{12}, and triangle motifs M_{13}–M_{16}. The motif counts of the undirected motifs are directly retrieved from those of the directed motifs in Figure 2. This is done by first resolving layer configuration equivalence for motifs with concurrent edges where equivalence was previously prevented by directionality, such as $M_{t,1,1}$. After that, counts for equivalent temporal configurations are summed. For star and triangle motifs this translates to adding together the rows as depicted in Figure 2b and 2c. Not accounting for equivalent layer configurations, this effectively reduces the number of motif configurations to categorize from 5,632 (88×4^3) to 1,024 (16×4^3).

6.2 Collaboration categories

The first set of motif categories that we define, consists of categories that capture the structural configuration (see Section 3.2.1). The most obvious distinction to be made is between edge, star, and triangle motifs. Each of these structural configurations has a distinct meaning in the context of co-authorship networks. As such we define three main categories:

CC. *Continued Collaboration* between two authors (edge motifs);
MC. *Multiple Collaborators*, i.e., an author that has multiple co-authors (star motifs); and
TC. *Team Collaboration*, i.e., three authors with each pair having co-authored a paper together (triangle motifs).

Recall that we enforce edge attribute exclusivity (see Section 3.2.2), which means that all motifs must consist of co-authorships on three different papers.

For the three node motifs (star and triangle), we define additional subcategories based on specific meaning derived from the layer configurations. Specifically, for star motifs we distinguish between layer configurations that indicate that the co-authors are *equidistant*, i.e., the organizations of the authors are equally far apart (same layer), possibly equidistant or not at all equidistant with respect to the central author (center node). This leads us to define two subcategories:

MEC. *Multiple Equidistant Collaborators*, i.e., an author that has multiple co-authors with the same proximity at the same time; and
MPEC. *Multiple Possibly Equidistant Collaborators*, i.e., an author that has multiple co-authors whose equal proximity may be prevented by a change in proximity for one of the co-authors.

For triangle motifs we define subcategories of category TC based on the same concept of equidistant co-authorships:

ETC. *Equidistant team collaboration*, i.e., three authors with each pair having co-authored a paper together with the same proximity;

https://doi.org/10.1017/nws.2021.12 Published online by Cambridge University Press

EP. *Equidistant Partner*, i.e., two authors, that have co-authored a paper at a local, national or international proximity, have both co-authored a paper with the same partner at the same proximity, which is equal or larger than their own proximity; and

OEP. *Organizational Equidistant Partner*, i.e., two authors, that have co-authored a paper at an organizational proximity, have both co-authored a paper with the same partner at the same proximity.

Note that ETC is entirely covered by EP and OEP, but that EP and OEP cover more motif configurations than ETC. For both the EP and OEP subcategories, we define two more subcategories: "cause" (EPC, OEPC), where the link between the two authors comes before the formation of the equidistant partnership in the temporal configuration and is likely the cause of the equidistant partner, and "effect" (EPE, OEPE), where the link between the two authors comes after the formation of the equidistant partnership and therefore likely follows from having the equidistant partner. An overview of all of these categories, with an example and short description, is given in Table 2a. Their mapping onto the full set of configurations is shown in Figure 10a.

6.3 International categories

Because we are interested in the relation between international collaboration and international mobility, the second categorization scheme we define deals with these concepts.

We define two main categories:

I. *International collaboration*, i.e., motifs with at least one edge indicating an international co-authorship; and

IM. *International mobility*, i.e., motifs where a mobility event is implied by the transition of an international collaboration to an organizational, local, or national collaboration, or vice versa.

From the perspective of individual countries, it is especially interesting whether the international mobility is incoming or outgoing. Therefore, we define the subcategories *international mobility incoming* (IMI) and *international mobility outgoing* (IMO). Indicating, respectively, whether we move from an international collaboration to a closer collaboration or move from a closer collaboration to an international one. Note that the direction of a mobility event cannot be determined when it is implied only by concurrent edges. Furthermore, for triangle motifs a mobility event can only be implied by a contradiction that occurs between all three edges and it can be associated with any of the authors. For example, if we have edges $(a,c,t_1,\mathrm{O}),(b,c,t_2,\mathrm{O}),(a,b,t_3,\mathrm{I})$ for authors a, b, c and assume a single affiliated organization per author at a time, then the first two edges indicate that all authors are associated with the same organization while the third edge indicates that authors a and b are associated with organizations in different countries. This contradiction indicates that an international mobility event has occurred. However, this mobility event can be associated with every author, including author c where author c first has the same affiliation as a and then moves to the same organization as b. In this case, the organization associated with either author a or b must be located in a different country, yet we can never be sure which. Therefore, we can never determine a direction for the mobility events implied by triangle motifs. Moreover, authors can have multiple affiliations, further complicating the inference of mobility, as we discuss in Section 6.4

An overview of the international categories, with an example and short description, is given in Table 2b. Their mapping onto the full set of motifs and layer configurations is shown in Figure 10b. Here, the "non-international" category indicates motif configurations that do not involve any international co-authorship links.

https://doi.org/10.1017/nws.2021.12 Published online by Cambridge University Press

