

An efficient counting method for the colored triad census

Jeffrey Lienert^{a,b,c}, Laura Koehly^a, Felix Reed-Tsochas^b, Christopher Steven Marcum^{a,c}

^a*Social Network Methods Section, National Human Genome Research Institute, National Institutes of Health*

^b*CABDyN Complexity Centre, Saïd Business School, University of Oxford*

^c*Corresponding Author*

Abstract

The triad census is an important approach to understand local structure in network science, providing comprehensive assessments of the observed relational configurations between triples of actors in a network. However, researchers are often interested in combinations of relational and categorical nodal attributes. In this case, it is desirable to account for the label, or color, of the nodes in the triad census. In this paper, we describe an efficient algorithm for constructing the colored triad census, based, in part, on existing methods for the classic triad census. We evaluate the performance of the algorithm using empirical and simulated data for both undirected and directed graphs. The results of the simulation demonstrate that the proposed algorithm reduces computational time many-fold over the naïve approach. We also apply the colored triad census to the Zachary karate club network dataset. We simultaneously show the efficiency of the algorithm, and a way to conduct a statistical test on the census by forming a null distribution from 1,000 realizations of a mixing-matrix conditioned graph and comparing the observed colored triad counts to the expected. From this, we demonstrate the method’s utility in our discussion of results about homophily, heterophily, and bridging, simultaneously gained via the colored triad census. In sum, the proposed algorithm for the colored triad census brings novel utility to social network analysis in an efficient package.

Keywords: triad census, labeled graphs, simulation

1. Introduction

The triad census is an important approach towards understanding local network structure. Holland and Leinhardt [1] first presented the 16 isomorphism classes of

4 structurally unique triads possible in a directed network without loops. To conduct
5 a triad census, one simply counts each occurrence of these structures, without re-
6 spect to the labeling of the nodes (here we use node label, color, characteristic, and
7 attribute interchangeably). This is useful insofar as specific triads, or combinations
8 thereof, may relate to underlying social processes giving rise to an observed network.
9 For example, bridges (triads with one null dyad and two non-null dyads) may be
10 important in navigating social networks [2], and certain triads may be more or less
11 favorable based on structural balance theory (e.g. the 300 is balanced but the 201
12 is not, see Figure 1) [3]. Moreover, a variant of the triad census, motif analysis, in-
13 vestigates the statistics of various triad configurations (motifs), and has found wide
14 application in biology [4].

15 Also important to network structure are nodal characteristics and how they relate
16 to tie formation or dissolution. This has been the subject of research on homophily
17 (individuals having similar attributes with those to whom they are connected) [5].
18 However, homophily is an observed phenomenon, not a process. The processes giving
19 rise to homophily are varied, often confound the relationship between networks and
20 outcomes, and are difficult to tease apart [6]. Methodological advances, such as
21 stochastic actor-oriented models can disentangle these effects to some extent [7].
22 Other analyses have attempted to disentangle the processes leading to homophily
23 from structural processes, such as triadic closure [8]. Additionally, the coloring of
24 nodes in a network has been an important question for many graph theorists and
25 indeed represents a major topic in this field [9].

26 Although nodal characteristics and the triad census are important, they have

27 rarely been examined fully in conjunction. Yet, there are a few cases where specific
28 colored triads have been studied. For example, Gould and Fernandez [10] study
29 brokerage based on triad structure and group membership simultaneously. This
30 same approach has been used to study brokerage in dynamic networks [11]. As
31 well, a study by Marcum and Koehly [12] examined specific colored triads based
32 on generational membership within families; in this work the authors showed that
33 inter-generational ties were observed in different quantities than expected based on
34 the underlying null model. None of the past research evaluated the full census of
35 colored triads, rather, researchers have focused instead on specific colored triads that
36 were *a priori* expected to be relevant to the processes at hand. As a result, these
37 foundational works were not exhaustive with respect to all alternatives. In other
38 words, previous research examining a subset of colored triads likely had an amount
39 of false negatives due to not examining every colored triad; this could be addressed
40 by censusing the colored triads.

41 The examination of node characteristics together with local structure is important
42 as it provides opportunity to simultaneously study the occurrence of triadic structure,
43 nodal attributes, and the interactions between them. For instance, certain colored
44 triads may be forbidden, such as three-cycles between strict heterosexuals in mixed-
45 orientation sexual contact networks [13]. Impermissible triads would be categorized
46 the same as those that were not observed due to chance in a triad census, potentially
47 missing important social processes or constraints at play in this type of network.
48 Only by incorporating node coloring into the triad census can this pattern be fully
49 elucidated.

50 Based on this methodological gap in the literature, we develop a method to census
51 the colored triads for any one-mode binary network with arbitrary number of colors.
52 Due to the large numbers of isomorphism classes of size 3 as the number of colors
53 increases, this method requires computational efficiency in addition to mathematical
54 accuracy. As well, one is often interested in forming a null distribution with which to
55 compare observed colored triad counts. If the null distribution cannot be analytically
56 solved, one would likely census the colored triads of many simulated networks, further
57 increasing the need for the algorithm to be computationally efficient.

58 Current efficient methods for the triad census exploit the sparseness of networks
59 [14], and scale sub-quadratically (as the number of edges increases the time to run
60 the algorithm is faster than the number of edges squared). However, methods that
61 exploit network sparseness by inferring the number of null triads do not work in
62 the colored case because they do not explicitly interrogate every triad, and there
63 are variations within the null triads due to the coloring. Therefore, we extend the
64 methodology of Moody [15], which is based on matrix algebra and interrogates every
65 triad; his method scales sub-quadratically with the number of nodes.

66 This paper (1) presents the colored triad census and its computational complexity,
67 (2) shows that this approach can be used on large networks (tested for up to 10,000
68 nodes) with up to 10 colors in relatively efficient time, and (3) uses the method many
69 times to create null distributions of colored triad censuses to form the basis of con-
70 ditional uniform graph tests. We illustrate the benefits of an analysis incorporating
71 the colored triad census using a well-known dataset, Zachary’s Karate Club [16].

