Using Triangles to Improve Community Detection in Directed Networks

C. Klymko[1], D. F. Gleich[2], T. G. Kolda[3]
[1]Lawrence Livermore National Laboratory
[2]Purdue University
[3]Sandia National Laboratories
klymko1@llnl.gov, dgleich@purdue.edu, tgkolda@sandia.gov

Abstract

In a graph, a community may be loosely defined as a group of nodes that are more closely connected to one another than to the rest of the graph. One common theme is many formalizations is that flows should tend to stay within communities. Hence, we expect short cycles to play an important role. For undirected graphs, the importance of triangles – an undirected 3-cycle – has been known for a long time and can be used to improve community detection. To identify communities in directed networks, we propose an undirected edge-weighting scheme based on directed triangles. We also propose a new metric on quality of the communities that is based on the number of 3-cycles that are split across communities. To demonstrate the impact of our new weighting, we use the standard METIS graph partitioning tool to determine communities and show experimentally that the resulting communities result in fewer 3-cycles being cut. The magnitude of the effect varies between a 10% and 50% reduction. We also find that this weighting scheme improves a task where plausible ground-truth communities are known.

keywords: community detection, directed networks, triangles, reciprocity, 3-cycles

1 Introduction

Many different systems can be viewed as complex networks made up of objects (nodes) and connections between them (links or edges). Over the past several decades the study of such networks has become important in many disciplines [5, 6, 9, 23]. A recurrent research theme is finding the communities or modules within these networks. These communities reveal important structures hidden within the network data.

Thus far, the majority of work in community detection has focused on undirected networks (see the survey [10]), although, recently, more research has focused on directed networks (see the survey [20]). Often, a community is a group of nodes that are more closely connected to each other than to the rest of the network. Community assignment methods use only topological features of the network unless additional information about the components of the network is known; thus the connectivity between nodes is often used alone to define metrics measuring the "quality" of the assigned groups. Common quality metrics measure (i) the density of links within a group (modularity) [21], (ii) the number (or weight) of cut edges relative to the group size (conductance), (iii) the stationary distribution of a random walk within the network (LinkRank) [17], (iv) and the probability of a (directed) link between two nodes [33].

In our paper, we propose a simple weighting scheme that converts a directed graph into a weighted undirected graph. This model enables us to utilize the richness and complexity of existing methods to find communities in undirected graphs. In particular, various schemes have been developed to optimize those four types of connectivity metrics, see [4, 12, 13, 24, 22, 32] among others.

Our specific weighting scheme is based on extending the idea that, within "good" communities, information can be shared within a community more easily than between communities. As information can flow along edges, short (directed) paths and (directed) cycles can be seen as important in the function of communities within networks. For instance, a short path between two nodes indicates that information can travel between them quickly—or can travel quickly from the source to the destination in the case of a directed path. A short cycle indicates that information can travel quickly among a group of nodes and, thus, is even more important in the indicating a community than a short path. The simplest cycle is a path that follows an undirected edge and then returns, and the second simplest cycle is a path that follows a triangle. Consequently, triangles are the basis of many community structures. They also arise because of homophily, the fact that friends of friends are likely to become friends. Directed networks, however, pose a more complicated problem since there are 7 different types of triangles and their contribution to the community structure can be interpreted differently. Here, we focus on the importance of reciprocated edges and directed 3-cycles. Our weighting schemes are designed to increase the weight associated with edges involved in both of these scenarios when a given directed graph is converted to a weighted undirected graph.

The specific contributions of our paper are:

- We introduce a scheme that uses information about directed triangles to improve community detection in directed networks.
- Our scheme involves creating an undirected but *weighted* version of the network, which allows us to utilize the wealth of existing community detection schemes for undirected networks.

 ISBN: 978-1-62561-000-3

- We propose a new metric on the quality of directed communities based on the number of edges contained in 3-cycles split across communities.
- We show up to a 50% reduction in the number of cycles cut in a partitioning into communities compared to simply ignoring the direction of each edge without a meaningful change to existing community metrics.

2 Background and Motivation.

2.1 Directed Triangles

In undirected networks, there is only one type of triangle. Directed networks have seven triangle types (as observed by, for instance, [28]). We show the difference between these seven types in Figure 1. Of these seven triangles, only a few are relevant for community detection. Recall that the reason triangles are important in community detection in undirected networks is that the presence of a triangle indicates a mutually close relationship and ability to share information between three nodes. This is not the case for all types of directed triangles. Only triangles containing a 3-cycle (closed path of length three) enable information sharing among three nodes. The directed triangles which contain a 3-cycle are types in the bottom row of the figure.

2.2 Notation

An undirected network $G = (V, E)$ consists of a set of $|V| = n$ nodes and $|E| = m$ edges consisting of unordered pairs of nodes. In a directed network, the edges are formed by ordered pairs of nodes. Let d_i denote the degree of node v_i in an undirected network. In a directed network, each node v_i has an in-degree, denoted d_i^{in}, which is the number of edges that point into node v_i, an out-degree, d_i^{out}, which is the number of edges pointing out of v_i, and a reciprocal degree, d_i^{rec}, consisting of the number of reciprocal pairs of links in which node v_i is involved. The reciprocal edges do not contribute to the in- and out- degrees of node v_i. The adjacency matrix of a network G is given by:

$$A = (a_{ij}); \quad a_{ij} = \begin{cases} 1 & \text{if } (v_i, v_j) \in E, \\ 0 & \text{otherwise.} \end{cases}$$

If G is undirected, A will be symmetric and if G is directed, A will not be. In the case of a directed network, let A_s be the symmetric part of A and A_{ns} be the nonsymmetric part. Then, the unweighted, undirected representation of G is given by the matrix $A_{ud} = A_s + A_{ns} + A_{ns}^T$. This is equivalent to simply dropping the direction information on each edge in G.

