AFRL-AFOSR-UK-TR-2015-0025

Detecting Statistically Significant Communities of Triangle Motifs in Undirected Networks

Marcus B. Perry

**IMPERIAL COLLEGE OF SCIENCE
TECHNOLOGY AND MEDICINE
EXHIBITION ROAD
LONDON, SW7 2AZ UNITED KINGDOM**

EOARD GRANT #FA9550-15-1-0019

Report Date: March 2015

Final Report from 15 October 2014 to 14 January 2015

**Air Force Research Laboratory
Air Force Office of Scientific Research
European Office of Aerospace Research and Development
Unit 4515, APO AE 09421-4515**

REPORT DOCUMENTATION PAGE

Form Approved OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From – To)*
16 March 2015	Final Report	15 October 2014 – 14 January 2015

4. TITLE AND SUBTITLE

Detecting Statistically Significant Communities of Triangle Motifs in Undirected Networks

5a. CONTRACT NUMBER

5b. GRANT NUMBER

FA9550-15-1-0019

5c. PROGRAM ELEMENT NUMBER

61102F

6. AUTHOR(S)

Marcus B. Perry

5d. PROJECT NUMBER

5d. TASK NUMBER

5e. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

IMPERIAL COLLEGE OF SCIENCE TECHNOLOGY AND MEDICINE
EXHIBITION ROAD
LONDON, SW7 2AZ UNITED KINGDOM

8. PERFORMING ORGANIZATION REPORT NUMBER

N/A

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

EOARD
Unit 4515
APO AE 09421-4515

10. SPONSOR/MONITOR'S ACRONYM(S)

AFRL/AFOSR/IOE (EOARD)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

AFRL-AFOSR-UK-TR-2015-0025

12. DISTRIBUTION/AVAILABILITY STATEMENT

Distribution A: Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The primary focus of this research was to extend the work of Perry et al. [6] by developing a statistical framework that supports the detection of triangle motif-based clusters in complex networks. The specific works accomplished over the 3-month period are as follows:
1. Developed a tractable hypothesis testing framework to assess, a priori, the need for triangle motif-based clustering.
2. Developed an algorithm for clustering undirected networks, where the triangle configuration was used as the basis for forming clusters.
3. Developed a C++ implementation of the proposed clustering framework.

15. SUBJECT TERMS

EOARD, Operations Research, Networks, Network Clustering, Statistics

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18, NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Jeremy Jordan
UNCLAS	UNCLAS	UNCLAS	SAR	30	19b. TELEPHONE NUMBER *(Include area code)* +44 (0)1895 616002

Grant No.: FA9550-15-1-0019

Detecting Statistically Significant Communities of Triangle Motifs in Undirected Networks

Marcus B. Perry[1]
Sponsored Researcher
Imperial College London
London, UK

Period of Performance
Oct 15, 2014 - Jan 14, 2015

[1]Associate Professor of Statistics, Department of Information Systems, Statistics & Management Science, University of Alabama, USA.

Contents

List of Figures

List of Tables

1 Summary

1.1 Summary of Tasks Completed

The primary focus of this research was to extend the work of Perry *et al.* [6] by developing a statistical framework that supports the detection of triangle motif-based clusters in complex networks. The specific works accomplished over the 3-month period are as follows:

1. Developed a tractable hypothesis testing framework to assess, *a priori*, the need for triangle motif-based clustering.
2. Developed an algorithm for clustering undirected networks, where the triangle configuration was used as the basis for forming clusters.
3. Developed a C++ implementation of the proposed clustering framework.

1.2 Deliverables

The deliverables for this grant include:

1. Technical report describing developmental aspects of the proposed clustering framework.
2. C++ implementation of the proposed clustering framework, to include all computer codes required for compilation.

2 Introduction

Clustering has a wide array of applications, from pattern recognition and spatial data analysis to data mining and military intelligence. Regardless of the application, clustering methodologies are often used to explore a data set where the goal is to partition the sample into distinct groups, or to provide new understanding about the underlying structure of the data. Although clustering algorithms are often applied to conventional data sets, they can also be applied to network data (e.g., social networks, biological networks, computer networks, etc.). In such a case, the goal is typically to assign each node in the network to one of several mutually exclusive groups based upon information contained in the edge set.

To date, most network clustering algorithms focus on finding groups of nodes that are densely intra-connected and sparsely inter-connected, where the dyad (or link between two nodes) serves as the building block for estimating clusters, e.g., see Perry *et al.* [6]. However, in many cases the minimal and functional structural entity of a network is not a simple dyad, but rather, a small sub-pattern (or motif) involving more than two nodes. Among possible motifs, the simplest involves three nodes (or triads), where the fully connected triad represents the basic unit of transitivity. For example, in a social network of friendship ties, transitivity might suggest that *friends of my friends are also my friends*. The importance of transitivity dates back to Watts and Strogatz [7], where the clustering coefficient was proposed and quantifies the total number of triangles in a network via the average likelihood that two neighbors of a vertex are also neighbors. Networks with a high level of transitivity are often more stable, balanced and harmonious. For social networks, Granovetter [3] in his work on "strength of weak ties" explains that strong social links are transitive and result in redundant social structures like cliques. On the other hand, bridging links that lie

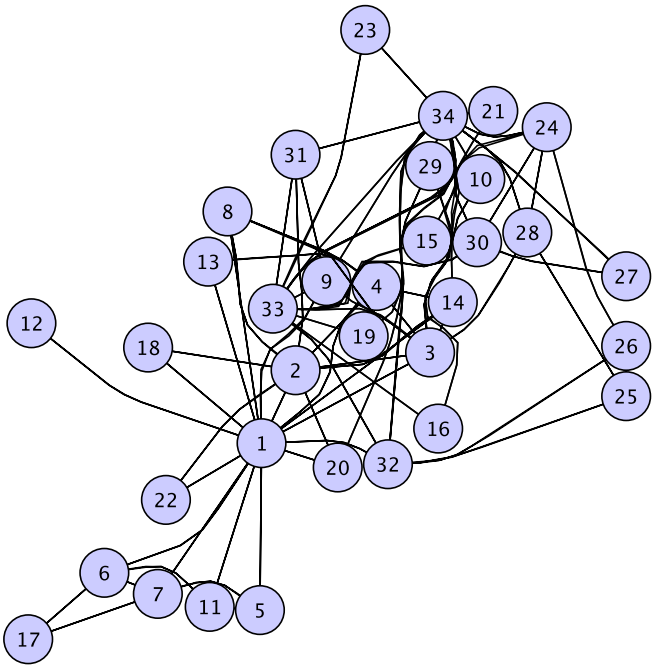

Figure 1: Zachary's karate club network.

between clusters are weak and do not follow the transitivity relations. Under this theory one can define a community (or cluster) to be a tightly knit and highly transitive group of nodes.

To motivate this work, consider the well-known 34 node karate club friendship network of Zachary [8] in Figure 1. Using the methodology given in Perry *et al.* [6], which uses the simple dyad as the basis for forming clusters, the network is clustered into 5 distinct groups, e.g., see Figure 2. Based on the idea that a community consists of tightly knit and highly transitive nodes, then the clustering exercise identified four communities and a *periphery* group. The periphery group consists of those nodes in the network with small degree, and very little influence, or nodes that are not tightly inwoven into any particular cluster. Notice that nodes 5 and 11 in Figure 2 are considered to be inwoven into the cluster containing nodes 6, 7, and 17; however, nodes 5 and 11 are only *weakly* connected to this cluster since they lack any transitive ties within this group. Thus, it would appear that nodes 5 and 11 should belong to the periphery group. If the triangle motif was used as the basis of forming clusters, as opposed to the simple dyad, it might be expected that nodes 5 and 11 would be placed in the periphery group, leaving nodes 6, 7, and 17 as one of the core communities. Such a placement would result in better agreement with Granovetter's definition of a community.

The primary objective of this work is to develop a statistical framework for clustering undirected networks by considering triangle motifs as building blocks for forming the clusters. Since triangle configurations are often overrepresented in many real-world networks, it is expected that the communities found based on this sub-pattern will improve upon those found based on the simple dyad, yielding more tightly knit and highly transitive groups of nodes. In the next sections, an approach to clustering based on the triangle motif is developed. Subsequently, using Monte Carlo simulation, clustering performance of the proposed algorithm is evaluated when applied to LFR benchmark graphs, relative to the method proposed by Perry *et al.* [6].

