2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining

Quantifying Triadic Closure in Multi-Edge Social Networks

Laurence Brandenberger
Chair of Systems Design
ETH Zürich
Zürich, Switzerland
lbrandenberger@ethz.ch

Giona Casiraghi
Chair of Systems Design
ETH Zürich
Zürich, Switzerland
gcasiraghi@ethz.ch

Vahan Nanumyan
Chair of Systems Design
ETH Zürich
Zürich, Switzerland
vnanumyan@ethz.ch

Frank Schweitzer
Chair of Systems Design
ETH Zürich
Zürich, Switzerland
fschweitzer@ethz.ch

***Abstract*—In social networks, edges often form closed triangles or triads. Standard approaches to measuring triadic closure, however, fail for multi-edge networks, because they do not consider that triads can be formed by edges of different multiplicity. We propose a novel measure of triadic closure for multi-edge networks based on a shared partner statistic and demonstrate that this measure can detect meaningful closure in synthetic and empirical multi-edge networks, where conventional approaches fail. This work is a cornerstone in driving inferential network analyses from the analysis of binary networks towards the analyses of multi-edge and weighted networks, which offer a more realistic representation of social interactions and relations.**

***Index Terms*—multi-edge networks, triadic closure, network inference, social networks, statistical learning**

I. Introduction

Triadic closure denotes the remarkable tendency observed in social networks to form triangles, or triads, between three individuals a, b and c (see Fig. 1) [1]. That means if an edge connects a and c, and b and c, there is a higher probability an edge also connects a and b [2].

The presence of closed triads can be quantified on the topological level by calculating the clustering coefficient. However, usually we want to examine whether triadic closure is an actual mechanism *driving* network formation, or whether observed triads are simply a *consequence* of other relational mechanisms. Similarly, it is essential to know to what extent other network properties of interest, e.g., communities or core-periphery structures, are already determined by triadic closure. For these reasons, triadic closure acts as a control variable in inferential network models. Achieving this, though, requires triadic closure to be included correctly in inferential models.

The concept of triadic closure is well-defined mainly for binary networks. In such networks, measuring triadic closure is done by counting triads. Most real networks, however, reflect *repeated interactions* between individuals, quantified by multi-edges, i.e., each interaction is represented by a separate edge, and multiple edges are incident to the same pair of nodes. For these networks, an operationalization by counting triads faces two problems. First, it neglects crucial information because the number of triads does not appropriately reflect the fact that nodes can repeatedly interact a different number of times. Second, because of these repeated interactions, multi-edge networks often have a high density. Such high density results in a maximum triad count for the network, overrating the fact that some of the nodes only have interacted once. The first investigation of this issue has been performed by [3], where the clustering coefficient has been extended to weighted networks. In this paper, we go beyond previous studies introducing a way of incorporating triadic closure in inferential models for repeated interaction networks.

ASONAM '19, August 27-30, 2019, Vancouver, Canada

ACM ISBN 978-1-4503-6868-1/19/08 $15.00
http://dx.doi.org/10.1145/3341161.3342926

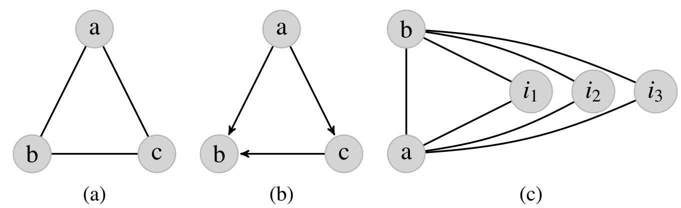


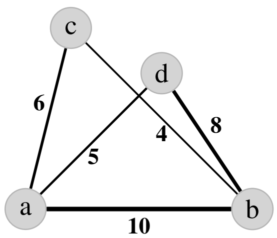
Fig. 1. Triadic closure: (a) undirected triangle, (b) transitive triplet, (c) edge-wise shared partners

II. Methods

A. Weighted Shared Partner Statistics for Multi-Edge Networks

Empirical studies operationalize triadic closure in different ways. In undirected binary networks, triadic closure can be measured by the percentage of triangles in the network. In directed networks, different constellations of incoming and outgoing edges give rise to different interpretations of triadic relationships. The most commonly used closed triad in directed networks is the transitive triplet, which indicates that if one node a has two friends b and c, then either b ties to c or c ties to b (see Figure 1b).

