The Local Closure Coefficient: A New Perspective On Network Clustering

Hao Yin
Stanford University
yinh@stanford.edu

Austin R. Benson
Cornell University
arb@cs.cornell.edu

Jure Leskovec
Stanford University
jure@cs.stanford.edu

ABSTRACT

The phenomenon of edge clustering in real-world networks is a fundamental property underlying many ideas and techniques in network science. Clustering is typically quantified by the clustering coefficient, which measures the fraction of pairs of neighbors of a given center node that are connected. However, many common explanations of edge clustering attribute the triadic closure to a "head" node instead of the center node of a length-2 path—for example, "a friend of my friend is also my friend." While such explanations are common in network analysis, there is no measurement for edge clustering that can be attributed to the head node.

Here we develop *local closure coefficients* as a metric quantifying head-node-based edge clustering. We define the local closure coefficient as the fraction of length-2 paths emanating from the head node that induce a triangle. This subtle difference in definition leads to remarkably different properties from traditional clustering coefficients. We analyze correlations with node degree, connect the closure coefficient to community detection, and show that closure coefficients as a feature can improve link prediction.

ACM Reference Format:
Hao Yin, Austin R. Benson, and Jure Leskovec. 2019. The Local Closure Coefficient: A New Perspective On Network Clustering. In *The Twelfth ACM International Conference on Web Search and Data Mining (WSDM '19), February 11–15, 2019, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3289600.3290991

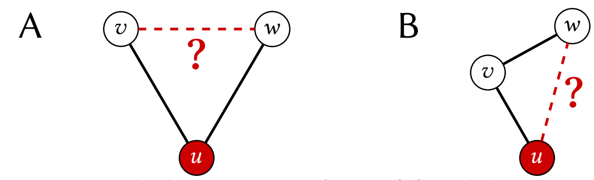

Figure 1: Local Clustering Coefficient (A) and the proposed Local Closure Coefficient (B) at node u. The Local Clustering Coefficient is the fraction of closed length-2 paths (wedges) with *center node* u. In contrast, we define the Local Closure Coefficient as the fraction of closed wedges where u is the *head* of the wedge.

1 INTRODUCTION

Networks are a fundamental tool for understanding and modeling complex physical, social, informational, and biological systems [8, 33]. Although network models of real-world systems are often sparse graphs, a recurring trait is that the edges tend to cluster [24, 39, 44, 50]. More specifically, the probability of a link between a pair of nodes sharing a common neighbor is much larger than one would expect compared to a random null model [50]. In many contexts, the edge clustering phenomenon is natural. For example, in a social network, two individuals with a common friend are more likely to become friends themselves [38]; in co-authorship networks, scientists with a mutual collaborator are more likely to collaborate in the future [20]; and in citation networks, two references appearing in the same publication are more likely to cite each other [51]. The clustering of edges underlies a number of areas of network analysis, including community detection algorithms [10, 12, 43], feature generation in machine learning tasks on networks [17, 23], and the development of generative models for networks [26, 41, 44].

The *clustering coefficient* is the standard metric for quantifying the extent to which edges of a network cluster. In particular, the *local* clustering coefficient of a node u is the fraction of pairs of neighbors of u that are connected by an edge (Figure 1A). The local clustering coefficient is a node attribute often used in machine learning pipelines utilizing network features for tasks like outlier detection [23] and role discovery [2, 17]. The local clustering coefficient has also been used as a covariate outside of computer science in, for example, psychological studies of suicide [5].

In many networks, large clustering coefficients can be explained by local evolutionary processes. In social networks, for example, local clustering is often explained by the old adage that "a friend of a friend is my friend," which is incorporated into network growth models [18, 27]. This intuitive explanation is an old idea in structural balance theory [15, 16]: the head node u trusts neighbor v, who in turn trusts its own neighbor w, leading u to trust w (Figure 1B).

However, in this explanation, closure is driven by an *end point* of the length-2 path (i.e., wedge), also called the *head* (this is node u in Figure 1B), who "creates" a link to a friend of a friend. The head-driven closure mechanism commonly appears in other network contexts as well. For example, such closure is sometimes attributed to status [14, 25]: the head node u thinks highly of a neighbor v, node v thinks highly of their neighbor w, and then the head node u thinks highly of w. And in citation networks, acknowledging and citing one paper usually leads to further reading on its references and subsequent citations [51].

Surprisingly, the above explanations on the mechanism underlying triadic closure and the clustering of edges are fundamentally

WSDM '19, February 11–15, 2019, Melbourne, VIC, Australia

ACM ISBN 978-1-4503-5940-5/19/02…$15.00
https://doi.org/10.1145/3289600.3290991

different from how clustering is actually measured today. Specifically, the clustering coefficient quantifies clustering from the *center* of the wedge, using edges that are not even adjacency to the center node u (Figure 1A). On the other hand, explanations for the emergence of clustering are often based on the head node u of a wedge, using edges actually adjacent to u (Figure 1B). Currently, there is no local metric quantifying clustering that is consistent with triadic closure processes driven by the head of a wedge. This poses a fundamental gap in how clustering is measured in networks.

Present work: Closure Coefficients. Here, we close this gap and propose a new metric for quantifying the level of clustering attributed to the head node of a wedge (Figure 1B). Our work stems from the observation that many processes leading to triadic closure are head-based but the standard metric for clustering (the clustering coefficient) is center-based (Figure 1A). We propose the *local closure coefficient* to measure clustering from the head of a wedge. We define the local closure coefficient of a node u as the fraction of length-2 paths (wedges) emanating from u that induce a triangle (Figure 1B). The traditional local clustering coefficient of a node u measures the edge density amongst the *neighbors* of u—thus u is not even adjacent to the edges that count towards its clustering (Figure 1A). In contrast, the local closure coefficient depends on edges adjacent to node u itself (Figure 1B). While naive computation of the closure coefficient would examine the node's 2-hop neighborhood, we show how to compute the local closure coefficient in the same time it takes to compute the local clustering coefficient.

Our goal is *not* to argue that the closure coefficient is an across-the-board better metric of edge clustering. Instead, we show that closure coefficients are a complementary metric and may be a useful measure of clustering in scenarios such as link prediction, role discovery, or outlier detection, which often use several node-based features [3, 4, 17, 23]. While the definition of the closure coefficient is only a subtle structural change from the clustering coefficient—measuring triadic closure from the head of a wedge rather than the center—it induces remarkably different empirical and theoretical properties. The first major difference is correlation with node degree. In many real-world networks, the local clustering coefficient decreases with node degree at a power-law-like rate [40, 48], and some models can explain this [44, 47]. We show empirically that the local closure coefficient *increases* with node degree and provide theoretical justification for this phenomenon.

We also show connections between the closure coefficient and the existence of communities in networks, as measured by conductance. Specifically, the conductance of the 1-hop neighborhood of *every* node is bounded by one minus the local closure coefficient. This is a generalization and strengthening of a prior result that relies on highly technical proof techniques and only bounds a *single* 1-hop neighborhood with minimal conductance [13]. Using closure coefficients, the proofs simplify while providing stronger and more general bounds. We also obtain analogous generalizations and strengthening of a similar result for "higher-order" clustering coefficients [53, 54] on a "motif conductance" metric [6]. Using our theory as motivation, we find that the closure coefficient can identify good seeds for personalized PageRank community detection.