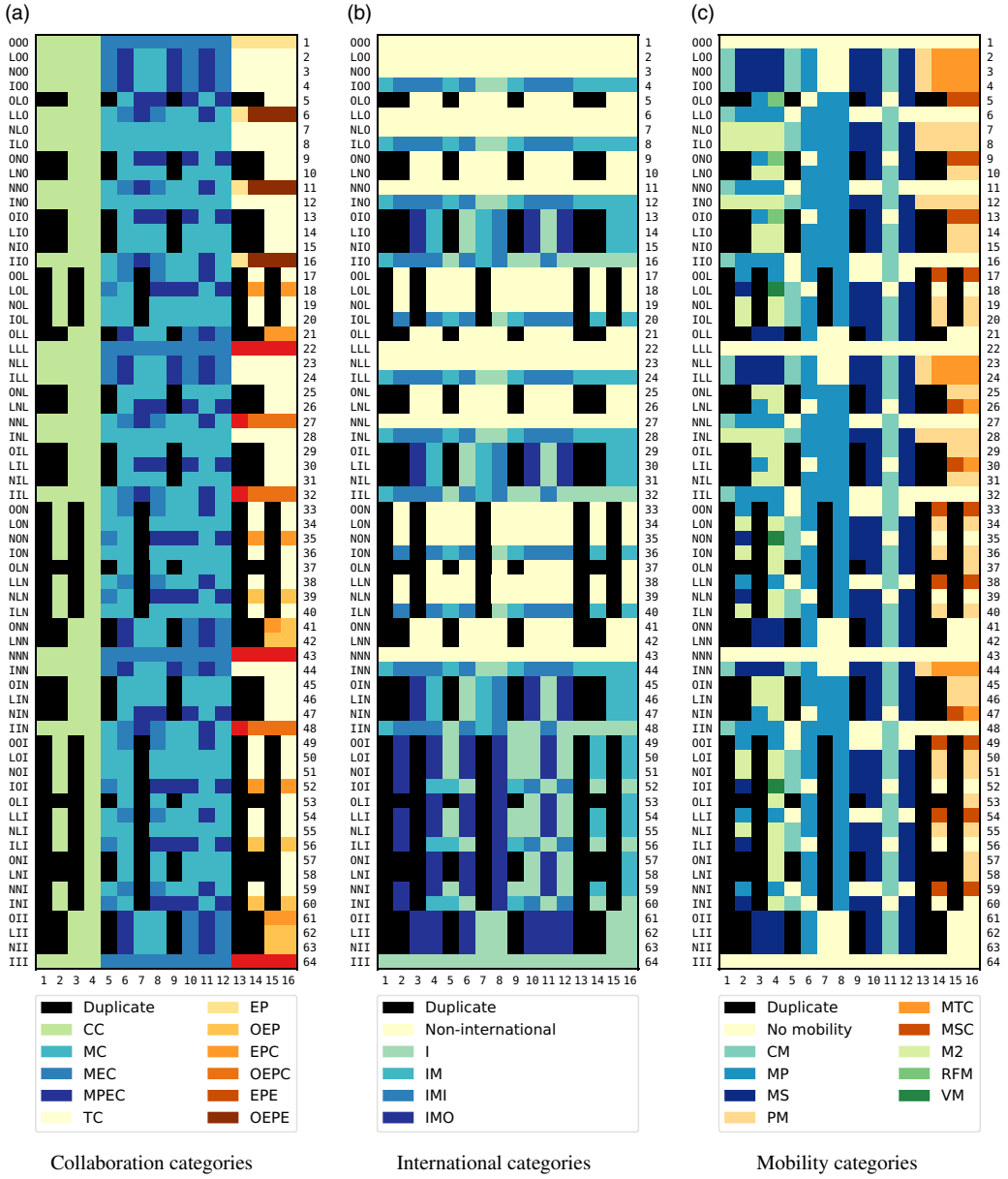

Figure 10. Mappings of the motif categories listed in Table 2 onto the full set of motifs. The "duplicate" categories indicate layer configurations that are equivalent to layer configurations listed above them. For configurations where categories overlap, subcategories take precedence. The full hierarchy of the categories is shown in Figure 11.

6.4 Mobility categories

The third and final set of motif categories we define are mobility categories. The mobility categories either describe a certain type of mobility or describe the context of the edges surrounding the mobility event. We define two main categories:

M. *Mobility*, i.e., a mobility event is implied by a contradiction in organizational proximity between co-authorship edges; and

https://doi.org/10.1017/nws.2021.12 Published online by Cambridge University Press

M2. *Duo-mobility*, i.e., two mobility events are implied.

In Section 6.3, we reasoned that we can never determine the direction of mobility events implied by triangle motifs. Moreover, for triangle motifs we cannot be sure a contradiction of collaboration distances even implies a mobility event or if it is an indicator that an author is affiliated with multiple organizations. For example, given the same set of edges as before, $(a,c,t_1,\text{O}),(b,c,t_2,\text{O}),(a,b,t_3,\text{I})$, we required the assumption of a single affiliation at a time to imply a mobility event. If we assume multiple affiliation are possible for an author, then the author organization affiliations $a \to \{A\}$, $b \to \{B\}$, and $c \to \{A, B\}$ would fit this motif configuration without implying any mobility event. As we cannot be sure that any mobility event implied by triangle motifs is not caused by an author being affiliated with multiple organizations, we divide category M into two subcategories:

CM. *Certain mobility*, i.e., a mobility event implied by an edge or star motif; and
PM. *Possible mobility*, i.e., a mobility event implied by a triangle motif.

Note that we are assuming that authors always list all their current affiliations for each paper. After all, for edge and star motifs, mobility is implied from a change in proximity between two co-authorship edges between the same two authors. Since this proximity is set to the minimum of all listed affiliations, for the proximity to change their list of affiliations must change. So, when the proximity changes, a mobility event must have occurred.

For certain mobility (CM), we know that the mobility event is implied by only two out of the three edges. This means we have either an additional preceding, succeeding, or concurrent edge and define two subcategories accordingly:

MP. *Mobility Preceding*, i.e., a mobility event is implied by a change in proximity between two co-authorship edges which are preceded by a third edge; and
MS. *Mobility Succeeding*, i.e., a mobility event is implied by a change in proximity between two co-authorship edges which are succeeded by a third edge.

Note that the additional edge can still be either a preceding or succeeding edge when it is concurrent with only one of the edges involved in the mobility event, because the mobility event itself will have occurred somewhere in the time between those two edges.

Even though we can never be certain about the direction of mobility for triangle motifs, we can infer some meaning on the possible mobility. We define two such subcategories for the PM category as follows:

MTC. *Mobility To Collaboration*, these motifs may imply collaboration as both a possible cause and effect of an incoming mobility event. For example, motif configuration $(a,b,t_1,\text{L}),(a,c,t_1,\text{O}),(b,c,t_2,\text{O})$ may imply that collaborations between authors a,b and a,c may have inspired author b or c to move to the same organization and start collaborating.
MSC. *Mobility Sustained Collaboration*, these motifs may imply that even after an author has moved further away, the ties to their previous organization may allow them to establish new collaborations through their former colleagues. For example, this may be implied by motif configuration $(a,b,t_1,\text{O}),(a,c,t_1,\text{O}),(b,c,t_2,\text{L})$.