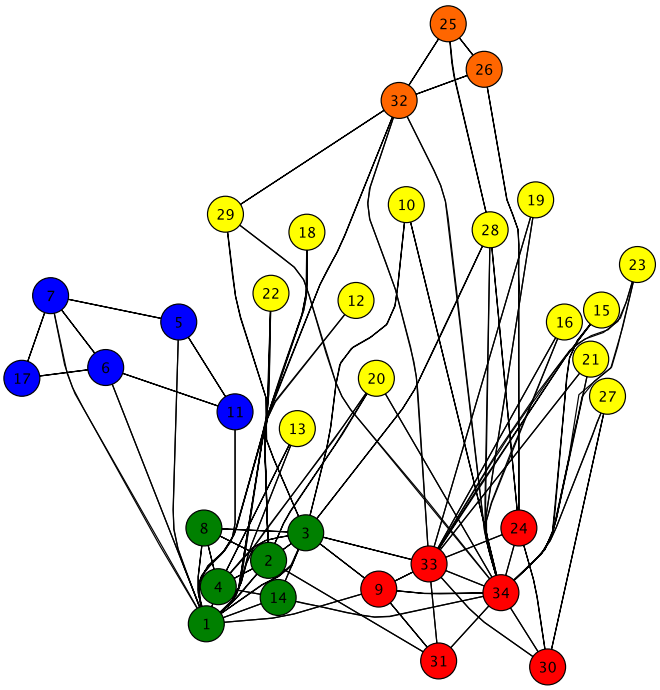

Figure 2: Karate club network clustered using the method given in Perry *et al.* [6].

Performance results of the proposed algorithm yield very promising results, and suggest that clustering on the basis of the triangle motif appears to outperform clustering on the basis of simple edges. The proposed algorithm is then applied to two real networks where the community structure is known. Finally, this report closes with a summary and discussion section.

3 Methodology

In this section, an approach to clustering undirected networks on the basis of the triangle motif will be discussed. Consider an undirected binary network with n nodes, and consider partitioning the n nodes into k mutually exclusive groups or clusters.

3.1 Technical Approach

Let $\mathbf{T} = [T_1, T_2, ..., T_k, T_b]'$ denote a $(k+1) \times 1$ vector of triangle counts, where T_m denotes the observed number of triangles between nodes assigned to group m, and T_b denotes the observed number of triangles formed between nodes belonging to different groups. In the paper by Perry *et al.* [6], the authors assume that the edges are sampled from independent Bernoulli trials, and thus they model edge counts as binomial random variables. Unfortunately, for the case of counting triangles, since each triangle shares an edge with $3n-9$ other triangles, the T_m's are not sums of independent Bernoulli trials. Specifically, let (X, Y) denote any two observed triangles, then for a Bernoulli(p) graph:

$$E(X) = E(Y) = p^3 \tag{1}$$

and

$$Cov(X,Y) = \begin{cases} 0, & if\ X\ and\ Y\ do\ NOT\ share\ an\ edge, \\ p^5(1-p), & if\ X\ and\ Y\ do\ share\ an\ edge, \\ p^3(1-p^3), & if\ X\ and\ Y\ are\ the\ same\ triangle, \end{cases}$$

so that the covariance between any two triangles that share an edge is $p^5(1-p)$. Thus, $X+Y$ cannot be a binomial random variable since X and Y are generally correlated. Note further that the vector $\mathbf{T}$ defined above is not a sufficient statistic since it contains no information about this covariance. That is, although the T_m's are uncorrelated, the T_m's are all correlated with T_b (see Appendix B), i.e.,

$$Cov(T_m, T_b) = 3 \begin{pmatrix} n_m \\ 3 \end{pmatrix} \left[(n-n_m)p^5(1-p) \right]. \tag{2}$$

Applying the methodology given in Perry *et al.* [6] to this problem would not be appropriate, since the likelihood approach they recommend requires sufficiency for the asymptotic results to hold. One could still use their method to cluster by simply replacing edge counts with triangle counts; however, their statistical test is less reliable. One alternative is to approximate sums of dependent Bernoulli trials as Poisson random variables, e.g., see Chen and Rollin [1]; however, it is not clear how to write the joint density of $\mathbf{T}$ in this case since the T_m's are correlated with T_b. Multivariate Poisson models do exist; however, the assumption is typically made that the correlation structure is *exchangable*, i.e., all pairs have same covariance. Unfortunately, this does not accurately describe the covariance structure of the vector $\mathbf{T}$. If an accurate *posterior* statistical test is to be derived on the estimated clusters, one would need to fully specify the joint likelihood function for the clustered network. This is a topic of ongoing research.

Instead of assessing the effectiveness of the clustering effort *after* the clustering operation is performed (such as that proposed in Perry *et al.* [6]), one alternative is to develop an *a priori* test to determine whether a clustering effort might be effective. If the test is significant, then this would suggest the need to cluster. An insignificant test would suggest the opposite. Let A denote the observed adjacency matrix and consider the null hypothesis H_0: number of triangles in A is consistent with Bernoulli graph with probability p versus the alternative H_1: number of triangles in A is NOT consistent with Bernoulli graph with probability p. If we define T as the total number of observed triangles in A, then one could test this hypothesis by computing

$$Z = \frac{T - E(T)}{\sqrt{Var(T)}}, \tag{3}$$

where for even moderately-sized networks Z is approximately standard normal under H_0, and thus, one can compare $|Z|$ to the upper $\alpha/2$ quantile of the standard normal distribution, where α is a user-specified type I risk. Note that $E(T)$ and $Var(T)$ in equation (3) are computed as given in equations (11) and (12) in Appendix A, respectively, by replacing p with $\hat{p}$, where $\hat{p}$ is computed from the observed adjacency matrix A and denotes the estimated density of the network. Thus, if $|Z| > z_{\alpha/2}$, the test concludes in favor of H_1. Note that $Z > z_{\alpha/2}$ suggests there are significantly more triangles in A than would otherwise be expected by a random graph with the same edge density[2], suggesting a need to cluster.

3.2 Quality Functions

Suppose that the *a priori* statistical test outlined above concludes in favor of H_1, and thus, a clustering effort is ensued. Then this section discusses a quality function to optimize in efforts to find the "best" set

[2]One could also use simulation and a configuration model to compute $E(T)$ and $Var(T)$ in equation (3), which will preserve both the density p, and the degree sequence of A. This approach, however, requires much more computational involvement.

of clusters. One possible approach to extending the method of Perry *et al.* [6] to the case of triangles is to assume that, for larger networks, the vector $\mathbf{T}$ follows a multivariate normal distribution with means, variances, and covariances given by the expressions derived in Appendix B. However, this is problematic since such an assumption would rely on the central limit theorem (CLT), and the CLT only applies to sums of *iid* random variables. Unfortunately, the elements in $\mathbf{T}$ are not sums of *iid* random variables; rather they are generally sums of correlated Bernoulli random variables.

The difficulty with modeling correlated Bernoulli trials is well known. In fact, it is not clear that a probability distribution even exists in some cases, as correlation coefficients must satisfy a stringent algebraic relationship, e.g., see Hisakado *et al.* [2]. One can use Poisson approximations to the marginal triangle counts in $\mathbf{T}$, and write out the likelihood function of the clustered network as if the elements in $\mathbf{T}$ were all uncorrelated. In doing so, we are essentially ignoring the correlation between the T_m's and T_b, and thus any statistical tests derived on the estimated clusters using this specification of the likelihood would be pseudo tests at best. However, if one's interest only lies in assessing the quality of the cluster, then such a specification might be sufficient. Specifically, one could assess the quality of a given group membership assignment vector $\mathbf{z}_k$ by evaluating

$$Q(\mathbf{z}_k|p,k) = \sum_{m=1}^{k} \log_e \frac{\exp\{T_m \log_e(\mu_{T_m}) - \mu_{T_m}\}}{T_m!} + \log_e \frac{\exp\{T_b \log_e(\mu_{T_b}) - \mu_{T_b}\}}{T_b!} \tag{4}$$

where $\mu_{T_m} = C(n_m)p^3$, $\mu_{T_b} = [C(n) - \sum_{m=1}^{k} C(n_m)]p^3$, and p denotes the overall density of the network. Note that Q is maximized when $T_m = \mu_{T_m}$ $(m = 1, ..., k)$ and $T_b = \mu_{T_b}$. Further, as the T's deviate from the μ's, Q gets smaller. Thus, one way to find the "best" group membership assignment vector is to find that $\mathbf{z}_k$ that yields the greatest discrepancy between what we observe and what we expect to see if all groups had density p. This can be accomplished by minimizing Q over all possible group membership assignments for a given p.