Finally, we show that including closure coefficients as a feature improves link prediction performance in many cases, especially in social networks. To explain these findings, we show that, compared to clustering coefficients, closure coefficients more closely match "true closures," where the edge in a triangle with the latest timestamp is the only one that counts towards closure.

In summary, we propose the local closure coefficient, a new metric of local clustering that is based on "a friend of my friend is my friend" triadic closure. The metric is subtly different from the classical local clustering coefficient but carries several interesting empirical and theoretical properties, and we anticipate that it will become part of the lexicon of basic node-level network statistics.

2 PRELIMINARIES AND BACKGROUND

Notation. We consider networks as undirected graphs $G = (V, E)$ without self-loops. We use $n = |V|$ as the number of nodes and $m = |E|$ as the number of edges. For any node $u \in V$, we denote its degree by d_u, which is the number of edges incident to node u. We denote the subset of nodes containing node u and all its neighbors as $N(u)$, which we refer to as the *(1-hop) neighborhood* of node u.

An ℓ-clique is an ℓ-node complete subgraph, and a triangle is a 3-clique. We denote the number of triangles in which node u participates by $T(u)$. For larger cliques, we use $K_\ell(u)$ where ℓ is the size of clique. Moreover, we use T and K_ℓ to denote the total number of triangles and ℓ-cliques in the entire network.

Background on clustering coefficients. The concept of the node-level clustering coefficient was initially proposed by Watts and Strogatz [50], although the notion of clustering in general has a longer history [38]. We say that a wedge is an *ordered* pair of edges that share exactly one common node, and the common node is called the *center* of the wedge (Figure 1A). A wedge is called *closed* if the edge between two ends of the wedge exists, inducing a triangle. The *local clustering coefficient* of a node u is defined as the fraction of wedges centered at node u that are closed,

$$C(u) = \frac{2T(u)}{d_u \cdot (d_u - 1)}.$$

The denominator $d_u \cdot (d_u - 1)$ is the number of wedges centered at node u, and the coefficient 2 in the numerator comes from the fact that each triangle at node u closes two wedges (two ordered pairs of neighbors). If there is no wedge centered at node u (i.e., the degree of node u is either 0 or 1), the local clustering coefficient is undefined. To measure the overall clustering of the entire network, the *average clustering coefficient* is the mean of local clustering coefficients at all nodes in the network,

$$\bar{C} = \tfrac{1}{|\widetilde{V}|} \textstyle\sum_{u \in \widetilde{V}} C(u),$$

where $\widetilde{V}$ is the set of nodes where the local clustering coefficient is well-defined. When undefined, the local clustering coefficient is sometimes assigned to be zero in calculations of the average clustering coefficient of the network [33]. We refer to Kaiser [21] for a discussion on how this can affect network analyses.

An alternative *global* version of the clustering coefficient is the fraction of closed wedges in the entire network [32, 49]

$$C = \frac{6T}{\sum_{u \in V} d_u \cdot (d_u - 1)}, \tag{1}$$

where the coefficient 6 in the numerator comes from the fact that each triangle contains 6 closed wedges (6 ordered pairs). We call this metric the *global clustering coefficient*.

Both the average and global clustering coefficient are weighted averages of local clustering coefficients. The weight in the global case is the number of wedges centered at each node, which is at the order of the degree squared, and thus places more weight on high-degree nodes. The weight in the average case is uniform amongst all nodes and thus (implicitly) places more weight on low-degree nodes as they outnumber high-degree nodes in real-world networks with heavy-tailed degree distributions. Therefore, global and average clustering coefficients often have substantially different values in real-world networks (including in the networks we will analyze; see Table 1).

3 THE LOCAL CLOSURE COEFFICIENT

Recall that a wedge is an ordered pair of edges that share exactly one common node (i.e., a length-2 path). The common node is called the center of the wedge, and here we define the *head* of this wedge as the other end point (i.e., other node) of the first edge. Now we give the formal definition of local closure coefficient.

Definition 3.1. The ***local closure coefficient*** of node u, denoted $H(u)$, is the fraction of wedges headed at u that are closed:

$$H(u) = \frac{2T(u)}{W^{(h)}(u)}, \tag{2}$$

where $W^{(h)}(u)$ is the number of wedges where u is the head and $T(u)$ is the number of triangles containing u. If there is no wedge with node u being the head, the local closure coefficient is undefined.

Figure 1B illustrates this definition. Each triangle at node u contains two closed wedges headed at node u, as combining either incident edge of u with the opposite edge gives a wedge headed at u. This gives the coefficient 2 in the numerator of Equation (2).

Even though both the local clustering and closure coefficients are natural metrics of local clustering, we find that they may be positively correlated, negatively correlated, or weakly correlated in real-world networks (as given by the Pearson correlation coefficient in Figure 2). Thus, the local closure coefficient captures complementary information on fundamental network clustering structure missed by the classical clustering coefficient, and we explore this further in the remainder of the paper.

Analogous to the local clustering coefficient, we also define an average closure coefficient as follows.

Definition 3.2. The ***average local closure coefficient*** of a graph, denoted by $\bar{H}$, is defined as the mean of the local closure coefficients of nodes with well-defined local closure coefficients,

$$\bar{H} = \frac{1}{|\widetilde{V}^{(h)}|} \sum_{u \in \widetilde{V}^{(h)}} H(u),$$

where $\widetilde{V}^{(h)}$ is the set of nodes with well-defined closure coefficients.

We can also define a *global closure coefficient* akin to the global clustering coefficient in Equation (1), but the two metrics are equivalent—both are simply the fraction of closed wedges in the entire network. The global measurement has in fact been described with a "head-based" closure definition [18], but this has not carried over to the local node-based metric that we study.

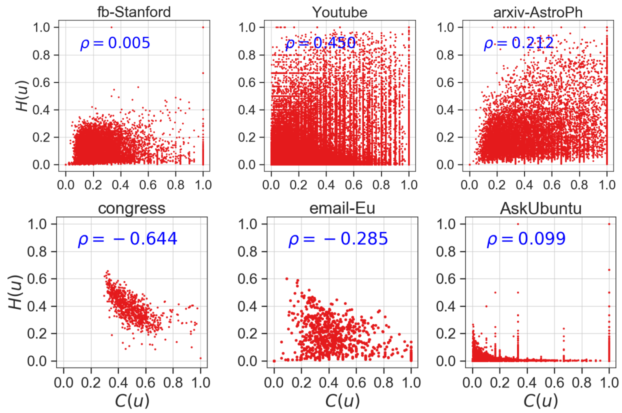

Figure 2: Scatter plots of local clustering coefficients $C(u)$ and local closure coefficients $H(u)$ in real-world networks from different domains, where each point in the plot represents a node in the network, along with the Pearson correlation ρ. While both metrics naturally quantify clustering, the correlation might be positive, negative, or weak, depending on the network, suggesting that the two metrics capture complementary clustering information.