To finish, we define two subcategories of duo-mobility:

RFM. *Return or Follow Mobility*, an author moving away from the same organization as their collaborating partner after which they either return to their old organization or the

https://doi.org/10.1017/nws.2021.12 Published online by Cambridge University Press

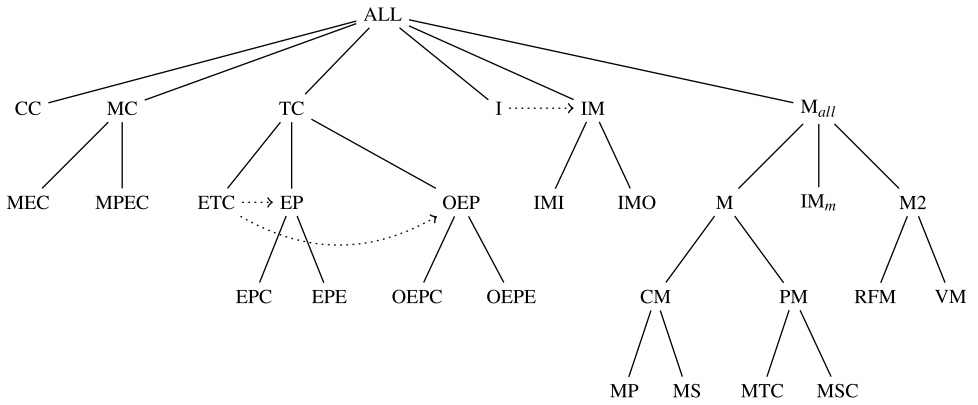

Figure 11. Category hierarchy. Note that motif counts for the subcategories do not have to add up to 100% of their parent category. Dotted lines indicate that categories overlap within their general classification but that no hierarchy is established between them. In Figure 10, the target category of a dotted line takes precedence over the source category.

collaborating partner follows them to the new organization, i.e., the proximity returns to organizational; and

VM. *Visit Mobility*, an author first moving to the same organization as the collaborating partner after which the author either moves back or moves to yet another organization at the same proximity as before, i.e., the proximity first changes to organizational and then returns to its old state.

An overview of the mobility categories, with an example and short description, is given in Table 2b. Their mapping onto the full set of motifs and layer configurations is shown in Figure 10c. Here, the "no mobility" category indicates motif configurations that do not appear to imply any mobility event.

7. Experiments and results

In this section we discuss our experiments and results. First, we discuss our experimental setup in Section 7.1. Then, we compare the five scientific fields by using the categories defined in the previous section to create a profile for each field in Section 7.2. Finally, in Section 7.3 we discuss the limitations of our data, methods and results. In Supplementary Material B, we compare for each field the 50 largest countries, where country size is based on its scientific output. In Supplementary Material C we analyze the performance of our new algorithms.

7.1 Experimental setup

For our experiments, we aim to use the categories defined in Section 6 to identify typical co-authorship behavior in the various scientific fields. Due to the difficulty of developing an underlying null hypothesis that correctly and exhaustively captures the basic mechanics of co-authorship networks (Artzy-Randrup *et al.*, 2004), we do not compare our motif count results with a null model but instead compare them with one another. Similar to the graphlet representativity measure by Charbey & Prieur (2019), we compute the difference between the relative frequency in one network and the relative "global", i.e., average, frequency. We do this not for individual motif configurations but for the motif categories.

The total motif count of a category is computed by summing the motif counts of all motif configurations assigned to that category. The relative importance (ri) of a category i in a given

https://doi.org/10.1017/nws.2021.12 Published online by Cambridge University Press

field j with respect to all fields is determined as:

$$ri_{i,j} = \frac{\frac{c_{i,j}}{c_{p(i),j}} - avg_i}{avg_i}, \tag{1}$$

with $c_{i,j}$ the total motif count of category i in field j, $p(i)$ denoting the parent category of category i, as depicted in Figure 11, and

$$avg_i = \frac{1}{|J|} \sum_{j \in J} \frac{c_{i,j}}{c_{p(i),j}}, \tag{2}$$

with J the set of scientific fields, in our case the five fields described in Section 5.

This means that a positive $ri_{i,j}$ indicates that in field j a relatively large proportion of the motifs of the parent category ($p(i)$) belong to category i and a negative $ri_{i,j}$ indicates a relatively small proportion of the motifs of the parent category belong to category i. Note that, due to the division by the average proportion (avg_i), we look at the difference in proportion relative to the size of the proportion. This means that we find only very small positive or negative ri for a category like MC, which encapsulates approximately 90% of all motifs (see first row of Tables S1–S5 in the Supplementary Material). Furthermore, note that the number of motif configurations assigned to a category does not play a role here and that only the summed motif counts of the motif configurations matters.

For each field, we analyze in detail the relative importance of all categories and their interplay to give insight into typical co-authorship and scientific mobility behavior. As such, we aim to identify what sets each field apart. Within each field, we examine outlier countries that represent unique co-authorship and mobility behavior and investigate commonalities between countries showing the same behavior in Supplementary Material B.2.

We analyze the ri computed for δ = 10 years, i.e., the full timespan of the datasets. A shorter timespan, such as $\delta = 3$ or 5 years, excludes motifs where the causal link between the co-authorships may be weaker due to the passing of time. Because a shorter timespan can impact the ri of a category, we investigate the robustness of ri and our conclusions in Supplementary Material D. We find that ri is robust for the larger datasets and categories and that conclusions drawn for $\delta = 10$ are representative for shorter timespans.

The multilayer temporal motif counting algorithms introduced in Section 4 were implemented as a component of the Stanford Network Analysis Project (SNAP, see (Leskovec & Sosič, 2016) for details). Our implementation can be found at (Boekhout, 2020). In Supplementary Materials C we investigate the empirical performance of our implementation with respect to the number of edges and the density and compare that to the theoretical complexities of the algorithms. We find that there is a linear relationship between the size of a dataset and the runtime of our implementation. Furthermore, we find that between 30 and 50,000 edges are processed per second, converging to around 40,000 edges per second as the number of edges increases. Thus, network datasets with millions of edges can be processed in a matter of minutes.