72 2. Algorithm

73 Since the original appearance of the triad census in 1976, a number of papers
74 have explored how to compute the triad census of a network in an efficient man-
75 ner. Although, for sparse networks, sub-quadratic methods (in terms of number of
76 nodes) exist for calculating the triad census (e.g. Batagelj and Mrvar [14]), we use
77 the quadratic algorithm presented by Moody [15] here. This is because the more
78 efficient methods avoid interrogating null triads directly by taking advantage of the
79 sparseness of graphs, the subsequent large number of null (003) triads, and the known
80 number of total triads. Instead, they interrogate all triads with at least one edge,
81 and then subtract that count from the total number of triads in the network to
82 arrive at the number of null triads. This is insufficient in the colored triad census
83 as there are differently-colored null triads, and the count of each cannot therefore
84 be algebraically determined. For example, if there are two colors, four different null
85 colored triads exist (0-3 nodes color A). The exact breakdown of the null triad into
86 the four colored triads cannot be determined without interrogating each null triad,
87 thereby losing the efficiency gained when not considering colors. Moody’s algorithm
88 does not employ this limiting shortcut, and we therefore use it as a basis for our
89 colored triad census algorithm. Additionally, because many networks are sparse, we
90 can leverage computational techniques for increasing the efficiency of sparse matrix
91 operations [17], further reducing the computational complexity of our method.

92 Moody [15] showed that the count of each of the 16 triad isomorphism classes
93 could be derived by using matrix algebra on the adjacency matrix of the graph and
94 its derivatives. To review, let $\mathbf{A}$ be the adjacency matrix of a network, and $A_{ij} = 1$

95 when a tie exists from node i to node j . Let E be the symmetrized matrix A , formed
 96 by making any edge in A reciprocal via $E_{ij} = \max(A_{ij}, A_{ji})$. The complement of E ,
 97 $\bar{E}$, is formed by subtracting the complete network adjacency matrix from E , so that
 98 $E_{ij} = 1$ if and only if there is neither a tie from i to j nor a tie from j to i . Next, we
 99 have M , the mutual matrix of A , and is made by removing any asymmetric edges
 100 from A , or $M_{ij} = \min(A_{ij}, A_{ji})$. Finally, C is the matrix of only asymmetric edges,
 101 and is calculated by $C = A - M$. Therefore, $C_{ij} = 1 \iff A_{ij} = 1 \ \& \ A_{ji} = 0$.
 102 Based on these matrices, Moody demonstrates how to calculate the number of each
 103 of the 16 isomorphism classes for the case of unlabeled graphs (or, equivalently, for
 104 a graph consisting of nodes of the same single color). Generally, this was done by
 105 multiplying (either through dot-product or element-wise multiplication) the three
 106 matrices corresponding to the relevant edges in the triad of interest. There were two
 107 triads ($111U$ and $111D$) that were not directly amenable to this process and were
 108 calculated via addition and subtraction of other triad types, respectively.

109 To extend this work to the case of multiple colors, we introduce the out-coloring
 110 and in-coloring matrices, K^r and $K^{r'}$, respectively, where r is the focal color of
 111 matrix K . Here, the in-coloring matrix is the transpose of the out-coloring matrix.
 112 The out-coloring matrix is calculated by evaluating the color of the nodes row-wise,
 113 such that rows indexing nodes of the focal color are composed in the following way:

$$K_{i\bullet}^r = \begin{cases} 1 & \text{if } R(i) = r \\ 0 & \text{if } R(i) \neq r \end{cases} \quad (1)$$

114 Where $R(i)$ is a function returning the color of node i . As above, the in-coloring

115 matrix is the transpose of the out-coloring matrix in Eq. 1.

116 Our algorithm works by using the in- and out-coloring matrices to evaluate and
 117 “switch on” edges that have nodes of the focal colors at the *ends* (or tails) of edges
 118 in the adjacency matrix A of the network. We adapt the triad census nomenclature
 119 of Holland and Leinhardt [1] by appending the colors after the name of the triad.
 120 The colors are ordered from the top node proceeding clockwise in Figure 1. We
 121 have arbitrarily adapted the orientation of the triads from the triad census figure in
 122 Holland and Leinhardt [1] for computational reasons. The orientation is important
 123 here because triads with the same orientation may no longer be isomorphic when
 124 color is introduced. Figure 1 makes it possible to count unambiguously and name
 125 only unique colored triads. Therefore, $T_{102-123}$ is the triad consisting of 1 symmetric
 126 dyad and 2 null dyads, where the top node is of color 1, the bottom-right node is
 127 of color 2 and the bottom-left node is of color 3. This is distinct from the $T_{102-312}$
 128 triad because the coloring of the nodes is not identical from the previous triad.

129 Following this, the general formula for an arbitrary triad “ T ” with an arbitrary
 130 coloring triplet is:

$$T = \text{Tr} \left((K^1 \times H(T, 1, 2) \times K^{2'}) (K^2 \times H(T, 2, 3) \times K^{3'}) (K^3 \times H(T, 3, 1) \times K^{1'}) \right) \quad (2)$$

131 In the above, “ $\times$ ” refers to element-wise multiplication, and “Tr” is the trace
 132 function. For an arbitrary triad, “ T ” has a color triplet r_1, r_2, r_3 . $H(T, i, j)$ is a
 133 function returning the matrix specific to the type of edge between nodes i and j
 134 in triad “ T ”. For example, in a 102 triad, the first edge from the top node going

135 clockwise is a symmetric edge from node one to node two (Figure 1). $H(T_{102}, 1, 2)$
 136 in this case would be the matrix E for the symmetric matrix, and the sandwiching
 137 color matrices would turn the proper edges on and off if nodes one and two were of
 138 the specified colors. If the edge is an asymmetric one, and the direction of the edge
 139 in Figure 1 is counter-clockwise, then C' is used instead of C to force the edge to go
 140 in the proper direction.