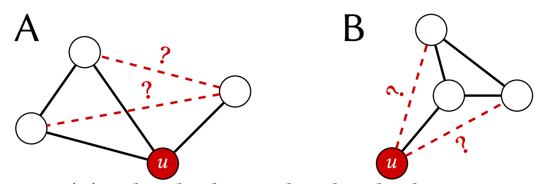

Figure 3: (A) The higher-order local clustering coefficient [53] examines the fraction of (ℓ-clique, edge) pairs with common center node u that induce an $(\ell+1)$-clique (in this figure, $\ell = 3$). (B) In contrast, we define the local higher-order closure coefficient as the fraction of closed wedges where u is the *head* of the wedge, i.e., the node in the (ℓ-clique, edge) pair that is not part of an ℓ-clique.

Efficiently computing the closure coefficient. At first glance, it may seem that computing closure coefficient is computationally expensive, since the number of wedges headed by a node u involves length-2 paths emanating from u. However, computing the local closure coefficient of a node u only requires examining the 1-hop neighborhood structure of node u, provided that one can efficiently access the degree of any node efficiently. For any neighbor $v \in N(u)$, each edge containing v (apart from the edge (u, v)) contributes one wedge headed at node u. Therefore,

$$W^{(h)}(u) = \sum_{v \in N(u)} (d_v - 1) = \sum_{v \in N(u)} (d_v - d_u). \tag{3}$$

Therefore, one can compute the local closure coefficient of every node in the graph in the time to enumerate all of the triangles in the graph. This is the same computational cost of computing the local clustering coefficient of every node.

Higher-order closure coefficients. The perspective of measuring local clustering from the head of each wedge naturally extends to our recent work on *higher-order* clustering coefficients [53]. We use this extension in our theory and experiments in Section 4.2. An *ℓ-wedge* is a pair of an ℓ-clique and an edge that share exactly one node (Figure 3A). The unique node intersecting the ℓ-clique and the edge is the *center* of the ℓ-wedge. An ℓ-wedge is called *closed* if the $\ell+1$ nodes in the ℓ-wedge induce an $(\ell+1)$-clique. The local ℓ^{th}-order clustering coefficient at a node u is defined as the fraction of ℓ-wedges centered at node u that are closed (Figure 3A):

$$C_\ell(u) = \frac{\ell K_{\ell+1}(u)}{K_\ell(u) \cdot (d_u - \ell + 1)}.$$

(Each $(\ell+1)$-clique contains ℓ different ℓ-wedges centered at u, and the total number of ℓ-wedges centered at u is $K_\ell(u) \cdot (d_u - \ell + 1)$.)

We define the head of an ℓ-wedge to be the node that is the other end point of the edge in the ℓ-wedge, and we define the *ℓ^{th}-order local closure coefficient of node u* as the fraction of ℓ-wedges headed at u that induce an $(\ell+1)$-clique (Figure 3B),

$$H_\ell(u) = \frac{\ell K_{\ell+1}(u)}{W_\ell^{(h)}(u)}.$$

When $\ell = 2$, the local higher-order closure coefficient is equivalent to the local closure coefficient defined earlier. For notational convenience, we omit the subscript ℓ in this case.

Next, we derive a formula similar to Equation (3) to show how to efficiently compute $W_\ell^{(h)}(u)$. For any ℓ-wedge headed at u, its center node, denote by v, must be a neighbor of node u. This ℓ-wedge corresponds to an ℓ-clique at node v in which u does not participate. Denote the number of ℓ-cliques containing u and v by $K_\ell(u,v)$. Then for any 1-hop neighbor v, there are $K_\ell(v) - K_\ell(u,v)$ ℓ-wedges of which u is the head and v is the center. Thus,

$$\begin{aligned} W_\ell^{(h)}(u) &= \textstyle\sum_{v \in N(u)} \left[K_\ell(v) - K_\ell(u,v)\right] \\ &= \textstyle\sum_{v \in N(u)} \left[K_\ell(v) - (\ell-1)K_\ell(u)\right]. \end{aligned} \qquad (4)$$

The last equality comes from the fact that $\sum_{v \in N(u)} K_\ell(u,v) = (\ell-1)K_\ell(u)$, since any ℓ-clique at node u contains $\ell-1$ nodes of u's neighbors. Therefore, the computational cost is the same as computing the ℓ^{th}-order clustering coefficient [53].

Summary. Thus far, we have defined (higher-order) local closure coefficients, given efficient ways of computing them, and provided evidence that they are complementary metrics to the classical clustering coefficients. The next section highlights interesting properties and applications of closure coefficients.

4 EXPERIMENTS AND ANALYSIS

In this section, we analyze closure coefficients and provide some case studies for their applications. We first study correlation with node degree, where closure coefficients have remarkably different characteristics than clustering coefficients, and we provide theoretical justification for this fact. We then show how closure coefficients explain the existence of community structure in graphs through a theorem connecting conductance of 1-hop neighborhoods to closure coefficients and use this as a principled heuristic for identifying good seed nodes for personalized PageRank community detection. Finally, we show that, compared to clustering coefficients, closure coefficients more strongly correlate with the temporal properties of triadic closure in networks and use this as motivation for improving link prediction with closure coefficient features.

Data. For our analysis and experiments, we examine the following real-world networks (Table 1 lists summary statistics):

Table 1: Summary statistics of networks: the number of nodes (n), the number of edges (m), the degree assortativity coefficient (r) [31], the global clustering coefficient (C), the average clustering coefficient ($\bar{C}$), and the average closure coefficient ($\bar{H}$) defined in this paper. Asterisks (*) mark datasets containing timestamps on edge creation.

Network	n	m	r	C	$\bar{C}$	$\bar{H}$
fb-Stanford	11,621	568K	0.102	0.157	0.253	0.103
Flickr*	584K	2,257K	-0.050	0.122	0.380	0.039
YahooAnswers*	598K	1,301K	-0.014	0.011	0.107	0.009
Youtube*	3,224K	9,377K	-0.064	0.001	0.169	0.013
LinkedIn*	6,881K	29,162K	-0.026	0.068	0.229	0.016
congress	871	79,886	-0.001	0.424	0.499	0.386
arxiv-AstroPh	18,772	198K	0.206	0.318	0.677	0.250
DBLP*	1,282K	5,180K	0.104	0.172	0.735	0.224
email-Enron*	141	1,414	-0.232	0.285	0.388	0.196
email-Eu*	986	16,064	-0.026	0.267	0.450	0.153
msg-College*	1,899	13,838	-0.188	0.057	0.138	0.022
AskUbuntu*	157K	456K	-0.155	0.011	0.184	0.001
SuperUser*	192K	715K	-0.101	0.011	0.188	0.001
StackOverflow*	2,584K	28,183K	-0.103	0.011	0.123	0.001

(i) *Online friendship networks.* fb-Stanford: the Facebook friendships at Stanford University [46]; Flickr: the friendships network at `flickr.com`, a photo-sharing website [24]; YahooAnswers: the friendship network of `answers.yahoo.com`, a Q&A website [24]; Youtube: the friendship network at `youtube.com` [30]; LinkedIn: the links in the LinkedIn professional network from 2003 to 2006 [24].
(ii) *Collaboration networks.* congress: a co-committee membership network of United States congresspersons [37]; arxiv-AstroPh: a co-authorship network derived from arXiv submissions on Astrophysics [27]; DBLP: a co-authorship network derived from DBLP [1].
(iii) *Communication networks.* msg-College: a Facebook-like messaging network between college students [35]; email-Eu and email-Enron: two email networks [22, 27].
(iv) *Online question-and-answer networks.* AskUbuntu, SuperUser, and StackOverflow [36]: networks derived from Stack Exchange websites where nodes represent users and edge represents interactions between two users (answer or comment).