7.2 Results—Field comparison

Based on the motif counting results from our experiments, for $\delta = 10$ years, we determined the relative importance of each category defined in Section 6 with the help of Equations 1 and 2. The results are shown in Table 3. By combining the relative importances of all categories, we create profiles for each field that can identify typical co-authorship and mobility behavior. Below we take a closer look at the profile of each of the five fields.

https://doi.org/10.1017/nws.2021.12 Published online by Cambridge University Press

Table 3. Field comparison of the relative importance of each category defined in Section 6

Collaboration categories												
field	CC	MC	MEC	MPEC	TC	ETC	EP	EPC	EPE	OEP	OEPC	OEPE
Social sciences & Humanities	0.11	-0.01	-0.07	-0.09	0.13	-0.13	0.99	-0.15	-0.25	-0.22	0.37	0.51
Biomedical & Health sciences	-0.42	-0.00	0.15	-0.37	0.12	0.31	-0.48	0.03	-0.07	0.17	-0.60	-0.60
Physical sciences & Engineering	-0.05	0.01	-0.03	0.19	-0.09	-0.12	0.03	-0.10	0.16	-0.05	0.22	0.15
Life & Earth sciences	-0.06	0.01	-0.07	0.33	-0.08	-0.05	-0.32	0.05	-0.02	0.02	0.02	-0.04
Mathematics & Computer science	0.43	0.00	0.02	-0.06	-0.08	-0.01	-0.22	0.18	0.18	0.08	-0.01	-0.02
International categories												
field	I	IM	IMI	IMO								
Social sciences & Humanities	-0.17	-0.25	-0.04	0.04								
Biomedical & Health sciences	-0.41	-0.48	-0.02	0.01								
Physical sciences & Engineering	0.21	0.29	0.01	-0.03								
Life & Earth sciences	0.18	0.42	0.03	0.01								
Mathematics & Computer science	0.19	0.03	0.01	-0.03								
Mobility categories												
field	M_{all}	IM_m	M	CM	MP	MS	PM	MTC	MSC	M2	RFM	VM
Social sciences & Humanities	-0.04	-0.20	-0.00	-0.01	-0.05	0.01	0.15	-0.06	-0.04	0.36	-0.21	0.02
Biomedical & Health sciences	-0.35	-0.19	0.00	-0.01	-0.01	-0.02	0.17	-0.03	0.07	-0.40	0.12	-0.07
Physical sciences & Engineering	0.15	0.15	0.00	0.00	0.03	-0.01	-0.05	0.14	-0.04	-0.11	0.12	0.07
Life & Earth sciences	0.28	0.13	0.00	0.01	-0.01	0.01	-0.12	0.07	-0.00	-0.08	0.14	0.07
Mathematics & Computer science	-0.05	0.11	-0.00	0.01	0.04	0.01	-0.15	-0.12	0.01	0.23	-0.16	-0.09

Social sciences & Humanities

For the field of Social sciences & Humanities (SSH) the equidistant partner categories EP, OEP and their subcategories clearly stand out (Table 3). In this field, a far greater proportion of teams, i.e., triangles of co-authorships, represents the formation of an equidistant partner with someone outside their own organization (EP +0.99). Note that EP +0.99 means that the proportion of EP motifs among all TC motifs (see Figure 11) is almost twice as high as the average proportion over all fields. Conversely, a much smaller proportion of teams represents the formation of equidistant partners with someone at their own organization (OEP -0.22). Furthermore, the negative *ri* for international categories I and IM, suggest that, although we noticed more equidistant partners outside the own organization, these partners are more likely to work within the same country. Results from Larivière *et al.* (2006) support this, showing a smaller proportion of international collaboration papers among inter-institutional collaboration papers for social sciences compared to natural sciences. We note that these observations can be explained, to some extent, by the lower proportion of organizational edges and higher proportion of national edges for this field, as shown in Table 1. The higher level of national collaboration can be considered an identifying trait for this field. The reason for this increased national collaboration remains an open question.

In Table 1, we see that the proportion of international edges for SSH is only slightly below that of fields with a positive *ri* for category I, while for SSH, we observed a negative *ri* for the international categories I and IM. This means that on average fewer motifs are formed per international edge. There are multiple possible explanations for the lower rate of motif formation. First, authors that collaborate internationally may have larger knowledge networks, i.e., collaborate with a greater number of different co-authors, thereby forming relatively fewer co-authorships per co-author. Second, authors linked to international co-authorships may publish papers that involve relatively fewer (non-international) co-authors per paper than on average. Third and least likely, authors linked to international co-authorships may be less productive than the average author in the field. For internationally active authors, each of these would result in relatively fewer international motifs per international co-authorship edge.

The third observation for SSH is that among mobility motifs a relatively large proportion of motifs consist of duo-mobility (M2 +0.36). In part, this can be attributed to an increased proportion of edge motifs (CC +0.11), the only motifs that can imply duo-mobility. However, for the most part, the relatively large proportion of M2 motifs implies that authors who continue to co-author through the years experience more changes of their proximity. In fact, because we see a neutral *ri* for VM, we can surmise that the additional duo-mobility also leads to a similar increase in visit mobility but does not result in additional return or follow mobility (RFM -0.21). As such, a relatively larger proportion of the additional M2 motifs have no apparent connection to the continued collaboration of the scholars.

The final observation we make for SSH is that among mobility motifs international mobility is underrepresented (IM_m -0.20), reenforcing our observation of a reduced proportion of motifs including international co-authorships.

Biomedical & Health sciences

The field of Biomedical & Health sciences (B&H) has, as can be seen in Table 3, the largest proportion of equidistant team collaborations (ETC +0.31). Additionally we see a relatively high proportion of team collaboration representing organizational equidistant partners, but with a very low proportion of the cause and effect subcategories OEPC and OEPE. Together these categories imply that a large proportion of ETC motifs are in fact three authors connected by only organizational links, because, when this occurs, neither a cause nor effect can be determined for the equidistant partnership and therefore relatively few OEPC and OEPE motifs are formed.

