Cycles in hypergraph-based networks: signal or noise, artefacts or processes?

Lionel Tabourier[1], Jean-Philippe Cointet[2] and Camille Roth[3]

[1] *SPEC, CEA, 91191 Gif-sur-Yvette,* lionel.tabourier@cea.fr
[2] *CREA, CNRS/Ecole Polytechnique, 1 rue Descartes, 75005 Paris,* cointet@poly.polytechnique.fr
[3] *LEREPS (U. Toulouse, 21 alle de Brienne, Toulouse) & CAMS (EHESS/CNRS, Paris),* camille.roth@polytechnique.edu

Les réseaux à structure de groupe sous-jacente induisent naturellement la création de cycles: chaque groupe peut être interprété comme un hyperlien connectant l'ensemble de ses noeuds les uns avec les autres, soit l'ajout d'une clique dans le réseau monoparti projeté. Nous nous intéresserons ici à l'origine des cycles de tailles n ($3 \geq n \geq 5$) associés à des coefficients de clustering généralisés jusqu'à l'ordre 5 (c_3, c_4 et c_5) dans des réseaux à structure de groupe (ou d'hypergraphe) sous-jacente. Ces paramètres topologiques peuvent-ils être expliquées uniquement par le processus spécifique de génération à base d'hyperliens, ou d'autres processus doivent-ils être invoqués? Nous mesurons ainsi ces motifs cycliques sur un ensemble de réseaux réels et distinguons deux catégories de cycles — "structurels" ou "séquentiels" — dont on évalue la part respective en fonction du type de réseau et de n, puis nous estimons la quantité de chaque type de motif obtenue à partir de différents modèles aléatoires de réseaux à base d'hypergraphes, en nous appuyant sur le cadre formel récemment introduit par Mahadevan [MKFV06] et en le prolongeant.

Keywords: réseau biparti, clustering, cycles, modèles de reconstruction, réseaux réels.

Introduction

We focus on networks featuring an *underlying group structure*, a.k.a. *group-based* or *event-based* networks. Affiliation networks, for instance, are such networks: nodes are affiliated with groups (or events), and the corresponding graph is such that links appear between all nodes belonging to a same group (or event). These networks may simply be described either (i) as the monopartite projection of a bipartite graph, where nodes on one side are linked to groups/affiliations on the other side, or (ii) as the projection of a hypergraph where hyperlinks gather nodes belonging to a same group or event.[†]

As such, a *group* or *event* induces *a clique* in the resulting graph. Its structural properties are plausibly influenced by this phenomenon: as a first effect, cliques of size 3 and more automatically inflate the number of cycles of size 3, or "triangles" — in other words, the presence of *clustering* is likely to be significantly influenced by the group-based nature of the network [NSW02]. It seems reasonable to expect that cycles of any length, in general, may simply be due to clique-generation processes, at least in a large part. More broadly, this process may also be responsible for numerous other patterns of interest, as suggested in [MIK$^+$04] — such exhaustivity, however, is beyond our scope in the present paper, and we address the following simple question: to what extent the cyclic structure observed in these networks could be explained by the underlying hypergraph structure?

This issue is strongly similar to the measure of clustering coefficients in graphs, for which several different definitions are also available. In the remainder, we distinguish the *monopartite graph* and its *underlying hypergraph*, the former being the projected graph of the latter; and we define *clustering* as the normalized ratio between the number of triangles $N_\triangle$ and the number of connected triples $N_\wedge$ in the monopartite graph, i.e.: $c_3 = \frac{3.N_\triangle}{N_\wedge}$. This definition can be generalized to longer cycles: we thus note $c_4 = \frac{4.N_\Diamond}{N_\sqcap}$, $c_5 = \frac{5.N_{\pentagon}}{N_{\wedge\wedge}}$, etc. More generally we define the n-order clustering coefficient c_n as the ratio between the number of cycles of length n, and n times the number of *broken cycles* of length n, where a *broken cycle* is defined as a cycle where at most one edge has been removed.

[†] Of course, any graph may be seen as a trivial event-based network where each link is considered as a two-sized group, but our interest lies in networks where groups can have any size, and most often sizes strictly over two and more.

network	N_a	N_g	k_a	k_g	k	$N_\triangle$	$N_{1gr\triangle}$	$N_{seq\triangle}$	$N_\wedge$	c_3
arXiv	16400	19885	2.80	2.31	3.60	17.82	16.31	1.51	231	0.23
Medline	13151	5916	1.77	3.94	6.43	94.17	92.82	1.35	526	0.54
TheyRule	4300	493	1.29	11.22	14.07	110.52	110.32	0.20	537	0.62
DutchElite	395	200	2.22	4.39	9.09	2.93	2.75	0.18	26.6	0.33

Table 1: For each network, number of actors N_a, number of groups N_g, avg. number of groups per actor k_a, avg. size of groups k_g, average degree k in the resulting monopartite graph — number of thousands of triangles N_Δ, of triangles due to a unique event $N_{1gr\Delta}$ or to several events $N_{seq\Delta}$; number of thousands of forks $N_\wedge$ and clustering coefficient c_3.