141 At this point, there are redundant triads due to certain colored triads being
 142 isomorphic. For instance, the $T_{003-122}$ is isomorphic with $T_{003-221}$ and $T_{003-212}$, and
 143 would be triple-counted. These are removed by checking for isomorphisms based
 144 on matrix row and column permutations of the triad. If two colored matrices are
 145 identical after such row and/or column permutations, then they are isomorphic, and
 146 one is removed. We arbitrarily decide to discard the triad whose coloring triplet
 147 name comes second alphanumerically. It should be noted that removing in this way
 148 is computationally expensive, particularly as the number of colors and nodes grows
 149 large. We therefore shorten this process by performing it once for 1 to 10 colors and
 150 storing the unique isomorphism classes. This leaves only unique isomorphism classes
 151 of colored triads, which can then be accessed in linear time.

152 The number of unique isomorphism classes for a given number of colors can be
 153 shown for each of the 16 isomorphism classes in the triad census. The 16 classes
 154 separate into four types of colored triads, depending on how many structurally-
 155 distinct positions there are in the triad (e.g. the two ends of the edge in a 102 triad
 156 are not structurally-distinct from one another, but are distinct from the node with
 157 no edges). The calculation for the number of each isomorphism class for arbitrary

number of colors (k) is shown in Table 1. Each combinatoric term in each row (together with their respective leading permutation coefficients) counts the number of colored triads when there are three, two, or one unique color(s), respectively. For example, in a network with three colors, the ‘300’ and ‘003’ classes have only one accessible permutation when there are three colors present in the triad (i.e. $\binom{3}{3}$), six ways when there are two colors (i.e. $2\binom{3}{2}$), and one way when there is one color in the triad (i.e. $\binom{3}{1}$).

Isomorphism classes	Number of colored triads
300 and 003	$\binom{k}{3} + 2\binom{k}{2} + \binom{k}{1}$
030C	$2\binom{k}{3} + 2\binom{k}{2} + \binom{k}{1}$
102 021D 021U 201 120D and 120U	$3\binom{k}{3} + 4\binom{k}{2} + \binom{k}{1}$
012 021C 111D 111U 030T 120C and 210	$6\binom{k}{3} + 6\binom{k}{2} + \binom{k}{1}$

Table 1: Expression for the number of isomorphism classes within a triad class. k is the number of colors

164

If these numbers are summed over the 16 isomorphism classes, the total number of colored isomorphism classes of triads for k colors is returned. Similarly, the same can be done for undirected triads, solely summing over the 4 triads observed in the undirected case. Table 2 reports the total number of colored triads for undirected and directed networks over a range of k . Clearly, the number of isomorphism classes grows quite quickly as k increases.

The algorithm implemented as an R package is publicly available and is linked to this paper via github: <https://github.com/jlienert/ColoredTriadCensus>.

172

Number of colors	Number of directed colored triads	Number of undirected colored triads
1	16	4
2	104	20
3	328	56
4	752	120
5	1440	220
6	2456	364
7	3864	560
8	5728	816
9	8112	1140
10	11080	1540

Table 2: The number of colored triad isomorphism classes for directed and undirected networks for k ranging from 1 to 10.

3. Algorithmic Performance

If a naïve implementation of matrix multiplication is used, this algorithm runs with computational complexity $O(N^3 * 3^k)$. It scales with the number of nodes cubed (N^3) because of the matrix multiplication involved in the algorithm. However, many software packages use algorithms that reduce the complexity of matrix multiplication to $O(N^{2.38})$ [18]. Furthermore, by taking advantage of methods for matrix multiplication using sparse matrices (as appropriate due to the sparse nature of most social networks), this complexity is reduced to something closer to $O(N^2)$ [19]. The exact benefit gained by using sparse matrix multiplication varies based on how sparse the matrix is. This ranges from the nearly-optimal $O(N^2)$ when very few edges exist, to worse than the optimized algorithm when many edges exist. The scaling with 3^k comes from the number of distinct colored triads the algorithm needs to evaluate, and the number of isomorphism classes scales in such a manner.

186 To test the efficiency of the algorithm, we apply it to networks ranging in size
 187 from $n = 10$ to $n = 10,000$ with the number of colors ranging from $k = 3$ to $k = 10$,
 188 all holding the average degree constant at 6 by creating Erdős-Rényi graphs with
 189 an edge probability of $\frac{6}{N-1}$. This reflects the average number of ties participants
 190 enumerate in social networks surveys [20]. The runtime of the algorithm with these
 191 parameters can be seen in Figure 2. In general, increasing K results in constant
 192 increases in $\log(runtime)$, which is what we expect based on the theoretical com-
 193 putational complexity. As expected, we also observe a super-quadratic increase in
 194 $\log(runtime)$ as N increases. Although it is super-linear, it is still below the curve
 195 that would exist if we used matrix multiplication not optimized for sparse matrices
 196 (dotted line in Figure 2). This difference shows the expected time saved by using
 197 sparse matrix methods. Finally, we observe changes in the rank-order and decreases
 198 in runtime going from 10 to 100 nodes. This is also due to the computational time
 199 involved in initializing the sparse matrices and storing and operating on sparse ma-
 200 trices, and as such is not unexpected. Additionally, because the average degree was
 201 held constant, the smaller networks are much more dense, and therefore are actually
 202 less efficient than if they used standard matrix multiplication methods. To be per-
 203 fectly optimized, therefore, the algorithm would use standard matrix multiplication
 204 for small networks, and switch to sparse methods for larger networks. However, the
 205 gains would be minimal, generally under 10 seconds, and would require additional
 206 logical steps to check for network size, further minimizing the gain. We therefore use
 207 sparse matrix methods for all network sizes.