The tendency for team formation within organizations forms an identifying trait for this field where the nature (of parts) of the research often requires larger groups of scientists, i.e., potential

https://doi.org/10.1017/nws.2021.12 Published online by Cambridge University Press

authors, to collaborate locally in, for example, a lab setting. Because larger groups of co-authors also create larger cliques of co-authorships, the phenomenon of large groups of locally collaborating scientists is reflected in the relative importance of many other categories for this field. For example, the strong negative *ri* for the EP category as well as the low proportion of international motifs (I -0.41) reflects the smaller proportion of inter-organization collaboration. Furthermore, the relatively low proportion of CC motifs and the high proportion of ETC motifs reflect the tendency for larger groups of authors, since larger groups form a greater proportion of triangle motifs than edge motifs over the course of several publications compared to smaller groups of authors. Note that the negative *ri* for the M2 category is directly linked to the negative *ri* of CC and does not provide a "new" observation.

As more and larger teams are formed within organizations, relatively fewer motifs are formed that imply a mobility event. B&H has by far the lowest *ri* of any field for the M_{all} category. In part this is yet another result of the larger and more teams at the organization level, but it might also suggest that scientists are less prone to move between organizations in this field. Additionally, we observe that among the mobility motifs a relatively small proportion of motifs represent international mobility (IM_m -0.20).

Physical sciences & Engineering

With respect to the collaboration categories, Table 3 shows that the field of Physical sciences & Engineering (P&E) appears to be associated with mostly intermediate *ri* values. Where other fields have strong positive or negative relative importances for a category, P&E is neutral. For example, where SSH has a very large proportion of EP motifs and the other three fields have very small proportions of EP motifs, P&E is around average (EP +0.03). However, unlike SSH, the percentage of organizational edges is roughly the same for P&E and the remaining three fields, B&H, L&E, and M&C. If we compare P&E only with these fields, we see a similar pattern emerge for the ETC, EP, and OEP categories as observed for SSH. This tells us that P&E forms comparably more equidistant partners with co-authors outside their own organization and less with co-authors within their organization. Notably, P&E is the only field where we can clearly observe a difference in the *ri* of the cause and effect subcategories of EP (EPC -0.10, EPE +0.16). This indicates that an inter-organizational co-authorship between two authors more often follows from them already having an equidistant partner than that their co-authorship predates, i.e., causes, the equidistant partnership.

For P&E we observe a positive *ri* for international motifs. Herein, it does not differ from the fields L&E and M&C, but shows an equal *ri* at around the same percentage of international links. At most we can state that P&E forms a greater number of motifs including international co-authorships than SSH per international edge.

Finally, we see that P&E has a relatively large proportion of mobility motifs (M_{all} +0.15). Only L&E shows a higher *ri* for mobility motifs. The proportion of international mobility (IM_m) follows a similar trend between the fields as observed for category I. We note a relatively high proportion of MTC mobility motifs in this field, possibly indicating an increased likelihood of incoming mobility having a direct cause or effect in the knowledge network of the authors. However, as we cannot be sure that triangle motifs even imply mobility, we cannot definitively conclude this.

Life & Earth sciences

Life & Earth sciences (L&H) shows, like B&H, a reduced proportion of equidistant partner motifs (EP -0.32) in Table 3. Unlike B&H, this is not associated with a greater proportion of ETC and OEP motifs. In other words, there is a relatively larger proportion of triangles of co-authorship that include edges with three different proximities. The occurrence of such motifs requires authors involved in them to either be mobile or be associated with multiple organizations, otherwise an

https://doi.org/10.1017/nws.2021.12 Published online by Cambridge University Press

equidistant partnership would be formed. For L&H we observe a positive *ri* for mobility motifs (M_{all} +0.28), indicating that mobility is likely the cause of the increased number of triangles with three different proximities.

Like for P&E, we see a positive *ri* for international motifs for L&H. More importantly, we see that the increased mobility and internationalism also translates to a larger proportion of international mobility motifs (IM +0.42, IM_m +0.13).

Mathematics & Computer science

It can be observed from Table 3 that the field of Mathematics & Computer science (M&C) has by far the greatest proportion of continued collaboration motifs (CC +0.43), i.e., edge motifs, which comes at the cost of team collaborations (TC -0.08). Furthermore, among the team collaborations we see more organizational equidistant partnerships than outside the organization (EP -0.22, OEP +0.08) and observe a greater likelihood for a clear cause or effect of EP motifs (EPC +0.18, EPE +0.18). Together this indicates that, in M&C, there is a greater trend to continue to co-author within the established knowledge network and organization and to expand the knowledge network through the sharing of contacts with scholars at the same organization rather than outside the organization. A possible cause, or symptom, of this behavior is the lower proportion of mobility motifs (M_{all} -0.05). Less mobility may cause authors to continue to collaborate with the same co-authors or as authors remain more set within their known knowledge network they may see less cause to become mobile. Note that the large proportion of duo-mobility motifs (M2) can be directly explained by the large proportion of CC motifs.

Despite the tendency to continue to co-author within the known knowledge network, we observe the same positive *ri* for international motifs as observed for P&E and L&E. Additionally, among the mobility motifs we also observe a positive *ri* for international mobility (IM_m +0.11).

The relationship between (international) mobility and collaboration

In Table 3 we have seen the same trend for all fields. A larger proportion of international motifs translating to a larger proportion of international mobility motifs among all mobility motifs (P&E, L&E, and M&C). At the same time, a smaller proportion of international motifs leads to a smaller proportion of international mobility motifs (SSH and B&H). This trend forms a good indicator of the existence of a relationship between international co-authorship, i.e., international collaboration, and international mobility, but it does not imply a direction for this relationship.