A common quality measure for community assignment is *modularity* [21, 22]. The modularity of a community assignment on an undirected network is given by

$$Q = \frac{1}{2m} \sum_{i,j} \left[A_{ij} - \frac{d_i d_j}{2m} \right] \delta(c_i, c_j)$$

where c_i is the community membership of node v_i and $\delta(c_i, c_j) = 1$ if $c_i = c_j$ and 0 otherwise. Modularity measures the difference between the observed density of edges within communities and the expectation in a random network with the same degree distribution. The concept of modularity can be extended to directed networks through *directed modularity* [2]. The directed modularity of a community assignment is:

$$Q_d = \frac{1}{m} \sum_{i,j} \left[A_{ij} - \frac{(d_i^{\text{out}} + d_i^{\text{rec}})(d_j^{\text{in}} + d_j^{\text{rec}})}{m} \right] \delta(c_i, c_j).$$

2.3 Related Work.

Triangle structure. The existence of triangles has been shown to be important in the formation of complex networks, especially those with an underlying community structure [8, 25, 27, 28]. Additionally, triangle structure has been shown to be important in community detection in undirected networks [3, 25, 26]. For example in [25], the authors define a "good" community to be a group of nodes that is dense in terms of triangles and introduce a community quality metric called *Weighted Community Clustering* (WCC). They experimentally show that a community with high WCC will be denser and contain more triangles than a community with high modularity and low WCC. The use of 3-node motifs (triangles and wedges) to identify communities of different types was introduced in [26]. The authors introduce a generalized version of modularity which takes these motifs into account and a spectral algorithm for approximating its maximum.

Weighting. It is well known that most community finding methods are heuristics and suffer from many potential faults. For instance, communities detected by these schemes can overlook important network characteristics [11]. Weighting schemes often enable simple algorithms to overcome these faults. In undirected networks, weighting edges based on the number of triangles in which the edge is involved has improved the quality of communities [3]. Additionally, this weighting scheme enables the Clauset, Newman, and Moore algorithm to discover communities smaller than the resolution limit of modularity. Other weighting schemes have also been used successfully [16]. In directed networks, the number of reciprocal (bidirectional) edges cut by community assignments has been used as a measure of the "goodness" of a community [19], and a weighting scheme based on the stationary distribution of a directed graph was also shown to generalize the notation of conductance in a network [7].

Algorithms. The majority of the quality optimization algorithms for community detection have been developed for undirected networks, especially those which are efficient enough to be applied to datasets of large size [4, 13, 24]. However, optimizing a community quality measurement (e.g. modularity) on the underlying unweighted, undirected network ignores important information about the direction of the links as we show via the next example.

3 The cycle cut ratio metric

We now introduce a new metric to measure the quality of a directed community assignment:

 ISBN: 978-1-62561-000-3

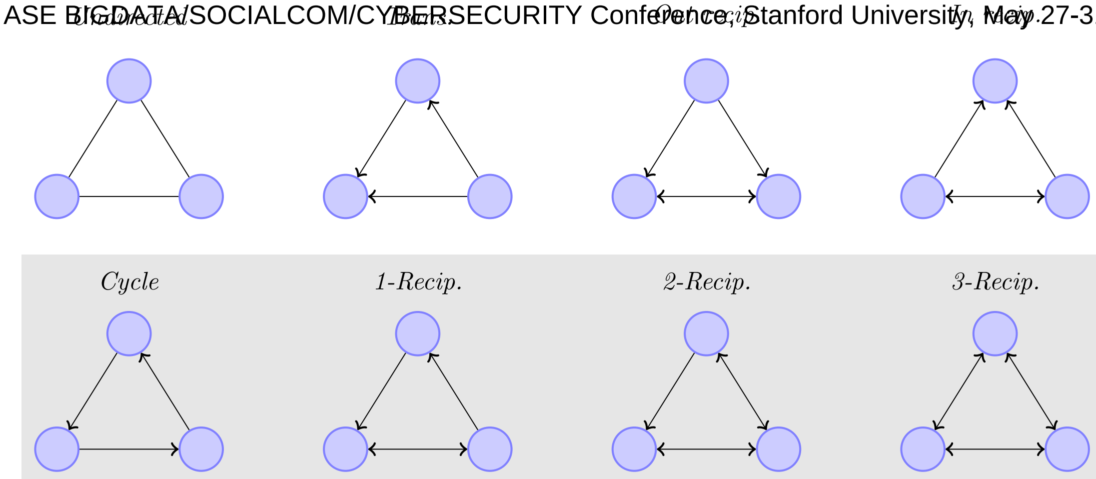


Figure 1: An undirected triangle and the seven types of directed triangles. The four triangles in the bottom row all contain a directed 3-cycle and form the basis of our weighting scheme to indicate community structure in directed networks; the three directed triangles in the top row only indicate partial information flow and we do not use them in our weightings.

Definition (k-cycle cut ratio) *The k-cycle edges of a graph are those that are contained within any directed length k-cycle. Given a partition of the vertices of the network, the k-cycle cut ratio is the fraction of k-cycle edges cut by the partition.*

The measure generalizes the number of reciprocal edges cut in a directed network, which is equivalent to the the 2-cycle cut ratio. Due to the importance of triangles in the formation of network communities, we propose that the 3-cycle cut ratio is an important new metric to evaluate directed communities.