Measures on real networks

Empirical datasets. We use four networks in our empirical evaluation[‡]. Two are collaboration networks, featuring scientists coauthoring papers (i.e. groups are paper authors): *arXiv*, extracted from preprints on the "arXiv cond-mat" database; and *Medline*, extracted from the "Pubmed" bibliographic archive, using the specific keyword "biomedicine". Two are interlock networks, produced by linking individuals belonging to the same board (i.e. groups are boards): *TheyRule* features the collection of U.S. top companies boards; *DutchElite* gathers affiliations of officials in the main national institutions of the Netherlands. Their basic features are given in Tab. 1.

Structural vs. sequential cycles. Hypergraph-based networks seem to be ubiquitous whenever social mechanisms are at work; in such networks indeed, groups (or events) gather agents thus induce cliques. Cycles in the monopartite graph may thus partly be a *mechanical* feature, in the sense that it is merely caused by the construction of the monopartite graph from an underlying hypergraph.

Nevertheless, non-mechanical processes may also account for the presence of cycles: for example in the case of 3-sized cycles, or triangles, A interacts with B in a group, B interacts with C in another group, and then A interacts with C in a later group — this is usually called "transitivity". In this setting, we thus distinguish two kinds of triangles in the monopartite graph. On one hand, "single-group" or "structural" triangles ($N_{1gr\triangle}$) result (at least) from one single group gathering 3 nodes (or more) *at once*. On the other hand, "sequential triangles" ($N_{seq\triangle}$) are created by a sequence of 3 events, none of them involving the entire triple of nodes (Fig. 1).[§]

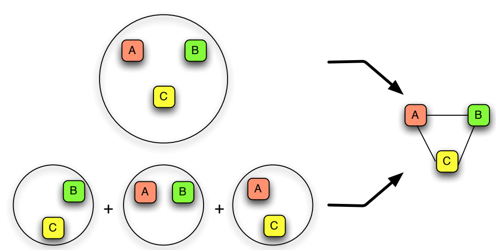

Figure 1: A triangle in the monopartite graph can arise from two kinds of configurations in the underlying hypergraph: on top, single-group triangle; at the bottom, sequential triangle made of three different groups.

In real networks, triangles are massively due to groups (Tab. 1): triangles stemming from a triad of groups are generally rare, and thus structural triangles are responsible for most of the clustering.

The notion of structural or sequential triangles can easily be extended to longer cycles: we may measure the number of diamonds or pentagons (cycles of length 4 or 5) produced by a single group and define sequential diamonds or pentagons as any cycle (of length 4 or 5) which is not based on a unique grouping. Contrary to triangles, results in Tab. 2 show that in most cases the proportion of higher-length sequential cycles is not negligible anymore — their presence may therefore not be explained only by the underlying clique aggregation process, i.e. by the fact that the monopartite graph is based on a hypergraph.

Morphogenesis models

Trivial underlying hypergraph. Since most triangles are structural, it seems plausible that a network model mimicking just the underlying hypergraph structure would lead to the same c_3 clustering coefficient.

[‡] Data available on, respectively: http://www.arxiv.org, http://www.ncbi.nlm.nih.gov/sites/entrez, http://www.theyrule.net, http://vlado.fmf.uni-lj.si/pub/networks/data/2mode/DutchElite.

[§] Here, triangles corresponding to both a single group and a sequence of groups are thus counted, by definition, as "structural", not "sequential", triangles ("there is at least one group involving the entire triple"). Empirically, this mixed case is negligibly rare.

network	$N_\Diamond$	$N_{1gr\Diamond}$	$N_{seq\Diamond}$	$N_\sqcup$	c_4	$N_{\pentagon}$	$N_{1gr\pentagon}$	$N_{seq\pentagon}$	$N_{\wedge\wedge}$	c_5
arXiv	43.5	15.4	28.1	2,060	0.084	159.8	13.0	146.8	20,347	0.039
Medline	717	545	172	7,265	0.39	7,091	4,260	2,831	114,280	0.31
TheyRule	930.8	904.5	26.3	10,374	0.36	8,698	8,095	603	194,680	0.22
DutchElite	14.86	9.89	4.97	375	0.16	103.9	40.2	63.7	5,274	0.10

Table 2: For each network, number of thousands of: diamonds (resp. pentagons) $N_\Diamond$ ($N_{\pentagon}$), diamonds (resp. pentagons) due to a unique event $N_{1gr\Diamond}$ ($N_{1gr\pentagon}$) or to several events $N_{seq\Diamond}$ ($N_{seq\pentagon}$); and broken diamonds (resp. broken pentagons) $N_\sqcup$ ($N_{\wedge\wedge}$) along with the clustering coefficient c_4 (resp. c_5).