Between categories IMI–IMO, MP–MS, and MTC–MSC, the categories that may imply some causation between collaboration and mobility, we see only minor variations in the relative importance. Table 4 shows that neither MP nor MS is more dominant, indicating that collaboration occurs before and after mobility to an equal degree. Additionally, Table 4 shows that a greater proportion of motifs that imply international mobility suggest outgoing international mobility (IMO). It also shows a greater proportion of MSC, mobility sustained collaboration, motifs compared to MTC. This means that more motifs are formed by authors sustaining their old knowledge network after moving abroad than motifs are formed by authors moving closer to scholars they have previously co-authored with.

Although one might interpret this as evidence of a relationship between international mobility and collaboration in one direction, namely that international mobility leads to international collaboration, as suggested by Kato & Ando (2017), it actually shows that the relationship is bidirectional. After all, a single international co-authorship preceding an incoming international mobility event may be sufficient to establish a causation between the collaboration and the mobility, whereas proof of maintaining the old knowledge network after an international mobility event requires international co-authorships with multiple co-authors from before the international mobility event. In other words, one international mobility event is likely to form international

https://doi.org/10.1017/nws.2021.12 Published online by Cambridge University Press

Table 4. The proportion of a subset of categories w.r.t. their parent category, for each field

Field	IMI	IMO	MP	MS	MTC	MSC
Social sciences & Humanities	0.32	0.48	0.36	0.41	0.24	0.56
Biomedical & Health sciences	0.32	0.46	0.38	0.40	0.25	0.63
Physical sciences & Engineering	0.33	0.44	0.39	0.40	0.29	0.56
Life & Earth sciences	0.34	0.46	0.37	0.41	0.27	0.58
Mathematics & Computer science	0.33	0.45	0.39	0.41	0.23	0.59

collaborations with more previously organizational co-authors (O→I), than it is to form organizational collaborations with previously international co-authors (I→O). Therefore, a greater proportion of IMO motifs over IMI motifs is to be expected. Thus, we conclude that the relationship between international mobility and collaboration appears to exist in both directions.

7.3 Limitations

Here we discuss limitations of the datasets and methodology that may affect the interpretation of our results.

First and foremost, we must acknowledge missing data. As previously mentioned in Section 5, authorships for which no affiliation information was present in WoS were excluded. In Table 1, we showed that this makes up around 20% of all authorships, which means our co-authorship networks are formed from only 80% of all authorships in WoS. Furthermore, WoS itself is not complete. For example, we know that conference papers play a big role in information diffusion in Computer science, but that conference papers are not included in the CWTS in-house version of WoS (which in particular does not include the conference proceedings citation index). Additionally, we know that some countries, such as Brazil, have their own internal publication system that is not included in WoS. The inclusion of this missing data could significantly alter the relative importances observed for the affected fields (and countries). However, our datasets still cover a significant number of papers and co-authorships in every field and we do not expect the missing data to substantially alter our main conclusions.

Second, because we infer mobility events from a change in co-authorship proximity, we may not be able to detect mobility which has no cause in previous co-authorships. As such, we may be underestimating the level of mobility in some fields (or countries). Because we have no way to speculate about the amount of undetected mobility for specific fields (or countries), we draw our conclusions based only on the mobility we are able to detect.

Third, a single mobility event of an author with a much larger than average knowledge network creates a greater potential number of pairwise mobility events. Therefore, observations of motifs demonstrating certain types of scientific mobility may be dominated by the mobility of these authors. We note that papers with many authors may facilitate these larger than average knowledge networks and thus may play a bigger role in the creation of observed tendencies in fields toward specific types of scientific mobility. However, since we excluded publications with more than 25 authors, we expect this effect to be marginal at the scope of global fields.

Fourth, categories that describe some causation, i.e., EPC, EPE, OEPC, OEPE, MTC, and MSC, ascribe a connection between the co-authorships within the motifs that may not exist. For example, an EPC motif may imply that two scholars that co-authored locally formed an equidistant partner nationally because of their earlier co-authorship, i.e., one of the scholars introduced the other to the equidistant partner, but they may very well both have been introduced to this equidistant partner directly or through a third party. In fact, the greater the proximity between the scholars, the less certain we can be of the causation the category defines for individual motifs.

https://doi.org/10.1017/nws.2021.12 Published online by Cambridge University Press

However, over an entire co-authorship network an increased proportion of motifs of one of these categories over their counterpart, i.e., EPC over EPE, does imply a greater likelihood of more of such causations occurring. Furthermore, we do not draw conclusions based on just one of these categories, but only based on their relation to other categories.

Fifth and last, in our conclusions we connect categories and interpret co-authorship or mobility behavior based purely on their relative importance, i.e., based on their motif counts. Because we do not study the individual motifs at the level of their nodes and edges, we cannot guarantee that categories that we connect utilize overlapping sets of nodes and edges. As such, the connections we have made throughout Section 7.2 and Supplementary Material B.2 may, although logically sound, not represent a real-world connection.

8. Conclusion and Future Work

In this paper we attempted to better understand scientific collaboration, scientific mobility, and how those relate. To this end, we first extended multilayer temporal motif counting algorithms from previous work to be able to count motifs that include concurrent edges. Second, we modified these algorithms to enforce edge attribute exclusivity, so that in each counted motif every edge has a unique attribute value. Theoretically, the extensions to the algorithms added only a small constant factor to the time complexity of the original algorithms, which had time complexities of, respectively, $O(m\lambda^2)$ and $O(m\sqrt{\tau}\lambda^2)$, where m is the number of links, λ the number of layers and τ the number of static triangles. Using experiments on large-scale co-authorship datasets extracted from Web of Science (WoS), we showed that the extended algorithms have execution runtimes linear with respect to the size of the datasets, processing between thirty and fifty thousand edges per second, meaning we can process network datasets with millions of edges in a matter of minutes.