Let us demonstrate this idea through an example. Consider the network in Figure 2. Nodes 1-5 form a clique as do nodes 6-10 and 11-13. Node 15 sits between the cliques. This node can both send and receive information to the first clique (nodes 1-5), but can only send information to the second (nodes 6-10) and can only receive information from the third (nodes 11-13).

Let community A be the community assignment where node 15 is grouped with nodes 1-5 and community B be where node 15 is grouped with nodes 6-10. Intuitively, node 15 should be grouped in a community with nodes 1 through 5 because those are the nodes that node 15 can both send information to and receive information from.

However, the quality of the community assignment measured under either undirected or directed modularity is the same whether node 15 is grouped with nodes 1 to 5 or with nodes 6 to 10. The values for these measures can be found in Table 1. This example shows the importance of measuring how many directed 3-cycles are cut in a community assignment. When only the number of reciprocal edges cut is considered, the two community assignments are the same. By minimizing the number of 3-cycles cut, it becomes clear that node 15 should be grouped with nodes 1-5.

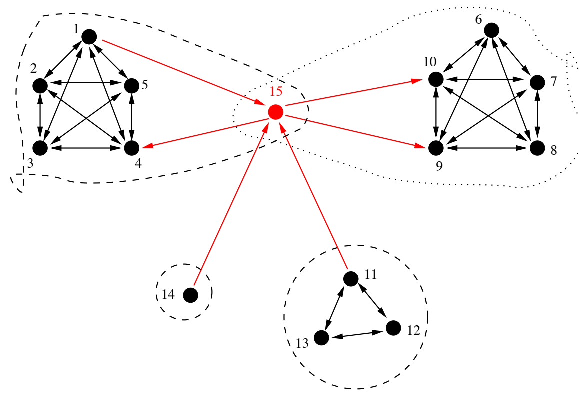


Figure 2: It is not clear whether node 15 should be grouped in a community with nodes 1-5 or with nodes 6-10.

	Q	Q_d
comm A	0.4703	0.5318
comm B	0.4703	0.5318

Table 1: Comparison community quality metrics for two community assignments of the nodes in the network in Figure 2 when edges are unweighted.

4 3-cycle weighting

Now that we have illustrated the importance of 3-cycles in the communities of a network, we wish to design a weighting scheme to turn a directed network into a weighted undirected network with large weights on the edges in the 3-cycles.

To calculate the weights for the 3-cycle weighting, we need to know the type of 3-cycles in which each edge participates. For each edge, we wish to know if it is:

- in a directed 3-cycle?
- in a directed 3-cycle with one reciprocal edge?
- in a directed 3-cycle with two reciprocal edges?
- in a directed 3-cycle with three reciprocal edges?

 ISBN: 978-1-62561-000-3

Let r_0, r_1, r_2, r_3 be indicator vectors over the edges of the graph denoting whether that edge is involved in any of these cases. Once we have this information, we compute a weighting vector over the edges:

$$w = \max(4r_3, 3r_2, 3r_1, 2r_0, 1).$$

Once we have this weight for each directed edge, we convert to an undirected weighted network by taking the maximum weight of edge (i, j) and (j, i). (Any edge that does not exist has weight 0.)

A simple strategy to compute these vectors begins by building a list of all directed cycles in the network. For each cycle, we then check on the number of reciprocal edges, and then update indicator vectors r_0, r_1, r_2, r_3 for each edge involved in the triangle. Our implementation uses this strategy, however, we walk through the list of directed 3-cycles algorithmically without writing out an explicit list. This simplification greatly accelerates the computation as there are often an incredibly large number of directed cycles, and building an explicit in-memory list is expensive. The strategy to walk the list implicitly starts with a directed edge (s, t), indexes the set of out-neighbors of t, and then searches the set of in-neighbors of s for any common vertex with the out-neighbors of t. We then check the reciprocal status for each edge in this cycle and update the appropriate vector.

5 Empirical Evaluation

Our goal is to evaluate if our new 3-cycle based weighting scheme yields improved communities of directed networks. Thus, we examined the effects of edge-weighting on the directed modularity, the percentage of reciprocal edges cut, and the percentage of edges contained in 3-cycles cut during community partitioning on a number of real world networks. The weightings we evaluate are:

- *unweighted*: The undirected, unweighted network is used:
$$A_{ud} = A_s + A_{ns} + A_{ns}^T.$$
- *reciprocal*: the underlying network is used, with edges that were reciprocal in the original network being given weight 2 and the remaining edges being given weight 1:
$$A_r = 2A_s + A_{ns} + A_{ns}^T.$$
- *3-cycle*: the 3-cycle weighting scheme we proposed in the last section.

We first review the method we use to identify communities in each network, then review the networks we study, and finally show the results of our weightings on each network.

5.1 Community detection

For the task of extracting communities from an undirected, weighted network, we use METIS, a well established and understood community detection method that is easy to adapt to edge-weights on a network.

The METIS software is a high-quality implementation of a multi-level graph partitioning method [15, 1]. It constructs a multi-level representation of an input graph by merging edges to form coarser representations, partitions the coarsest graph, and then propagates and refines the partitioning as it un-coarsens. It was originally designed to yield balanced, minimum edge-cut-style partitions suitable for distributed computing; yet it also produces useful sets for clustering [14] and community detection. For community detection, in particular, METIS is often used as a benchmark or baseline method [31].

Each network was partitioned into 5, 10, 25, 50, and 100 communities using the three weighting schemes.