4.1 Correlation with node degree

A natural first question is how metrics of local clustering vary with node degree. This is fundamental in network analysis; for example, it is well-known that the average clustering coefficient tends to decrease with node degree [24, 40, 44, 45] and the same is true for higher-order clustering coefficients [53].

We start our analysis with empirical evaluations, which motivates our subsequent theoretical analysis. In contrast with the clustering coefficient, we observe the opposite behavior with the closure coefficient. Specifically, the average closure coefficient tends to *increase* as the degree of the node increases, even when the clustering coefficient decreases (Figure 4).

Both the local clustering and closure coefficients are normalized number of triangles at each node, and the major difference in the

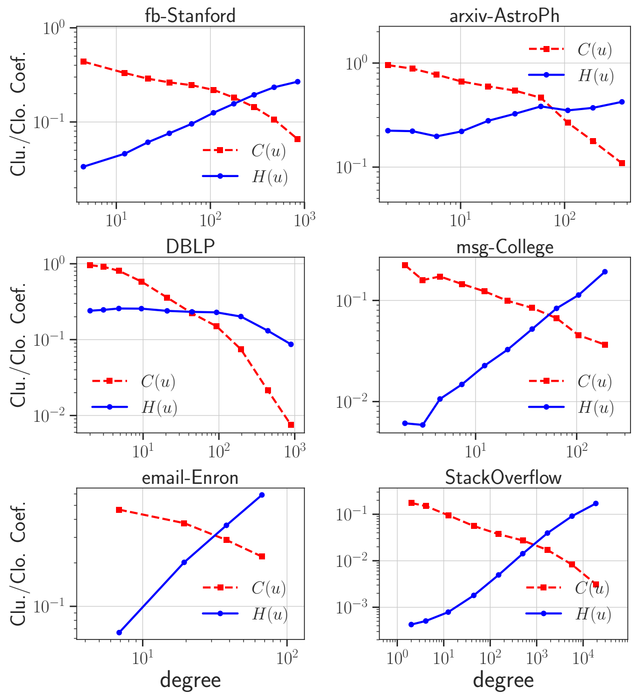

Figure 4: Correlation of local clustering and closure coefficients with node degree. We group nodes by logarithmic binning and show the average degree with average clustering and closure coefficients in each bin. We also include the Pearson correlation ρ between degree and the local clustering and closure coefficients.

correlation with node degree comes from how we apply the normalization. With the local clustering coefficient, normalization is by the number of wedges centered at the node, which is on the order the square of degree. On the other hand, normalization with the local closure coefficient is by the number of wedges headed at the node, which is the sum of degrees of its neighbors. Since the degree distribution of many real-world networks are heavy-tailed (including the ones we are studying), the neighbors of large degree nodes cannot all have large degree, and the degree of a high-degree node's neighbor is likely smaller than the degree of the node itself. Thus, the number of wedges headed at a node grows more slowly than the square of the degree, leading to larger closure coefficients at high degree nodes and smaller ones at low degree nodes.

To better understand why the local closure coefficient increases with node degree, we provide a theoretical justification under the configuration model. The configuration model uniformly at random sample graphs with a prescribed degree sequence, and is a standard tool to analyze the behavior of measures and patterns on networks [11, 29]. One way to implement the configuration model is via stub matching [11]: first generate stubs (half-edges) at each node, where the number of stubs is the same as the node's degree in the given degree sequence, and then match these stubs uniformly at random. In the configuration model, any stub pair is matched with the same probability, so the probability of forming an edge between nodes u and v is of the order of $d_u \cdot d_v/(2m)$. The following result says that as graphs grow large, the closure coefficient in the configuration model increases with degree.

Theorem 4.1. *Let $S = [d_u]_{u\in V}$ be a given degree sequence and G be a random graph generated from the configuration model with S, the local closure coefficient of any node u satisfies*

$$\mathbb{E}[H(u)] = \frac{\bar{k}-1}{2m} \cdot (d_u - 1) \cdot (1 + o(1))$$

as $n \to \infty$, where $\bar{k} = (\sum_v d_v^2)/(\sum_v d_v)$ is the expected degree if we randomly choose a node with probability proportional to its degree.

Proof. For any wedge headed at node u, we denote the node on the other end by v. This wedge is closed if there is an edge between u and v, and thus the probability is $(d_u - 1)(d_v - 1)/2m \cdot (1 + o(1))$. Note that here we need to subtract by 1 because one stub of node u (and v) has already been used in forming the wedge.

Now we show that $\mathbb{E}[d_v] = \bar{k}$. Since any node v has d_v stubs to match in forming the second edge of the wedge, the probability of v being the other end of wedge is proportional to d_v, and thus being $\frac{d_v}{\sum_{v\in V} d_v}$. Therefore, $\mathbb{E}[d_v] = \sum_{v\in V} d_v \cdot \frac{d_v}{\sum_v d_v} = \bar{k}$.

In summary, we have

$$\begin{aligned}\mathbb{E}[H(u)] &= \mathbb{E}\left[(d_u - 1)(d_v - 1)/(2m)\right] \cdot (1 + o(1)) \\ &\sim \frac{\mathbb{E}[d_v] - 1}{2m} \cdot (d_u - 1) = \frac{\bar{k}-1}{2m} \cdot (d_u - 1).\end{aligned}$$

□

Theorem 4.1 shows that the expected value of the local closure coefficient under the configuration model increases with node degree. Quantitatively, it states that expected value is proportional to the node degree, and thus it increases linearly at slope 1 under the log-log axes scaling (as is in Figure 4). In real-world networks, even though we observe the increasing relationship (Figure 4), the slope of the line is sometimes smaller than 1 (such as in the Facebook friendship network and the two co-authorship networks). Indeed, the configuration model is a simple null model that preserves the degree distribution but does not necessarily preserve other properties which may affect the correlation between degree and closure coefficient, such as the degree assortativity [31]. Many real-world networks, such as friendship networks and collaboration networks, exhibit degree assortativity (the degree assortativity coefficient $r > 0$ as is listed in Table 1), meaning that large-degree nodes are more likely to connect with large-degree nodes. Consequently, large-degree nodes are heads at more wedges than expected, and thus have lower local closure coefficient than expected under the configuration model. In comparison, in two communication networks and the StackOverflow network (Figure 4) with degree dissortativity ($r < 0$), the slopes of the lines are close to 1.