208 4. Empirical Use and Example

209 To show the empirical value of this algorithm, we use the Zachary karate club
210 social network [16]. This is a well-known historical network that describes the social
211 relationships between 34 members of a university karate club. Ties exist between
212 members if they overlapped in at least one of eight contexts representing undirected
213 relations. These relations varied in terms of likely strength of the association. Likely
214 at the weak end of the spectrum is being enrolled in the same class at the university,
215 while likely at the strong end is being a student-teacher at the studio. Additionally,
216 three ties are specific to activities with a part-time instructor. Member “factions”
217 were identified as a node attribute, taking one of five mutually exclusive values:
218 strongly associated with the president, weakly associated with the president, neutral,
219 weakly associated with the part-time instructor, or strongly associated with the part-
220 time instructor. These are labeled “Zs”, “Zw”, “N”, “Hw”, and “Hs”, respectively.
221 These labels can be placed on an ordinal scale from -2 (Zs) to 2 (Hs) to quantify
222 members’ direction and strength of alignment. This undirected network with five
223 colors represents a case that is rich in the number of colored triads (220) for detailed
224 conclusions to be drawn using the proposed algorithm (which is general to both
225 undirected and directed networks).

226 We initially ran the colored triad census on the social network using the faction
227 as the nodal attribute. This provided the basis for our empirical observed colored
228 triad census. To determine whether these triads were observed more or less often
229 than expected by chance, we constructed a null model. As the choice of null model
230 can have important ramifications for the null distribution of triads, we chose a model

231 where edge formation is a function of the probability of ties between nodes of specific
 232 attributes [21]. The null model is a mixing-matrix conditioned uniform random
 233 graph distribution based on probabilities of edges between nodes of particular color
 234 combinations [22]. This matrix comprises empirical probabilities of ties between
 235 groups, with the diagonal representing within-group tie probabilities. Observations
 236 of significantly over- or under-represented colored triads are the result of network
 237 effects beyond homophily and heterophily. Networks are then generated from this
 238 matrix via a Bernoulli random graph process á la Erdős and Rényi [23]. This null
 239 model therefore conditions on graph size, the distribution of node factions, and the
 240 probability of ties within and between factions. By generating networks from the
 241 null model, we can observe whether colored triad counts deviate from that expected
 242 based on the marginal distribution of faction mixing. Because we condition on the
 243 above parameters, if we observe statistical deviations in our colored triad census, it
 244 indicates that the structure of the network is dependent on parameters other than
 245 those on which we conditioned.

246 Moreover, for any triad, the expected number and variance can be calculated
 247 assuming each tie follows a Binomial distribution (which is a reasonable assumption
 248 for most binary social network data). The observed number can then be compared
 249 to these numerical results and a p-value extracted from an exact Binomial test. This
 250 equates to the following probability, expectation, and variance for an example colored
 251 triad:

$$\begin{aligned}
P(T) = & P(A_{ij} = 1 | R(i) = r_1, R(j) = r_2) \times P(A_{ij} = 1 | R(i) = r_2, R(j) = r_3) \\
& \times P(A_{ij} = 1 | R(i) = r_3, R(j) = r_1)
\end{aligned} \tag{3}$$

$$E(T) = P(T) \times \prod_{r=1}^{L(T)} \binom{\sum K_{\bullet 1}^r}{S(T, r)} \tag{4}$$

$$V(T) = E(T) \times (1 - P(T)) \tag{5}$$

The probability of T , $P(T)$ in Equation 3 is based on the mixing-matrix of the three colors (r) involved in the triad T . As is standard for the mixing-matrix approach, this continues to assume that all edges in the graph are independent. For the expected value of a specific triad, we multiply the probability of a single one of those triads by the total number of colored triplets that exist in the graph. In Equation 4, the expectation of the triad, $L(T)$ returns the number of unique colors in T and $\sum K_{\bullet 1}^r$ is the number of nodes of color r in the graph. Also, we take the nodes one, two, or three at a time depending on how many times that color repeats in T , represented by $S(T, r)$. This expectation therefore follows a binomial distribution, and it's variance follows accordingly in Equation 5.

However, to show that this method also works for null distributions that are not analytically solvable, we construct a null distribution based on simulated draws from the null model. As the number of trials increases, the simulated null distribution of the colored triad census should asymptotically approach the analytical solution shown above. For each of 1,000 trials, we draw random networks from the null

distribution, and run the triad census on all these networks. Comparing our observed count to the null distribution then allows us to get an approximate p-value for a conditional uniform graph test, and test the over- or under-representation of each colored triad. We now turn to these results.

4.1. Results

Figure 3 is a heatmap of the approximate p-values associated with each binomial exact test against the null for each triad, clustered by the triad and the colored triplet as returned by the proposed algorithm. We use a clustering algorithm to group color triplets with similar profiles across the types of triads. This assists with identifying trends across different colored triads, leading to conclusions that would likely be missed if all the colored triads were individually examined. We find particular importance in three branch cutpoints in the clustering algorithm on the color triplets. The first branch in the clustering algorithm (A in Figure 3) separates four color triplets, comprising 16 colored triads, with a pattern of over-observed 003 and 102 triads, and under-observed 201 and 300 triads. These results show that these color triplets are those that are less clustered than expected by chance. The color triplets all contain nodes of two factions with the first two nodes being *Hs*, that is, those strongly aligned with the part-time instructor. This indicates that those who are so aligned are likely to form ties to one another, but not to members of other factions. The only exception in this group is that two *Hs* nodes are more likely to form a tie from one of the *Hs* members to a *Hw* member, but even in this case the complete triad (003) is still observed less than expected by chance. This particular result is, perhaps, unsurprising, since *Hs* and *Hw* members are close in alignment,

292 more so than with those aligning with the president. Therefore, given the tendency
293 towards homophily they are likely to overlap, though less strongly than members of
294 the same faction; hence, the under-observed $T_{003-HsHsHw}$.