For our experiments, we extracted five large global co-authorship datasets from WoS, each covering one scientific field in the period 2007–2016. Using our algorithms, motif counts were computed for each of those datasets. Next, we introduced a new systematic categorization that assigns meaning to the motifs in the domain of co-authorship and scientific mobility. By determining the relative importance of each of the categories in specific fields (or countries) based on the computed motif counts, we were able to infer characteristic co-authorship and mobility behavior. The inferred characteristic co-authorship and mobility behavior for the various fields include:

- For Social sciences & Humanities (SSH), we found that authors co-author to a greater level with scholars outside their own organization than in other fields. Additionally, they establish more equidistant partners with scholars outside their own organization and continue to collaborate throughout multiple mobility events to a greater degree. Fewer motifs are formed per international co-authorship, indicating that internationally active authors in SSH display different co-authorship behavior than in other fields.
- For Biomedical & Health sciences (B&H), we found that our results reflected the nature of the type of research conducted in this field, which often lends itself more to large team collaborations within an organization resulting in fewer inter-organizational collaborations.
- For Physical sciences & Engineering (P&E), we found that authors in this field form comparably more equidistant partners with co-authors outside their own organization and less with co-authors within their organization. Notably, we found that in P&E, an inter-organizational co-authorship between two authors more often follows from them already having an equidistant partner than that their co-authorship predates, i.e., causes, the equidistant partnership.
- For Life & Earth sciences (L&E), we found that relatively more mobility has lead to entirely inter-organizational team formations (triangle co-authorships). Additionally, we found that

https://doi.org/10.1017/nws.2021.12 Published online by Cambridge University Press

increased mobility and internationalism also translates to relatively more international mobility in L&E.
- For Mathematics & Computer science (M&C), we found that there is a greater trend to continue to co-author within the established knowledge network and organization and to expand the knowledge network through the sharing of contacts with people at the same organization rather than outside the organization. Although this is associated with relatively less overall mobility, we still observe a relatively high proportion of international mobility for M&C.

The inferred characteristic co-authorship and mobility behavior found for specific countries in specific fields are summarized in Supplementary Material B.3. In short, throughout all fields, we found that countries with increased team formation within organizations display relatively less international collaboration and (international) mobility. Conversely, countries that display an increased amount of inter-organizational team formation show relatively more international collaboration and (international) mobility.

Finally, we weighed in on the discussion in literature on the relationship between international collaboration and international mobility. We found evidence that supports the existence of this relationship in both directions, from collaboration to mobility and from mobility to collaboration.

In future work we would like to consider motifs larger than three nodes and three edges. While these larger motifs can not be counted as efficiently (Boekhout *et al.*, 2019), we could study them on smaller networks where they may give us further insight into more complex co-authorship and mobility behavior. Additionally, larger motifs could provide a greater insight into the evolution of knowledge networks. We also want to apply these algorithms to different types of networks to show the versatility of a multilayer temporal motif counting approach to gain insight into complex networks, as we have shown for co-authorship networks in this work. We hope that the proposed approach and framework will pave the way for a new line of research for understanding higher-order patterns in the dynamics of scientific collaboration networks.

Availability of data and materials

The code implementing the algorithms introduced and used in this work, the datasets of the five extracted global co-authorship networks used in this work, as well as the script used to preprocess these networks for use with the motif counting code are openly available online at `https://bitbucket.org/Fractals-/count_mult_temp_motifs` (Boekhout, 2020).

Acknowledgments. We are grateful for the feedback and suggestions made by the anonymous reviewers.

Funding. This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

Conflicts of interest. None.

Supplementary. For supplementary material for this article, please visit http://dx.doi.org/10.1017/nws.2021.12.

References

Aman, V. (2018). Does the Scopus author ID suffice to track scientific international mobility? A case study based on Leibniz laureates. *Scientometrics*, *117*(2), 705–720.

Appelt, S., van Beuzekom, B., Galindo-Rueda, F., & de Pinho, R. (2015). Which factors influence the international mobility of research scientists? In *Global mobility of research scientists* (pp. 177–213). Elsevier.

Artzy-Randrup, Y., Fleishman, S. J., Ben-Tal, N., & Stone, L. (2004). Comment on "Network motifs: Simple building blocks of complex networks" and "Superfamilies of evolved and designed networks". *Science*, *305*(5687), 1107c.

Barabási, A.-L. (2016). *Network science*. Cambridge University Press.

Barabási, A.-L., Jeong, H., Néda, Z., Ravasz, E., Schubert, A., & Vicsek, T. (2002). Evolution of the social network of scientific collaborations. *Physica A: Statistical Mechanics and its Applications*, *311*(3–4), 590–614.

Baruffaldi, S. H., & Landoni, P. (2010). Effects and determinants of the scientific international mobility: the cases of foreign researchers in Italy and Portugal. In *Paper for the Triple Helix VIII conference*.

Benson, A. R., Gleich, D. F., & Leskovec, J. (2016). Higher-order organization of complex networks. *Science*, *353*(6295), 163–166.

Birnholtz, J. P. (2006). What does it mean to be an author? The intersection of credit, contribution, and collaboration in science. *Journal of the American Society for Information Science and Technology*, *57*(13), 1758–1770.

Boekhout, H. D. (2020). *Counting multilayer temporal motifs*. Retrieved from https://bitbucket.org/Fractals-/count_mult_temp_motifs, June 4, 2020.

Boekhout, H. D., Kosters, W. A., & Takes, F. W. (2019). Efficiently counting complex multilayer temporal motifs in large-scale networks. *Computational Social Networks*, *6*(1), 1–34.

Bordons, M., Aparicio, J., González-Albo, B., & Díaz-Faes, A. A. (2015). The relationship between the research performance of scientists and their position in co-authorship networks in three fields. *Journal of Informetrics*, *9*(1), 135–144.

Caron, E., & van Eck, N. Jan. (2014). Large scale author name disambiguation using rule-based scoring and clustering. In *Proceedings of the 19th international conference on science and technology indicators* (pp. 79–86). CWTS-Leiden University.

Chakrabarty, B., & Parekh, N. (2016). NAPS: Network analysis of protein structures. *Nucleic Acids Research*, *44*(W1), W375–W382.

Charbey, R., & Prieur, C. (2019). Stars, holes, or paths across your facebook friends: A graphlet-based characterization of many networks. *Network Science*, *7*(4), 476–497.

Chinchilla-Rodríguez, Z., Miao, L., Murray, D., Robinson-García, N., Costas, R., & Sugimoto, C. R. (2017). Networks of international collaboration and mobility: A comparative study. In *Proceedings of the 16th international conference on scientometrics & infometrics*. ISSI.

