TANGLES IN NETWORKS AND THEIR APPLICATION TO DESCRIBING COMMUNITIES

Michal Salter-Duke, Geoffrey Whittle and Stephen Marsland

SIAM Workshop on Network Science 2020
July 9–10 · Toronto

Summary

A tangle is a structure that defines a highly cohesive region of a graph in a flexible manner [8]. As communities in networks are also highly cohesive regions, tangles may correspond to communities. Here, we describe an algorithm for finding tangles and use it to investigate how well these tangles represent communities.

Tangles

Graph theory defines many cohesive structures, such as grids, cliques and blocks. These structures are defined crisply, and it can be difficult to describe regions that *almost* meet the definitions; tangles are an attempt to give a mathematically rigorous description of cohesive structures that allows for some 'fuzziness' of the boundaries [8]. Our hypothesis is that tangles correspond to communities in networks, in the context of overlapping communities.

Tangles were originally described by separations of edges [8], and also exist in other contexts, but separations of vertices are more natural for describing communities and are used here. Separations are then bipartitions of vertices, the number of edges connecting the parts giving the order of the separation. They are denoted by an ordered pair of their parts, (A, B), whose union is the entire vertex set.

(A, B) is *oriented towards* B and is distinct from (B, A). Oriented separations can be selected so that they are all oriented towards overlapping parts of a single highly cohesive region, and such a selection, obeying certain properties, is called a tangle.

Formally, a tangle $\mathcal{T}$ of order θ in a graph is a set of oriented vertex separations, all of order $< \theta$, conforming to the following three axioms:

1. the tangle contains exactly one orientation of every separation of order $< \theta$.
2. any three separations must be oriented consistently towards some part of the graph: for every $(A_1, B_1), (A_2, B_2), (A_3, B_3) \in \mathcal{T}$ we have $B_1 \cap B_2 \cap B_3 \neq \emptyset$. This provides most of the descriptive power.
3. for any $(A, B) \in \mathcal{T}, |B| > 1$. This prevents trivial tangles.

An algorithm for finding tangles

The current method of finding all the separations is an exhaustive search. This task is specific to separations of vertices, and must be modified for other types of tangles. Once the separations of a given order are identified, they must be oriented, a task which is general to all types of tangles. An empty tree is created to record these orientations. Each separation is tested in sequence against the three axioms, and a branch is added to the tree for each orientation that conforms to all axioms.

Since a tangle is a collection of oriented separations, potentially representing a community, it is necessary to identify the vertices with communities. We have assigned a vertex to a given tangle community if some proportion (the vertex inclusion threshold, either 0.95 or 1) of the separations are oriented towards that vertex.

As lower-order tangles identify more clearly distinct regions, they are more useful for identifying communities, so we only compute tangles of each order up to a predefined maximum (6).

Our current implementation scales poorly with the number of edges, and limits the results presented here.

Results

To assess how closely tangles correspond to communities, we use three criteria for assessing the quality of divisions of a network. These require metadata that reflects the community memberships of each node. In protein-protein interaction networks, the Gene Ontology (GO) annotations [2, 9] provide this. We thus present results from two small protein-protein interaction networks (labelled A and B) from *Saccharomyces cerevisiae* [10].

The three quality metrics we use are:

- Community Similarity: the average similarity of all pairs of nodes in a community divided by average similarity [1]. We use Total Ancestry Measure [11], the probability that two proteins share common GO ancestors.
- Normalised Mutual Information: The mutual information between the GO annotations for each node

and the communities it is assigned to [4].

- Community Coverage: The fraction of nodes assigned to at least one non-trivial community (≥ 3 nodes) [1].

For all metrics, larger values reflect better performance.

In order to judge the tangle algorithm in context, we additionally calculated these metrics for a number of existing overlapping community detection algorithms:

- Line graph methods: a line graph of the original graph is created, then a disjoint community detection method is used [3], with the following options:
  - The line graph was either unweighted (UW) or weighted (W), where the edge weights were scaled by the degrees of the original vertices
  - The disjoint community detection method was either edge-betweenness hierarchical clustering (EB) [6], or modularity optimisation (Mod) [5]
- Clique percolation method (CPM): communities are defined as the unions of adjacent overlapping cliques [7]. Clique sizes of 3 to 6 were used, but for the larger clique sizes, no communities were detected.

Network A			
Algorithm	Sim	NMI	Cover
Tangles (3, 1.0)	3.1328	0.4276	0.55
Tangles (4, 1.0)	3.1304	0.43	0.55
EB, UW	1.6143	0.4777	1
Mod, W	1.6913	0.2967	1
CPM, k=3	3.3709	0.419	0.49

Network B			
Algorithm	Sim	NMI	Cover
Tangles (3, 1.0)	0.9596	0.3107	0.57
Tangles (6, 0.95)	1.5297	0.2839	0.98
EB, UW	1.6454	0.4197	1
CPM, k=3	2.3158	0.4241	0.38

Quality metrics for two protein networks, for the tangle algorithm, parameters shown as (maximum tangle order, vertex inclusion threshold), and for the comparison methods grouped by class. Only the best results for each metric in each class are reported.

Discussion

The metrics for the tangle algorithm are generally comparable to those found using existing methods, although it performs better on the first network. A vertex inclusion threshold of 1 generally gives better results than 0.95, with the exception of similarity for network B, but classifies only part of the network, similarly to clique percolation. The poor performance on similarity for network B appears to be due to a single order 2 tangle which is composed of the entire network excluding the leaf nodes, which has lower average similarity than the network itself. Since the network is small with only a few tangles, this low-similarity tangle has a disproportionate effect, particularly at threshold 1 as there are fewer non-trivial tangles. This effect should disappear for larger networks

While the tangle algorithm is too computationally expensive to provide an practical method for finding communities, these results suggest that communities can indeed be described by tangles, and tangles may thus give insight into the nature of communities. We are undertaking further work to explore the relationship and to improve the efficiency of the algorithm.

References

[1] Y.-Y. Ahn, J. P. Bagrow, and S. Lehmann. Link communities reveal multiscale complexity in networks. *Nature*, 466(7307):761–764, 2010.

[2] M. Ashburner et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, 25(1):25–29, 2000.

[3] T. S. Evans and R. Lambiotte. Line graphs, link partitions, and overlapping communities. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.*, 80(1 Pt 2):016105, 2009.

[4] A. Lancichinetti, S. Fortunato, and J. Kertész. Detecting the overlapping and hierarchical community structure in complex networks. *New J. Phys.*, 11(3):033015, 2009.

[5] M. E. J. Newman. Modularity and community structure in networks. *Proc. Natl. Acad. Sci. U. S. A.*, 103(23):8577–8582, 2006.

[6] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Phys. Rev. E*, 69(2):026113, 2004.

[7] G. Palla, I. Derényi, I. Farkas, and T. Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043):814–818, 2005.

[8] N. Robertson and P. D. Seymour. Graph minors. X. Obstructions to tree-decomposition. *J. Combin. Theory Ser. B*, 52(2):153–190, 1991.

[9] The Gene Ontology Consortium. The gene ontology resource: 20 years and still GOing strong. *Nucleic Acids Res.*, 47(D1):D330–D338, 2019.

[10] H. Yu et al. High-quality binary protein interaction map of the yeast interactome network. *Science*, 322(5898):104–110, 2008.

[11] H. Yu, R. Jansen, G. Stolovitzky, and M. Gerstein. Total ancestry measure: quantifying the similarity in tree-like classification, with genomic applications. *Bioinformatics*, 23(16):2163–2173, 2007.