

ORIGINAL ARTICLE

# Measuring directed triadic closure with closure coefficients

Hao Yin<sup>1</sup> , Austin R. Benson<sup>2</sup> and Johan Ugander<sup>3\*</sup>

<sup>1</sup>Institute for Computation and Mathematical Engineering, Stanford University, Stanford, CA, USA (e-mail: [yinh@stanford.edu](mailto:yinh@stanford.edu)), <sup>2</sup>Department of Computer Science, Cornell University, Ithaca, NY, USA (e-mail: [arb@cs.cornell.edu](mailto:arb@cs.cornell.edu)) and <sup>3</sup>Department of Management Science and Engineering, Stanford University, Stanford, CA, USA  
\*Corresponding author. Email: [jugander@stanford.edu](mailto:jugander@stanford.edu)

Action Editor: Ulrik Brandes

## Abstract

Recent work studying triadic closure in undirected graphs has drawn attention to the distinction between measures that focus on the “center” node of a wedge (i.e., length-2 path) versus measures that focus on the “initiator,” a distinction with considerable consequences. Existing measures in directed graphs, meanwhile, have all been center-focused. In this work, we propose a family of eight *directed closure coefficients* that measure the frequency of triadic closure in directed graphs from the perspective of the node initiating closure. The eight coefficients correspond to different labeled wedges, where the initiator and center nodes are labeled, and we observe dramatic empirical variation in these coefficients on real-world networks, even in cases when the induced directed triangles are isomorphic. To understand this phenomenon, we examine the theoretical behavior of our closure coefficients under a directed configuration model. Our analysis illustrates an underlying connection between the closure coefficients and moments of the joint in- and out-degree distributions of the network, offering an explanation of the observed asymmetries. We also use our directed closure coefficients as predictors in two machine learning tasks. We find interpretable models with AUC scores above 0.92 in class-balanced binary prediction, substantially outperforming models that use traditional center-focused measures.

**Keywords:** directed networks; triadic closure; closure coefficients; configuration model

## 1. Introduction

A fundamental property of networks across domains is the increased probability of edges existing between nodes that share a common neighbor, a phenomenon known as triadic closure (Simmel, 1908; Rapoport, 1953; Watts & Strogatz, 1998). This concept underpins various ideas in the study of networks—especially in undirected network models with symmetric relationships—including the development of generative models (Leskovec et al., 2005; Jackson & Rogers, 2007; Seshadhri et al., 2012; Robles et al., 2016), community detection methods (Fortunato, 2010; Gleich & Seshadhri, 2012), and feature extraction for network-based machine learning tasks (Henderson et al., 2012; LaFond et al., 2014).

A standard measure for the frequency of triadic closure on undirected networks is the *clustering coefficient* (Watts & Strogatz, 1998; Barrat & Weigt, 2000; Newman et al., 2001). At the node level, the *local clustering coefficient* of a node  $u$  is defined as the fraction of wedges (i.e., length-2 paths) with center  $u$  that are *closed*, meaning that there is an edge connecting the two ends of the wedge, inducing a triangle. At the network level, the *average clustering coefficient* is the mean of the local clustering coefficients (Watts & Strogatz, 1998), and the *global clustering coefficient*, also known

as *transitivity* (Barrat & Weigt, 2000; Newman et al., 2001), is the fraction of wedges in the entire network that are closed.

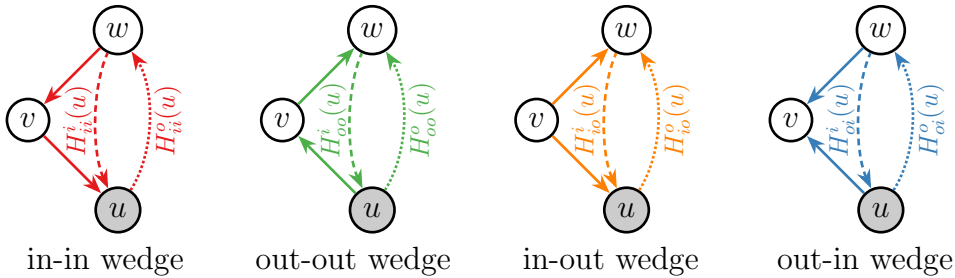
Recent research has pointed out a fundamental gap between how triadic closure is measured by the clustering coefficient and how it is usually explained (Yin et al., 2019). Local triangle formation is usually explained by some transitive property of the relationships that edges represent; for social networks, this is embodied in the idea that “a friend of my friend is my friend.” In these explanations, however, triadic closure is driven not by the center of a length-2 path but rather by an end node (which we refer to as the *head*), who initiates a new connection. In contrast, the local clustering coefficient that measures triadic closure from the center of a wedge implicitly accredits the closure to the center node. The recently proposed *local closure coefficient* closes this definitional gap for undirected graphs by measuring closure with respect to the fraction of length-2 paths starting from a specified head node that are closed (Yin et al., 2019).

These closure coefficients were only defined on undirected networks, but the interactions in many real-world networks are more accurately modeled with an associated orientation or direction. Examples of such networks include food webs, where the direction of edges represents carbon or energy flow from one ecological compartment to another; hyperlink graphs, where edges represent which web pages link to which others; and certain online social networks such as Twitter, where “following” relationships are often not reciprocated (Cheng et al., 2011). The direction of edges may reveal underlying hierarchical structure in a network (Homans, 1950; Davis & Leinhardt, 1971; Ball & Newman, 2013), and we should expect the direction to play a role in local triadic closure.

Extensions of clustering coefficients have been proposed for directed networks (Fagiolo, 2007; Seshadhri et al., 2016), which are center-based at the node level. However, formulating directed triadic measures from the center of a wedge is even less natural in the directed case, while measuring from the head is a more common description of directed closure relationships. For example, in citation networks, paper *A* may cite *B*, which cites *C* and leads *A* to also cite *C*. In this scenario, the initiator of this triadic closure is really paper *A*. Similarly, in directed social networks, outgoing edges may represent differential status (Leskovec et al., 2010; Ball & Newman, 2013), where if person *A* thinks highly of *B* and *B* thinks highly of *C*, then *A* is likely to think highly of person *C* and consequently initiate an outbound link.

In the above examples, measuring triadic closure from *A* would be the analog of the closure coefficient for directed networks, which is what we develop in this paper. More specifically, we propose a family of directed closure coefficients, which are natural generalizations of the closure coefficient for undirected networks. Like the undirected version of closure coefficients, these measures are based on the head node of a length-2 path, in agreement with common mechanistic interpretations of directed triadic closure and fundamentally different from the center-based clustering coefficient. Specifically, the *directed clustering coefficients* proposed by Fagiolo (2007) are not to be confused with the *directed closure coefficients* introduced in this work.

Our measurements are based on the notion of a *directed wedge* as an ordered pair of directed edges that share a common “center” node, and the “non-center” end nodes of this wedge on the first and second edge are called the *head* and *tail* nodes, respectively (in Figure 1, solid lines mark the wedge, where node *u* is the head and node *w* is the tail). Since each edge may be in either direction, there are four directed wedge types. When considering triadic closure for each wedge type, the closing edge between the head and tail nodes may also take either direction. Therefore, at each node, there are eight local directed closure coefficients, each representing the frequency of directed triadic closure with a certain wedge type and closure direction (Figure 1). Analogous to the undirected case, we also define the average and global directed closure coefficients to measure the overall frequency of triadic closure in the entire network. These statistics provide a natural and intuitive way to study the frequency of directed triadic closure in detail, including how directions of the incident and second edge influence a node’s tendency to initiate or receive directed triadic closure.



**Figure 1.** Illustration of four wedge types and eight local directed closure coefficients at node  $u$ . The type of wedge is denoted by two letters, each representing an edge direction ( $i$  for incoming or  $o$  for outgoing). The first letter represents the direction of the edge between the head node  $u$  and the center node  $v$  with respect to  $u$ . The second letter represents the direction of the edge between  $v$  and the tail node  $w$  with respect to  $v$ . A wedge is  $i$ -closed if there is an incoming edge to the head node from the tail, and  $o$ -closed if there is an outgoing edge from the head node to the tail. There are eight local directed closure coefficients at node  $u$ , denoted as  $H_{xy}^z(u)$  with  $x, y, z \in \{i, o\}$ . Each local directed closure coefficient measures the frequency of triadic closure of a certain wedge type (denoted by subscript  $xy$ ) and closing direction (denoted by superscript  $z$ ).

Our empirical evaluation of the directed closure coefficients on real-world networks in Section 3.2 reveals several interesting patterns. At the node level, we find clear evidence of a 2-block correlation structure among the eight local directed closure coefficients, where coefficients within one block are positively (but not perfectly) correlated while coefficients from distinct blocks are nearly uncorrelated. The block separation coincides with the direction of the closing edge in the closure coefficients. We also provide theoretical justification for this observation, gleaned from studying the expected behavior of the closure coefficients for directed configuration model random graphs. Specifically, we show that the expected value (under this model) of each local directed closure coefficient increases with the node degree in the closing edge direction, and thus coefficients with the same closure direction and directed degree are correlated.

From empirical network measurements, we also find surprising asymmetry among average closure coefficients. Consider the in-out wedge in Figure 1, where the coefficients  $H_{io}^i(u)$  and  $H_{io}^o(u)$  correspond to the same directed induced subgraph. For such symmetric wedges, the likelihood for outbound closure can be substantially higher than for inbound, even though the two induced subgraphs are structurally identical. On the other hand, we show in Section 4.1 that networks from the same domain exhibit the same asymmetries.

With extremal analysis, we show in Section 4.1 that there is in fact no positive lower or upper bound on the ratio between types of directed average closure coefficients. Additional probabilistic analysis under the configuration model shows that the expected values of the directed closure coefficients depend on various second-order moments of the joint in- and out-degree distribution of the network. This result partly explains the significant difference in values between a pair of seemingly related average closure coefficients: their expected behaviors correspond to different second-order moments of the degree distribution.

Beyond our intrinsic study on the structure of directed closure coefficients, we illustrate in Section 5 how these coefficients can be powerful features for network-based machine learning. In a lawyer advisory network where every node (lawyer) is labeled with a status level (partner or associate) and directed edges correspond to who talks to whom for profession advice (Lazega, 2001), we show that local directed closure coefficients are much better predictors of status compared to other structural features such as degree or Fagiolo’s directed clustering coefficients. Analysis of the regularization path of the predictive model yields the insight that it is not *how many one advises* but rather *who one advises* that is predictive of partner status. We conduct a similar network classification task in an entirely different domain using a food web from an ecological study. Using the same tools, we find that directed closure coefficients are good predictors of whether or not

a species is a fish. This highlights how our proposed measurements are potentially useful across many domains.