5.2 Network data

We examined a total of 9 networks from a variety of real-world sources. Basic information on these networks can be found in Table 2; more detail is given in the extended version [18]. If any network had edge weights, we remove them before running our methods. All of the networks (other than the Wikipedia network) can be found in SNAP [29].

The Wikipedia network is made up of the largest strongly connected component of the Wikipedia article-link graph, restricted to pages in categories containing at least 100 pages. We used the Wikipedia article dump from 2011-09-01 [30]. For each page, we also have category annotations from Wikipedia (these indicate topics within the encyclopedia) that we use as a surrogate for communities. We will make this data publicly available when this paper is published. In total, there are 17,364 categories. The average category contains 274 pages and the median category contains 149 pages. The largest category has 418k pages and includes all living people with pages on Wikipedia.

Table 2: Basic network data for the 9 networks that we use in our empirical evaluation. We report the number of vertices n, number of directed edges m, the number of reciprocal edges, and the fraction of reciprocal edges. Then we also list the number of directed 3-cycles with 0, 1, 2, and 3 reciprocal edges.

Network	r	3-cycle	1-recip.	2-recip.	3-recip.
amazon0505	0.547	623	67k	809k	837k
Celegans	0.168	72	179	148	16
soc-Epinions1	0.405	7.7k	84k	328k	160k
soc-Slashdot0902	0.854	92	10k	77k	406k
wb-cs-Stanford	0.476	185	470	2197	6898
web-BerkStan	0.250	177	2185	12k	72k
web-NotreDame	0.517	9.5k	41k	107k	6,780k
wiki-Vote	0.057	6.8k	18k	15k	2.1k
Wikipedia	0.215	553k	2,529k	3,929k	1,091k

5.3 Results

The directed modularity, number of reciprocal edges cut, and number of edges in 3-cycles cut in the networks under examination were calculated with the three weighting schemes. Detailed results can be found in the extended version of this paper [18]. Table 3 reports the percentage decrease in the number of 3-cycles cut under the reciprocal and 3-cycle based weighting schemes when compared to the unweighted partition for the number of communities that had the best overall weighting scheme (as measured by directed modularity); Figure 3 displays the change in 3-cycle ratio visually.

The first aspect of our data that we wish to highlight is that the weight schemes do not appreciably change the directed modularity scores. These indicate that the partitions we produce through this scheme have not greatly reduced the quality of the communities by traditional community detection metrics. The largest drop in directed modularity between the two sets of communities occurs in the soc-Epinions1 network, where the directed modularity decreases by 0.0130. This decrease is not especially significant.

Table 3: The percentage decrease in the number of 3-cycles cut under reciprocal and 3-cycle weighting as compared to the unweighted case for the community partition with the highest directed modularity. The results show that our weighting scheme can be highly effective in networks such as web-NotreDame.

	Network	# comm	reciprocal weighting	3-cycle weighting
1	amazon0505	50	15.28%	32.77%
2	Celegans	5	11.20%	40.00%
3	soc-Epinions1	25	2.78%	14.44%
4	soc-Slashdot0902	25	2.37%	0.64%
5	wb-cs-Stanford	50	5.69%	20.06%
6	web-BerkStan	25	19.55%	41.50%
7	web-NotreDame	100	53.78%	82.91%
8	wiki-Vote	10	-2.25%	9.63%
9	Wikipedia	25	5.24%	12.34%

The second aspect we wish to note is that, for each of the networks examined, weighting the edges of the network based on participation in 3-cycles significantly decreases the 3-cycle cut ratio. Table 3 presents the relative effects of the reciprocal and 3-cycle based weighting schemes on the 3-cycle cut ratio compared to the unweighted case. For all the networks considered, with the exception of wiki-Vote network, reciprocal weighting also increases the 3-cycle cut ratio, although often not nearly as significantly as 3-cycle weighting; and, for all of the networks considered, 3-cycle weighting improves the 3-cycle cut ratio, often by a very significant margin. In the web-NotreDame network, the 3-cut cycle ratio is improved by over 80% compared to the unweighted case. The only network where the improvement is not significant is the web-BerkStan network, which shows an improvement of only 0.64%. For all the networks examined (with the exception of the wiki-Vote network), the reciprocal edge-based weighting scheme also increases the number of 3-cycle edges which are preserved, although not as significantly as under the 3-cycle based weighting.

We refine our understanding of this change in Figure 3. The results presented in Figure 3 demonstrate that weighting edges based on participation in 3-cycles improves the ratio of 3-cycles cut in almost all partitions. The exceptions are the wb-cs-Stanford network when 5 and 10 communities are considered and the web-BerkStan and wiki-Vote networks when 100 communities are considered. In all other cases, weighting edges by participation in 3-cycles reduces the 3-cycle cut ratio, often very significantly.

The third, and final aspect, we wish to mention that in the majority of cases (except wb-cs-Stanford), weighting edges based on participation in 3-cycles also significantly reduces the number of reciprocal edges cut. For the Celegans, web-BerkStan, web-NotreDame, and wiki-Vote networks, the 3-cycle based weighting scheme results in the greatest number of reciprocal edges being preserved. For the amazon0505, soc-Epinions1, soc-Slashdot0902, and Wikipedia networks, the number of reciprocal edges preserved increases from the unweighted case but is lower than under the reciprocal edge-based weighting scheme. The wb-cs-Stanford network is the only network where the number of reciprocal edges preserved decreases from the unweighted case to the case where weights are based on 3-cycles.