Chinchilla-Rodríguez, Z., Miao, L., Murray, D., Robinson-García, N., Costas, R., & Sugimoto, C. R. (2018). A global comparison of scientific mobility and collaboration according to national scientific capacities. *Frontiers in Research Metrics and Analytics*, *3*, 17.

Choobdar, S., Ribeiro, P., Bugla, S., & Silva, F. (2012). Comparison of co-authorship networks across scientific fields using motifs. In *Proceedings of the international conference on advances in social networks analysis and mining (ASONAM)* (pp. 147–152). IEEE Computer Society.

Czaika, M., & Orazbayev, S. (2018). The globalisation of scientific mobility, 1970–2014. *Applied Geography*, *96*, 1–10.

Das, K., Samanta, S., & Pal, M. (2018). Study on centrality measures in social networks: a survey. *Social Network Analysis and Mining*, *8*(1), 13.

Gaillard, J., & Gaillard, A. M. (1997). Introduction: the international mobility of brains: exodus or circulation? *Science, Technology and Society*, *2*(2), 195–228.

Glänzel, W., & Schubert, A. (2005). Domesticity and internationality in co-authorship, references and citations. *Scientometrics*, *65*(3), 323–342.

Guth, J., & Gill, B. (2008). Motivations in East–West doctoral mobility: Revisiting the question of brain drain. *Journal of Ethnic and Migration Studies*, *34*(5), 825–841.

Holme, P., & Saramäki, J. (2019). *Temporal network theory*. Springer.

Hu, X., Li, O. Z., & Pei, S. (2019). Of stars and galaxies – Co-authorship network and research. *China Journal of Accounting Research*.

Jazayeri, A., & Yang, C. C. (2020). Motif discovery algorithms in static and temporal networks: A survey. *Journal of Complex Networks*, *8*(4), cnaa031.

Kato, M., & Ando, A. (2017). National ties of international scientific collaboration and researcher mobility found in Nature and Science. *Scientometrics*, *110*(2), 673–694.

Krumov, L., Fretter, C., Müller-Hannemann, M., Weihe, K., & Hütt, M-T. (2011). Motifs in co-authorship networks and their relation to the impact of scientific publications. *The European Physical Journal B*, *84*(4), 535–540.

Kumar, S. (2015). Co-authorship networks: a review of the literature. *Aslib Journal of Information Management*, *67*(1), 55–73.

Larivière, V., Gingras, Y., & Archambault, É. (2006). Canadian collaboration networks: A comparative analysis of the natural sciences, social sciences and the humanities. *Scientometrics*, *68*(3), 519–533.

Laudel, G. (2003). Studying the brain drain: Can bibliometric methods help? *Scientometrics*, *57*(2), 215–237.

Leskovec, J., & Sosič, R. (2016). SNAP: A general-purpose network analysis and graph-mining library. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *8*(1), 1.

Leyman, A. (2009). Home sweet home? International mobility among Flemish doctoral researchers. In *Higher education, partnership, innovation* (pp. 67–74). Budapest: IHEPI.

Mali, F., Kronegger, L., Doreian, P., & Ferligoj, A. (2012). Dynamic scientific co-authorship networks. In: *Models of science dynamics* (pp. 195–232). Springer.

Melin, G., & Persson, O. (1996). Studying research collaboration using co-authorships. *Scientometrics*, *36*(3), 363–377.

Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., & Alon, U. (2002). Network motifs: Simple building blocks of complex networks. *Science*, *298*(5594), 824–827.

Mingers, J., & Leydesdorff, L. (2015). A review of theory and practice in scientometrics. *European Journal of Operational Research*, *246*(1), 1–19.

https://doi.org/10.1017/nws.2021.12 Published online by Cambridge University Press

Moed, H. F., Aisati, M., & Plume, A. (2013). Studying scientific migration in Scopus. *Scientometrics*, *94*(3), 929–942.
Molontay, R., & Nagy, M. (2019). Two decades of network science: as seen through the co-authorship network of network scientists. In *Proceedings of the 2019 IEEE/ACM international conference on advances in social networks analysis and mining* (pp. 578–583).
Newman, M. E. J. (2004). Coauthorship networks and patterns of scientific collaboration. *Proceedings of the National Academy of Sciences*, *101*(1), 5200–5205.
Paranjape, A., Benson, A. R., & Leskovec, J. (2017). Motifs in temporal networks. In *Proceedings of the tenth ACM international conference on web search and data mining (WSDM)* (pp. 601–610). ACM.
Paul-Hus, A., Mongeon, P., Sainte-Marie, M., & Larivière, V. (2017). The sum of it all: Revealing collaboration patterns by combining authorship and acknowledgements. *Journal of Informetrics*, *11*(1), 80–87.
Ribeiro, P., Paredes, P., Silva, M.E. P., Aparicio, D., & Silva, F. (2021). A survey on subgraph counting: Concepts, algorithms, and applications to network motifs and graphlets. *ACM Computing Surveys (CSUR)*, *54*(2), 1–36.
Stark, O., Helmenstein, C., & Prskawetz, A. (1997). A brain gain with a brain drain. *Economics Letters*, *55*(2), 227–234.
Takes, F. W., Kosters, W. A., Witte, B., & Heemskerk, E. M. (2018). Multiplex network motifs as building blocks of corporate networks. *Applied Network Science*, *3*(1), 39.
Wagner, C. S., & Leydesdorff, L. (2005). Network structure, self-organization, and the growth of international collaboration in science. *Research Policy*, *34*(10), 1608–1618.
Waltman, L., Calero-Medina, C., Kosten, J., Noyons, Ed. C. M., Tijssen, R. J. W., van Eck, N. J., . . . Wouters, P. (2012). The Leiden Ranking 2011/2012: Data collection, indicators, and interpretation. *Journal of the American society for Information Science and Technology*, *63*(12), 2419–2432.

Cite this article: Boekhout H. D., Traag V. A. and Takes F. W. (2021). Investigating scientific mobility in co-authorship networks using multilayer temporal motifs. *Network Science* **9**, 354–386. https://doi.org/10.1017/nws.2021.12