It should be noted that for any given p, $Var(T_m) > E(T_m)$ (and $Var(T_b) > E(T_b)$), suggesting that perhaps a better quality function to minimize is

$$\begin{aligned} Q(\mathbf{z}_k|p,k) = \sum_{m=1}^{k} \log_e \frac{\Gamma(T_m + r_m)}{\Gamma(T_m + 1)} \left(\frac{\mu_{T_m}}{\mu_{T_m} + r_m}\right)^{T_m} \left(\frac{r_m}{r_m + \mu_{T_m}}\right)^{r_m} \\ + \log_e \frac{\Gamma(T_b + r_b)}{\Gamma(T_b + 1)} \left(\frac{\mu_{T_b}}{\mu_{T_b} + r_b}\right)^{T_b} \left(\frac{r_b}{r_b + \mu_{T_b}}\right)^{r_b} \end{aligned} \tag{5}$$

where equation (5) is based on a Poisson-Gamma mixture parameterization of the negative binomial distribution, and accounts for the over dispersion observed in the triangle counts. A limiting form of the function in equation (5) is that in equation (4) when the over-dispersion parameters (i.e., r_m's and r_b) approach infinity. One can determine the r's via the relationship $Var(T_m) = \mu_{T_m} + \frac{\mu_{T_m}^2}{r_m}$ (and similarly for r_b), where μ_{T_m} and $Var(T_m)$ can be computed from equations (16) and (17) in Appendix B, respectively.

In practice, one can compute $\hat{p}$ from the observed adjacency matrix of the network, and then use a meta-heuristic (e.g., simulated annealing) to minimize Q over the set of possible group membership assignments. As we shall see in the next section of this report, this approach yields better clustering performance when applied to LFR benchmark graphs, relative to the method proposed by Perry *et al.* [6].

3.3 Clustering Algorithm

Using information provided in the discussions above, as well as results derived in the appendices, a practical algorithm for clustering undirected networks on the basis of the triangle motif is described in this section.

In addition to the *a priori* statistical test described above (which serves to justify the clustering effort), a *posterior* statistical test will also be described and will be used to establish a stopping criteria for the algorithm when the number of groups is unknown, thus, providing an estimate for k.

For a given network density p and number of groups k, the optimization problem can be written formally as

$$\mathbf{z}_k^* = \arg\min_{\mathbf{z_k} \in \mathbf{Z_k}} [Q(\mathbf{z}_k | p, k)], \tag{6}$$

where $\mathbf{z}_k^*$ is the $n \times 1$ group membership assignment vector with elements $z_k^*(i) \in [1, 2, ..., k]$ that minimizes Q, and $\mathbf{Z}_k$ is the set of all possible membership assignment vectors for a given k. The optimization problem given in equation (6) is challenging, particularly due to the combinatorial explosion for large n and k. In general, the problem is NP hard even for moderately-sized networks. As a consequence, throughout this effort, a simulated annealing (SA) algorithm will be employed to effectively search the vast set $\mathbf{Z}_k$ in attempts to locate $\mathbf{z}_k^*$. Although there are several heuristics that could be used, the SA was chosen due to its ease of programming.

Often in practice the number of groups is unknown, and thus the problem involves finding $\mathbf{z}_{k^*}^*$, or the "best" group membership assignment vector corresponding to the "best" k. The optimization problem for k can be written formally as

$$k^* = \arg\min_k \{ \min_{\mathbf{z_k} \in \mathbf{Z_k}} [Q(\mathbf{z}_k | p, k)] \} \tag{7}$$

where k^* denotes an estimate for the number of groups. Since, for any k, only one parameter is estimated (i.e., p), one can simply minimize equation (6) over a range of k, and return that integer k that yields the smallest Q. Unfortunately, it is not often the case that the practitioner can define an appropriate range for k, and even more problematic, is that the computation time required to find k^* can be significant if the range for k is large.

In attempts to reduce the computation time required when k is unknown, we suggest the following strategy. Suppose that for a given k one finds $\mathbf{z}_k^*$ via equation (6). Then define the null hypothesis for group m as H_{0_m}: number of triangles in group m is consistent with Bernoulli graph with probability p_m. Then for each m, one can compute

$$Z_m = \frac{T_m - E(T_m)}{\sqrt{Var(T_m)}}, \tag{8}$$

where T_m was defined earlier and denotes the number of observed triangles in group m, and $E(T_m)$ and $Var(T_m)$ are defined in Appendix B. Note that in equation (8) the mean and variance of T_m are computed using the observed *within-group* density, or $\hat{p}_m$. If the null hypothesis is true for each $m = 1, ..., k$, then

$$W = \frac{\sum_{m=1}^k n_m Z_m}{\sqrt{\sum_{m=1}^k n_m^2}} \tag{9}$$

approximately follows a standard normal distribution, where n_m denotes the number of nodes assigned to group m. The test statistic defined in equation (9) is known as Stouffer's test, and follows closely in theory to Fisher's combined probability test. Thus, one can compute W for a given group membership assignment vector and compare to the upper δ quantile of the standard normal distribution, or z_δ. If $W < z_\delta$, then this would suggest little evidence to support the need to cluster further since the m subgraphs found via the clustering effort appear to be consistent with Bernoulli graphs with corresponding probability p_m. Obviously, if $p_m = 1$, this implies a *complete* subgraph, or "clique", and if $p_m = 0$, this denotes an *empty* subgraph. In these two cases, the variance of T_m is zero and equation (8) is undefined. Therefore, components of the

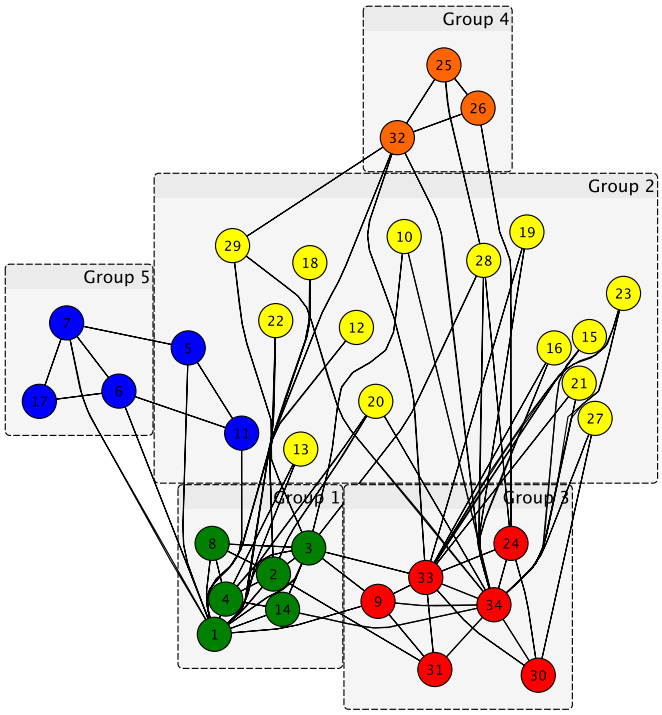

Figure 3: Karate club network network clustered using the proposed methodology.

sum in equation (9) that involve these cases are ignored. For example, suppose that $k = 3$ and $p_1 = 0.4$, $p_2 = 0.6$, and $p_3 = 1$. Then the summations in equation (9) will only include the components corresponding to groups 1 and 2, since Z_3 is undefined.

Using equation (9), one can then define a stopping scheme in the search for the "best" k. Suppose that the search begins with $k = 2$, and upon completion of the clustering effort one obtains a reasonable estimate for $\mathbf{z}_2^*$. At this point one can compute W and compare to the z_δ critical value. If $W > z_\delta$, then increment k by 1 and repeat the search to find $\mathbf{z}_3^*$. Once can continue to increment k until $W < z_\delta$, at which point the algorithm will stop and return $\mathbf{z}_{k^*}^*$. Although the proposed strategy is simple, it is also shown to be quite effective.

At this point the karate club network of Zachary [8] can be revisited using the proposed clustering framework with the objective function given in equation (5). Figure 3 shows the results, where groups of nodes contained in the large rectangles were identified using the proposed approach. The proposed method assigned nodes to groups similarly to the method of Perry *et al.* [6]; however, as anticipated, nodes 5 and 11 were assigned to the periphery group since they lack any transitive ties to nodes 6, 7, and 17. Thus, the four main clusters of nodes identified using the proposed framework consist primarily of "strong ties", and therefore, are in better agreement with Granovetter's definition of a community.

Although application of the proposed clustering framework to the karate club network appears to perform quite well, it reveals nothing about how the proposed approach performs on the average. Therefore, in the next section results of Monte Carlo simulation studies used to assess the expected clustering performance of the proposed algorithm when applied to LFR benchmark graphs are presented. Clustering performances using both forms of Q given in equations (4) and (5) are evaluated, and for both cases of known and unknown

k, relative to the method suggested by Perry *et al.* [6].