In summary, we propose the directed closure coefficients, a family of eight new metrics for directed triadic closure on directed networks. We provide extensive theoretical analysis which helps explain some counterintuitive empirical observations on real-world networks. Through two case studies, we demonstrate that our proposed measurements are good predictors in network-based machine learning tasks.

## 2. Background and preliminaries

An undirected network (graph)  $G = (V, E)$  is a node set  $V$  and an edge set  $E$ , where an edge  $e \in E$  connects two nodes  $u$  and  $v$ . We use  $d(u)$  to denote the degree of node  $u \in V$ , that is, the number of edges adjacent to  $u$ . A *wedge* is an ordered pair of edges that share exactly one node; the shared node is the *center* of the wedge. A wedge is *closed* if there is an edge connecting the two non-center nodes (i.e., the nodes in the wedge induce a triangle in the graph).

Although the notion of triadic closure in general has a long history (Rapoport, 1953; Wasserman & Faust, 1994), perhaps the most common metric for measuring triadic closure in undirected networks is the average clustering coefficient (Watts & Strogatz, 1998). This metric is the mean of the set of *local clustering coefficients* of the nodes, where the local clustering coefficient of a node  $u$ ,  $C(u)$ , is the fraction of wedges centered at node  $u$  that are closed:

$$C(u) = \frac{2T(u)}{d(u) \cdot (d(u) - 1)},$$

where  $T(u)$  denotes the number of triangles in which node  $u$  participates. The denominator  $d(u) \cdot (d(u) - 1)$  is the number of wedges centered at  $u$ , and the coefficient 2 corresponds to the two wedges (two ordered pairs of neighbors) centered at  $u$  that the triangle closes. If there is no wedge centered at  $u$  (i.e.,  $d(u) \leq 1$ ), the local clustering coefficient is undefined.

Again, to measure the overall triadic closure of the entire network, the *average clustering coefficient* is defined as the mean of the local clustering coefficients of all nodes:

$$\bar{C} = \frac{1}{|V|} \sum_{u \in V} C(u).$$

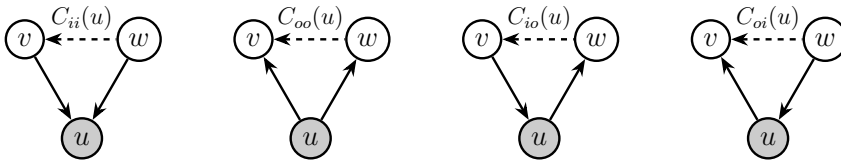
When undefined, the local clustering coefficient is treated as zero in this average (Newman, 2003), although there are other ways to handle these cases (Kaiser, 2008). An alternative network-level version of the clustering coefficient is the *global clustering coefficient*, which is the fraction of closed wedges in the entire network (Barrat & Weigt, 2000; Newman et al., 2001),

$$C = \frac{2 \sum_{u \in V} T(u)}{\sum_{u \in V} d(u) \cdot (d(u) - 1)}.$$

This measure is also sometimes called *transitivity* (Boccaletti et al., 2006).

Recent research has exposed fundamental differences in how triadic closure is interpreted and measured (Yin et al., 2019). For example, social network triadic closure is often explained by the old adage that “a friend of my friend is my friend,” which accredits the creation of the third edge to the end node (also called the *head*) of the wedge. This interpretation, however, is fundamentally at odds with how triadic closure is measured by the clustering coefficient, which is from the perspective of the center node. To close this gap, Yin et al. (2019) proposed the *local closure coefficient* that measures triadic closure from the head node of wedges. Formally, they define this as

$$H(u) = \frac{2T(u)}{\sum_{v \in N(u)} [d(v) - 1]}$$



**Figure 2.** Illustration of the local directed clustering coefficients at node  $u$ , due to Fagiolo (2007). The definition is a direct extension of the local clustering coefficient (Watts & Strogatz, 1998), which measures triadic closure from the center of each wedge.

where  $N(u)$  is the set of neighbors of  $u$ . In this case, the denominator is the number of length-2 paths emanating from node  $u$ . Thus, in social networks, the closure coefficient of a node  $u$  can be interpreted as the fraction of friends of friends of  $u$  that are themselves friends with  $u$ . The closure coefficient has since been investigated under scale-free random graph models (Stegheuis, 2019).

**Extensions to directed networks.** The focus of this paper is on measuring triadic closure in directed networks. The only definitional difference from undirected networks is that the edges are equipped with an orientation, and  $(u, v) \in E$  denotes a directed edge pointing from  $u$  to  $v$ .<sup>1</sup> We assume that  $G$  does not contain multi-edges or self-loops and denote the number of nodes by  $n = |V|$  and the number of edges by  $m = |E|$ .

When an end node  $u$  of an edge is specified, we denote the direction of an edge as  $i$  (for incoming to  $u$ ) or  $o$  (for outgoing from  $u$ ). For any node  $u \in V$ , we use  $d_i(u)$  and  $d_o(u)$  to denote its *in-degree* and *out-degree*, that is, the number of edges incoming to and outgoing from node  $u$ , respectively. For a sequence of joint in- and out-degrees  $[(d_i(u), d_o(u))]_{u \in V}$ , we use  $M_{xy}$ , with  $x, y \in \{i, o\}$  being the direction indicators, to denote the different second-order moments of the degree sequence, that is,

$$M_{xy} = \frac{1}{n} \sum_{u \in V} d_x(u)d_y(u).$$

There are three second-order moments:  $M_{ii}$ ,  $M_{oo}$ , and  $M_{io} = M_{oi}$ .

Fagiolo (2007) proposed a generalization of the clustering coefficient to directed networks. Similar to the undirected case, a *directed wedge* is an *ordered* pair of edges that share a common node, and the common node is called the *center* of this wedge. The wedge is then called *closed* if there is an edge from the opposite end point of the second edge to the opposite end point of the first edge (this constraint, along with the ordering of the two edges, covers the symmetries in the problem). In total, there are four directed clustering coefficients, each defined by the fraction of certain types of wedges that are closed (Figure 2). Seshadhri *et al.* (2016) extended Fagiolo’s definition by explicitly accounting for bidirected (reciprocal) edges, with a focus on network-level (as opposed to node-level) metrics. While we could also explicitly differentiate bidirected links, here we focus on bidirected links as counting toward wedges and closure in both directions.

Directed clustering coefficients have found applications in analyzing functional magnetic resonance imaging (fMRI) data (Liao *et al.*, 2011), financial relationships (Minoiu & Reyes, 2013), and social networks (Ahnert & Fink, 2008). However, as discussed above, existing directed clustering coefficients measure clustering from the center of a wedge, a limited perspective. Our head-based directed closure coefficients, which we define formally in the next section, thus enhance the toolkit for these diverse applications using triadic closure patterns in directed networks.

**Additional related work.** The first research on directed triadic closure is due to Davis & Leinhardt (1971) who studied the relative frequency of each three-node directed subgraph pattern and compared the frequencies with random graph models. Milo *et al.* (2002) later examined significantly

recurring patterns of connected directed subgraphs as “network motifs,” with a particular emphasis on the role of so-called “feed-forward loops” in biology (Mangan et al., 2003). Similar to the case of directed clustering coefficients (Fagiolo, 2007), prior research has studied the ratio of closed wedges at the global (network) level (Onnela et al., 2005; Brzozowski & Romero, 2011; Seshadhri et al., 2016), which is sometimes called “motif intensity” (Onnela et al., 2005). Most generally, we also note that the language of “directed graphlets” (Sarajlić et al., 2016), which can be used to quantify node-level subgraph participation counts within directed networks, provides an expansive characterization of 128 automorphism orbits from which our closure coefficients can be thought of as specific, derived quantities. The key differences in our definitions of directed closure coefficients in the next section are that (i) we measure closure at the node level; and (ii) they are head-node-based metrics which are more agreeable with traditional explanations of directed triadic closure. We will show later that our measures also have considerably different behavior than previous measures.

Directed triadic closure also appears in dynamic network analysis. Lou et al. (2013) proposed a graphical model to predict the formation of a certain type of directed triadic closure: closing an  $oo$ -type wedge with outbound link. This model was later generalized to predict the closure of any type of wedge based on node attributes (Huang et al., 2014). Similarly, the notion of a “closure ratio” has been used to analyze copying phenomena in directed networks (Romero & Kleinberg, 2010). This is also an end-node-based metric that measures a closure of in-in wedges with an incoming edge. Our definitions of directed closure coefficients are different in that they (i) are defined on static networks, (ii) measure diverse types of triadic closure, and (iii) are closely connected to undirected measures of closure and the traditional perspective of triadic closure. Connecting our static measures of directed closure and temporal counterparts is an interesting avenue for future research.

### 3. Directed closure coefficients

In this section, we provide our formal definition of directed closure coefficients and measure them on some representative real-world networks to demonstrate how they provide empirical insights. These insights provide direction and motivation for our theoretical analysis in Section 4. We then show in Section 5 how directed closure coefficients are useful features in machine learning tasks.

#### 3.1 Definitions

With the same motivation as the undirected closure coefficient, we propose to measure directed triadic closure from the end point of a directed wedge. Recall that a directed wedge is an ordered pair of edges that share exactly one common node. The common node is called the center of the wedge, and here we define the *head* of this wedge as the other end of the first edge, and the *tail* as the other end of the second edge. Regardless of the direction of the edges, we denote a wedge by an ordered node triple  $(u, v, w)$ , where  $u$  is the head,  $v$  is the center, and  $w$  is the tail.

Since each edge is directed, there are four types of directed wedges.<sup>2</sup> We denote the type of wedge with two variables, say  $x$  and  $y$ , each taking a value in  $\{i, o\}$  to denote incoming or outgoing. Specifically, a wedge is of *type*  $xy$  (an  $xy$ -wedge) if the first edge is of direction  $x$  to the *head*, and the second edge is of direction  $y$  to the *center* node. Figure 1 shows the four types of directed wedges.

We say that a wedge is *i-closed* if there is an incoming edge from the tail to the head node, and analogously, it is *o-closed* if there is an outgoing edge from head to the tail node. For any  $u \in V$  and  $x, y, z \in \{i, o\}$ , we denote  $W_{xy}(u)$  as the number of wedges of type  $xy$  where node  $u$  is the head, and  $T_{xy}^z(u)$  as the number of  $z$ -closed wedges of type  $xy$  where node  $u$  is the head.

Now we give our formal definition of local directed closure coefficients, which is also illustrated in Figure 1.



**Definition 3.1.** The local directed closure coefficients of node  $u$  are eight scalars, denoted by  $H_{xy}^z(u)$  with  $x, y, z \in \{i, o\}$ , where

$$H_{xy}^z(u) = \frac{T_{xy}^z(u)}{W_{xy}(u)}. \tag{1}$$

If node  $u$  is not the head of any  $xy$ -wedge, then the two corresponding closure coefficients are undefined.