5.4 Wikipedia categories

Our final empirical evaluation consists of comparing the communities that result from our three weighting schemes to the categories in Wikipedia. We are primarily concerned with the number of intra-category edges that were cut by each of the schemes. These are the edges that are within the ground-truth communities in Wikipedia, but are separated by our methods. We use the partition of Wikipedia into 25 communities because that had the highest directed modularity. The results are shown in the following table, which lists the number of edges cut by each of the methods, the number of within-community edges cut, the fraction of total cut edges that are within a community, and the improvement in that fraction relative to the unweighted scheme.

	cut edges	cuts within	ratio	
unweighted	10030459	768484	0.0766	0%
reciprocal	9958426	730711	0.0734	4%
3-cycle	10249256	743340	0.0725	5%

Both weighting schemes reduce the number of within-category edges cut and improve the fraction of within-community edges cut. Although the 3-cycle scheme actually cuts more within community edges than the reciprocal scheme, it also makes many more cuts in general, and thus its ratio of within community edges cut is lower. We see these results as evidence that these weighting schemes are effective at improving detection of real world communities.

6 Conclusions.

We have described a simple weighting scheme to improve the detection of communities in directed networks. We have also described a new metric for directed communities, the 3-cycle cut ratio. When we use our weighting scheme to convert a directed network into a weighted undirected network and apply a standard network partitioning tool, we find a substantial reduction in the 3-cycle cut ratio, without any appreciable change in the traditional community detection metrics such as modularity. Due to the simplicity of this approach, and the property that it reduces to a weighted, undirected network that can be analyzed with any new method, we are optimistic that this scheme will be used by others to study directed community detection.

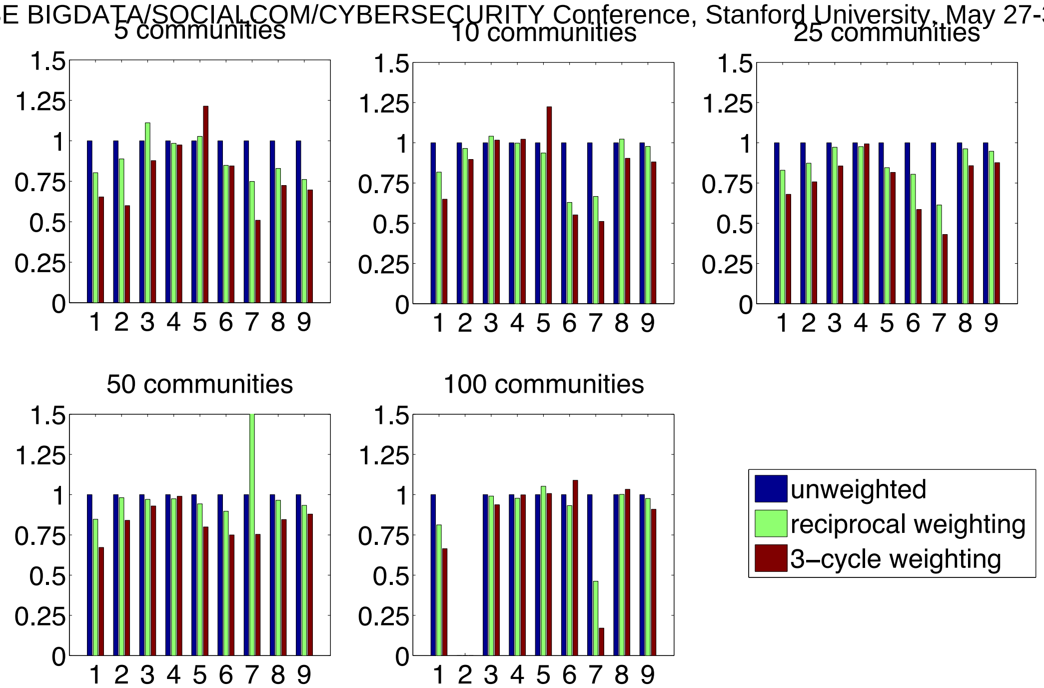


Figure 3: The 3-cycle cut ratio, relative to the unweighted case, for the nine networks considered (labeled according to their number in Table 2) for various numbers of communities. For network number 7 (the web-NotreDame network), the reciprocal weighting doubles 3-cycle cut ratio, which is cut off in the plot. Overall, these results show that we are able to reduce the 3-cycle cut ratio by 10-50%.

There are a few ways to continue investigating this idea. First, it is known that real world communities often overlap. Thus, it would likely be fruitful to extend this method to overlapping community detection. Second, the current set of weights assigned to each edge was not optimized at all; we conjecture that it will be possible to further improve upon our results by optimizing these weights for specific types of graphs. This step, however, requires care not to overtune the weights to a particular type of network. Third, it is possible that using easy-to-compute graph structures such as biconnected components, k-cores, and other well-studied features may enable faster community detection in light of these weights and the directed triangle structure.

Acknwoledgements: The Sandia National Laboratories work was funded by the DARPA GRAPHS program. Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

Gleich's work was supported by NSF CAREER award 1149756-CCF.

References

[1] A. Abou-Rjeili and G. Karypis. Multilevel algorithms for partitioning power-law graphs. In *IEEE International Parallel & Distributed Processing Symposium (IPDPS)*, pages 10–27, 2006.

[2] A. Arenas, J. Duch, A. Fernández, and S. Gómez. Size reduction of complex networks preserving modularity. *New Journal of Physics*, 9:176–190, 2007.

[3] J. W. Berry, B. Hendrickson, R. A. LaViolette, and C. A. Phillips. Tolerating the community detection resolution limit with edge weighting. *Phys. Rev. E*, 83(5):056119, May 2011.

[4] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008.

[5] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang. Complex networks: Structure and dynamics. *Physics Reports*, 424:175–308, 2006.