4 Performance Results

In this section, results of Monte Carlo simulation studies used to assess the performance of the proposed clustering framework are presented. Due to their non-homogeneous degree and community size distributions, the LFR benchmark graphs developed by Lancichinetti *et al.* [4] are used in assessing the performances. For a simple review of these benchmark graphs, see Section 5.1 of Perry *et al.* [6]. The proposed clustering framework developed in the previous section is evaluated and subsequently compared to the performances achieved by using the framework in Perry *et al.* [6]. For the proposed method, both objective functions shown in equations (4) and (5) are considered.

Clustering performance for any given method is measured using the adjusted mutual information (AMI), and to make *relative* comparisons amongst the three competing approaches, the relative mean index (RMI) is employed. Refer to Perry *et al.* [6] for explicit detail on these performance measures. LFR benchmark graphs of size $n = 100$ nodes with a maximum degree of 20 and three different values of average degree, or $AveDeg$ =8, 10, and 12 are considered. For each value of $AveDeg$, four different combinations for the parameters γ and β are considered; namely, (γ, β)=$(2, 1)$, $(2, 2)$, $(3, 1)$, and $(3, 2)$, which encompasses the extremes of the ranges of these LFR benchmark parameters. Note that γ is the parameter for the power law on the degree distribution in the LFR benchmark graphs, and similarly, β for the community size distribution. Finally, values of the mixing parameter μ are considered between 0.1 and 0.5, in increments of 0.1, where small values of μ suggest more dense intra-group and less dense inter-group connections.

4.1 Cooling Schedule for SA Algorithm

For all methods evaluated, a simulated annealing algorithm was used to optimize the corresponding objective function. For more details about simulated annealing and its theoretical underpinnings, see Kirkpatrick *et al.* [5]. The cooling schedule for the SA algorithm was set equal for all simulations and is shown in Table 1. Note that the parameters *Max # of successes* and *Max # consecutive rejections* represent the maximum number of successes before the SA temperature is reduced and the maximum number of consecutive rejections before the SA algorithm quits, respectively.

Table 1: Cooling schedule for SA algorithm.

Parameter	Value
Initial Temperature	1
Cooling Rate	0.99
Temperature Length	300
Stop Temperature	1.0×10^{-8}
Max # of successes	30
Max # consecutive rejections	2000

4.2 Clustering Performance when k is Known

Clustering performance is first investigated for the case where the number of groups k is known *a priori*, and although this is a less likely scenario relative to the unknown k case, it is studied for completeness. In what follows, the simulation model is given in detail.

For any given simulation run, an LFR benchmark graph with parameters μ, γ, β and $AveDeg$ was generated and subsequently clustered using three different approaches: 1) the method of Perry *et al.* [6], 2) the proposed method with objective function given in equation (4), and 3) the proposed method with objective function given in equation (5). This process is repeated $N = 100$ times, and the average AMI values were then computed over the N independent Monte Carlo simulations. Figure 4 illustrates the results from the method given in Perry *et al.* [6], which uses edges as the basis for building clusters, while Figures 5 and 6 illustrate results from the proposed approach. Since an AMI of unity would suggest perfect clustering, it would appear from these figures that forming clusters on the basis of the triangle motif is effective. In general, as the mixing parameter μ increases, the clustering performance decreases. This is intuitive since smaller μ indicates denser clusters. It also appears that as the average degree of the network approaches the maximum degree, the clustering performance increases. This is also intuitive since as the average degree approaches the maximum degree, the degree distribution becomes more homogeneous. Further, general clustering performance increases as the parameter γ increases. This result is also intuitive as an increase in γ results in a decrease in the variance of the degree distribution of the graph. There also appears to be a small effect on clustering performance due to changes in β, i.e., as β increases, so does the AMI. This is particularly true for larger values of the mixing parameter μ. In order to make relative performance comparisons, we compute the RMI for each method being compared across the mixing parameter μ. The method that achieves the best performance then yields the smallest RMI. Results are shown in Table 2, and suggest that clustering using the proposed method with the objective function in equation (5) generally achieves the better relative performance.

It should be pointed out that performances of all thee methods could likely be improved with appropriate changes to the cooling schedule of the SA algorithm. However, even if results are suboptimal, those given here still provide a good idea of the performances of the methods as functions of LFR benchmark parameters, as well as relative performance between the competing methods.

4.3 Clustering Performance when k is Unknown

In this section, results are presented for the k unknown case. For the method given in Perry *et al.* [6], the value of $k \geq 2$ that yields the minimum Bayesian information criterion (BIC) over a predetermined range of k is then denoted as k^*. For the proposed method, the "best" k is the smallest value of $k \geq 2$ such that $W > z_\delta$, where $\delta = 0.001$ and W is Stouffer's test statistic given in equation (9).

Figure 7 shows average AMI results obtained from the method given in Perry *et al.* [6], while Figures 8 and 9 show average AMI results obtained from the proposed method. Table 3 shows corresponding RMI values for the three methods. General results are similar to the k known case given in the previous subsection and suggest that the proposed methodology should be considered as an alternative to the method given in Perry *et al.* [6], especially when the clusters are less densely connected (i.e., $\mu > 0.3$).

Since the k unknown case is considered here, estimates for the root mean square error (RMSE) of k^* (i.e., the estimated number of groups) were also obtained from the simulation model over the $N = 100$ Monte Carlo runs, and are illustrated graphically in Figures 10, 11, and 12, for the methodology outlined in Perry *et*

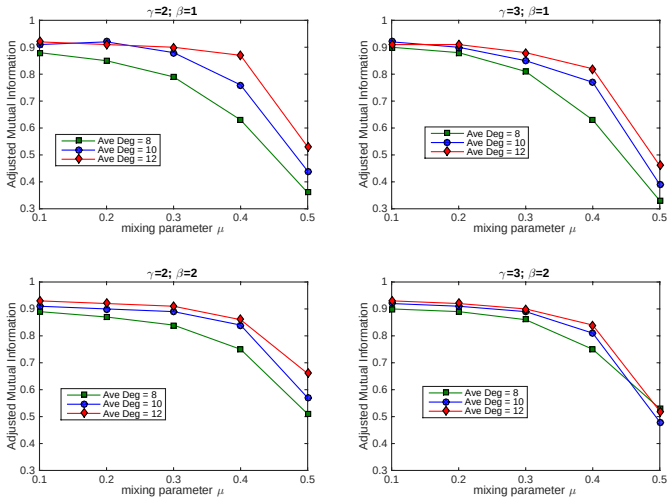

Figure 4: Average AMI results for methodology in Perry *et al.* [6] for the case where k is known.

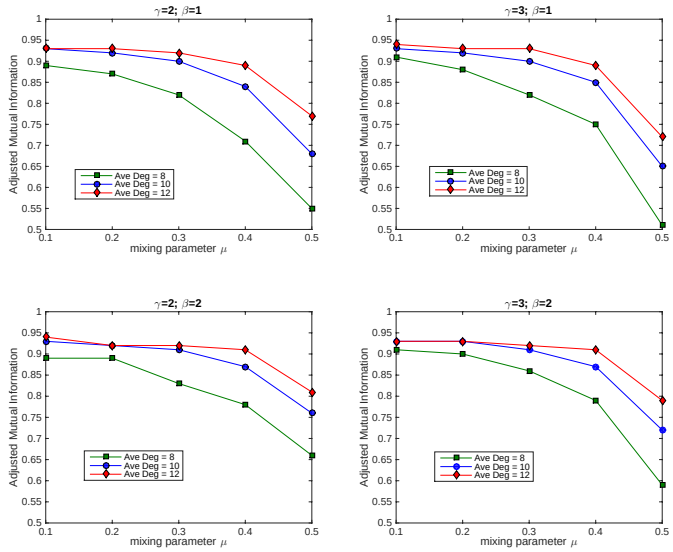

Figure 5: Average AMI results for proposed methodology using Q_1 in equation (4) for the case where k is known.

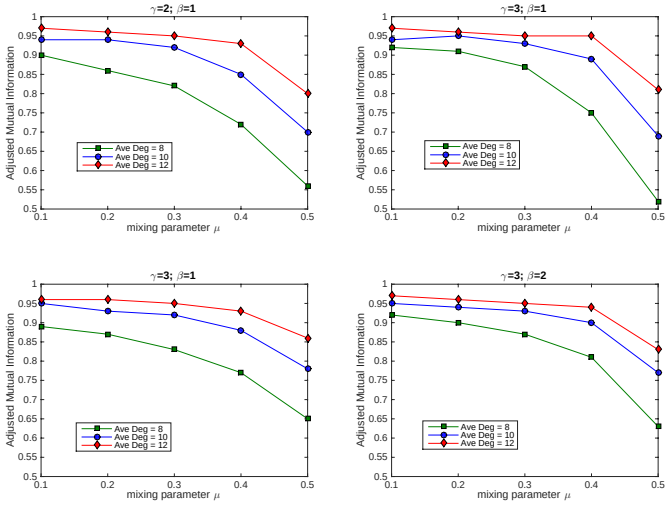

Figure 6: Average AMI results for proposed methodology using Q_2 in equation (5) for the case where k is known.