Here, we highlight again the fundamental difference between the local directed closure coefficients we proposed and the local directed clustering coefficients proposed by Fagiolo (2007): the closure coefficients measure triadic closure from the head of wedges, which agrees with natural initiator-driven explanations of triadic closure, while the clustering coefficients measure from the center of wedges. We will show that this small definitional difference yields substantial empirical and theoretical disparity.

Analogous to the undirected clustering coefficient, we also define average and global directed closure coefficients to measure overall directed triadic closure tendencies of a network

**Definition 3.2.** The average directed closure coefficients of a graph are eight scalars, denoted by  $\bar{H}_{xy}^z$  with  $x, y, z \in \{i, o\}$ , each being the mean of the corresponding local directed closure coefficient across the network:

$$\bar{H}_{xy}^z = \frac{1}{n} \sum_{u \in V} H_{xy}^z(u).$$

We treat local closure coefficients that are undefined as taking the value 0 in this average, though most nodes in the datasets we analyze have eight well-defined closure coefficients.

**Definition 3.3.** The global directed closure coefficients of a graph are eight scalars, denoted by  $H_{xy}^z$  with  $x, y, z \in \{i, o\}$ , each being the fraction of closed directed wedges in the entire network:

$$H_{xy}^z = \frac{T_{xy}^z}{W_{xy}} \tag{2}$$

where  $W_{xy} = \sum_{u \in V} W_{xy}(u)$  and  $T_{xy}^z = \sum_{u \in V} T_{xy}^z(u)$  are the total number of  $xy$ -wedges and closed  $xy$ -wedges.

The global directed closure coefficients are equivalent to some global metrics of directed clustering coefficients (Onnela et al., 2005; Seshadhri et al., 2016), since the difference in measuring from head or center does not surface.

### 3.2 Empirical analysis

To obtain intuition and empirical insights before diving into theoretical analysis, we evaluate the directed closure coefficients on 11 networks from 5 different domains:

- (1) Three social networks. SOC-LAWYER (Lazega, 2001): a professional advisory network between lawyers in a law firm; SOC-EPINIONS (Richardson et al., 2003): an online network of who-trusts-whom relationships; and SOC-LIVEJOURNAL (Backstrom et al., 2006): an online social friendship network.
- (2) Two communication networks. MSG-COLLEGE (Panzarasa et al., 2009): an online messaging network between college students; and EMAIL-EU (Yin et al., 2017): an email network between researchers at a European institute.

**Table 1.** Summary statistics of networks.

Network	$n$	$m$	$M_{ii}$	$M_{io}$	$M_{oo}$	$r$	$\Delta_c$	$\Delta_{ac}$
SOC-LAWYER	71	892	227.41	166.15	208.65	0.39	880	5075
SOC-EPINIONS	75.9K	509K	1179.40	526.15	721.82	0.41	740K	3.59M
SOC-LIVEJOURNAL	4.85M	69.0M	2091.52	1220.33	1504.35	0.75	244M	946M
MSG-COLLEGE	1899	20.3K	347.80	391.99	592.42	0.64	11K	40K
EMAIL-EU	1005	25.6K	1428.97	1509.56	1756.77	0.72	132K	433K
CIT-HEPTh	27.8K	353K	1746.72	269.14	416.35	0.00	572	1.49M
CIT-HEPPH	34.5K	422K	790.63	189.62	380.70	0.00	555	1.29M
FW-EVERGLADES	69	916	394.12	136.52	257.16	0.07	538	4781
FW-FLORIDA	128	2106	493.08	201.92	451.62	0.03	357	8688
WEB-GOOGLE	876K	5.11M	1572.90	69.30	77.46	0.31	3.89M	28.2M
WEB-BERKSTAN	685K	7.60M	62430.80	324.71	390.55	0.25	13.8M	131M

Number of nodes  $n$ ; number of edges  $m$ ; second-order moments of the degree sequence  $M_{ii}$ ,  $M_{io}$ , and  $M_{oo}$ ; fraction  $r$  of edges that are reciprocal (i.e., reciprocity); and number of cyclic and acyclic triangles ( $\Delta_c$  and  $\Delta_{ac}$ ).

- (3) Two citation networks. CIT-HEPTh and CIT-HEPPH (Gehrke et al., 2003): constructed from arXiv submission in two categories.
- (4) Two food webs. FW-FLORIDA and FW-EVERGLADES (Ulanowicz & DeAngelis, 2005): carbon exchange relationships collected from the Florida Bay and the Everglades.
- (5) Two web graphs. WEB-GOOGLE and WEB-BERKSTAN (Leskovec et al., 2009): hyperlink networks from a Google competition as well as a crawl of [berkeley.edu](http://berkeley.edu) and [stanford.edu](http://stanford.edu) domains.

Table 1 lists some basic statistics of the networks. We emphasize that the reciprocity of these networks vary substantially. For example, the citation networks and food webs contain mostly unidirectional edges, and the communication networks have many bidirected (reciprocal) edges.

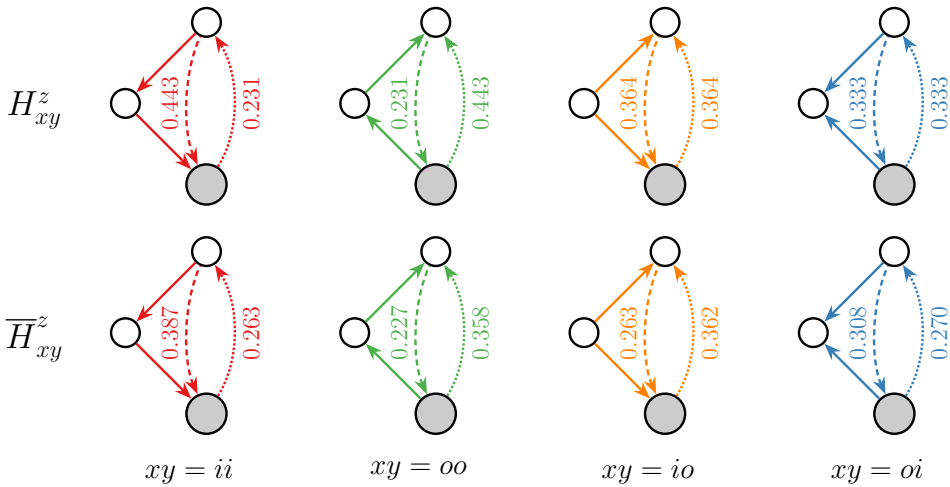
Figure 3 shows the global and average directed closure coefficients of the SOC-LAWYER dataset. From the first row, we see that the eight global closure coefficients can be grouped into four pairs,

$$\{(H_{ii}^i, H_{oo}^o), (H_{ii}^o, H_{oo}^i), (H_{io}^i, H_{io}^o), (H_{oi}^i, H_{oi}^o)\},$$

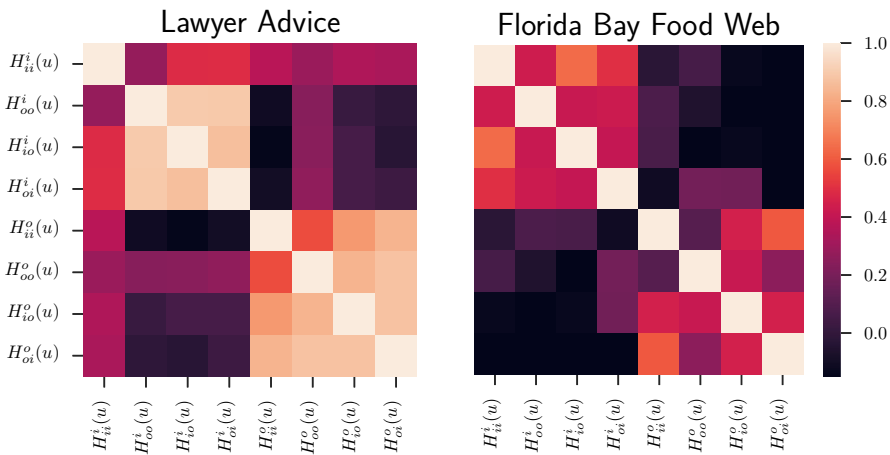
with each pair of coefficients taking the same value. This observation is expected due to the symmetry in the wedge structure, which we study in more detail in Section 4.1. In contrast, these groupings do not take the same value in the case of the average closure coefficients (the second row of Figure 3): we observe an *a priori* unexpected asymmetry. For example,  $\bar{H}_{io}^o = 0.362 \gg \bar{H}_{io}^i = 0.263$  (in orange, Figure 3). When an in-out wedge is closed with either an incoming or outgoing edge, the induced triangle is actually the same: both are feedforward loops (Milo et al., 2002). It is not obvious why closure with an outgoing edge is so much more likely that with an incoming edge. We develop some theoretical explanations for this asymmetry in Section 4.2.

We also explore the correlations between the eight average directed closure coefficients in the SOC-LAWYER and FW-FLORIDA networks (Figure 4). Each network has a clear separation among the eight local closure coefficients: the coefficients in the first four rows and columns (with incoming closure edge), and the coefficients in the last four rows and columns (with outgoing closure edge). Within each group, the coefficients are strongly correlated. In SOC-LAWYER, the coefficients in different groups are nearly uncorrelated, whereas in FW-FLORIDA, the coefficients in different groups are negatively correlated. This correlation pattern is representative across the





**Figure 3.** Global (top) and average (bottom) directed closure coefficients in SOC-LAWYER, with head nodes in gray. The global closure coefficients exhibit symmetry (e.g.,  $H_{io}^i = H_{io}^o$ ), while the average closure coefficients exhibit counterintuitive asymmetry between pairs of coefficients, for example,  $\bar{H}_{io}^o = 0.263 \ll \bar{H}_{io}^o = 0.362$  (in orange, second row). The induced structure is the same in both closure coefficients (a feedforward loop or acyclic triangle). We explain this phenomenon in Section 4.



**Figure 4.** Heatmap of the correlation matrix of the eight local directed closure coefficients in SOC-LAWYER (left) and FW-FLORIDA (right). There is a clear separation on the eight local closure coefficients: the ones for *i*-closed and the ones for *o*-closed. Coefficients within each group are highly correlated while between groups are almost uncorrelated.

networks that we have studied with directed closure coefficients, and we explain this correlation separation as part of the next section.