Some authors indeed already suggested [NSW02, GL04] that this very feature could be reconstructed by a traditional null-model of bipartite graph (or hypergraph), the Molloy-Reed (MR) model [MR95]. MR generates a random bipartite graph with the same connectivity distributions from one side to the other side of the bipartite graph — in other words, MR generates a hypergraph made of as many hyperlinks of a given size as in the original hypergraph, with nodes belonging to as many hyperlinks as well.

In order to assess how a trivial underlying hypergraph structure may account for the monopartite topological features, we therefore first perform simple MR reconstructions of our 4 empirical cases — in other words, we thus preserve the degree distribution of nodes to groups and the size distribution of groups. Table 3 gathers results concerning both structural and sequential length n cycles (for $n = 3, 4, 5$) on 20 distinct MR realizations, to be compared to original graph values on Tab. 1 and 2.

network	$N_\triangle$	$N_{1gr\triangle}$	$N_\wedge$	c_3	$N_\Diamond$	$N_{1gr\Diamond}$	$N_\sqcup$	c_4	$N_{\pentagon}$	$N_{1gr\pentagon}$	$N_{\wedge\wedge}$	c_5
arXiv	18.7	18.5	518	0.11	19.6	16.2	6,685	0.012	48.9	13.3	86,132	0.0028
Medline	105.3	103.9	1,459	0.22	625	575	38,775	0.064	5,586	4,365	1,031,746	0.027
TheyRule	110.4	110.3	541	0.61	908	905	9,612	0.38	8,175	8,095	171,539	0.24
DutchElite	3.03	2.76	30.2	0.30	16.36	9.89	484	0.14	136.1	40.2	7,633	0.089

Table 3: For each MR-reconstructed network, number of thousands of (i) 3-node patterns: total triangles ($N_\triangle$) and triangles coming from a unique group ($N_{1gr\triangle}$), broken triangles, or forks ($N_\wedge$) and clustering coefficient c_3; (ii) 4-node patterns ($N_\Diamond$, $N_{1gr\Diamond}$, $N_\sqcup$) and c_4; and (iii) 5-node patterns ($N_{\pentagon}$, $N_{1gr\pentagon}$, $N_{\wedge\wedge}$) and c_5.

Because *structural* triangles, diamonds and pentagons are directly induced by groups (which size distribution is the same as in the original network), these values are unsurprisingly acceptably reconstructed by MR graphs. The story is much different for *sequential* cycles, and two classes of networks are exhibited. Interlock networks, on one hand, display acceptable fits for cycles, and broken cycles as well, (in the vicinity of 10% around the empirical value), consistently with partial previous research [RA04]; in this case, these features are plausibly an artefact of the underlying hypergraph structure. Collaboration networks, on the other hand, are not properly reconstructed by a simple hypergraph structure: be it either (i) for cycles of length $n \geq 4$,[¶] often under-estimated by MR-graphs, or (ii) for broken cycles of any length, often over-estimated by MR-graphs (e.g. the number of broken triangles, or forks, is at least twice larger). Consequently, clustering coefficients are not correctly reproduced for these graphs, because of reconstruction failures for both cycles and broken cycles. The limitation of a simple hypergraph-based model for collaboration networks may be typical of non-artefactual, complex social processes. For instance, some kind of social transitivity (transitive creation of new relationships among "friends of friends") may be needed if we want to account for the large value of the clustering coefficients when compared to random networks.

Enhanced underlying hypergraph. We propose to extend the MR model to constraints stronger than just bipartite degree distributions, yet still pertaining to the underlying hypergraph structure, and particularly to distributions of grouping and affilition sizes. In a recent paper, Mahadevan *et al.* [MKFV06] introduced random graph generation methods aiming at reconstructing increasingly more properties of an original input graph G by fitting increasingly detailed correlations on degrees in the original graph. This reconstruction is based on the notion of "dK distributions", where a larger value of d corresponds to more constraining degree correlations. For example, $0K$-graphs only reconstruct the mean connectivity of G, $1K$-graphs reconstruct the original degree distribution, while $2K$-graphs reconstruct the joint distribution of degrees of G, etc.

[¶] Remember that sequential cycles of length 3 are few, hence the correct overall reconstruction of cycles of length 3 (triangles, $N_\triangle$) in collaboration networks might actually also be coincidental.

We will thus elaborate upon this approach by adapting it to bipartite graphs. Note that in this framework, MR is equivalent to a bipartite version of the $1K$-graph, constraining both degree distribution P and Q; where $P = \{p_i\}_i$ is the node degree distribution (i.e. a node has probability p_i to be connected to i groups) and $Q = \{q_j\}_j$ is the group size distribution (i.e. a group has probability q_j to be of size j). By analogy with a bipartite $1K$ reconstruction, we shall denote MR as an "$1BK$-graph".