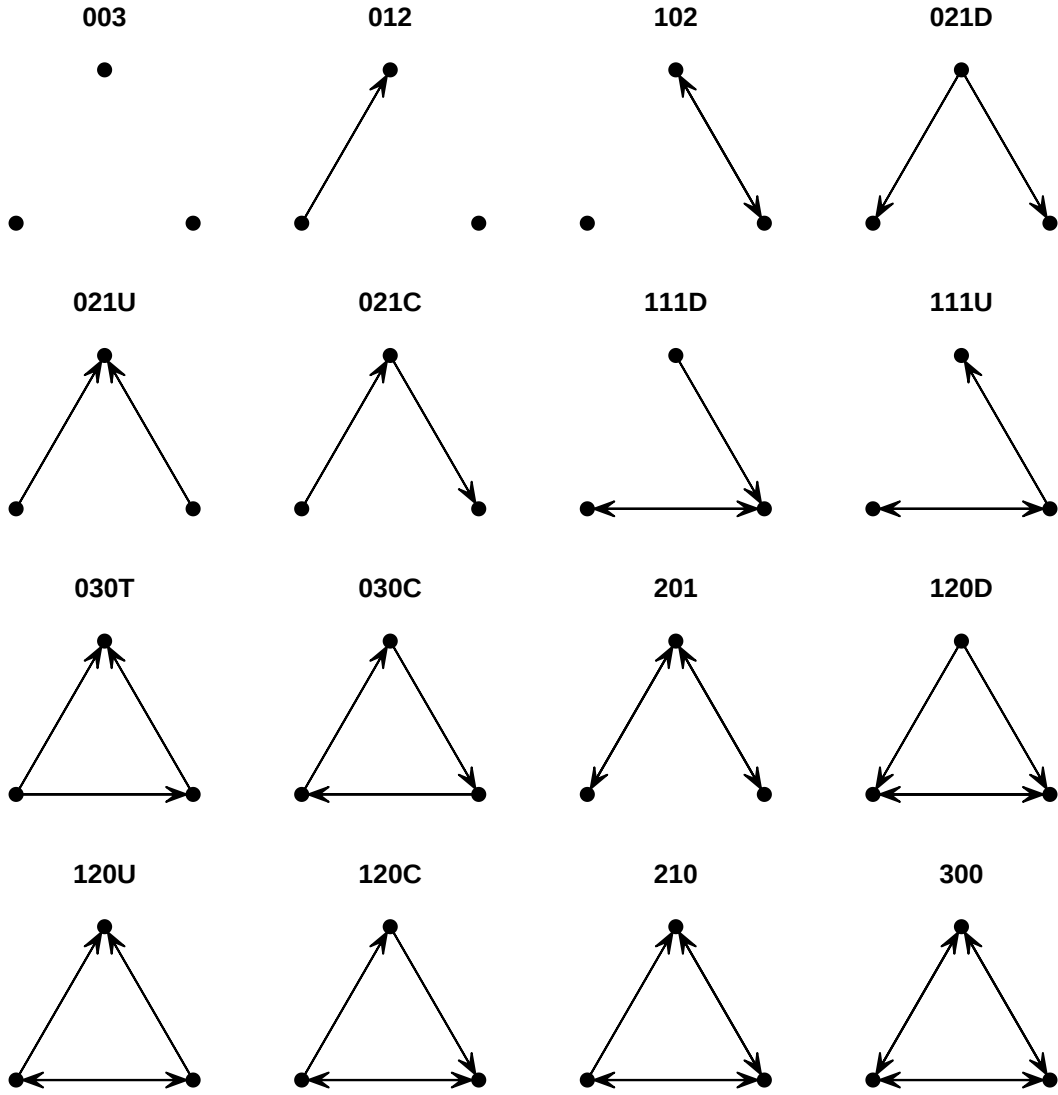


Figure 1: The 16 isomorphism classes of triads and their orientation used here with respect to the color numbering. When colors are added to these triads, they are labeled starting from the top node and proceeding clockwise.