To study the difference in frequencies of directed triadic closure, we visualize the 8 average directed closure coefficients of 10 networks in Figure 5, where each row contains two networks within the same domain. We find that each domain of networks has their own directed triadic closure patterns, even though the absolute closure rates can be different (see varying *y*-axis scales) due to the size and average degree of each network. In social networks, different wedge types have similar closure frequencies, likely due to the abundance of reciprocal edges (Newman et al., 2002). In communication networks, the tall blue bars associated with out-in wedge type means one is more likely to connect to people with whom they both send communications; this might

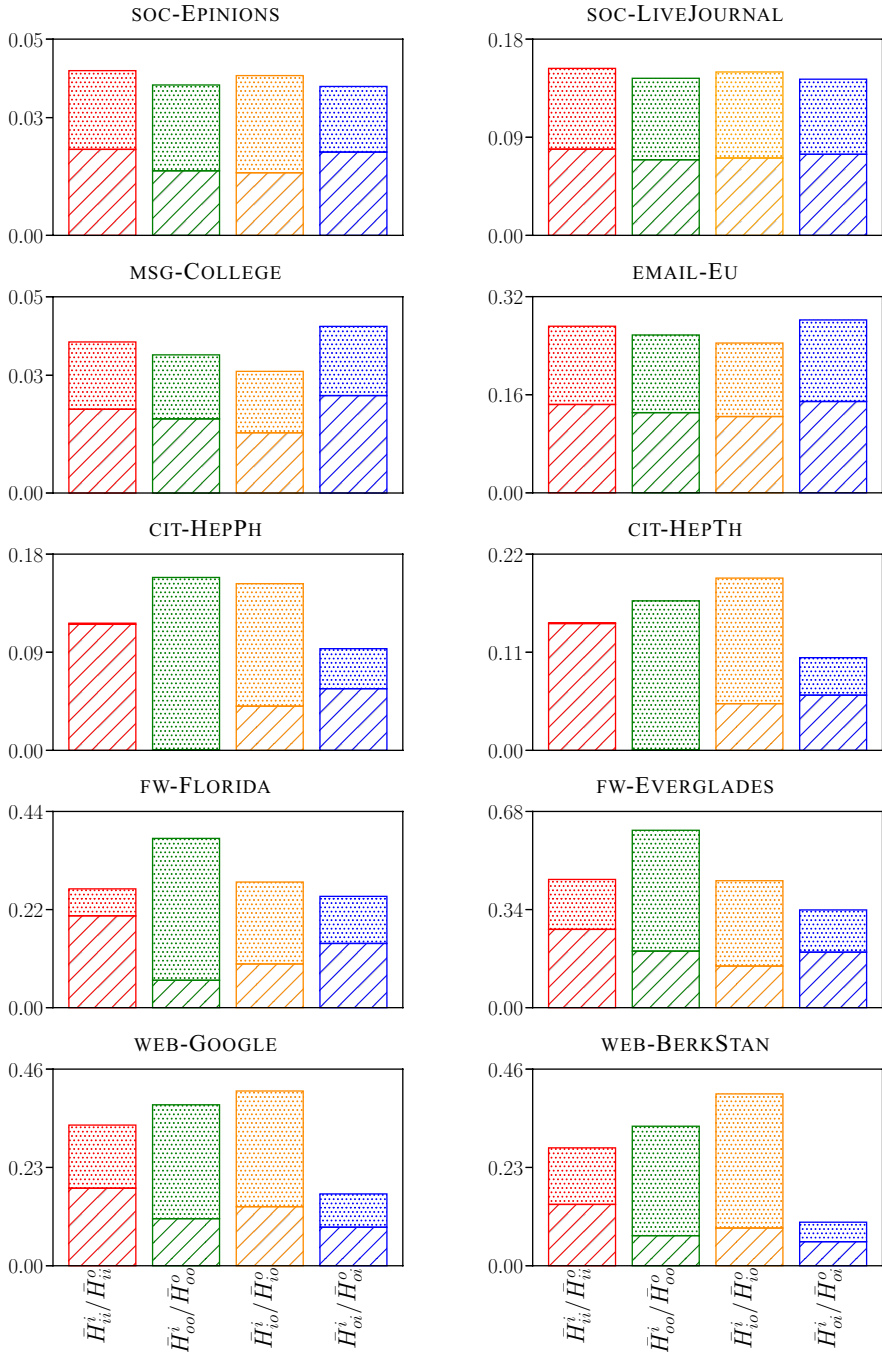


Figure 5. Average directed closure coefficients of networks from five domains. Wedge types are colored in the same way as in Figure 3, with incoming closure represented by slashed-bars and outgoing closure represented by dotted-bars. Although networks from the same domain (each row) can have different absolute closure rates (see varying y-axis scales), their relative rate in closure direction of each wedge type are similar. Such relative closure pattern can be quite different across domains.

be a result of shared interest. In contrast, citation networks have low closure coefficients for out-in wedges (short blue bars), meaning that one is not likely to cite or be cited by papers with the same reference: this phenomenon might come from a conflict of interest; moreover, due to near nonexistence of cycles, in-in wedges and out-out wedges are each only closed in one direction. Similar patterns appear in the food webs and web graphs, where there is a hierarchical structure and few cycles. Lastly, we observe similar asymmetry in all citation, food web, and web graphs, namely that  $\bar{H}_{io}^o > \bar{H}_{io}^i$ , as the orange bars show significantly higher outbound closure rate than inbound rate.

#### 4. Theoretical analysis

We now provide theoretical analysis of our directed closure coefficients. We first prove the symmetry between the four pairs of global directed closure coefficients. Motivated by the empirical asymmetry among average directed closure coefficients, we first prove that this asymmetry can be unboundedly large. Finally, to explain the asymmetry, we study how the in- and out-degree distributions influence the expected value of each average closure coefficients under a directed configuration model with a fixed joint degree distribution.

##### 4.1 Symmetry and asymmetry

Recall that each global directed closure coefficient is the fraction of certain types of wedges that are closed in the entire network. We observed in Section 3 that the eight global directed closure coefficients can be grouped into four pairs, with each pair of coefficients having the same value. The following proposition shows that these values must be the same in any network.

**Proposition 4.1.** *In any directed network,  $H_{ii}^i = H_{oo}^o$ ,  $H_{ii}^o = H_{oo}^i$ ,  $H_{io}^i = H_{io}^o$ , and  $H_{oi}^i = H_{oi}^o$ .*

*Proof.* Here we only prove  $H_{ii}^i = H_{oo}^o$ , and the other three identities can be shown analogously. By counting wedges from the center node,  $W_{ii} = \sum_u d_i(u) \cdot d_o(u) = W_{oo}$ . Next, there is a one-to-one correspondence between a closed in-in wedge and a closed out-out wedge by flipping the roles of the head and tail nodes. Thus,  $T_{ii}^i = T_{oo}^o$  and  $H_{ii}^i = H_{oo}^o$ , according to Definition 3.3.  $\square$

Proposition 4.1 illustrates the fundamental symmetry among the eight global directed closure coefficients. The four pairs of global closure coefficients




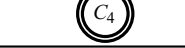
$$\{(H_{ii}^i, H_{oo}^o), (H_{ii}^o, H_{oo}^i), (H_{io}^i, H_{io}^o), (H_{oi}^i, H_{oi}^o)\}$$

correspond to the same structure and triadic closure pattern in the entire network, so their values have to be the same.

As an alternative global measure of directed triadic closure, we might expect the average closure coefficients to have a similar symmetric pattern. Specifically, by pairing up the average closure coefficients in the same way,

$$\{(\bar{H}_{ii}^i, \bar{H}_{oo}^o), (\bar{H}_{ii}^o, \bar{H}_{oo}^i), (\bar{H}_{io}^i, \bar{H}_{io}^o), (\bar{H}_{oi}^i, \bar{H}_{oi}^o)\}, \tag{3}$$

one might initially guess that the two values in a pair would be close. However, our empirical evaluation on the SOC-LAWYER datasets above, as well as all of the citation, food webs, and web graphs, showed *asymmetry* in these metrics; for example,  $\bar{H}_{io}^o \gg \bar{H}_{io}^i$  in the SOC-LAWYER dataset. Here, we study how large such difference can be and find that there is no nontrivial upper or lower bound on  $\bar{H}_{io}^i$  based on  $\bar{H}_{io}^o$  and vice versa. Furthermore, this same flavor of unboundedness is valid for the other three pairs of average directed closure coefficients.

Class	#nodes	$W_{io}(u)$	$T_{io}^i(u)$	$T_{io}^o(u)$
	$n_1$	$n_3n_2 + n_3n_4$	$n_3n_2$	0
	$n_2$	$n_3n_1 + n_3n_4$	0	$n_3n_1$
	$n_3$	0	0	0
	$n_4$	$n_3n_1 + n_3n_2$	0	0

**Figure 6.** An example graph used in the proof of Theorem 4.2, showing maximal differences between directed closure coefficients  $\bar{H}_{io}^i$  and  $\bar{H}_{io}^o$ . Each double circle  $C_j$  represents a class of nodes and an edge  $C_j \rightarrow C_k$  means that  $u_j \rightarrow u_k$  for all  $u_j \in C_j$  and  $u_k \in C_k$ .

**Theorem 4.2.** For any  $\varepsilon > 0$ , and any pair of average directed closure coefficients from Equation (3), denoted as  $(\bar{H}_a, \bar{H}_b)$ , there is a finite graph such that  $\bar{H}_a < \varepsilon$  and  $\bar{H}_b > 1 - \varepsilon$ , and another finite graph such that  $\bar{H}_a > 1 - \varepsilon$  and  $\bar{H}_b < \varepsilon$ .

*Proof.* Here we give a constructive proof for the pair  $(\bar{H}_{io}^i, \bar{H}_{io}^o)$ ; the same technique works for the other three pairs. We use the example graph in Figure 6. Each double circle in the figure, denoted by  $C_j$  with  $j \in \{1, 2, 3, 4\}$ , represents a set of nodes, and we let  $n_j$  denote the number of nodes in each class. A directed edge from class  $C_j$  to  $C_k$  means that for any node  $u_j \in C_j$  and any node  $u_k \in C_k$ , there is an edge  $u_j \rightarrow u_k$ . The number of in-out wedges as well as closed wedges are listed in the last three columns of the table. We have that  $H_{io}^i(u) = \frac{n_2}{n_2+n_4}$  for any node  $u \in C_1$ ,  $H_{io}^i(u) = 0$  for  $u \in C_2$  or  $C_4$ , and  $H_{io}^i(u)$  undefined for  $u \in C_3$ . Now,

$$\bar{H}_{io}^i = \frac{n_1n_2}{(n_2 + n_4)(n_1 + n_2 + n_3 + n_4)}, \quad \bar{H}_{io}^o = \frac{n_1n_2}{(n_1 + n_4)(n_1 + n_2 + n_3 + n_4)}.$$

The  $n_j$ 's can take any integer value. We first fix  $n_3 = n_4 = 1$ . If  $n_1 = k^2$  and  $n_2 = k$  for any integer  $k > 3/\varepsilon$ ,  $\bar{H}_{io}^i > 1 - \varepsilon$  and  $\bar{H}_{io}^o < \varepsilon$ . And if  $n_1 = k$  and  $n_2 = k^2$  for any integer  $k > 3/\varepsilon$ ,  $\bar{H}_{io}^i < \varepsilon$  and  $\bar{H}_{io}^o > 1 - \varepsilon$ . □

In contrast, the directed clustering coefficients due to Fagiolo (2007) are based on the center of wedges, so the two edges are naturally symmetric and consequently, the metric is always symmetric. Therefore, there are four directed clustering coefficients and eight directed closure coefficients.