4.2 Connections to community detection

In this section, we connect local closure coefficient to graph clustering and community detection. We generalize and strengthen recent results on the connection between the global (higher-order) clustering coefficient and the existence of 1-hop node neighborhoods with small (clique motif) conductance. Our proofs are also much simpler than technical results from prior work. Our new theoretical results explain why previous empirical results work much better than predicted by prior theory. We also use our theory to find good seeds for personalized PageRank community detection.

Background: community detection, (motif) conductance, and 1-hop neighborhoods. The community detection problem aims to find a subset of nodes $S \subset V$ that is densely connected inside while separated from the rest of the network [10]. This problem is often formulated as finding a subset of nodes with small *conductance* [43]:

$$\phi(S) = \frac{\text{cut}(S)}{\min(\text{vol}(S), \text{vol}(\bar{S}))}. \tag{5}$$

Here cut(S) is the number of edges with one end point in S and the other end point in the complement set $\bar{S} = V \backslash S$, and $\text{vol}(S) = \sum_{u \in S} d_u$ is the number of edge end points in S. When $\text{vol}(S) \le \text{vol}(\bar{S})$, the conductance measures the ratio of the number of edges leaving S to the number of edges in S. Small conductance indicates a tightly knit community and is recognized as one of the most important graph clustering criterions [43]. Conductance minimization is also effective at capturing ground-truth communities [52].

Some of our recent work generalized the conductance measure to account for network motifs [6]. A network motif M is any chosen small connected graph (such as a triangle), and an *instance* of M in a graph G is some induced subgraph of G that is isomorphic to M. The *motif conductance* for a given motif M is defined as

$$\phi_M(S) = \frac{\text{cut}_M(S)}{\min(\text{vol}_M(S), \text{vol}_M(\bar{S}))}.$$

where $\text{cut}_M(S)$ is the number of instances of M that have at least one end point (i.e., node) in S and at least one end point in $\bar{S}$, and $\text{vol}_M(S)$ is the number of motif instance end points in S (i.e., the sum—over all instances of M—of the number of nodes in each motif instance that are in S). When M is an edge, this definition reduces to the original conductance measure described above. For undirected networks, common motif examples are cliques, and we use $\phi_\ell(S)$ to denote the motif conductance when the motif M is an ℓ-clique.

Prior work upper bounds the smallest conductance (Equation (5)) over all 1-hop neighborhoods in a network by a function of the global clustering coefficient [13]. The bound is low when global clustering is large, formalizing that clustering leads to community structure. Our past work extended this result to motif-based community detection, where the motif is a clique [54]: for any graph with global ℓ^{th}-order clustering coefficient C_ℓ,

$$\min_{u \in V} \phi_\ell(N(u)) \le c(C_\ell) \cdot (1 - C_\ell), \tag{6}$$

where $c(C_\ell)$ is a function that takes value between 1 and 2.

The upper bound is weak but provids motivation for studying 1-hop neighborhoods (or egonets) as communities. Empirically, both studies found *many* 1-hop neighborhoods with small (motif) conductance. In other words, the theory was much weaker than what was observed in practice. We next provide new theory that subsumes Equation (6) and also explains why many 1-hop neighborhoods have small (motif) conductance. Furthermore, our proofs are much simpler than those from prior work [13, 54].

Connecting closure coefficients to low-conductance sets. Prior results upper bounding the neighborhood conductance with the clustering coefficient are proved with a complex probabilistic method. Furthermore, the upper bound only applies to the 1-hop neighborhood with smallest conductance. Moreover, the upper bound far from empirical observations—there are often many 1-hop neighborhoods with small conductance [13, 54].

Using the local closure coefficient, we give a local version of the previous results, showing that the node neighborhood conductance of *any* node is upper bounded by $1-H_\ell(u)$. The upper bound is tight in practice and thus closes the previous theory–practice gap [13, 54]. Furthermore, the prior results follow as a simple corollary to our theorem (see Corollary 4.3). The following theorem provides an upper bound on the motif conductance for any node u in the network, in terms of the local closure coefficient.

Theorem 4.2. *Let u be a node in a graph with $\text{vol}_\ell(N(u)) \le \text{vol}_\ell(V)/2$. Then the motif conductance for the ℓ-clique motif of the 1-hop neighborhood of node u is bounded by one minus the local closure coefficient. Formally,*

$$\phi_\ell(N(u)) \le 1 - H_\ell(u). \tag{7}$$

Proof. For every ℓ-clique that is cut by $N(u)$, it must correspond to an open ℓ-wedge headed at node u. Note that no two ℓ-cliques will correspond to the same open ℓ-wedge, thus

$$\text{cut}_\ell(N(u)) \le (1 - H_\ell(u)) W_\ell^{(h)}(u).$$

Next, we give a lower bound on $\text{vol}_\ell(N(u))$, which counts three types of ℓ-cliques: (i) cliques that are cut by $N(u)$; (ii) cliques in $N(u)$ that do not contain u; and (iii) cliques in $N(u)$ containing u. For the first type, note that each ℓ-clique cut by $N(u)$ will contribute at least 1 into $\text{vol}_\ell(N(u))$. For the second type, each of clique contributes ℓ to $\text{vol}_\ell(N(u))$, and also corresponds to ℓ different closed ℓ-wedges headed at node u. Therefore,

$$\begin{aligned} \text{vol}_\ell(N(u)) &\ge \text{cut}_\ell(N(u)) + \tfrac{1}{\ell} \cdot \ell \cdot H_\ell(u) W_\ell^{(h)}(u) + \ell \cdot K_\ell(u) \\ &\ge \text{cut}_\ell(N(u)) + H_\ell(u) W_\ell^{(h)}(u), \end{aligned}$$

and consequently,

$$\begin{aligned} \phi_\ell(N(u)) &\le \frac{\text{cut}_\ell(N(u))}{\text{cut}_\ell(N(u)) + H_\ell(u) W_\ell^{(h)}(u)} \\ &\le \frac{(1-H_\ell(u)) W_\ell^{(h)}(u)}{(1-H_\ell(u)) W_\ell^{(h)}(u) + H_\ell(u) W_\ell^{(h)}(u)} = 1 - H_\ell(u). \end{aligned}$$

□

Prior results (Equation (6); [13, 54]) are now a simple corollary of Theorem 4.2.

Corollary 4.3. *For any graph with global ℓ^{th}-order clustering coefficient C_ℓ, $\min_{u \in V} \phi_\ell(N(u)) \le 1 - \max_{u \in V} H_\ell(u) \le 1 - C_\ell$.*

Proof. Note that $\max_{v \in V} H_\ell(u) \ge C_\ell$ because the global clustering coefficient is a weighted average of the local closure coefficient at all nodes (i.e., weighted by the number of wedges headed at each node). Therefore, it must be no greater than the largest local closure coefficient. Let u^* be the node with the largest closure coefficient. Then $\min_{u \in V} \phi_\ell(N(u)) \le \phi_\ell(N(u^*)) \le 1 - H_\ell(u^*) \le 1 - C_\ell$. □

Our result using the local closure coefficient contrasts with the prior results in following ways: (i) the bound is stronger, (ii) there is a bound for every node, and (iii) the proof is much simpler and more informative. From part (ii), we can understand why many 1-hop node neighborhoods have small (motif) conductance. Simply put, there are several nodes u with large values of $H_\ell(u)$.