Table 2: RMI results of AMIs for k known case.

	Ave Degree	Edges	Q_1	Q_2
$\gamma = 2,\ \beta = 1$	8	0.111	0.011	0.001
	10	0.118	0.018	0.000
	12	0.111	0.039	0.000
$\gamma = 3,\ \beta = 1$	8	0.129	0.023	0.000
	10	0.149	0.037	0.000
	12	0.153	0.049	0.000
$\gamma = 2,\ \beta = 2$	8	0.056	0.001	0.000
	10	0.086	0.014	0.000
	12	0.085	0.034	0.000
$\gamma = 3,\ \beta = 2$	8	0.058	0.025	0.000
	10	0.118	0.030	0.000
	12	0.126	0.038	0.000

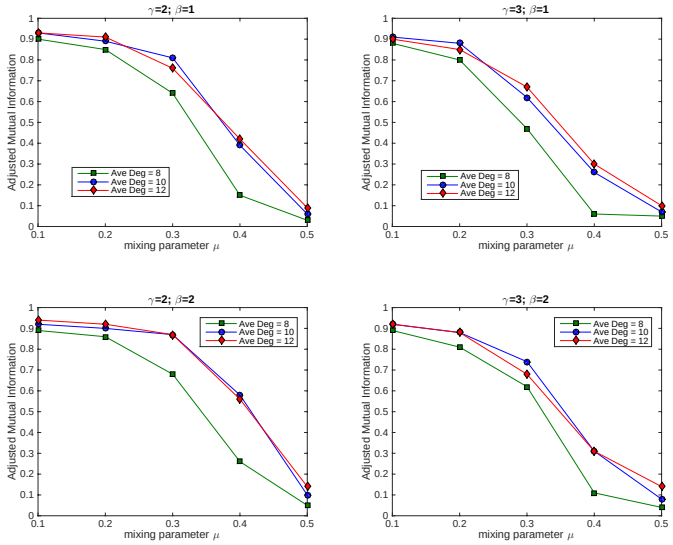

Figure 7: Average AMI results for methodology in Perry *et al.* [6] for the case where k is unknown.

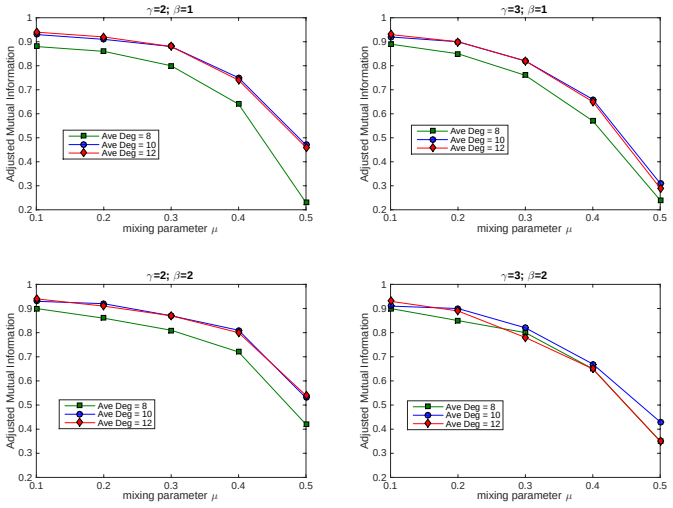

Figure 8: Average AMI results for proposed methodology using Q_1 for the case where k is unknown.

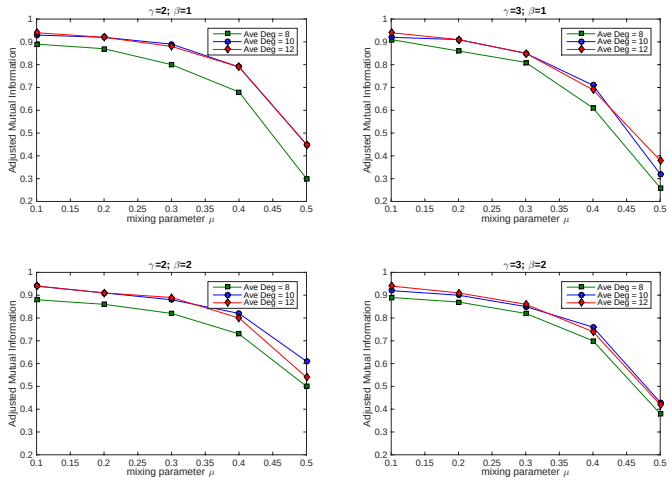

Figure 9: Average AMI results for proposed methodology using Q_2 for the case where k is unknown.

Table 3: RMI results of AMIs for k unknown case.

	Ave Degree	Edges	Q_1	Q_2
$\gamma = 2,\ \beta = 1$	8	0.379	0.060	0.001
	10	0.298	0.018	0.011
	12	0.284	0.012	0.005
$\gamma = 3,\ \beta = 1$	8	0.446	0.043	0.000
	10	0.349	0.031	0.000
	12	0.322	0.070	0.000
$\gamma = 2,\ \beta = 2$	8	0.351	0.038	0.003
	10	0.230	0.036	0.003
	12	0.210	0.007	0.002
$\gamma = 3,\ \beta = 2$	8	0.409	0.038	0.001
	10	0.312	0.034	0.000
	12	0.300	0.083	0.000

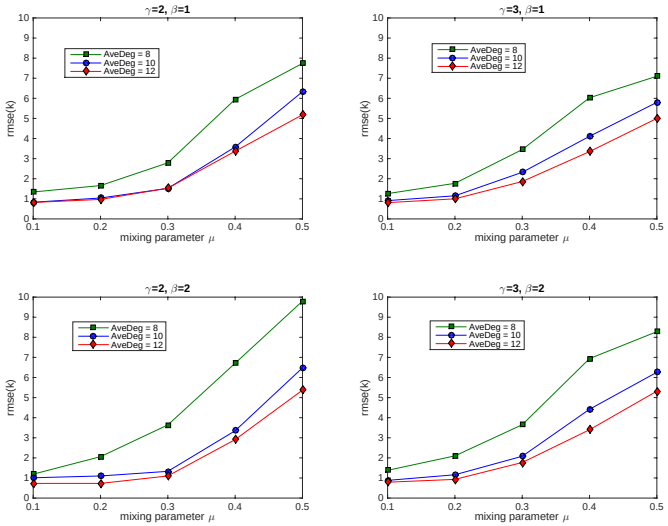

Figure 10: Root mean square estimates for k using methodology in Perry *et al.* [6].

al. [6], and the proposed methodology with objective functions given by equations (4) and (5), respectively. Results are quite intuitive as an increase in the RMSEs is observed as the mixing parameter increases, and a decrease in the RMSE is observed as the degree distribution becomes more homogeneous. Also, there does not appear to be any evidence that the RMSE performance is significantly affected by the parameters γ and β; however, more simulation runs may reveal a small effect. The RMIs in Table 4 summarize the RMSEs across the mixing parameter μ and suggest that the proposed method, using either of the objective functions given in equations (4) and (5), consistently outperforms the method suggested by Perry *et al.* [6]. It should be noted that the parameter δ likely has a significant effect on the RMSE of the estimated number of groups. However, due to the short interval of time allotted for this research effort, and the fact that primary interest was in *relative* performance comparisons, it was decided to set $\delta = 0.001$ in the simulations and leave performance evaluations of the proposed method as a function of δ for future research.

5 Application to Real Networks

In the section the proposed clustering methodology is applied to two real networks, one corresponding to an approximately homogeneous degree sequence, and the other corresponding to a highly skewed degree sequence. In particular, the 2012 college Football Bowl Series (FBS) game schedules network will be considered, as well as the author's personal Facebook network. The reasoning behind choosing these networks lies in the fact that the community structure is known, and thus, a meaningful evaluation of our method can ensue.