[6] U. Brandes and T. Erlebach, editors. *Network Analysis: Methodological Foundations, Lecture Notes in Computer Science Vol. 3418.* Springer, New York, 2005.

[7] F. Chung. Laplacians and the Cheeger inequality for directed graphs. *Annals of Combinatorics*, 9(1):1–19, 2005. 10.1007/s00026-005-0237-z.

[8] N. Durak, A. Pinar, T. G. Kolda, and C. Seshadhri. Degree relations of triangles in real-world networks and

graph models. In *CIKM '12*, CIKM '12, pages 1712–1716. ACM, 2012.

[9] E. Estrada. *The Structure of Complex Networks*. Oxford University Press, 2011.

[10] S. Fortunato. Community detection in graphs. *Physics Reports*, 486(3):75–174, 2010.

[11] S. Fortunato and M. Barthélemy. Resolution limit in community detection. *Proc. Natl. Acad. Sci.*, 104 (1):36–41, 2007.

[12] D. F. Gleich and C. Seshadhri. Vertex neighborhoods, low conductance cuts, and good seeds for local community methods. In *KDD'12*, pages 597–605, 2012.

[13] B. H. Good, Y. A. de Montjoye, and A. Clauset. The performance of modularity maximization in practical contexts. *Phys. Rev. E*, 81:046106, 2010.

[14] G. Karypis. CLUTO – a clustering toolkit. Technical report, University of Minnesota, Department of Computer Science, 2002.

[15] G. Karypis and V. Kumar. A fast and high quality multilevel scheme for partitioning irregular graphs. *SIAM J. Sci. Comput.*, 20(1):369–392, 1998.

[16] A. Khadivi, A. A. Rad, and M. Hasler. Network community-detection enhancement by proper weighting. *Phys. Rev. E*, 83:046104, 2011.

[17] Y. Kim, S. W. Son, and H. Jeong. Finding communities in directed networks. *Phys. Rev. E*, 81:016103, 2010.

[18] C. Klymko, D. Gleich, and T. G. Kolda. Using triangles to improve community detection in directed networks. arXiv:1404.5874v1, April 2014.

[19] Y. Li, Z.-L. Zhang, and J. Bao. Mutual or unrequited love: Identifying stable clusters in social networks with uni-and bi-directional links. In *WAW'12: Algorithms and Models for the Web Graph*, pages 113–125. Springer, 2012.

[20] F. D. Malliaros and M. Vazirgiannis. Clustering and community detection in directed networks: A survey. arXiv:1308.0971, Aug. 2013.

[21] M. E. J. Newman. Mixing patterns in networks. *Phys. Rev. E*, 67:026126, 2003.

[22] M. E. J. Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577–8582, June 2006.

[23] M. E. J. Newman. *Networks: An Introduction*. Cambridge University Press, Cambridge, UK, 2010.

[24] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Phys. Rev. E*, 69:026113, 2004.

[25] A. Prat-Pérez, D. Dominguez-Sal, J. M. Brunat, and J. L. Larriba-Pey. Shaping communities out of triangles. In *CIKM'12*, pages 1677–1681, 2012.

[26] B. Serrour, A. Arenas, and S. Gómez. Detecting communities of triangles in complex networks using spectral optimization. *Computer Communications*, 34:629–634, 2011.

[27] C. Seshadhri, T. G. Kolda, and A. Pinar. Community structure and scale-free collections of Erdős-Rényi graphs. *Physical Review E*, 85(5):056109, May 2012.

[28] C. Seshadhri, A. Pinar, N. Durak, and T. G. Kolda. Directed closure measures for networks with reciprocity. arXiv:1302.6220, Feb. 2013. revised Sept. 2013.

[29] SNAP. Stanford network analysis project.

[30] Various. Wikipedia XML database dump from September 1, 2011. Accessed from http://dumps.wikimedia.org/enwiki/, September 2011.

[31] L. Wang, T. Lou, J. Tang, and J. E. Hopcroft. Detecting community kernels in large social networks. Technical report, Cornell University, Tsinghua University, 2011.

[32] J. J. Whang, D. F. Gleich, and I. S. Dhillion. Overlapping community detection using seed set expansion. In *CIKM'13*, 2013.

[33] T. Yang, Y. Chi, S. Zhu, Y. Gong, and R. Jin. Directed network community detection: A popularity and productivity link model. In *SIAM Data Mining'10*, 2010.

2014 ASE BIGDATA/SOCIALCOM/CYBERSECURITY Conference, Stanford University, May 27-31, 2014

Using Triangles to Improve Community Detection in Directed Networks

Christine Klymko[1], David F. Gleich[2], Tamara G. Kolda[3]