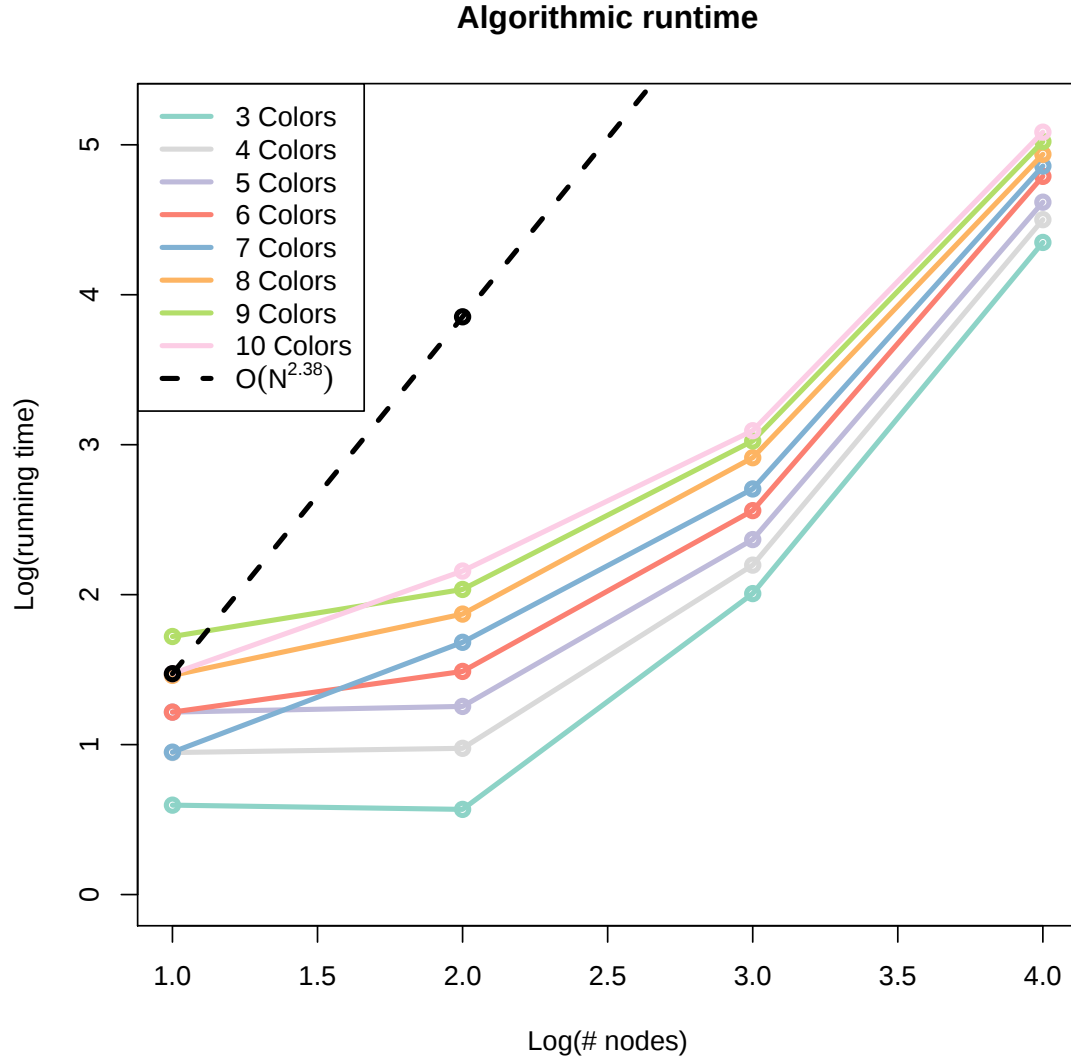


Figure 2: Runtime of the algorithm on networks ranging from size 10 to 10,000 nodes in orders of magnitude, and from one to ten colors. Additionally, the dashed line represents the computational time that would be expected using standard matrix multiplication methods for $k = 10$. These runtimes were generated using virtual PCs including 1 dual core CPU and 10GBs of RAM.

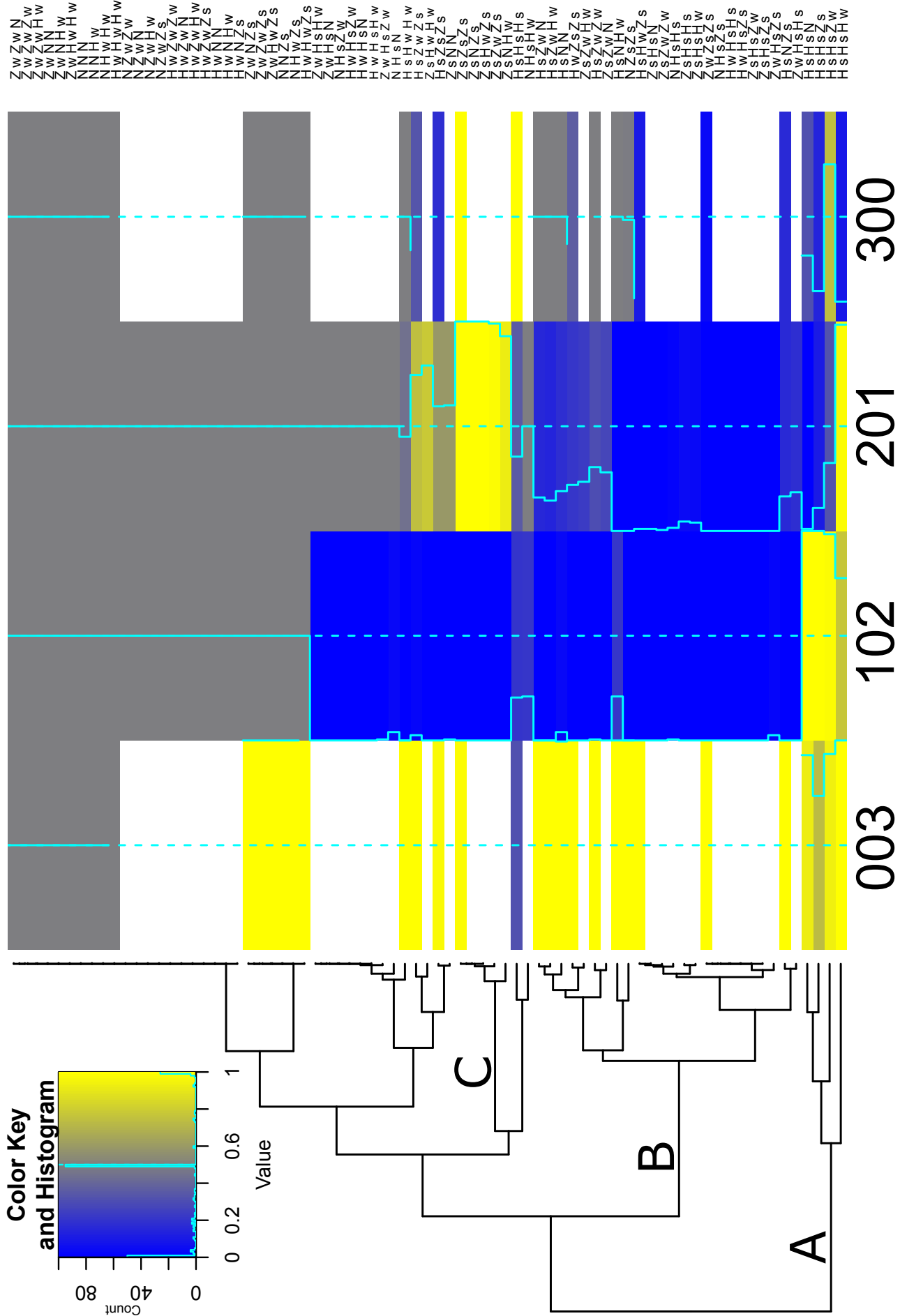


Figure 3: Heatmap of colored triads and their corresponding p-value of how often they were observed in the empirical networks relative to the null distribution. The columns separate triads based on the MAN configuration, and the rows separate triads based on the triplet of colors. Standard clustering algorithms were used to create the dendrograms. White space indicates redundant isomorphism classes. Gray boxes are either those with 0 triads observed in the network or in any of the networks of the null distribution, and therefore have an undefined pseudo p-value, or those with a pseudo p-value of 0.5. The three labels correspond to three breakpoints in the clustering that separate meaningful groups. (A) is a group of four color triplets exhibiting homophily between *Hs* nodes. (B) is a group of 21 colored triplets exhibiting low clustering between heterogeneous nodes. (C) is a group of 6 colored triplets that show potential significant amounts of bridging.