Empirical validation. We first validate our theoretical results by computing the conductance and 3-clique motif conductance for Flickr and DBLP (Figure 5). The upper bound in Theorem 4.2 is

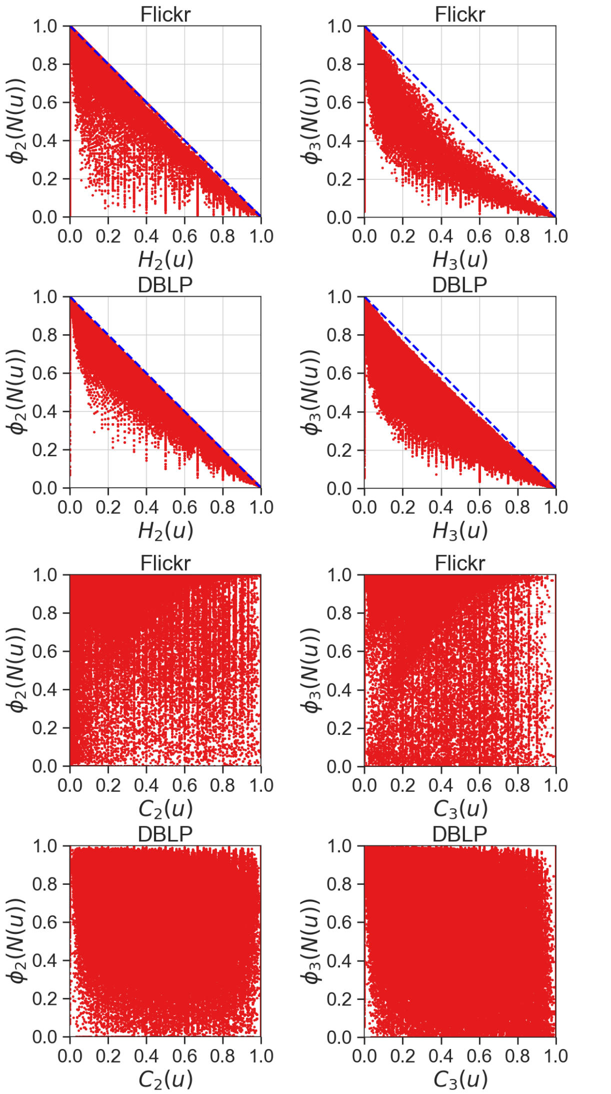

Figure 5: (First two rows) Scatter plots of the local closure coefficient and the conductance (left) or 3-clique motif conductance (right) of the 1-hop neighborhood. Each red dot is one node in the network, and the blue line is the upper bound from Theorem 4.2. The upper bound is tight for many nodes in the network. (Last two rows) Scatter plots of the classical clustering coefficient and 1-hop neighborhood conductance, from which we see no connection between the two metrics.

tight for many values of $H_\ell(u)$ in practice. The last two rows of Figure 5 also show the relationship between the local clustering coefficient of nodes and the conductances of their 1-hop neighborhoods. In contrast to the closure coefficient, we see little correlation between these features.

One implication and application of Theorem 4.2 is the identification of good seeds in local clustering, which is the problem of finding communities in the network that contains a specified seed node [19]. A key problem in local clustering is how to find many good seed nodes that find clusters of low (motif) conductance. Previous work proposes to *local minima*, which are nodes whose 1-hop neighborhood has conductance smaller than all of its neighors. Here we propose a simpler method, which is to use nodes whose closure coefficient are higher than all of its neighbors; we call these *local maxima*. We use our motif-based personalized PageRank (MAPPR) algorithm [54] seeded at local maxima in terms of 2nd- and 3rd-order closure coefficient and compare the (motif) conductance of the communities with those resulting from random seeds. A one-sided Mann Whitney U test rejected the null hypothesis, in all networks and at tiny significance level ($< 10^{-16}$), that the motif conductance from local maxima is no less than the motif conductance from random seed, showing that local maxima are better seeds than non-local maxima.

Table 2: Ratio of positive samples (second column) and the test set AUC-PR values of the regularized logistic regression model with and without the local closure coefficient covariate for link prediction (last two columns). Significant improvement/decrease (more than 2%) is put in bold type. We find significant improvement in most networks after adding the closure coefficient feature, especially social friendship networks.

Network	positive ratio	only similarity	add $H(u)$
Flickr	0.0063	0.0487	**0.0498**
YahooAnswers	0.0017	0.0240	**0.0249**
Youtube	0.0012	0.0028	**0.0039**
LinkedIn	0.0060	0.0425	**0.0434**
DBLP	0.0063	0.0191	**0.0197**
email-Enron	0.1108	0.1687	0.1719
email-Eu	0.1069	0.3441	0.3434
msg-College	0.0076	0.0175	**0.0201**
AskUbuntu	0.0004	**0.0248**	0.0241
StackOverflow	0.0005	0.0159	0.0158
SuperUser	0.0007	0.0283	**0.0299**

4.3 Temporal Closure and Link Prediction

As a new metric on local clustering, the closure coefficient provides an additional feature for node-level graph analysis and inference. As an example, we exhibits its utility in the link prediction problem, where we see significant improvement on test set performance, and we explain this improvement by examining the connection of triadic closure with the graph evolution process.

Link prediction is a fundamental problem in network science [28]. Given a network up to time t, one uses the network structure to predict new links that will appear after time t. Traditional methods look for pairs of nodes that are "close" with respect to some measure. Common examples are the number of common neighbors, the Jaccard similarity, or Adamic-Adar similarity [28]. Such methods are "center-based" and do not explicitly look at closing length-2 paths via friend-of-friend closure. The closure coefficient provides a complementary signal, and our experiments below show that including the closure coefficient as a covariate leads to better prediction, especially on friendship networks.

Our experiment procedure is as follows. For each temporal network, we first obtain the snapshot of network at the time when 50% of edges in the final static graph have been created, and then we use

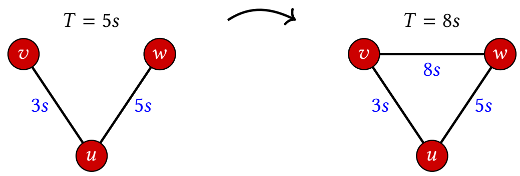

Figure 6: Difference between a wedge in the static graph and a temporal wedge in a network with edge creation timestamps. In the final static graph ($T = 8$, right), any length-2 path, such as (v, u, w) and (u, w, v), is a wedge. Only (v, u, w) is a temporal wedge, since there is a time point ($T = 5$) when the length-2 path (v, u, w) exists without inducing a triangle.

this snapshot to predict the appearance of new edges in the future. Specifically, we obtain the link prediction dataset by, at each node in the snapshot, randomly sampling 20 other nodes that are two hops away in the snapshot. The dataset is unbalanced, and the fraction of positive samples (pairs of nodes between which new edges are created) is listed in the second column of Table 2. After randomly split the dataset into training (80%) and testing (20%), on the training set we fit a regularized logistic regression model with all three similarity measures (number of common neighbors, Jaccard, and Adamic-Adar) and the local closure coefficients to predict the appearance of new edges, and evaluate the prediction performance on the test set. To reduce variance from random train/test split, we repeat the above step 500 times and report the mean AUC-PR value. As a baseline, we use regularized logistic model that uses only the three similarity measures (i.e., without the closure coefficient). Note that here we use the regularized logistic regression model and evaluate the performance on the test set, thus our comparison and analysis is not affected by the increased dimension of features.