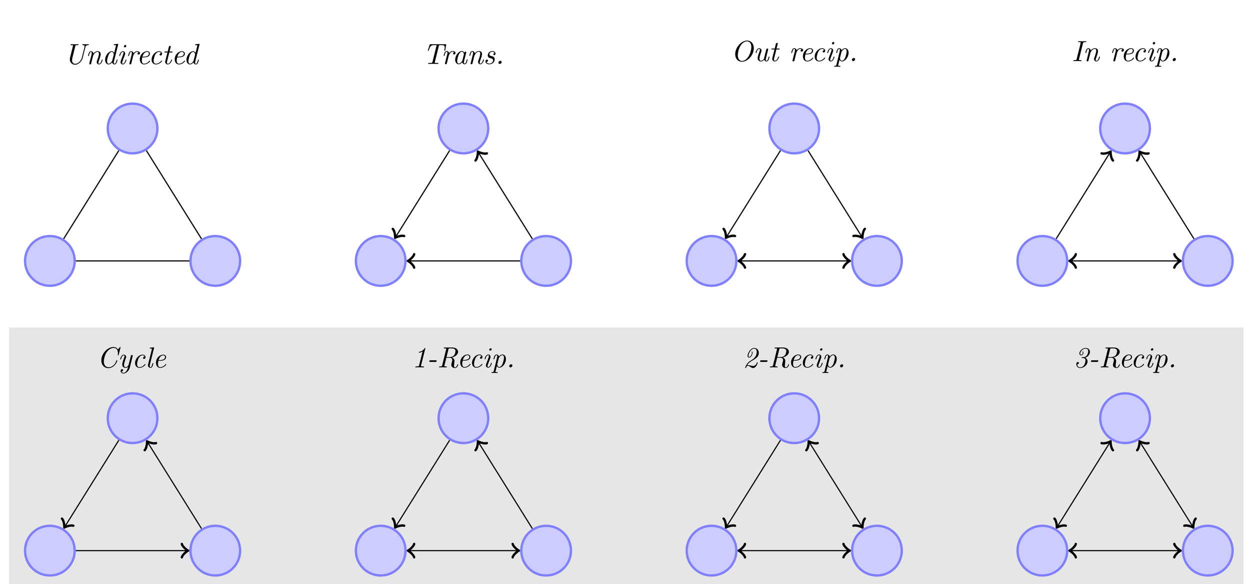
[1] Lawrence Livermore National Laboratory, [2] Purdue University, [3] Sandia National Laboratories

ABSTRACT

In a graph, a community may be loosely defined as a group of nodes that are more closely connected to one another than to the rest of the graph. While there are a variety of metrics that can be used to specify the quality of a given community, one common theme is that flows tend to stay within communities. For undirected graphs, the importance of triangles -- an undirected 3-cycle -- has been known for a long time and can be used to improve community detection. In directed graphs, the situation is more nuanced. The smallest cycle is simply two nodes with a reciprocal connection. Using information about reciprocation has proven to improve community detection. Our new idea is based on the four types of directed triangles that contain cycles. To identify communities in directed networks, then, we propose an undirected edge-weighting scheme based on the type of the directed triangles in which edges are involved. We also propose a new metric on quality of the communities that is based on the number of 3-cycles that are split across communities.

DIRECTED TRIANGLES AND CYCLE WEIGHTING

In a directed network, there are seven types of directed triangles, 4 of which contain 3-cycles. We use a weighting scheme in which edges which participate in these 3-cycles are given large weights.



For each edge, we need the following information:

- is it in a directed 3-cycle?
- is it in a 1-reciprocal cycle?
- is it in a 2-reciprocal cycle?
- is it in a 3-reciprocal cycle?

Let r_0, r_1, r_2, and r_3 be indicator vectors answering these questions. These are used to compute a weighting vector on the edges:

$$w = \max(4r_3, 3r_2, 3r_1, 2r_0, 1)$$

One we have the weights for each directed edge, the network is converted to an undirected network by taking the maximum weight on *(i,j)* and *(j,i)*.

COMMUNITY DETECTION

We evaluate three types of edge weighting:

- **unweighted**: $A_{ud} = A_s + A_{ns} + A_{ns}^T$
- **reciprocal**: $A_r = 2A_s + A_{ns} + A_{ns}^T$
- **3-cycle**: the 3-cycle weighting scheme is used

To extract communities, we use the METIS software[1]. Each network was partitioned into 5, 10, 25, 50, and 100 communities.

CYCLE CUT RATIO METRIC

Definition (*k*-cycle cut ratio): *The k-cycle edges of a graph are those that are contained within any directed length k-cycle. Given a partition of the vertices of the network, the k-cycle cut ration is the fraction of k-cycle edges cut by the partition.*

This measure is a generalization of the number of reciprocal edges cut in a community partition.

The quality of the found communities was measured using: directed modularity[2], LinkRank[3], and the 3-cycle cut ratio.