Inferential network models are employed to statistically test whether a social network exhibits significant tendencies for triadic closure. Conventional inferential network models,



	a	b	c	d
a	-	2	1	1
b	2	-	1	1
c	1	1	-	2
d	1	1	2	-

	a	b	c	d
a	-	9	4	8
b	9	-	6	5
c	4	6	-	9
d	8	5	9	-

Fig. 2. *Left:* Example multi-edge network. *Right Top:* Shared partner statistic. *Right Bottom:* Weighted shared partner statistic.

though, build on binary representations of social networks. Triadic closure can then be operationalized through the concept of shared partner statistics (see Figure 1c). This is an edge-based measure that calculates for each edge between nodes a and b how many *shared partners* they have in common [4]. Empirical evidence, in fact, shows that having mutual friends has a cumulative effect on friendship formation and preservation [5], [6], and the statistic, therefore, captures clustering tendencies in social networks more accurately. This conventional approach through shared partner statistics, though, poses a problem for multi-edge networks. The shared partner statistic has to be adapted to incorporate edge-counts in order to detect meaningful triadic closure in multi-edge networks.

Repeated interactions are often considered an indicator of the strength of the relation linking two nodes. [7] postulates that "the more frequently persons interact with one another, the stronger their sentiments of friendship for one another are apt to be" (p. 133). Hence, if repeated interactions are meaningful and indicate edge strength, this strength should be adequately reflected in the measurement of triadic closure. We propose to measure triadic closure in multi-edge networks by employing an edge-based approach. For each dyad (a, b) in network N, we calculate whether a two-path with dyads (a, i) and (b, i) exists; i.e., we calculate whether nodes a and b have shared partners i. It is important to note that regardless of whether or not nodes a and b have interacted (repeatedly), we calculate whether this dyad (a, b) could potentially close a triad. In the ERGM-family, this approach is summarized in so-called change statistics [8], [9]. The resultant matrix holds the values of this endogenous network statistic for each dyad. This way, it captures complex patterns between nodes in a network, without including the state of the focal dyad (a, b). This independence of the dyad state (a, b) allows testing whether the values of the pattern correlate with the number of edges in the multi-edge network.

Since the edge-counts $v(a, i)$, $v(b, i)$ give the number of edges for each two-path (a, i)-(b, i), we incorporate them into the shared partner statistic. Failing to do so would treat triangles with equal numbers of edge-counts and triangles with very different numbers of edge-counts the same, which precisely should be avoided. Furthermore, by incorporating edge counts, the variance of the shared partner statistics becomes broader, which allows for more accurate parameter estimates of the effect of triadic closure on the network structure.

Figure 2 shows a simple multi-edge network with four nodes. The tables on the right side report its binary (top), and weighted shared partner statistics (bottom). For each dyad in the network, the number of shared partners is counted and noted in the top matrix. Instead of counting shared partners, the weighted statistic weights each shared partner by the minimum edge count in the two-path (a, i)-(b, i). Because the sample network is nearly complete (only dyad (c, d) is absent), the unweighted shared partner statistic shows little variance compared to the weighted statistic that accounts for edge-counts. When regressing both shared partner statistics against the edge counts of the multi-edge network, the unweighted statistic explains 1% of the variance, the weighted statistic 4%[1], indicating that more information is stored in the weighted shared partner statistic.

In the next section, we use two different inference models for multi-edge networks (ERGM-count and gHypEG) to assess the degree to which social networks are structured by triadic closure.

B. ERGM

Exponential random graph models (ERGMs) are generative models to estimate the effects of endogenous and exogenous covariates on network formation [10]. The endogenous covariates typically include features of the network, such as degree distributions or triadic closure effects, and exogenous covariates are used to estimate homophily effects or effects of nodal attributes on the activity or popularity of nodes. [11] expanded the ERGM to fit count data by allowing counts to populate dyads (i, j). The model can be expressed as the probability of observing the given network N over all possible permutations $\mathcal{N}$ of the network:

$$P(N, \boldsymbol{\theta}) = \frac{\boldsymbol{l}(N) \exp\{\boldsymbol{\theta}^T \boldsymbol{h}(N)\}}{\sum_{N^* \in \mathcal{N}} \boldsymbol{l}(N) \exp\{\boldsymbol{\theta}^T \boldsymbol{h}(N^*)\}}, \tag{1}$$

where $\boldsymbol{l}(N)$ sets the shape of the dyad distributions, for instance modeling a poisson distribution.