In the next section, we study how we expect the directed closure coefficients to behave in a configuration model, which provides additional insight into why asymmetries in the directed closure coefficients might be unsurprising.

**4.2 Expectations under configuration model**

The previous section showed that pairs of average directed closure coefficients can have significantly different values; in fact, our extremal analysis showed that their ratio can be unbounded in theory. However, we have not yet provided any intuition for asymmetry in real-world networks. Here, we provide further theoretical analysis to show that the structure of the joint in- and out-degree distribution of a network provides one explanation of this asymmetry. When considering

random graphs generated under a directed configuration model with a fixed joint degree sequence, the coefficients are generally asymmetric even in their expectations.

The configuration model (Molloy & Reed, 1995; Fosdick et al., 2018) is a standard tool for analyzing the behavior of patterns and measures on networks. The model is typically studied for undirected graphs with a specified degree sequence, but the idea cleanly generalizes to directed graphs with a specified joint degree sequence (Chen & Olvera-Cravioto, 2013). It is often hard to understand the determinants of unintuitive observations on networks. What aspect of the specific network under examination leads to a given observation? As one specific angle on this question, does the observation hold for typical graphs with the observed joint degree sequence, and if so, what are the determinants of the behavior? Analyses using the configuration model can thus be used to investigate the expected behavior of a measure, in our case the directed closure coefficients, under this random graph distribution.

If a degree sequence  $S$  satisfies the condition that the maximum degree is upper bounded by  $\sqrt{n}$ , then under the configuration model with this degree sequence, the probability of forming an edge  $u \rightarrow v$  is

$$\mathbb{P}[(u, v) \in E \mid S] = \frac{d_o(u) \cdot d_i(v)}{m} \cdot (1 + o(1)), \tag{4}$$

where the  $o(1)$  term is with respect to large graphs (i.e.,  $n \rightarrow \infty$ ) (Newman, 2003). As further notation for this section, for an event denoted by  $A$ , we use  $\mathbf{1}_{[A]}$  as the indicator function for event  $A$ , that is,  $\mathbf{1}_{[A]} = 1$  when event  $A$  happens and 0 otherwise. Moreover, we use the symbol “ $\sim$ ” between two quantities  $X \sim Y$  if  $X = Y \cdot (1 + o(1))$ . For any direction variable  $x \in \{i, o\}$ , we use  $\bar{x}$  to denote the opposite direction of  $x$ .

Before presenting the main theoretical results, we first provide a useful lemma. The error term  $o(1)$  here, as well as those in subsequent theorems in this subsection, vanishes as the size of network grows to infinity, which is the scenario when the probability of an edge between two nodes in the directed configuration model is proportional to the product of the nodes’ degrees [Equation (4)].

**Lemma 4.3.** *Suppose  $G$  is a random directed graph sampled from the directed configuration model with joint degree sequence  $S$  and let  $u$  be any node. Let  $(u, v, w)$  be a random type- $xy$  wedge with head node  $u$ . Then for either direction  $z \in \{i, o\}$ ,*

$$\mathbb{E}[d_z(w) \mid S, u] = (n/m) \cdot M_{\bar{y}z} \cdot (1 + o(1)).$$

*Proof.* Conditional on the degree sequence, for any node pair  $v^*$  and  $w^*$ ,  $(u, v^*, w^*)$  forms an  $xy$ -wedge with probability

$$\frac{d_x(u)d_{\bar{x}}(v^*)}{m} \cdot \frac{d_y(v^*)d_{\bar{y}}(w^*)}{m} \cdot (1 + o(1)) \sim C \cdot d_{\bar{y}}(w^*)$$

where  $C$  is a constant independent of  $w^*$ . Therefore, for any node  $w^*$ , it is the other end of a random wedge with probability  $\mathbb{P}[w = w^* \mid S, u] \propto d_{\bar{y}}(w^*) \cdot (1 + o(1))$ , and thus

$$\mathbb{P}[w = w^* \mid S, u] \sim \frac{d_{\bar{y}}(w^*)}{\sum_{w \in V} d_{\bar{y}}(w)} = \frac{d_{\bar{y}}(w^*)}{m}.$$

Consequently, we have

$$\mathbb{E}[d_z(w) \mid S, u] = \sum_{w^* \in V} d_z(w^*) \cdot \mathbb{P}[w = w^* \mid S, u] \sim \sum_{w^* \in V} \frac{d_{\bar{y}}(w^*)d_z(w^*)}{m} = \frac{nM_{\bar{y}z}}{m}.$$

□

Now we present the following theoretical results on the expected value of local directed closure coefficients under the directed configuration model, which relates the expected closure coefficient of node  $u$  with closing direction  $i$  and  $o$  to the in- and out-degrees  $d_i(u)$  and  $d_o(u)$  of  $u$ .

**Theorem 4.4.** *Let  $S$  be a joint degree sequence and  $G$  a random directed graph sampled from the directed configuration model with  $S$ . For any node  $u$  and any local directed closure coefficient  $H_{xy}^z(u)$ , we have*

$$\mathbb{E}[H_{xy}^z(u) | S] = \frac{n(d_z(u) - \mathbf{1}_{[x=z]})}{m^2} \cdot \left( M_{\bar{y}\bar{z}} - \mathbf{1}_{[y=z]} \cdot \frac{m}{n} \right) \cdot (1 + o(1))$$

where  $M_{\bar{y}\bar{z}}$  is the second-order moment of degree sequence  $S$ .

*Proof.* Note that  $H_{xy}^z(u)$  can be directly interpreted as the probability that a random type- $xy$  wedge  $(u, v, w)$  is  $z$ -closed, where node  $u$  is the head of this wedge. This is the case if there is an edge between  $u$  and  $w$  of direction  $z$  (with respect to  $u$ ): a  $z$ -stub from node  $u$  is matched to a  $\bar{z}$ -stub from node  $w$ . Note that the number of  $z$ -stubs of node  $u$  that are not used in wedge  $(u, v, w)$  is  $(d_z(u) - \mathbf{1}_{[x=z]})$ , where we need to subtract the indicator function because one  $z$ -stub is already used in wedge  $(u, v, w)$  if  $x = z$ . Similarly, the number of  $\bar{z}$ -stubs of node  $w$  that are not used in wedge  $(u, v, w)$  is  $(d_{\bar{z}}(w) - \mathbf{1}_{[\bar{y}=\bar{z}]})$ . According to the setup of the directed configuration model [Equation (4)], this probability is

$$(d_z(u) - \mathbf{1}_{[x=z]}) \cdot (d_{\bar{z}}(w) - \mathbf{1}_{[\bar{y}=\bar{z}]}) / m \cdot (1 + o(1))$$

with the given joint degree sequence, and consequently

$$\begin{aligned} \mathbb{E}[H_{xy}^z(u) | S] &\sim \mathbb{E}[(d_z(u) - \mathbf{1}_{[x=z]}) \cdot (d_{\bar{z}}(w) - \mathbf{1}_{[\bar{y}=\bar{z}]}) / m | S] \\ &= \frac{d_z(u) - \mathbf{1}_{[x=z]}}{m} \cdot (\mathbb{E}[d_{\bar{z}}(w) | S] - \mathbf{1}_{[\bar{y}=\bar{z}]}) \\ &\sim \frac{d_z(u) - \mathbf{1}_{[x=z]}}{m} \cdot ((n/m) \cdot M_{\bar{y}\bar{z}} - \mathbf{1}_{[\bar{y}=\bar{z}]}) \\ &= \frac{n(d_z(u) - \mathbf{1}_{[x=z]})}{m^2} \cdot \left( M_{\bar{y}\bar{z}} - \mathbf{1}_{[y=z]} \cdot \frac{m}{n} \right) \end{aligned}$$

where the second step follows from the fact that the only random variable is the degree of a random tail node  $w$ , and the third step is due to Lemma 4.3. □

Theorem 4.4 shows that the expected value of the local directed closure coefficient  $H_{xy}^z(u)$  increases with  $d_z(u)$ , the degree in the direction of closure. One corollary of this result is that under the configuration model, the expected values of the local closure coefficient with the same closure direction are all monotonic with the same corresponding degree and thus they should be correlated themselves. This result provides one intuition for the block structure of the correlations between coefficients found in Figure 4.

We can easily aggregate the results of Theorem 4.4 to give expected values of the average directed closure coefficients.

**Theorem 4.5.** *Let  $S$  be a joint degree sequence and  $G$  be a random directed graph generated from the directed configuration model with  $S$ . For any average directed closure coefficient  $\bar{H}_{xy}^z$ ,*

$$\mathbb{E}[\bar{H}_{xy}^z | S] = \frac{m - n \cdot \mathbf{1}_{[x=z]}}{m^2} \cdot \left( M_{\bar{y}\bar{z}} - \mathbf{1}_{[y=z]} \cdot \frac{m}{n} \right) \cdot (1 + o(1)).$$



*Proof.* We have

$$\begin{aligned} \mathbb{E}[\overline{H}_{xy}^z | S] &= \frac{1}{n} \sum_u \mathbb{E}[H_{xy}^z(u) | S] \\ &\sim \left( M_{\overline{y\bar{z}}} - \mathbf{1}_{[y=z]} \cdot \frac{m}{n} \right) \cdot \frac{1}{m^2} \sum_u [d_z(u) - \mathbf{1}_{[x=z]}] \\ &= \left( M_{\overline{y\bar{z}}} - \mathbf{1}_{[y=z]} \cdot \frac{m}{n} \right) \cdot \frac{m - n \cdot \mathbf{1}_{[x=z]}}{m^2} \end{aligned}$$

where the second line is due to Theorem 4.4. □

Theorem 4.5 shows that the expected value of any average closure coefficient  $\overline{H}_{xy}^z$  is mainly determined by  $M_{\overline{y\bar{z}}}$ , a second-order moment of the degree sequence. In the SOC-LAWYER dataset, we have  $M_{i_0} = 166.15 \ll 227.41 = M_{ii}$ , meaning that  $\mathbb{E}[\overline{H}_{i_0}^i] \ll \mathbb{E}[\overline{H}_{i_0}^0]$ . This result (partly) explains the asymmetry observed in Figure 3: the different coefficients are related to different moments of the joint degree sequence of the network, at least for graphs sampled from a configuration model with different empirical joint degree sequences.

Finally, we can also determine the expected value of global directed closure coefficients under the configuration model, as given in Theorem 4.7. Again we first present a useful lemma, which is analogous to Lemma 4.3.