5.1 2012 FBS Football Schedule Network

In this subsection the 2012 college Football Bowl Series (FBS) game schedules network is clustered using the proposed method. In this network, each node denotes an FBS college or university, and if an edge exists between two nodes, then these two colleges played each other. The network is illustrated in Figure 13, while

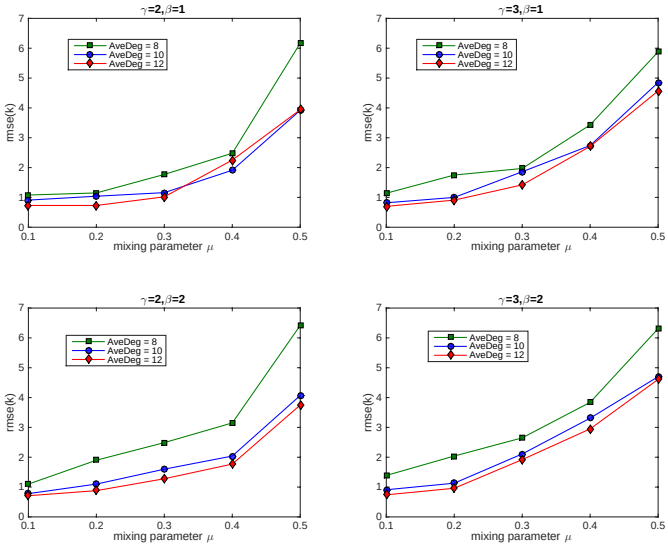

Figure 11: Root mean square estimates for k using proposed methodology with Q_1.

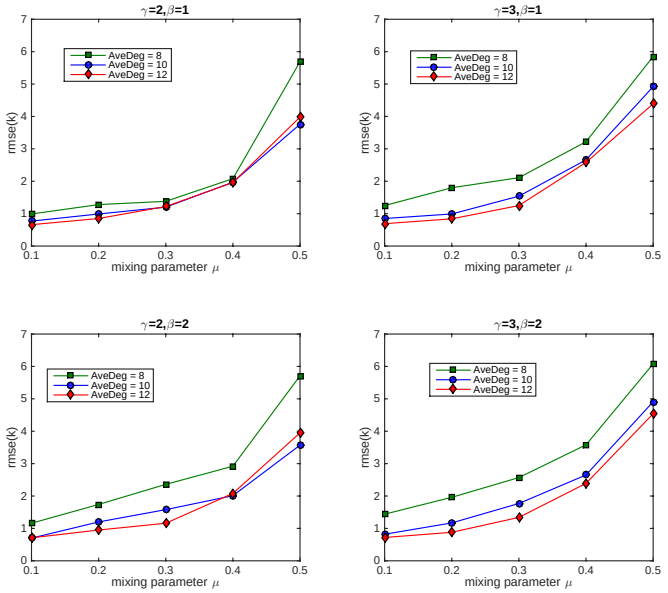

Figure 12: Root mean square estimates for k using proposed methodology with Q_2.

Table 4: RMIs results of RMSEs for three different estimators of k.

	Ave Degree	Edges	Q_1	Q_2
$\gamma = 2$, $\beta = 1$	8	0.807	0.131	0.021
	10	0.402	0.056	0.014
	12	0.416	0.049	0.071
$\gamma = 3$, $\beta = 1$	8	0.392	0.014	0.039
	10	0.302	0.050	0.009
	12	0.258	0.065	0.000
$\gamma = 2$, $\beta = 2$	8	0.564	0.071	0.010
	10	0.383	0.095	0.056
	12	0.232	0.065	0.117
$\gamma = 3$, $\beta = 2$	8	0.358	0.038	0.008
	10	0.254	0.107	0.016
	12	0.216	0.162	0.000

Figure 14 shows a stem plot of of the degree-ordered vertices versus degree. Notice that this network has a fairly constant degree sequence, and thus the degree distribution is near homogeneous.

The *a priori* statistical test on the triangle count was highly significant, with $Z = 24.5$ and corresponding p-value of practically zero, suggesting an overrepresentation of triangles, and thus, the presence of highly transitive subgraphs. Note that the actual clusters are known for this network. In particular, there are 11 primary clusters, each representing one of the major conferences in the FBS. Additionally, there are four "independent" colleges or vertices, namely, Notre Dame, BYU, Army, and Navy. An interesting application of the proposed clustering algorithm to this network is assessing the strength of schedules for the independent teams, relative to the other conferences in the FBS. Applying the proposed clustering algorithm to the FBS network using the objective function given in equation (5) and $\delta = 0.1$ yields the results shown in Figures 15 and 16. Notice that the proposed algorithm correctly identified all 11 conferences, as well as those teams that belong to those conferences. The "independent" teams were also assigned to a conference, with Notre Dame assigned to the *Big 10*, BYU and Navy assigned to the *WAC*, and Army assigned to the *MAC*. This would seem to suggest that these independent teams have strength of schedules on par with the conferences to which they were assigned.

5.2 Author's Personal Facebook Network

In this subsection the author's personal Facebook network is clustered using the proposed framework. This network is illustrated graphically in Figure 17, and a stem plot of the degree-ordered vertices versus the degree is shown in Figure 18. Notice that for this network, the degree sequence is much more skewed than that of the college football network. The *a priori* test on the triangle count yields $Z = 49.7$ with a p-value of practically zero. Since the degree sequence is skewed, a configuration model was also used to perform the test, where the degree sequence of the observed network was preserved. The test also resulted in a p-value of practically zero, suggesting the presence of highly transitive subgraphs. Figures 19 and 20 show output results of the algorithm, where a total of 9 groups were estimated. Notice that group 6 represents a periphery group, and the other 8 groups consist of tightly knit and highly transitive communities.

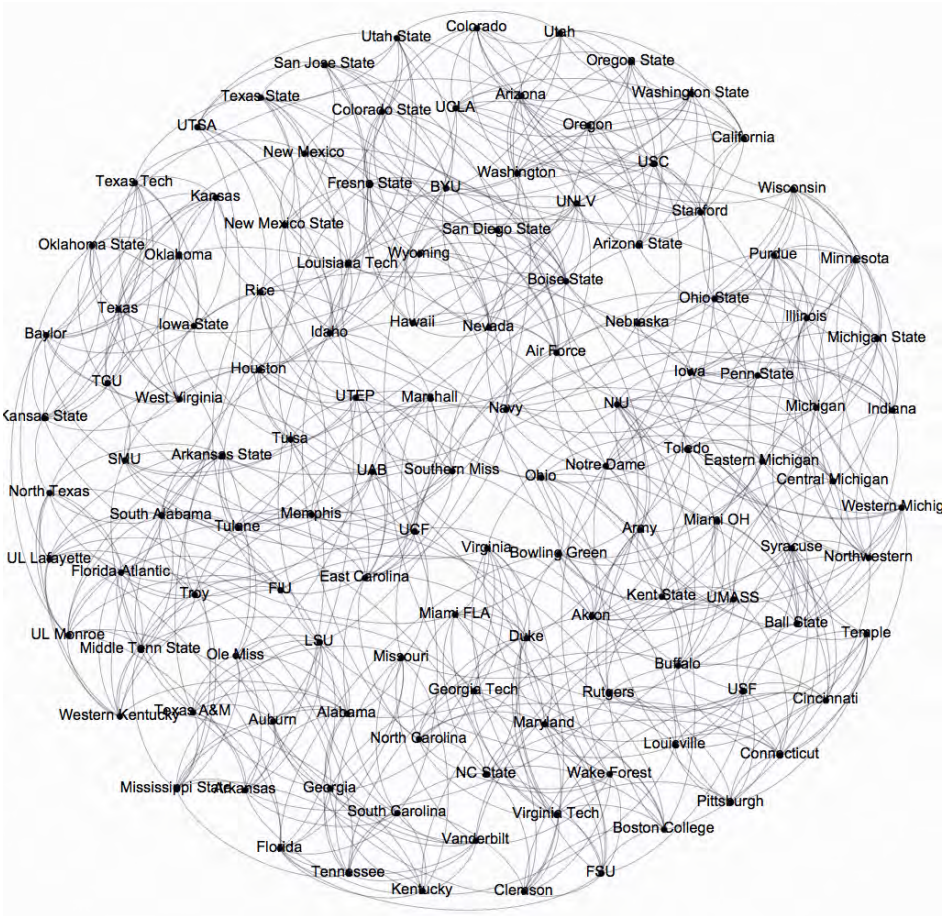

Figure 13: Unclustered FBS 2012 college football schedule network.