Estimation of parameter values θ in ERGM is carried out with Markov Chain Monte Carlo Maximum Likelihood Estimation (MCMC MLE). We use the `ergm.count` package [12] for `R` to estimate count ERGMs.

C. GHypEG

Generalized hypergeometric ensembles of random graphs (gHypEG) are network models specifically developed to deal with multi-edge networks [13]. Like ERGMs, they can be used to estimate the effects of endogenous and exogenous covariates on network formation [14]. The model is defined as the probability of observing the given network N by sampling its edges from an urn containing all possible combinations of edges. By specifying the number of possible combinations of edges between each pair of nodes $\Xi_{ij} = k_i^{\text{out}} \cdot k_j^{\text{in}}$ as a function

[1]Estimates obtained as the R^2 of the linear regressions *network* $\sim$ *unweighted statistic* and *network* $\sim$ *weighted statistic*, respectively.

TABLE I
MODELS' FIT ON SYNTHETIC NETWORKS

ERGM				gHypEG		
network	(a)	(b)	(c)	(a)	(b)	(c)
closure	0.001	0.03$^{(*)}$	0.025$^{(*)}$	.30	1.13$^{(*)}$	1.76$^{(*)}$
nonzero	-0.07	-1.67$^{(*)}$	-1.67$^{(*)}$	–	–	–
sum	0.53$^{(*)}$	0.13	0.13	–	–	–

$^{(*)}$ indicates p-value < 0.001

of their in and out degrees $k_i^{\text{out/in}}$, the probability for a directed network is given as follows:

$$P(N, \boldsymbol{\theta}) = \prod_{ij} \binom{\Xi_{ij}}{A_{ij}} \int_0^1 \prod_{ij} (1 - z^{\frac{\Omega_{ij}}{S_{\boldsymbol{\Omega}}}})^{A_{ij}} dz, \quad (2)$$

where $S_{\boldsymbol{\Omega}} = \sum_{ij} \Omega_{ij}(\Xi_{ij} - A_{ij})$. In Eq.2, $\boldsymbol{A}$ is the adjacency matrix of the observed network N. $\Omega_{ij} = \exp\{\boldsymbol{\theta}^T \log[\boldsymbol{h}_{ij}(N)]\}$ is the relative propensity of two nodes i, j to be connected in terms of the estimated parameters $\boldsymbol{\theta}$ and the vector of statistics $\boldsymbol{h}_{ij}(N)$ that contains exogenous and endogenous dyadic covariates [14]. Importantly, the parameters $\boldsymbol{\theta}$ estimate the "degree-corrected" effect of the covariates as a propensity to connect two nodes beyond what prescribed by degrees.

The estimation of parameter values θ in a gHypEG is carried out with numerical Maximum Likelihood Estimation (MLE) [14], thus forgoing computationally intense simulations and allowing the analysis of large networks. We use the `ghypernets` package [15] to estimate the gHypEGs.

III. RESULTS

A. Validation with Synthetic Data

We first validate our approach to quantify triadic closure in multi-edge networks by simulating three different scenarios. Network (a) is a small network (34 nodes) made up of $1,000$ randomly sampled edges. Network (b) has the same number of nodes, but the $1,000$ edges are assigned to a set of randomly selected triangles ($n_{tri} = 26$). Both networks represent two extremes and do not reflect the structure of real-world social networks. Instead, they report instances of networks, where triadic closure is the driving factor (b) and where triadic closure is present but meaningless (a). Network (c) represents a combination of (a) and (b) (34 nodes, 2000 edges). It captures a maximally dense network where half of the edges only form triangles. We use this network to test if the weighted change statistic can capture meaningful triadic closure effects in dense multi-edge networks.

Table I reports the coefficient estimates θ (cf. Eq.1, 2) for triadic closure for both inferential models, applied on the three networks. The larger the coefficient, the stronger the estimated effect of triadic closure. For the ERGM, two additional model terms are reported. The `sum`-term represents an intercept term for edge counts. It controls for the expected number of interactions (i.e., edge counts) in the multi-edge network, which is fixed in the gHypEG. The `nonzero`-term reports a negative coefficient, indicating that there is some zero-inflation in the data.

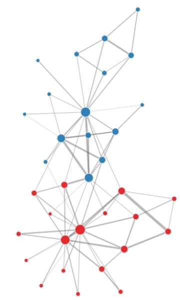


Fig. 3. ZKC network. Nodes are colored by factions.