**Lemma 4.6.** *Suppose  $G$  is a random directed graph sampled from the directed configuration model with joint degree sequence  $S$ . Let  $(u, v, w)$  be a random type- $xy$  wedge, then*

- (1)  $\mathbb{E} [d_z(w) | S] = (n/m) \cdot M_{\overline{y\bar{z}}} \cdot (1 + o(1));$
- (2)  $\mathbb{E} [d_z(u) | S] = (n/m) \cdot M_{xz} \cdot (1 + o(1));$
- (3)  $\mathbb{E} [d_z(u)d_z(w) | S] = \mathbb{E} [d_z(u) | S] \cdot \mathbb{E} [d_z(w) | S] \cdot (1 + o(1)).$

*Proof.* Result 1 is a corollary of Lemma 4.3:

$$\mathbb{E}[d_z(w) | S] = \mathbb{E} [\mathbb{E}[d_z(w) | S, u] | S] \sim \frac{nM_{\overline{y\bar{z}}}}{m}.$$

Result 2 is a corollary of result 1:  $(u, v, w)$  being a type- $xy$  wedge is equivalent to  $(w, v, u)$  being a type- $\overline{y\bar{x}}$  wedge.

Now, we show the last result. Conditional on the degree sequence, for any node triple  $(u^*, v^*, w^*)$ , it forms an  $xy$ -wedge with probability

$$\frac{d_x(u^*)d_{\overline{x}}(v^*)}{m} \cdot \frac{d_y(v^*)d_{\overline{y}}(w^*)}{m} \cdot (1 + o(1)) \sim C \cdot d_x(u^*)d_{\overline{y}}(w^*)$$

where  $C$  is a constant independent of  $u^*$  and  $w^*$ . Therefore, for any node pair  $u^*$  and  $w^*$ , they are the two ends of a random wedge with probability  $\mathbb{P} [u = u^*, w = w^* | S] \propto d_x(u^*)d_{\overline{y}}(w^*) \cdot (1 + o(1))$ , and thus

$$\mathbb{P} [u = u^*, w = w^* | S] \sim \frac{d_x(u^*)d_{\overline{y}}(w^*)}{\sum_{u,w \in V} d_x(u)d_{\overline{y}}(w)} = \frac{d_x(u^*)d_{\overline{y}}(w^*)}{m^2}.$$

Consequently, we have

$$\begin{aligned} \mathbb{E} [d_z(u)d_z(w) | S] &= \sum_{u^*, w^* \in V} d_z(u^*)d_z(w^*) \cdot \mathbb{P} [u = u^*, w = w^* | S] \\ &\sim \sum_{u^*, w^* \in V} d_z(u^*)d_z(w^*) \cdot \frac{d_x(u^*)d_{\overline{y}}(w^*)}{m^2} \end{aligned}$$

$$\begin{aligned}
 &= \frac{nM_{xz}}{m} \cdot \frac{nM_{\bar{y}\bar{z}}}{m} \\
 &\sim \mathbb{E} [d_z(u) | S] \cdot \mathbb{E} [d_{\bar{z}}(w) | S]
 \end{aligned}$$

which completes the proof. □

**Theorem 4.7.** *Let  $S$  be a joint degree sequence and  $G$  be a random directed graph generated from the directed configuration model with  $S$ . For any global directed closure coefficient  $H_{xy}^z$ ,*

$$\mathbb{E}[H_{xy}^z | S] = \left(M_{\bar{y}\bar{z}} - \mathbf{1}_{[y=z]} \cdot \frac{m}{n}\right) \cdot \left(M_{xz} - \mathbf{1}_{[x=z]} \cdot \frac{m}{n}\right) \cdot \frac{n^2}{m^3} \cdot (1 + o(1)).$$

*Proof.* For a random type- $xy$  wedge  $(u, v, w)$ , we have shown that the probability of it being  $z$ -closed is of the order  $(d_z(u) - \mathbf{1}_{[x=z]}) \cdot (d_{\bar{z}}(w) - \mathbf{1}_{[\bar{y}=\bar{z}]})/m$ . Different from the proof of Theorem 4.4 where node  $u$  is fixed, here we do not fix node  $u$ , meaning that both node  $u$  and node  $w$  are random.

$$\begin{aligned}
 \mathbb{E} [H_{xy}^z | S] &\sim \frac{1}{m} \cdot \mathbb{E} [(d_z(u) - \mathbf{1}_{[x=z]}) \cdot (d_{\bar{z}}(w) - \mathbf{1}_{[\bar{y}=\bar{z}]}) | S] \\
 &\sim \frac{1}{m} \cdot (\mathbb{E}[d_z(u) | S] - \mathbf{1}_{[x=z]}) \cdot (\mathbb{E}[d_{\bar{z}}(w) | S] - \mathbf{1}_{[\bar{y}=\bar{z}]}) \\
 &\sim \frac{1}{m} \cdot \left(\frac{n}{m} \cdot M_{xz} - \mathbf{1}_{[x=z]}\right) \cdot \left(\frac{n}{m} \cdot M_{\bar{y}\bar{z}} - \mathbf{1}_{[\bar{y}=\bar{z}]}\right)
 \end{aligned}$$

where the second line is due to the last result in Lemma 4.6, and the last line is due to the first two results in Lemma 4.6. □

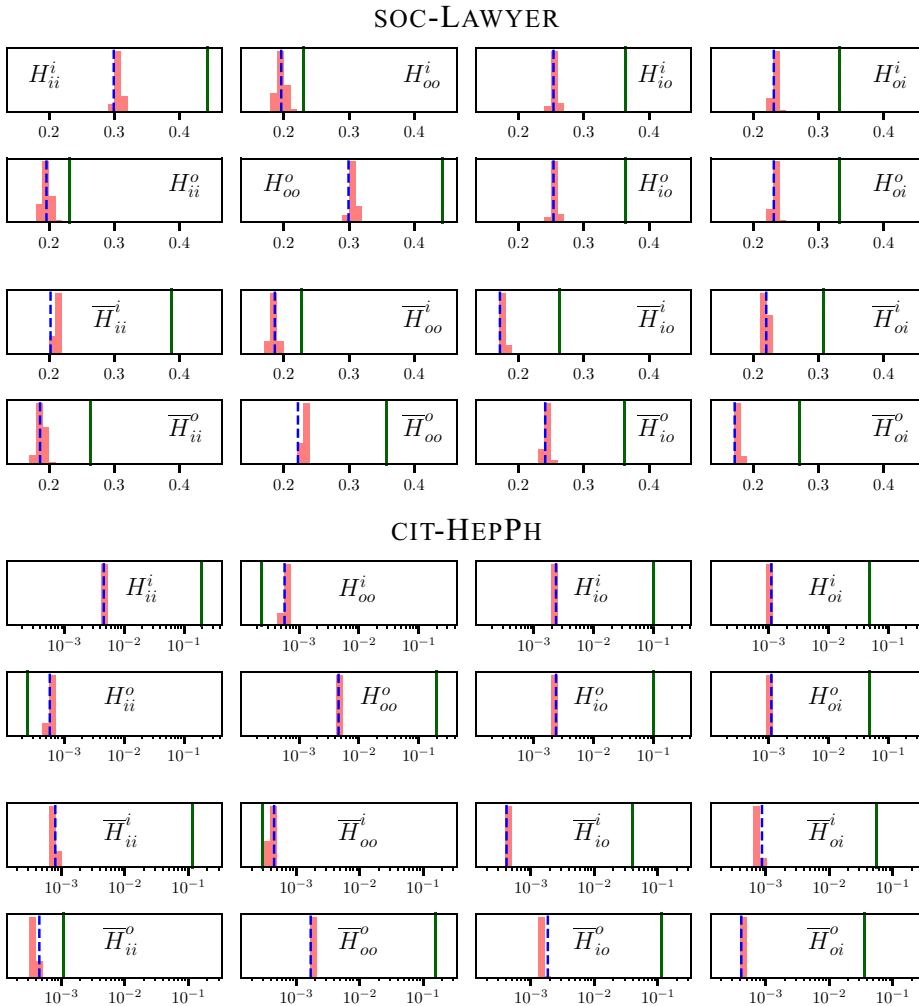
As a byproduct of our analysis, the proof of Theorem 4.7 also shows that, under the directed configuration model, the probability that a wedge is closed is independent of the center node, and thus equal to the network-level average. This observation gives us the expected value of Fagiolo’s (2007) directed local clustering coefficients under this random graph model as well.

**Proposition 4.8.** *Let  $S$  be a joint degree sequence and  $G$  be a random directed graph generated from the directed configuration model with  $S$ . For local directed clustering coefficient  $C_{xy}(u)$ ,*

$$\mathbb{E}[C_{xy}(u) | S] = \mathbb{E}[H_{xy}^i | S].$$

*Proof.* Consider a random wedge with  $u$  being the center node,  $(v, u, w)$ , which is an  $\bar{x}y$ -wedge to node  $v$ . From the definition of  $C_{xy}(u)$ , this wedge is closed if it is  $i$ -closed to node  $v$ . Since the probability of this wedge being  $i$ -closed is independent of node  $u$ , it is the same as if we randomly choose a wedge without constraining node  $u$  as the center, and thus  $\mathbb{E}[C_{xy}(u) | S] = \mathbb{E}[H_{xy}^i | S]$ . □

Next, we study the accuracy of the theoretical expected values of the average and global closure coefficients under the directed configuration model. The directed configuration model can be sampled using double-edge swaps (Rao et al., 1996). To sample graphs from the model, we begin with an empirical graph (the graph of interest) with joint degree sequence  $S$ . We then select a pair of random directed edges to swap, which changes the graph slightly but notably preserves the degree sequence. Taking care to avoid self-loops and multi-edges (Fosdick et al., 2018), the double-edge swap can be interpreted as a random walk in the space of simple graphs with the same degree sequence, and the stationary distribution of this random walk is the uniform distribution over the network space. The swap is then repeated many times to generate graphs that are sampled from the stationary distribution. The mixing time of these random walks are generally believed to be well behaved, but few rigorous results are known (Greenhill, 2014).



**Figure 7.** Histogram of each global closure coefficient (first two rows) and average closure coefficient (last two rows) in 1,000 directed configuration model random graphs with the joint degree sequence of the SOC-LAWYER and CIT-HEPPH networks. The  $x$ -axis is the value of various directed closure coefficients and the  $y$ -axis is the frequency. Besides the histogram, we also plot the expected value of closure coefficients from Theorems 4.5 and 4.7 (blue dashed) as well as the actual value in the original network (green solid).