On the whole, constraints induced by $1BK$-graphs seem too weak to yield a proper reconstruction of collaboration networks. We suspected higher-level correlations between actor degrees and group sizes to play a role in the observed discrepancy, and therefore introduced bipartite versions of $2K$ and $3K$ models, as $2BK$ and $3BK$ models — in the $2BK$ case, e.g., we thus fit degrees at the end of bipartite links.

Nonetheless, these models still failed to account for the number of cycles and broken cycles of collaboration networks. Put differently, the observed cyclic structure seems independent of strictly structural constraints, at least those induced by the first dBK reconstructions only respecting degree correlations ($0BK$, $1BK$, $2BK$ and $3BK$). To be sure, yet, we introduced a slightly modified alternative to $2BK$, called "$2BK'$ reconstruction", aiming at conserving the following probability distribution: $P(\sum_{l\in\mathcal{V}_i} K_l, k_i, K_j, \sum_{l\in\mathcal{V}_j} k_l)$ (where $\mathcal{V}_\rangle$ denotes the (bipartite) neighborhood of node i, k_i the number of groups in which node i takes part, K_j the size of group j). In other words, a $2BK'$-graph preserves the original joint probability distributions $P(p_i, q_j)$ — like in the $2K$ case — and preserves, for each bipartite link (node $v \leftrightarrow$ group g), the sum of degrees of nodes connected to group g and the sum of sizes of groups connected to actor v.

Results of the $2BK'$ reconstruction fit much more satisfyingly original values of collaboration networks — the amount of 3-, 4- and 5-node cycles and broken cycles is now suitable (see Fig. 2), while the ratio of sequential vs. structural cycles is also correctly reproduced. In short, the corresponding topology may well be explained, still, by a simple kind of degree correlations in the underlying hypergraph.

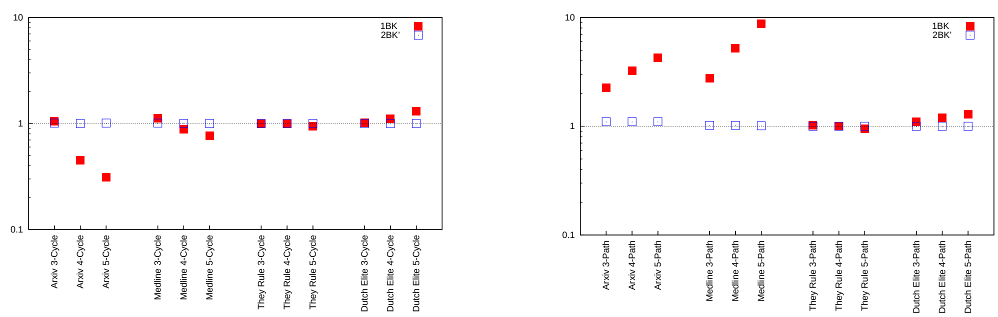

Figure 2: For each network and each pattern (*at left:* 3, 4 and 5-nodes cycles; *at right:* 3, 4 and 5-nodes broken cycles), we compare the ratios between their real value and their $1K$ and $2BK'$ reconstructions (resp., dark and empty boxes).

Conclusion

Focusing on graphs with an underlying hypergraph structure, and using measures on cycles and broken cycles of various length, thus incidentally on clustering coefficients, we show that *most of the corresponding topological features are likely to stem first from structural phenomena*, rather than peculiar processes and interaction behaviors proper to the particular real-world context of the graph. Extensions to other empirical settings would be most fruitful to assess the generality of these results.

References

[GL04] J.-L. Guillaume and M Latapy. Bipartite structure of all complex networks. *IPL*, 90(5):215–221, 2004.

[MIK+04] R Milo, S Itzkovitz, N Kashtan, R Levitt, S Shen-Orr, I Ayzenshtat, M Sheffer, and U Alon. Superfamilies of evolved and designed networks. *Science*, 303(5663):1538–1542, Mar 2004.

[MKFV06] P Mahadevan, D Krioukov, K Fall, and A Vahdat. Systematic topology analysis and generation using degree correlations. *arxiv*, cs.NI, May 2006.

[MR95] M Molloy and B Reed. A critical point for random graphs with a given degree sequence. *Random Structures and Algorithms*, 161(6):161–179, 1995.

[NSW02] M Newman, S Strogatz, and D Watts. Random graphs models of social networks. *PNAS*, 99:2566, 2002.

[RA04] G Robins and M Alexander. Small worlds among interlocking directors: Network structure and distance in bipartite graphs. *Computational & Mathematical Organization Theory*, Jan 2004.