295 The second branching point in the clustering (B in Figure 3) separates the group
 296 of color triplets that are over-observed for the 003 triad, under-observed for the 102
 297 and 201 triads, and observed about as much as expected for the 300 triads. All the
 298 triplets in question have nodes of different factions in the first and second position.
 299 Because the edge in the 102 triad is between the first and second node in the triplet
 300 (Figure 1), this means that these are all triplets where the first edge is less likely
 301 than expected by chance, and the lack of formation of the first edge subsequently
 302 hampers the formation of the edge between the second and third nodes in the triplet
 303 (201 triad). The first two nodes of these triplets are often (e.g., 16 out of 21) two
 304 factions at least a distance of two away (e.g. N and Hs), indicating members of
 305 a faction are not likely to overlap with members who are too disparate from their
 306 faction. Put another way, this pattern of triads shows a lack of faction heterophily.

307 The third branch point (unlabeled) is primarily singling out the group of color
 308 triplets that were not observed in the network, and we cannot draw conclusions about
 309 their prevalence. The fourth branch point (C in Figure 3), however, distinguishes a
 310 group of five triplets that are under-observed for the 102 triad and over-observed for
 311 the 201 triad. This means that the edge between the first two nodes is less likely
 312 than expected by chance, but once that edge does occur, the second edge occurs
 313 more often than expected by chance. All these triplets begin with a Zs member,
 314 and the 201 triad in this case is effectively a bridging tie between it and another.
 315 Interestingly, the bridging node is anything other than an Hs (whom are primarily
 316 consigned to this role in branch A , as discussed above). The third node was another
 317 Zs member in four of five triplets. This indicates that Zs members of the karate

318 club did not often overlap members of other factions, but when they did, provided
319 it was not with an *Hs*, that second person also often overlapped with another *Zs*.

320 Although the above examples show homophily and bridging, analyzing the full
321 colored triad census allows us to draw further conclusions by looking at other colored
322 triads. In particular, the homophily has mostly been a story of the *Hs* nodes, and the
323 bridging primarily about the *Zs* nodes. The 300 triad of both of these factions, when
324 comprising three nodes of the same faction, are observed more often than expected
325 by chance in both cases, which has different implications on the previously-noted
326 results. For the *Hs* nodes, homophily is strengthened, as not only do *Hs* nodes not
327 often overlap with members of other factions, they also very strongly overlap with
328 one another. This may partially be an artifact of the types of overlap, as stated
329 before, three of the overlap activities involve direct participation in the part-time
330 instructor's studio, but there are no corresponding groups for the president. This
331 means that those who are *Hs* or *Hw* may have more opportunity to overlap with
332 one another due solely to the structure of the data. On the other hand, the triplet
333 of all *Zs* members also has an over-observed 300 triad. Although there are other
334 triads that seem to indicate bridging between *Zs* members (C in Figure 3), given
335 that *Zs* members are also densely connected to one-another, the practical effect of
336 these potential bridging ties is reduced. Observing this joint effect of homophily and
337 bridging ties was possible only through the complete colored triad census. Neither
338 a standard triad census nor a brokerage analysis would have revealed the intricacies
339 of these results.

340 In sum, it is clear from these results that the colored triad census allows one

341 to examine multiple trends simultaneously that are often done in isolated analyses,
342 including homophily, heterophily, and brokerage. Importantly, it also allows for
343 generalizations based on the clustering of various triads or color triplets, as well as
344 specific results based on individual triads. In this manner, the colored triad census
345 can yield results on multiple structural levels simultaneously, all while examining
346 local structure, nodal attributes, and their interaction—that is, net of all alternatives
347 involving mixtures of node coloring and triadic configurations.

348 **5. Limitations**

349 There are some limitations to this method. First, it is only computationally ef-
350 ficient relative to existing methods (including brute force counting). Networks of
351 10,000 nodes or more will take over a day to run using the proposed algorithm for
352 the colored triad census. However, this is an easily parallelizable process (by par-
353 titioning the separate algebraic steps, for example), and so the real time necessary
354 to run the analysis can be greatly reduced by taking advantage of this feature. The
355 time needed for the parallelized colored triad census is approximately inversely pro-
356 portional to the number of computational cores used in the calculation (plus some
357 overhead). Second, the interpretation and visualization of these results is compli-
358 cated, particularly as the number of colors increases. Examining all of the triads
359 simultaneously reduces the likelihood of missing interesting results because a specific
360 colored triad was excluded. However, the sheer number of colored triads means that
361 making complete sense of results can be difficult. Even if the results are carefully
362 examined for all colored triads, it is conceivable that one might miss an important

363 result out of the 11,080 colored triads in a directed, 10-color network, no matter how
 364 meticulous the examiner's eye. However, use of standard clustering algorithms and
 365 heatmaps (as in Figure 3) may help to ease interpretation of the results at both a
 366 coarse- (general groups of triads) or fine-grained (individual colored triads) perspec-
 367 tive. That said, we recognize that the interpretation is not straightforward and that
 368 this is a first effort at understanding these results, but we believe that having an
 369 algorithm to efficiently calculate the colored triad census will spur additional work
 370 towards interpreting and using the results. As a result better approaches therein will
 371 emerge with time and use.

372 **6. Conclusions**

373 In this paper, we have extended the matrix algebra methods of Moody [15] to
 374 calculate the colored triad census for any network, directed or undirected, with an
 375 arbitrary number of colors in a relatively computationally efficient manner. We have
 376 shown a number of mathematical results regarding the colored triad census, including
 377 a generalized equation for an arbitrary colored triad, the number of isomorphism
 378 classes for arbitrary numbers of colors, and the expectation and variances for colored
 379 triads. We analyzed an empirical social network using our algorithm, and calculated
 380 approximate p-values for each colored triad, based on an analytic exact binomial
 381 test for less complex null distributions, or approximately through simulation for
 382 more complex null distributions. We have also shown the type of conclusions that
 383 can be drawn from these results, observing results that would not be feasible with
 384 many other currently available methods.