We compare the test set AUC-PR value of the baseline model as well as the model with additional closure coefficient covariate in Table 2. We observe significant improvement (at least 2%) in 7 out of 11 networks after introducing the closure coefficient, while in only one of the online Q&A networks we see significant decrease. We observe better improvements in friendship networks where head-driven triadic closure process takes place.

To understand why the closure coefficient improves link prediction performance, we examine the connection of the closure coefficient with the graph evolution process. Both the clustering and closure coefficient are motivated from the perspective of graph evolution—the third edge appears between the two end points of an existing length-2 path to form a triangle.

We define a *temporal wedge* to be a length-2 path such that there exists a time point when the two edges have been created while the two endpoints of the length-2 path does not have an edge between them (Figure 6). We say that a temporal wedge is *closed* if the third edge is created afterwards. Now we define the *temporal clustering coefficient* of a node u, denoted by $\widetilde{C}(u)$, as the fraction of closed temporal wedges centered at u, and the *temporal closure coefficient* $\widetilde{H}(u)$ as the fraction of closed temporal wedges where u is the head.

We compute the local temporal clustering and closure coefficient of each node, and compare them with the clustering and closure coefficient on the final static graph (Figure 7). The local closure coefficient in the static graph is more strongly correlated with the temporal closure coefficient than the local clustering coefficient with the temporal clustering coefficient. For most nodes in every

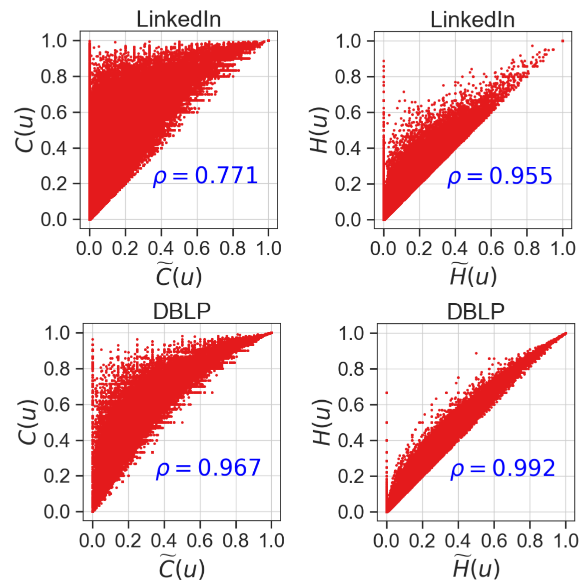

Figure 7: Correlation between the temporal and static clustering coefficient ($\widetilde{C}(u)$ and $C(u)$, left) and the temporal and static closure coefficient ($\widetilde{H}(u)$ and $H(u)$, right) in two networks. Each red dot corresponds to a node, and each scatter plot includes the Pearson correlation ρ. In each dataset, there is a stronger correlation between the temporal and static closure coefficients than that between the temporal and static clustering coefficients. In the scatter plots for closure coefficients, most nodes lie close to the diagonal, indicating that the local closure coefficient is an accurate approximation to the temporal closure coefficient, even though we require no temporal information in calculating the local closure coefficient in the final static graph.

network, the local closure coefficient is almost the same as the temporal closure coefficient even if we have no temporal information in calculating the local closure coefficient in the static graph. In other words, the closure coefficient, as measured from a static graph, more closely captures the temporal dynamics of triadic closure than the clustering coefficient, and thus is potentially useful in network evolution related task such as link prediction.

5 ADDITIONAL RELATED WORK

We now summarize some additional related work, focusing on other metrics of local clustering. Since local clustering coefficient tends to decrease with node degree in most real-world networks (as we saw in Section 4.1), Soffer and Vázquez proposed a modified definition of the local clustering coefficient to remove the degree bias [45]. Local clustering has also been extended to weighted [34], directed [9], and multiplex [7] networks. However, these extensions still measure clustering from the center node, rather than the head node. Most similar to our work is a "closure ratio" used to analyze a copying phenomenon in directed information networks [42]. While, our proposed local closure coefficient is similar in that they both consider the triadic closure processes from the head of a length-2 path, our closure coefficients (i) are defined on static and undirected networks and (ii) are closely connected to the traditional perspective of triadic closure.

6 CONCLUSION

We have introduced the closure coefficient, a simple metric for local clustering that is based on the head, rather than the center, of a length-2 path. The definition closes a gap in the network science literature, which often *describes* triadic closure from the *head* of a length-2 path but *measures* clustering from the *center*. We demonstrated that the local closure coefficient is a complementary feature to the classical local clustering coefficient while also possessing a number of useful and interesting properties.

First, local closure coefficients tend to increase with node degree, providing an explanation for how "popular" (i.e., high-degree) nodes are well-connected locally—a fact which is missed by the local clustering coefficient. We also provided theoretical justification for why this is true. Next, the closure coefficient is also a useful theoretical tool, letting us prove a strong connection between clustering and community detection (as measured by (motif) conductance). Finally, the closure coefficient is also strongly correlated with the temporal dynamics of the graph, which gives more credence to its motivation of capturing local evolutionary processes of triadic closure. This explains why closure coefficients can improve link prediction tasks. These examples demonstrate how closure coefficients are a useful tool for network analysis, and the simple definition and interpretation of closure coefficients should make it easy to incorporate the metric into other graph mining tasks and tools.

ACKNOWLEDGMENTS

This research has been supported in part by ARO MURI, JD, Stanford Data Science Initiative, and Chan Zuckerberg Biohub. ARB was supported by NSF Award DMS-1830274. We thank Johan Ugander for insightful comments, and Mason Porter and Peter Mucha for access to the Congress data.