TABLE II
RESULTS FOR THE ZKC NETWORK

	ERGM	gHypEG
nonzero	$-3.281^{(*)}$	–
sum	$-1.166^{(*)}$	–
degree dist.	$0.028^{(*)}$	–
closure	-0.016	-0.160
faction	$1.123^{(*)}$	$1.090^{(*)}$
AIC	-869.4	674.7
Null AIC	0	869.1

$^{(*)}$ indicates p-value < 0.001

For network (a), both models show a near-zero, non-significant effect of triadic closure in the random network. The analysis of network (b), instead, confirms that the weighted shared partner statistic significantly correlates with the dyadic counts of the multi-edge network. Finally, both models report a positive and highly significant effect of triadic closure for network (c), as expected. The results in Table I are stable across 1,000 random generations of the synthetic networks. The gHypEG reports an average coefficient of -0.09 (min $=$ -1.55, max $= 0.73$, sd $= .34$), 1.16 (min $= .88$, max $= 1.55$, sd $= .09$), and 1.81 (min $= 0.94$, max $= 2.44$, sd $= 0.20$) for the three cases, respectively.

For network (c), in particular, the unweighted shared partner statistic cannot capture triadic closure effects (neither the gHypEG, nor the ERGM): the gHypEG reports a near-zero and non-significant coefficient $\beta = 0.08$, $SE = 0.20$, p-value $= 0.694$). Conventional approaches fail here because the network is too dense to distinguish meaningful triadic closure. By neglecting edge counts, triadic closure cannot be adequately estimated or controlled for in complete, or near complete, networks. The weighted shared partner statistic is thus necessary to estimate whether triadic closure is a meaningful driving force for the network structure.

B. Case Studies

a) Zachary's Karate Club Network: In our first case study, we examine whether the interactions between the 34 nodes in Zachary's Karate Club [16] (ZKC) show signs of triadic closure. [16] recorded different social contexts in which the 34 nodes interacted, for instance, attending a university bar, or participating in the same tournament. Out of these different interactions, a network was constructed, counting for each dyad (i, j) how often two nodes i and j were part of the same social context.

In addition to triadic closure, we control for the known separation into two distinct 'factions' (colors in Figure 3). Furthermore, we control for degree distributions to examine whether triadic closure is a simple artifact of the degree distribution. In ZKC (repeated) social interactions have a precise meaning. I.e., instead of accounting for direct interactions, they quantify the co-occurrence of two members in the same social contexts. Therefore, we assume that triadic closure does *not* play a substantial role in this network. It is relatively unlikely,

Fig. 4. Contact network of students colored by classes.

TABLE III
RESULTS FOR HIGH SCHOOL NETWORK

	ERGM	gHypEG
nonzero	$-4.033^{(*)}$	–
sum	$0.239^{(*)}$	–
degree dist.	$0.00002^{(*)}$	–
closure	$0.008^{(*)}$	$0.819^{(*)}$
class	$0.162^{(*)}$	$0.879^{(*)}$
AIC	−47,511.5	593,853.5
Null AIC	0	1,346,750

$^{(*)}$ indicates p-value < 0.001

that if member a attends an academic class with b, and b visits a weekend karate lesson with c, a and c interact in a third context (e.g., at a bar). Such a closed triad is only plausible if co-occurrence in the same social contexts also means direct social interactions among all members present.

The results of the two inferential network models are reported in Table II. Both models report a negative, non-significant coefficient of triadic closure. That means, holding node degrees and faction membership constant, triadic closure does not add to the explanation of the network structure. On the other hand, both models report a positive and significant effect of faction homophily. Nodes are more likely to share the same social contexts with nodes from the same faction. For the count ERGM, three additional model terms are reported. Together with the `sum`-term and the `nonzero`-term explained above, we control for degree distributions to ensure that degree-based effects do not mask the triadic closure effect. The gHypEG already reports degree-corrected parameters.

b) Friendship Network in a French High School: In our second case study, we examine a multi-edge friendship network, reported by [17]. The network consists of 327 nodes, each indicating a student from the same French high school, and 188,508 undirected ties, measured using contact sensors. For each dyad, we know whether or not they attended classes together. Classroom homophily, moreover, is an inherent factor in friendship networks as they provide increased opportunities to meet. Figure 4 depicts the network, where nodes are colored according to school classes. We included the information about school classes in the inference models for a more severe test of triadic closure in this network, as triadic closure has been theorized to be the product of increased opportunity to interact [1]. Attending classes together increases the opportunity for students to interact and become friends. Hence, by controlling for this potential exogenous influencing factor, we can test whether triadic closure drives friendship formation patterns beyond class-based interactions.