We generate 1,000 random graphs with the same joint degree sequence as the SOC-LAWYER and CIT-HEPPH network; to generate each graph, we repeat the edge-swapping procedure 10,000 times. Figure 7 shows histograms of the distribution of each average and global closure coefficient under this configuration model. We see that our approximate formulas from Theorems 4.5 and 4.7 are very accurate even when the network is only moderate in size ( $n = 71$  for SOC-LAWYER). The theoretical formulas are only guaranteed to be accurate on large sparse networks, and we do observe a small difference between the expected and simulated means (e.g.,  $\bar{H}_{ii}^i$ ).

The simulation shows that the average and global closure coefficients have low variance under this configuration model, and the values in the original network deviate significantly from these distributions. More specifically, the values in the citation network are mostly larger than the distribution by orders of magnitude, and the exceptions are  $H_{oo}^i$  and  $H_{ii}^o$ , as well as their average closure

coefficient counterparts, where the real-world values are low due to the natural lack of cycles in citation networks. This provides evidence that the directed closure coefficients of real-world networks capture interesting empirical structure beyond what one would expect from a graph drawn uniformly at random from the space of graphs with the same joint degree sequence.

## 5. Case studies in node-type predictions

Now that we have a theoretical understanding of our directed closure coefficients, we turn to applications. Directed closure coefficients are a new measurement for directed triadic closure and thus can serve as a feature for network analysis and inference. In this section, we present two illustrative examples to exhibit the strong predictive potential of directed closure coefficients. Specifically, we present two case studies of node-type classification tasks, where we show the utility of local directed closure coefficients in predicting the node type in the SOC-LAWYER and FW-FLORIDA datasets analyzed above. Using an interpretable regularized model, we are able to identify the salient directed closure coefficients that are useful for prediction. This analysis reveals new social status patterns in the social network and also automatically identifies previously studied triadic patterns in food webs as good predictors.

### 5.1 Case study I: Identifying lawyer status in an advisory network

The SOC-LAWYER dataset collected by Lazega (2001) is a social network of lawyers at a corporate firm. There is a node for each of the 71 lawyers, and each is labeled with a status level—*partner* or *associate*. Of the 71 lawyers in the dataset, 36 are partners and 35 are associates. The edges come from survey responses on who individuals go to for professional advice: there is an edge from  $i$  to  $j$  if person  $i$  went to person  $j$  for professional advice. Of the edges, there are 395 between 2 partners; 196 between 2 associates; 59 from partner to associate; and 242 from associate to partner.

In this case study, our goal is to predict the status of the lawyers (associate or partner) with predictors extracted from the advice network. We consider the following six sets of network attributes as predictors:

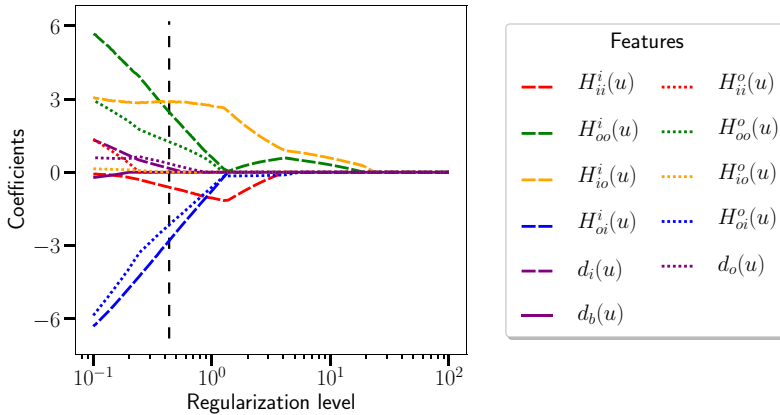
1. *degree*: the in- and out-degree, and the number of reciprocal edges at each node;
2. *degree + 1-hop*: the union of the degree predictors and four neighbor degrees: the average in- and out-degree of all in- and out-neighbors of each node;
3. *closure*: the eight local directed closure coefficients defined in this paper;
4. *closure + degree*: the union of the closure coefficients and the degree predictors;
5. *clustering*: the four local directed clustering coefficients as defined by Fagiolo (2007); and
6. *clustering + degree*: the union of the local directed clustering coefficients and the degree predictors.

For each predictor set, we use 100 random instances of threefold cross-validation to select an  $\ell_1$ -regularized logistic regression model for predicting whether or not a node is a partner (i.e., the positive label is for partner). Table 2 reports validation set accuracy and the area under the curve (AUC) metric of the Receiver Operating Characteristic (ROC). Even though different predictor sets have different dimensions, evaluating the performance in this way makes them comparable. The predictors that include our local directed closure coefficients substantially outperform the other predictor sets. The predictor set that includes both degrees and closure coefficients slightly underperforms the one with only closure coefficients, indicating slight overfitting in the training data, which implies that the degree attributes provide redundant and noisy information in addition to the closure coefficient attributes in this prediction task. In contrast, adding the *1-hop degrees* does not give as much improvement as they do not consider the triadic closure factors.

**Table 2.** Validation set accuracy and AUC in classifying node types in the soc-LAWYER dataset (partner vs. associate).

	Degree	Degree + 1-hop	Closure	Closure + degree	Clustering	Clustering + degree
Accuracy	0.7884	0.8270	<b>0.8743</b>	0.8585	0.6255	0.7884
AUC	0.8763	0.8978	<b>0.9235</b>	0.9183	0.6362	0.8765

Our proposed local directed closure coefficients (in bold) are the best set of predictors, illustrating the utility of directed closure coefficients in node-level prediction tasks. In contrast, the local directed clustering coefficients (Fagiolo, 2007) are not as effective.



**Figure 8.** Regularization path of the  $\ell_1$ -regularized logistic regression model with predictor set *closure + degree* for the model of the soc-LAWYER dataset. The x-axis is the regularization level, and the y-axis is the regression coefficient for each predictor. The vertical black dashed line represents the optimal regularization level obtained from cross-validation. The degree attributes are only selected at very low regularization levels, and various local directed closure coefficients dominate the prediction model.

To understand how the directed local closure coefficients improve prediction performance, we analyze the regularization path of our model, a standard method in sparse regression to visualize the predictors at each regularization level (Friedman et al., 2010). Figure 8 shows the regularization path for the predictor set that includes both the local directed closure coefficients and the degree predictors.

We highlight a few important observations. First, as regularization decreases, directed local closure coefficients are selected before the degree predictors, indicating that the closure coefficients are more relevant in prediction than degrees. Second, the two predictors with largest positive coefficients at the optimal level of regularization are  $H_{io}^i(u)$  and  $H_{oo}^i(u)$ , meaning that lawyers with partner status are more likely to advise people who also advise others. In contrast, the in-degree  $d_i(u)$  predictor is not one of the first selected, which implies that it is not *how many one advises* but rather *who one advises* that is correlated with partner status. Finally, the two predictors with the largest negative coefficients at the optimal regularization are  $H_{oi}^i(u)$  and  $H_{oi}^o(u)$ , meaning that partner-status lawyers are less likely to interact with other lawyers with whom they share an advisor.

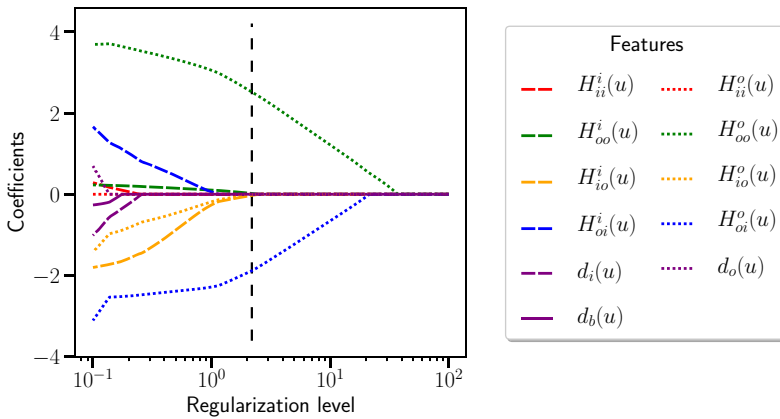
### 5.2 Case study II: Identifying fish in a food web

We now perform a similar network prediction task. Here, the data come from an entirely different domain (ecology), but we still find that our local directed closure coefficients are effective predictors for identifying node type.

**Table 3.** Validation set accuracy and AUC in classifying node types in the FW-FLORIDA dataset (fish vs. non-fish).

	Degree	Degree + 1-hop	Closure	Closure + degree	Clustering	Clustering + degree
Accuracy	0.6250	0.8466	<b>0.8735</b>	0.8700	0.6875	0.7366
AUC	0.6772	0.9127	<b>0.9538</b>	0.9529	0.7472	0.7834

Our proposed local directed closure coefficients (in bold) are again the best set of predictors (see also Table 2), illustrating the utility of directed closure coefficients in node-level prediction tasks outside of social network analysis.



**Figure 9.** Regularization path of the  $\ell_1$ -regularized logistic regression model with predictor set *closure + degree* for the model of the FW-FLORIDA dataset. The  $x$ -axis is the regularization level, and the  $y$ -axis is the regression coefficient for each predictor. The vertical black dashed line represents the optimal regularization level obtained from cross-validation. The degree attributes are only selected at very low regularization levels, and various local directed closure coefficients dominate the prediction model.

More specifically, we study a food web collected from the Florida Bay (Ulanowicz & DeAngelis, 2005). In this dataset, nodes correspond to ecological compartments (roughly, species) and edges represent directed carbon exchange (roughly, who-eats-whom). There is an edge from  $i$  to  $j$  if energy flows from compartment  $i$  to compartment  $j$ . There are 128 total compartments, of which 48 correspond to fish. Our prediction task in this case study is to identify which nodes are fish using basic node-level features. The dataset contains 2,106 edges, of which 268 are between fish; 699 are between non-fish; 648 are from a fish to a non-fish; and 491 are from a non-fish to a fish.

We used the same model selection procedure as in the first case study on the SOC-LAWYER dataset described above. Table 3 lists the accuracy and AUC of the  $\ell_1$ -regularized logistic regression model. We again find that our proposed directed closure coefficients form the best set of predictors for this task. We also find minimal difference in prediction accuracy when including degree features, indicating that the degree features provide little predictive information beyond the directed closure coefficients.

In fact, the regularization path shows that the two closure coefficients  $H_{oo}^o(u)$  and  $H_{oi}^o(u)$  are the most important predictors for identifying fish (Figure 9), the former being positively correlated with the fish type and the latter positively correlated with the non-fish type. The type of closure associated with the coefficient  $H_{oo}^o(u)$  has previously been identified as important for the network dynamics of overfishing (Bascompte et al., 2005), so it is reasonable that this predictor is important.