REFERENCES

[1] Dblp network dataset – KONECT, Apr. 2017.
[2] N. K. Ahmed, R. A. Rossi, J. B. Lee, X. Kong, T. L. Willke, R. Zhou, and H. Eldardiry. Learning role-based graph embeddings. *arXiv*, 2018.
[3] L. Akoglu, M. McGlohon, and C. Faloutsos. Oddball: Spotting anomalies in weighted graphs. In *PAKDD*, 2010.
[4] L. Akoglu, H. Tong, and D. Koutra. Graph based anomaly detection and description: a survey. *Data Mining and Knowledge Discovery*, 2015.
[5] P. S. Bearman and J. Moody. Suicide and friendships among american adolescents. *Am. J. Public Health*, 2004.
[6] A. R. Benson, D. F. Gleich, and J. Leskovec. Higher-order organization of complex networks. *Science*, 2016.
[7] E. Cozzo, M. Kivelä, M. De Domenico, A. Solé-Ribalta, A. Arenas, S. Gómez, M. A. Porter, and Y. Moreno. Structure of triadic relations in multiplex networks. *New Journal of Physics*, 17(7):073029, 2015.
[8] D. Easley and J. Kleinberg. *Networks, crowds, and markets: Reasoning about a highly connected world*. Cambridge University Press, 2010.
[9] G. Fagiolo. Clustering in complex directed networks. *Physical Review E*, 76(2):026107, 2007.
[10] S. Fortunato. Community detection in graphs. *Phys. Rep*, 2010.
[11] B. K. Fosdick, D. B. Larremore, J. Nishimura, and J. Ugander. Configuring random graph models with fixed degree sequences. *SIAM Review*, 60(2):315–355, 2018.
[12] M. Girvan and M. E. Newman. Community structure in social and biological networks. *PNAS*, 2002.
[13] D. F. Gleich and C. Seshadhri. Vertex neighborhoods, low conductance cuts, and good seeds for local community methods. In *KDD*, 2012.
[14] R. Guha, R. Kumar, P. Raghavan, and A. Tomkins. Propagation of trust and distrust. In *WWW*, 2004.
[15] F. Harary et al. On the notion of balance of a signed graph. *The Michigan Mathematical Journal*, 1953.
[16] F. Heider. Attitudes and cognitive organization. *The J. of Psychology*, 1946.
[17] K. Henderson, B. Gallagher, T. Eliassi-Rad, H. Tong, S. Basu, L. Akoglu, D. Koutra, C. Faloutsos, and L. Li. Rolx: structural role extraction & mining in large graphs. In *KDD*, 2012.
[18] M. O. Jackson and B. W. Rogers. Meeting strangers and friends of friends: How random are social networks? *American Economic Review*, 97(3):890–915, 2007.
[19] L. G. Jeub, P. Balachandran, M. A. Porter, P. J. Mucha, and M. W. Mahoney. Think locally, act locally: Detection of small, medium-sized, and large communities in large networks. *Physical Review E*, 91(1):012821, 2015.
[20] E. M. Jin, M. Girvan, and M. E. J. Newman. Structure of growing social networks. *PRE*, 2001.
[21] M. Kaiser. Mean clustering coefficients: the role of isolated nodes and leafs on clustering measures for small-world networks. *New J. Phys.*, 2008.
[22] B. Klimt and Y. Yang. Introducing the enron corpus. In *CEAS*, 2004.
[23] T. LaFond, J. Neville, and B. Gallagher. Anomaly detection in networks with changing trends. In *ODD*^{2} *Workshop*, 2014.
[24] J. Leskovec, L. Backstrom, R. Kumar, and A. Tomkins. Microscopic evolution of social networks. In *KDD*, 2008.
[25] J. Leskovec, D. Huttenlocher, and J. Kleinberg. Signed networks in social media. In *CHI*, 2010.
[26] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graphs over time: densification laws, shrinking diameters and possible explanations. In *KDD*, 2005.
[27] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graph evolution: Densification and shrinking diameters. *TKDD*, 2007.
[28] D. Liben-Nowell and J. Kleinberg. The link-prediction problem for social networks. *Journal of the American society for information science and technology*, 2007.
[29] R. Milo, N. Kashtan, S. Itzkovitz, M. E. J. Newman, and U. Alon. On the uniform generation of random graphs with prescribed degree sequences. *arXiv*, 2003.
[30] A. Mislove, H. S. Koppula, K. P. Gummadi, P. Druschel, and B. Bhattacharjee. Growth of the flickr social network. In *WOSN*, 2008.
[31] M. E. Newman. Assortative mixing in networks. *PRL*, 2002.
[32] M. E. Newman, S. H. Strogatz, and D. J. Watts. Random graphs with arbitrary degree distributions and their applications. *PRE*, 2001.
[33] M. E. J. Newman. The structure and function of complex networks. *SIAM Rev.*, 2003.
[34] T. Opsahl and P. Panzarasa. Clustering in weighted networks. *Social networks*, 31(2):155–163, 2009.
[35] P. Panzarasa, T. Opsahl, and K. M. Carley. Patterns and dynamics of users' behavior and interaction: Network analysis of an online community. *J. Assoc. Inf. Sci. Technol.*, 2009.
[36] A. Paranjape, A. R. Benson, and J. Leskovec. Motifs in temporal networks. In *WSDM*, 2017.
[37] M. A. Porter, P. J. Mucha, M. E. J. Newman, and C. M. Warmbrand. A network analysis of committees in the U.S. House of Representatives. *PNAS*, 2005.
[38] A. Rapoport. Spread of information through a population with socio-structural bias: I. assumption of transitivity. *The Bull. of Math. Biophysics*, 1953.
[39] E. Ravasz and A.-L. Barabási. Hierarchical organization in complex networks. *PRE*, 2003.
[40] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A.-L. Barabási. Hierarchical organization of modularity in metabolic networks. *Science*, 2002.
[41] P. Robles, S. Moreno, and J. Neville. Sampling of attributed networks from hierarchical generative models. In *KDD*, 2016.
[42] D. M. Romero and J. M. Kleinberg. The directed closure process in hybrid social-information networks, with an analysis of link formation on twitter. In *ICWSM*, 2010.
[43] S. E. Schaeffer. Graph clustering. *Computer Science Review*, 2007.
[44] C. Seshadhri, T. G. Kolda, and A. Pinar. Community structure and scale-free collections of erdős-rényi graphs. *PRE*, 2012.
[45] S. N. Soffer and A. Vázquez. Network clustering coefficient without degree-correlation biases. *PRE*, 2005.
[46] A. L. Traud, P. J. Mucha, and M. A. Porter. Social structure of facebook networks. *Physica A: Statistical Mechanics and its Applications*, 2012.
[47] A. Vázquez. Growing network with local rules: Preferential attachment, clustering hierarchy, and degree correlations. *PRE*, 2003.
[48] A. Vázquez, R. Pastor-Satorras, and A. Vespignani. Large-scale topological and dynamical properties of the internet. *PRE*, 2002.
[49] S. Wasserman and K. Faust. *Social network analysis: Methods and applications*, volume 8. Cambridge university press, 1994.
[50] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 1998.
[51] Z.-X. Wu and P. Holme. Modeling scientific-citation patterns and other triangle-rich acyclic networks. *PRE*, 2009.
[52] J. Yang and J. Leskovec. Defining and evaluating network communities based on ground-truth. *Knowledge and Information Systems*, 2015.
[53] H. Yin, A. R. Benson, and J. Leskovec. Higher-order clustering in networks. *Physical Review E*, 97(5):052306, 2018.
[54] H. Yin, A. R. Benson, J. Leskovec, and D. F. Gleich. Local higher-order graph clustering. In *KDD*, 2017.