RESULTS

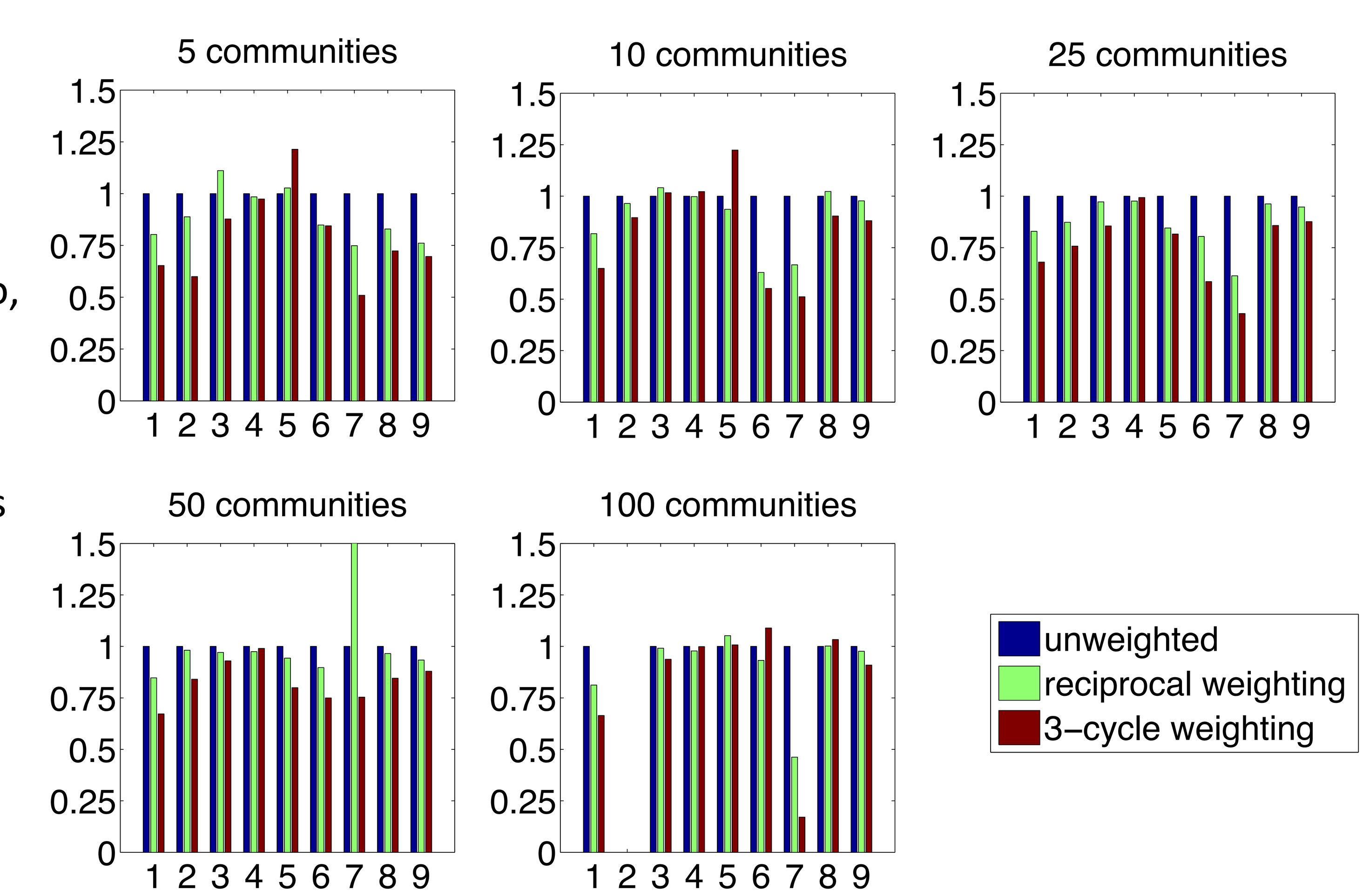
	Network	n	m	recip.	r	3-cycle	1-recip.	2-recip.	3-recip.
1	amazon0505	410K	3,357K	1,835K	0.547	623	67k	809k	837k
2	Celegans	297	2.3K	394	0.168	72	179	148	16
3	soc-Epinions1	76K	509K	206K	0.405	7.7k	84k	328k	160k
4	soc-Slashdot0902	82K	948K	810K	0.854	92	10k	77k	406k
5	wb-cs-Stanford	9.9K	37K	18K	0.476	185	470	2197	6898
6	web-BerkStan	685K	7601K	1902K	0.250	177	2185	12k	72k
7	web-NotreDame	326K	1,470K	759K	0.517	9.5k	41k	107k	6,780k
8	wiki-Vote	8.3K	104K	5.9K	0.057	6.8k	18k	15k	2.1k
9	Wikipedia	2118K	28511K	6131K	0.215	553k	2,529k	3,929k	1,091k

We used 9 directed networks from a variety of real-world sources in our empirical evaluation. Basic data on these networks can be found in the table to the left.

The different weighting schemes do not appreciably change the directed modularity or LinkRank scores of the output communities.

The figure to the right displays the 3-cut ratio, relative to the unweighted case, for the nine networks examined.

Overall, our 3-cycle weighting scheme results in a reduction of the 3-cycle cut ration by 10-50%.



	cut edges	cuts within	ratio	
unweighted	10030459	768484	0.0766	0%
reciprocal	9958426	730711	0.0734	4%
3-cycle	10249256	743340	0.0725	5%

When the weighting schemes are applied to the Wikipedia dataset, we can measure how many intra-category edges which are cut. That is, how many edges within ground-truth communities are cut by each weighting scheme.

The 3-cycle weighting scheme results in the lowest ratio of cut edges being intra-community edges.

CONCLUSIONS

We have introduced a scheme that uses information about directed triangles to improve community detection in directed network. This method creates an undirected, but *weighted*, version of the network that can be input into already existing community detection algorithms for undirected networks.

Additionally, we propose a new metric for measuring the quality of directed networks: the *3-cycle cut ratio*.

Finally, we show up to a 50% reduction on the number of 3-cycles cut in a partitioning into communities of real networks with our weighting scheme compared with simply ignoring edge directionality without compromising other metrics.

FUTURE WORK

- The examination of additional weighting schemes and experimentation to see how this weighting affects other community detection algorithms.
- The extension of this idea to overlapping communities.

REFERENCES

1. A. Abou-Rjeili and G. Karypis. Multilevel algorithms for partitioning power-law graphs. In IEEE International Parallel & Distributed Processing Symposium (IPDPS), pages 10–27, 2006.
2. A. Arenas, J. Duch, A. Fernández, and S. Gómez. Size reduction of complex networks preserving modularity. New Journal of Physics, 9:176–190, 2007.
3. Y. Kim, S. W. Son, and H. Jeong. Finding communities in directed networks. Phys. Rev. E, 81:016103, 2010.

ACKNOWLEDGEMENTS

The Sandia National Laboratories work was funded by the DARPA GRAPHS program. Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

Gleich's work was supported by NSF CAREER award 1149756-CCF.

ISBN: 978-1-62561-000-3