Table III reports the results of the two inferential network models[2]. Both models show a positive and significant effect of triadic closure beyond degree-based clustering, confirming triadic closure to be one of the essential driving factors of friendship formation in social networks.

[2]Note that the fit for the ERGM did not converge in reasonable time, due to the size of the dataset. Results after 40 iterations of MCMCMLE are reported.

IV. DISCUSSION

In this paper, we have presented a new operationalization of triadic closure for social networks that considers information on repeated interactions among nodes.

We can detect meaningful closure in multi-edge networks where standard operationalizations of triadic closure fail. Using synthetic data, we show that even in dense networks, meaningful triadic closure can be detected and controlled for.

Our operationalization allows to test or to control for one of the most critical relational mechanisms in social networks. It an be applied to inferential network models of multi-edge networks, such as the Exponential Random Graph Model (ERGM) for valued networks or the generalized hypergeometric ensembles of random graphs (gHypEG). Thus, with this article we facilitate inferential network analyses to move away from the study of binary networks, towards the analyses of multi-edge and weighted networks, which offer a more realistic representation of social interactions and relations.

REFERENCES

[1] G. Simmel, *The Sociology of Georg Simmel.* New York: Free Press, 1908 (1950).
[2] M. S. Granovetter, "The strength of weak ties," *American journal of sociology*, vol. 78, no. 6, pp. 1360–1380, 1973.
[3] T. Opsahl and P. Panzarasa, "Clustering in weighted networks," *Social Networks*, vol. 31, no. 2, pp. 155–163, may 2009. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0378873309000070
[4] D. R. Hunter and M. S. Handcock, "Inference in curved exponential family models for networks," *Journal of Computational and Graphical Statistics*, vol. 15, no. 3, pp. 565–583, 2006.
[5] R. S. Burt and M. Knez, "Kinds of third-party effects on trust," *Rationality and society*, vol. 7, no. 3, pp. 255–292, 1995.
[6] J. Martin and K.-T. Yeung, "Persistence of close personal ties over a 12-year period." *Social Networks*, vol. 28, pp. 331–62, 2006.
[7] G. C. Homans, *The Human Group.* New York: Hardcourt, Brace and World, 1950.
[8] T. A. Snijders, P. E. Pattison, G. L. Robins, and M. S. Handcock, "New specifications for exponential random graph models," *Sociological methodology*, vol. 36, no. 1, pp. 99–153, 2006.
[9] P. N. Krivitsky, M. S. Handcock, and M. Morris, "Adjusting for network size and composition effects in exponential-family random graph models," *Statistical methodology*, vol. 8, no. 4, pp. 319–339, 2011.
[10] G. Robins, P. Pattison, Y. Kalish, and D. Lusher, "An introduction to exponential random graph (p*) models for social networks," *Social Networks*, vol. 29, no. 2, pp. 173–191, 2007.
[11] P. N. Krivitsky, "Exponential-family random graph models for valued networks," *Electronic Journal of Statistics*, vol. 6, p. 1100, 2012.
[12] ——, *ergm.count: Fit, Simulate and Diagnose Exponential-Family Models for Networks with Count Edges. The Statnet Project.*, 2018, r package version 3.3.0. https://CRAN.R-project.org/package=ergm.count.
[13] G. Casiraghi, V. Nanumyan, I. Scholtes, and F. Schweitzer, "Generalized Hypergeometric Ensembles: Statistical Hypothesis Testing in Complex Networks," *arXiv preprint arXiv:1607.02441*, jul 2016.
[14] G. Casiraghi, "Multiplex Network Regression: How do relations drive interactions?" *arXiv preprint arXiv:1702.02048*, feb 2017.
[15] G. Casiraghi and V. Nanumyan, *ghypernet: R package for Generalised Hypergeometric Ensembles of Random Graphs (gHypEG)*, 2019, r package version 0.5.0. https://github.com/gi0na/r-ghypernet/.
[16] W. W. Zachary, "An information flow model for conflict and fission in small groups," *Journal of anthropological research*, vol. 33, no. 4, pp. 452–473, 1977.
[17] R. Mastrandrea, J. Fournet, and A. Barrat, "Contact patterns in a high school: A comparison between data collected using wearable sensors, contact diaries and friendship surveys," 2015.