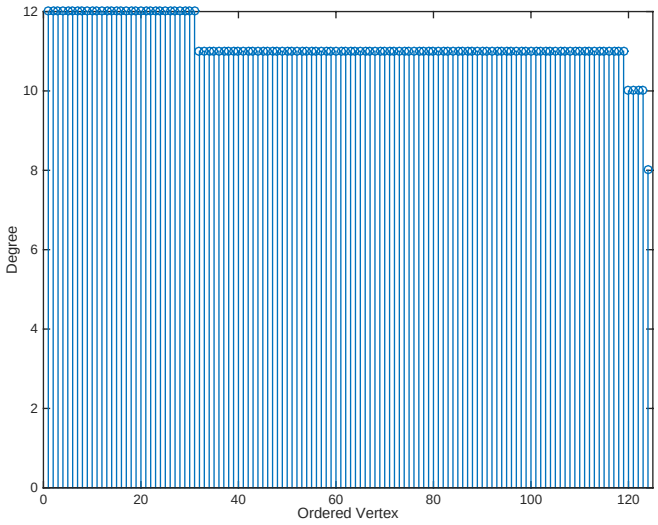

Figure 14: Stem plot of degree-ordered vertices versus the degree for college football network.

```
************************************
Estimated clusters= 11
Energy= -1725.09

{p-value}[density](group size)=
{0.265928}[0.75](9)
{0.538151}[0.653846](13)
{0.561376}[0.626374](14)
{0.670656}[0.727273](12)
{0.656283}[0.742424](12)
{No Test}[1](10)
{No Test}[1](8)
{0.624991}[0.888889](10)
{0.624991}[0.888889](10)
{0.65555}[0.818182](12)
{0.525925}[0.626374](14)

Group membership assignment vectors output to bestztriangles.txt file

Stouffer's test statistic for triangles= -0.560147
P-value =0.712388

Execution time = 0.65 minutes
************************************
```

Figure 15: Output of proposed algorithm implemented in C++ and applied to 2012 FBS college football network.

Since this is the author's Facebook network, the true community structure is known and the algorithm can thus be validated. All of the groups to the left of the red line in Figure 20 are groups of individuals associated with the author during his childhood and up through high school. All of the groups to the right of the red line correspond to individuals who became associated with the author through marriage. Essentially there are three main clusters: "Salvadoran Family", "Perry Family", and "High School", with smaller peripheral clusters attached to these main groups. The "High School" group was actually split into two communities, and appears to correspond to a difference in the ages of the individuals. Overall, and based upon the author's knowledge of this network, the proposed algorithm appears to correctly cluster the individuals into the appropriate communities, thus providing more validity in support of the proposed method.

6 Conclusions

In this effort an algorithm was developed for detecting clusters in undirected networks, where the triangle motif was used as the basis for forming clusters. It was shown that the likelihood ratio test outlined in Perry *et al.* [6] cannot be directly extended to the case of triangles since the vector of triangle counts $\mathbf{T}$ is not a complete sufficient statistic, i.e., it does not contain information about the correlation between T_m's and T_b. Therefore, in this effort, an *a priori* statistical test on the observed triangle count was developed, and was used as a means for deciding whether one should ensue a clustering exercise. It was also used as a basis for establishing stopping criteria for the proposed algorithm in the likely case that the number of groups k is unknown.

Given that sums of dependent Bernoulli random variables can often be well approximated by Poisson random variables, two different objective functions were proposed, i.e., one based on the Poisson mass function, and the other based on the negative binomial mass function. For a given observed network density,

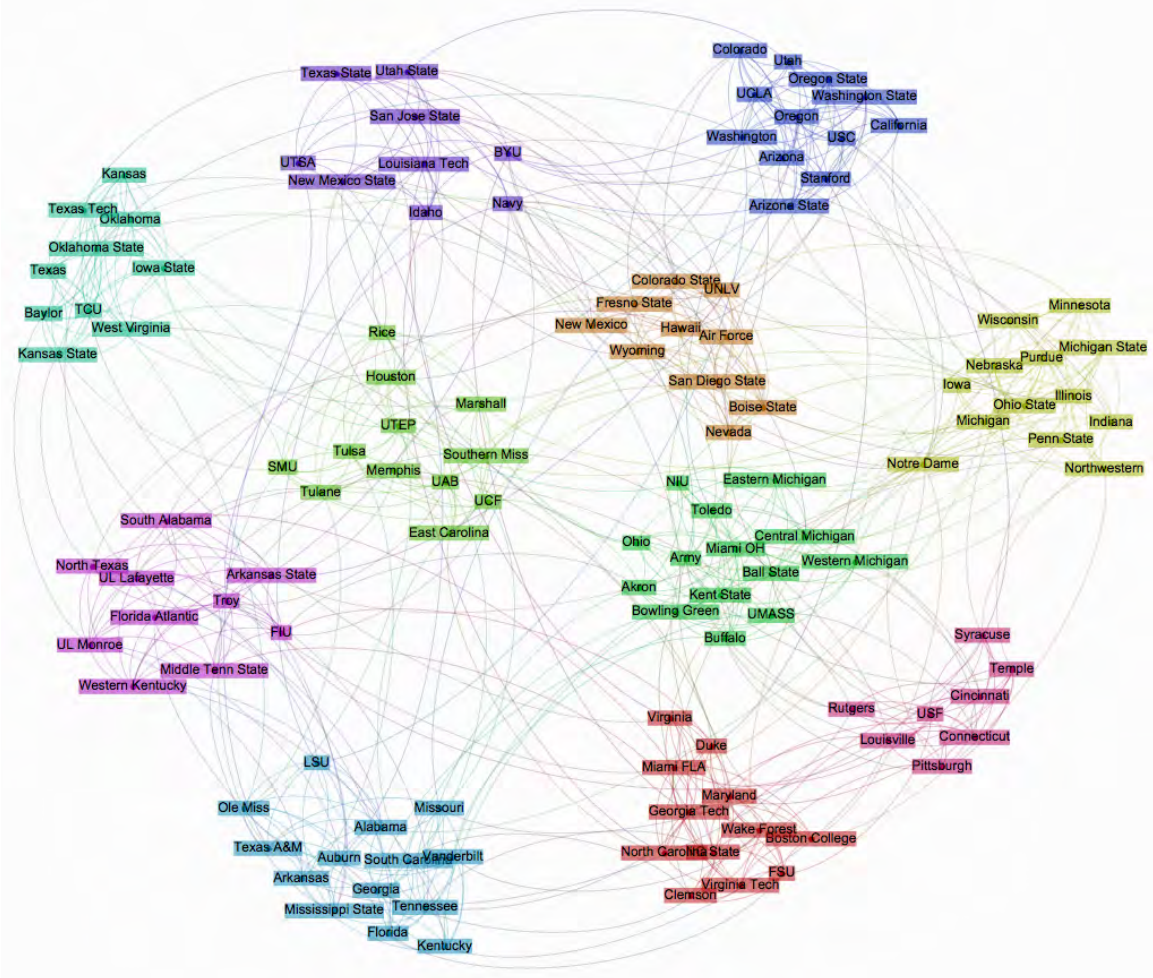

Figure 16: Clustered FBS 2012 college football schedule network.

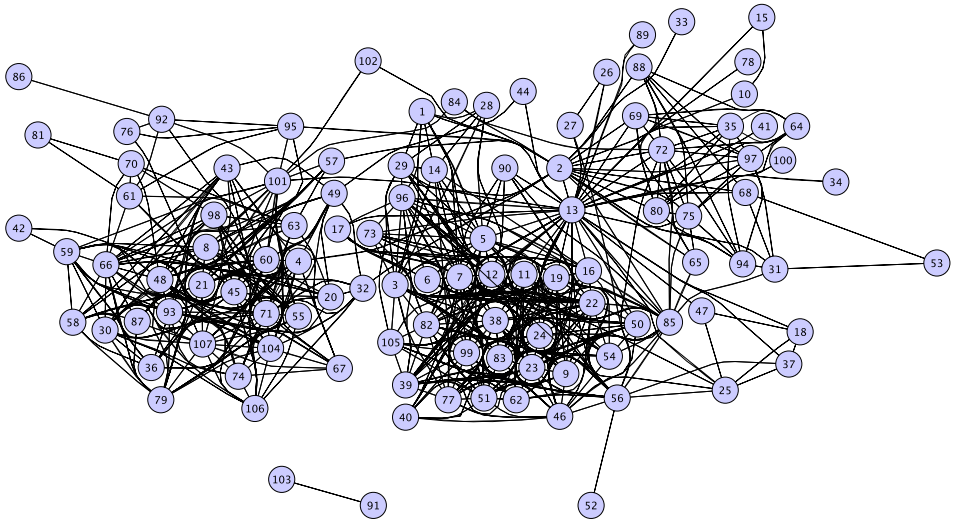

Figure 17: Unclustered Facebook network.