385 One additional benefit of this method is that it can be directly used as a counting
386 tool for sufficient statistics in network inference models, such as exponential random
387 graphs (ERGM). The colored triad census essentially allows one to simultaneously
388 evaluate the effect of local structure and node attribute on network structure in an
389 ERGM, building off previous work where researchers explicated the ERGMs capacity
390 for including the triad census [24]. We believe that the colored triad census is a useful
391 technique with an efficient implementation that can be widely-applicable in social
392 networks research, showing the continued importance of the triad census even in this
393 era of stochastic models for complex networks.

394 7. Acknowledgements

395 The authors would like to thank the two anonymous reviewers and the editors
396 for their contributions to this manuscript. This work utilized the computational
397 resources of the NIH HPC Biowulf cluster (<http://hpc.nih.gov>). This research was
398 funded via National Human Genome Research Institute, National Institutes of Health
399 (Grant number ZIA HG200335) and the Oxford Martin School, University of Oxford
400 (Grant number LC1213-006).

401 8. References

- 402 [1] P. W. Holland, S. Leinhardt, Local Structure in Social Networks, *Sociological*
403 *Methodology* 7 (1976) 1–45.
- 404 [2] M. S. Granovetter, The strength of weak ties, *American Journal of Sociology*
405 78 (1973) 1360–1380.

- 406 [3] D. Cartwright, F. Harary, Structural balance: a generalization of Heider's the-
407 ory., *Psychological Review* 63 (1956) 277–293.
- 408 [4] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, U. Alon, Network
409 motifs: simple building blocks of complex networks, *Science* 298 (2002) 824–827.
- 410 [5] M. McPherson, L. Smith-Lovin, J. M. Cook, BIRDS OF A FEATHER : Ho-
411 mophily in Social Networks, *Annual Review of Sociology* 27 (2014) 415–444.
- 412 [6] C. R. Shalizi, A. C. Thomas, Homophily and contagion are generically con-
413 founded in observational social network studies, *SMR* 40 (2012) 211–39.
- 414 [7] T. A. Snijders, Stochastic actor-oriented models for network change, *Journal of*
415 *mathematical sociology* 21 (1996) 149–172.
- 416 [8] S. M. Goodreau, J. A. Kitts, M. Morris, Birds of a feather, or friend of a
417 friend? using exponential random graph models to investigate adolescent social
418 networks, *Demography* 46 (2009) 103–125.
- 419 [9] T. R. Jensen, T. Bjarne, Graph coloring problems, John Wiley and Sons, 2011.
- 420 [10] R. V. Gould, R. M. Fernandez, Structures of mediation: A formal approach to
421 brokerage in transaction networks, *Sociological methodology* (1989) 89–126.
- 422 [11] E. S. Spiro, R. M. Acton, C. T. Butts, Extended structures of mediation: Re-
423 examining brokerage in dynamic networks, *Social Networks* 35 (2013) 130 –
424 143.

- 425 [12] C. S. Marcum, L. M. Koehly, Inter-generational contact from a network per-
426 spective, *Advances in Life Course Research* 24 (2015) 10–20.
- 427 [13] C. S. Marcum, J. Lin, L. Koehly, End note growing-up and coming-out: Are
428 4-cycles present in adult hetero/gay hook-ups? (2016).
- 429 [14] V. Batagelj, A. Mrvar, A subquadratic triad census algorithm for large sparse
430 networks with small maximum degree, *Social Networks* 23 (2001) 237–243.
- 431 [15] J. Moody, Matrix methods for calculating the triad census, *Social Networks* 20
432 (1998) 291–299.
- 433 [16] W. W. Zachary, An information flow model for conflict and fission in small
434 groups, *Journal of anthropological research* 33 (1977) 452–473.
- 435 [17] I. S. Duff, A. M. Erisman, J. K. Reid, *Direct methods for sparse matrices*,
436 Oxford University Press, 2017.
- 437 [18] A. M. Davie, A. J. Stothers, Improved bound for complexity of matrix multipli-
438 cation, *Proceedings of the Royal Society of Edinburgh: Section A Mathematics*
439 143 (2013) 351369.
- 440 [19] R. Yuster, U. Zwick, Fast sparse matrix multiplication, *ACM Transactions on*
441 *Algorithms (TALG)* 1 (2005) 2–13.
- 442 [20] P. V. Marsden, Interviewer effects in measuring network size using a single name
443 generator, *Social networks* 25 (2003) 1–16.

- 444 [21] K. Faust, A puzzle concerning triads in social networks: Graph constraints and
445 the triad census, *Social Networks* 32 (2010) 221–233.
- 446 [22] M. E. Newman, Mixing patterns in networks, *Physical Review E* 67 (2003)
447 026126.
- 448 [23] P. Erdős, A. Rényi, On random graphs i, *Publicationes Mathematicae* 6 (1959)
449 290–297.
- 450 [24] O. N. Yaveroglu, S. M. Fitzhugh, M. Kurant, A. Markopoulou, C. T. Butts,
451 N. Przulj, *ergm . graphlets : A Package for ERG Modeling Based on Graphlet*
452 *Statistics, Journal of Statistical Software* 65 (2015) 1–29.

Appendix A. Variable and Functional Definitions

Variable or function notation	Description of variable or function
A	Adjacency matrix
E	Symmetrized adjacency matrix
$\bar{E}$	Complement of symmetrized adjacency matrix
M	Adjacency matrix including only mutual ties
C	Adjacency matrix including only asymmetric ties
K^r	Coloring matrix for color r
$R(i)$	Function returning the color of node i
$H(T, i, j)$	Function returning the matrix of the edge in triad T between nodes i and j
T	An arbitrary colored triad, with a MAN configuration, and colored triplet r_1, r_2, r_3
$L(T)$	A function returning the number of unique colors for a given colored triad
$S(T, r)$	Function returning the number of times color r appears in colored triad T
$P(T)$	The probability of observing triad T
$E(T)$	The expectation of triad T under a binomial model
$V(T)$	The variance of triad T under a binomial model

Table A.3: List of variables, constants, and functions defined in this manuscript.