## 6. Discussion

Triadic closure and local clustering are fundamental properties of complex networks. Although these concepts have a storied history, only recently have there been local closure measurements (for undirected graphs) that accurately reflect the “friend of friend” mechanism pervasive in discussions of closure. In this paper, we have extended the subtle definitional difference of initiator-based versus center-based clustering to directed networks, where clustering in general has received relatively little attention. We observed a seemingly counterintuitive result that the same induced triadic structure can produce two different average directed closure coefficients; however, this asymmetry is understandable through our analysis of closure coefficients within a configuration model, which points to the role of moments of the in- and out-degree distributions. Additional analysis showed that this asymmetry can be arbitrarily large.

One of the benefits of these new local network measurements is that they can be used as predictors for statistical inference on networks. Two case studies showed that our directed closure coefficients are good predictors at identifying node types in two starkly different domains—social networks and ecology—with simple models using these features achieving over 92% mean AUC in both cases. Furthermore, directed closure coefficients are much better predictors than generalizations of clustering coefficients to directed graphs for these tasks. We anticipate that closure coefficients will become a useful tool for understanding the basic local structure of directed complex networks.

**Acknowledgments.** The authors would like to thank two anonymous reviewers as well as the editor, Ulrik Brandes, for their insightful comments. This research has been supported in part by an ARO Young Investigator Award, NSF Award DMS-1830274, and ARO Award W911NF-19-1-0057.

**Conflict of interest.** Hao Yin, Austin R. Benson, and Johan Ugander have nothing to disclose.

## Notes

1 One nuance in directed networks is that an edge might be reciprocal:  $(u, v) \in E$  and  $(v, u) \in E$ . A pair of reciprocal edges are sometimes treated as a single bidirected edge (Garlaschelli & Loffredo, 2004; Seshadhri et al., 2016) and sometimes treated as two distinct edges (Sarajlić et al., 2016). For readability purposes, in this paper, we treat a pair of reciprocal edges as two separate edges. Extensions for special considerations of reciprocal edges are straightforward and similar theoretical and empirical results can be found.

2 Again, for readability purposes, we do not consider reciprocal edges separately; instead, a reciprocal edge is treated as two separate directed edges. Our definitions and analyses can easily be extended to study reciprocal edges, though there would be 9 types of directed wedges and 27 closure coefficients.

## References

- Ahnert, S. E., & Fink, T. M. A. (2008). Clustering signatures classify directed networks. *Physical Review E*, 78(3), 036112.
- Backstrom, L., Huttenlocher, D., Kleinberg, J., & Lan, X. (2006). Group formation in large social networks: Membership, growth, and evolution. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 44–54). ACM.
- Ball, B., & Newman, M. E. J. (2013). Friendship networks and social status. *Network Science*, 1(1), 16–30.
- Barrat, A., & Weigt, M. (2000). On the properties of small-world network models. *The European Physical Journal B-Condensed Matter and Complex Systems*, 13(3), 547–560.
- Bascompte, J., Melián, C. J., & Sala, E. (2005). Interaction strength combinations and the overfishing of a marine food web. *Proceedings of the National Academy of Sciences*, 102(15), 5443–5447.
- Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., & Hwang, D.-U. (2006). Complex networks: Structure and dynamics. *Physics Reports*, 424(4–5), 175–308.
- Brzozowski, M. J., & Romero, D. M. (2011). Who should I follow? Recommending people in directed social networks. In *Fifth International AAAI Conference on Weblogs and Social Media*.
- Chen, N., & Olvera-Cravioto, M. (2013). Directed random graphs with given degree distributions. *Stochastic Systems*, 3(1), 147–186.

- Cheng, J., Romero, D. M., Meeder, B., & Kleinberg, J. (2011). Predicting reciprocity in social networks. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing* (pp. 49–56). IEEE.
- Davis, J. A., & Leinhardt, S. (1972). The structure of positive relations in small groups. In J. Berger, M. Zelditch, & B. Anderson (Eds.), *Sociological theories in progress* (vol. 2, pp. 218–251). Boston, MA: Houghton Mifflin.
- Fagiolo, G. (2007). Clustering in complex directed networks. *Physical Review E*, 76(2), 026107.
- Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, 486(3–5), 75–174.
- Fosdick, B. K., Larremore, D. B., Nishimura, J., & Ugander, J. (2018). Configuring random graph models with fixed degree sequences. *SIAM Review*, 60(2), 315–355.
- Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1), 1.
- Garlaschelli, D., & Loffredo, M. I. (2004). Patterns of link reciprocity in directed networks. *Physical Review Letters*, 93(26), 268701.
- Gehrke, J., Ginsparg, P., & Kleinberg, J. (2003). Overview of the 2003 KDD Cup. *ACM SIGKDD Explorations Newsletter*, 5(2), 149–151.
- Gleich, D. F., & Seshadhri, C. (2012). Vertex neighborhoods, low conductance cuts, and good seeds for local community methods. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 597–605). ACM.
- Greenhill, C. (2014). The switch Markov chain for sampling irregular graphs. In *Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms* (pp. 1564–1572). SIAM.
- Henderson, K., Gallagher, B., Eliassi-Rad, T., Tong, H., Basu, S., Akoglu, L., Koutra, D., Faloutsos, C., & Li, L. (2012). RoIX: Structural role extraction & mining in large graphs. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1231–1239). ACM.
- Homans, G. C. (1950). *The human group*. Harcourt: Brace & World.
- Huang, H., Tang, J., Wu, S., & Liu, L. (2014). Mining triadic closure patterns in social networks. In *Proceedings of the Twenty-Third International Conference on World Wide Web* (pp. 499–504). ACM.
- Jackson, M. O., & Rogers, B. W. (2007). Meeting strangers and friends of friends: How random are social networks? *American Economic Review*, 97(3), 890–915.
- Kaiser, M. (2008). Mean clustering coefficients: The role of isolated nodes and leafs on clustering measures for small-world networks. *New Journal of Physics*, 10(8), 083042.
- LaFond, T., Neville, J., & Gallagher, B. (2014). Anomaly detection in networks with changing trends. In *Outlier Detection and Description Under Data Diversity at the International Conference on Knowledge Discovery and Data Mining*.
- Lazega, E. (2001). *The collegial phenomenon: The social mechanisms of cooperation among peers in a corporate law partnership*. Oxford, UK: Oxford University Press.
- Leskovec, J., Kleinberg, J., & Faloutsos, C. (2005). Graphs over time: Densification laws, shrinking diameters and possible explanations. In *Proceedings of the eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining* (pp. 177–187). ACM.
- Leskovec, J., Lang, K. J., Dasgupta, A., & Mahoney, M. W. (2009). Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters. *Internet Mathematics*, 6(1), 29–123.
- Leskovec, J., Huttenlocher, D., & Kleinberg, J. (2010). Signed networks in social media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1361–1370). ACM.
- Liao, W., Ding, J., Marinazzo, D., Xu, Q., Wang, Z., Yuan, C., Zhang, Z., Lu, G., & Chen, H. (2011). Small-world directed networks in the human brain: Multivariate Granger causality analysis of resting-state fMRI. *Neuroimage*, 54(4), 2683–2694.
- Lou, T., Tang, J., Hopcroft, J., Fang, Z., & Ding, X. (2013). Learning to predict reciprocity and triadic closure in social networks. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 7(2), 5.
- Mangan, S., Zaslaver, A., & Alon, U. (2003). The coherent feedforward loop serves as a sign-sensitive delay element in transcription networks. *Journal of Molecular Biology*, 334(2), 197–204.
- 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.
- Minoiu, C., & Reyes, J. A. (2013). A network analysis of global banking: 1978–2010. *Journal of Financial Stability*, 9(2), 168–184.
- Molloy, M., & Reed, B. (1995). A critical point for random graphs with a given degree sequence. *Random Structures & Algorithms*, 6(2–3), 161–180.
- Newman, M. E. J. (2003). The structure and function of complex networks. *SIAM Review*, 45(2), 167–256.
- Newman, M. E. J., Strogatz, S. H., & Watts, D. J. (2001). Random graphs with arbitrary degree distributions and their applications. *Physical Review E*, 64(2), 026118.
- Newman, M. E. J., Forrest, S., & Balthrop, J. (2002). Email networks and the spread of computer viruses. *Physical Review E*, 66(3), 035101.

- Onnela, J.-P., Saramäki, J., Kertész, J., & Kaski, K. (2005). Intensity and coherence of motifs in weighted complex networks. *Physical Review E*, 71(6), 065103.
- Panzarasa, P., Opsahl, T., & Carley, K. M. (2009). Patterns and dynamics of users' behavior and interaction: Network analysis of an online community. *Journal of the American Society for Information Science and Technology*, 60(5), 911–932.
- Rao, A. R., Jana, R., & Bandyopadhyay, S. (1996). A Markov chain Monte Carlo method for generating random (0, 1)-matrices with given marginals. *Sankhyā: The Indian Journal of Statistics, Series A*, 58, 225–242.
- Rapoport, A. (1953). Spread of information through a population with socio-structural bias: I. Assumption of transitivity. *The Bulletin of Mathematical Biophysics*, 15(4), 523–533.
- Richardson, M., Agrawal, R., & Domingos, P. (2003). Trust management for the semantic web. In *International Semantic Web Conference* (pp. 351–368). Springer.
- Robles, P., Moreno, S., & Neville, J. (2016). Sampling of attributed networks from hierarchical generative models. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1155–1164). ACM.
- Romero, D. M., & Kleinberg, J. (2010). The directed closure process in hybrid social-information networks, with an analysis of link formation on Twitter. In *Fourth International AAAI Conference on Weblogs and Social Media*.
- Sarajlić, A., Malod-Dognin, N., Yaveroğlu, Ö. N., & Pržulj, N. (2016). Graphlet-based characterization of directed networks. *Scientific Reports*, 6, 35098.
- Seshadhri, C., Pinar, A., Durak, N., & Kolda, T. G. (2016). Directed closure measures for networks with reciprocity. *Journal of Complex Networks*, 5(1), 32–47.
- Seshadhri, C., Kolda, T. G., & Pinar, A. (2012). Community structure and scale-free collections of Erdős-Rényi graphs. *Physical Review E*, 85(5), 056109.
- Simmel, G. (1908). *Soziologie: Untersuchungen über die formen der vergesellschaftung*. Leipzig, Germany: Duncker & Humblot.
- Stegehuis, C. (2019). Closure coefficients in scale-free complex networks. *arxiv preprint arxiv:1911.11410*.
- Ulanowicz, R. E., & DeAngelis, D. L. (2005). Network analysis of trophic dynamics in South Florida ecosystems. *US Geological Survey Program on the South Florida Ecosystem*, 114, 45.
- Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications*. Cambridge, UK: Cambridge University Press.
- Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, 393(6684), 440.
- Yin, H., Benson, A. R., Leskovec, J., & Gleich, D. F. (2017). Local higher-order graph clustering. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 555–564). ACM.
- Yin, H., Benson, A. R., & Leskovec, J. (2019). The local closure coefficient: A new perspective on network clustering. *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining* (pp. 303–311). ACM.