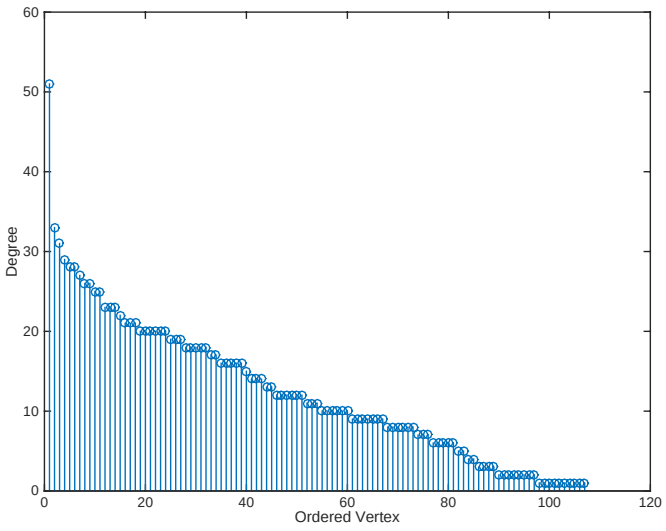

Figure 18: Stem plot of degree-ordered vertices versus the degree for Facebook network.

```
**************************************
Estimated clusters= 9
Energy= -1507.02

{p-value}[density](group size)=
{No Test}[1](4)
{0.284333}[0.720588](17)
{0.593301}[0.833333](4)
{0.593301}[0.833333](4)
{No Test}[1](3)
{0.536372}[0.014245](27)
{0.370397}[0.709091](11)
{0.153146}[0.729345](27)
{0.507372}[0.844444](10)

Group membership assignment vectors output to bestztriangles.txt file

Stouffer's test statistic for triangles= 0.814218
P-value =0.20776

Execution time = 0.883333 minutes
**************************************
```

Figure 19: Output of proposed algorithm implemented in C++ and applied to Facebook network.

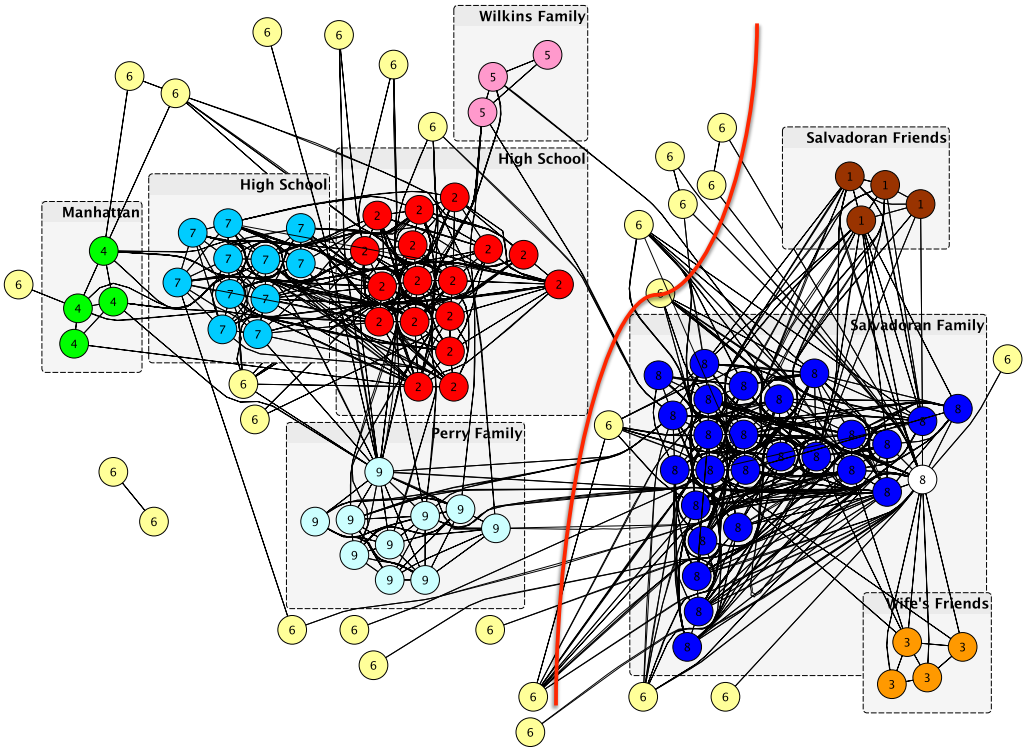

Figure 20: Clustered Facebook network.

say $\hat{p}$, the objective is to find the k-group partition that yields the greatest discrepancy between that observed and that expected if all k groups had density $\hat{p}$. This was accomplished by minimizing the proposed objective functions for a given $\hat{p}$. Using a statistical test for triangle counts, in conjunction with Stouffer's test, a stopping criteria for the algorithm was developed so that an estimate for k is made available. Thus, the resulting partition is one that maximizes discrepancy, *such that the k resulting subgraphs can be viewed as independent Bernoulli graphs.* It was shown that this approach appears to work quite well on both synthetic and real networks.

Monte Carlo simulation was used to assess the expected performance of the proposed algorithm when applied to the LFR benchmark graphs, and comparisons were made with the methodology outlined in Perry *et al.* [6]. General results suggest that clustering on the basis of the triangle motif, as opposed to the simple edge configuration, seems to yield significantly better clustering performance, especially when the mixing parameter μ is larger (say, $\mu \geq 0.3$). Additionally, when the number of groups k is unknown, the proposed algorithm will produce a smaller RMSE of the estimated k. Finally, the proposed method with either of the objective functions in equations (4) and (5) will outperform the methodology in Perry *et al.* [6]; however, use of the objective function in equation (5) generally achieves the best relative performance for all measures considered.

The work outlined in this report provides some significant opportunities for future research. One area involves deriving the exact joint distribution of the vector of triangle counts, $\mathbf{T}$ (or at least a reasonable approximation to this distribution). This could potentially permit the development of an approximate log-likelihood ratio test on the detected clusters, similar to that given in Perry *et al.* [6]. Further, the methodology outlined in this report might be a reasonable approach to clustering directed networks under particular motifs of interest. For example, one might be interested in partitioning a directed network, such that resulting groups have strong *within-group* and weak *between-group* transitive ties. This is a topic of the author's on-going research.

References

[1] Chen, L. H. Y. and Rollin, A. (2013), "Approximating dependent rare events," *Bernoulli* 19(4) pp. 1243-1267.

[2] Hisakado, M., Kitsukawa, K. and Mori, S. (2006), "Correlated binomial models and correlation structure," *J. Phys. A: Math. Gen.* 39 15365.

[3] Granovetter, M. (1983), "The strength of weak ties: A network theory revisited," *Sociological Theory* 1 pp. 201-233.

[4] Lancichinetti, A., Fortunato, S., Radicchi, F., (2008), "Benchmark graphs for testing community detection algorithms," *Physical Review E* 78 (046110).

[5] Kirkpatrick, S., Gelatt, C., Vecchi, M., (1983), "Optimization by simulated annealing," *Science* 4598 (220), 671?680.

[6] Perry, M. B., Michaelson, G. V., and Ballard, M. A. (2013). "On the statistical detection of clusters in undirected networks," *Computational Statistics and Data Analysis*, 68, pp. 170-189.

[7] Watts, D. J. and Strogats, S. H., (1998), "Collective dynamics of small world networks," *Nature*, 393, pp. 440-442.

[8] Zachary, W., 1977. "An information flow model for conflict and fission in small groups," *Journal of Anthropological Research* 33, pp. 452473.

7 Appendices

A: Mean and Variance of Triangle Count for Bernoulli Graph

Let A denote an $n \times n$ adjacency matrix with elements $a_{ij} \in [0,1]$ independently sampled from a Bernoulli(p) distribution for all $i < j = 1, ..., n$, and define

$$T = \sum_{i=1}^{n-2} \sum_{j=i+1}^{n-1} \sum_{\ell=j+1}^{n} a_{ij} a_{i\ell} a_{j\ell} \tag{10}$$

as the total number of triangles observed in A. Let $C(n) = \begin{pmatrix} n \\ 3 \end{pmatrix}$, then for a given p,

$$E(T) = C(n)p^3 \tag{11}$$

and

$$Var(T) = C(n)[(3n-9)p^5 - (3n-8)p^6 + p^3]. \tag{12}$$

Proof. The expected value in equation (11) is straight-forward; thus, I will prove the expression for the variance in equation (12). Note that in general $Var(T) = E(T^2) - E(T)^2$, where

$$E(T^2) = \sum_{i=1}^{n-2} \sum_{j=i+1}^{n-1} \sum_{\ell=j+1}^{n} \sum_{r=1}^{n-2} \sum_{s=r+1}^{n-1} \sum_{t=s+1}^{n} E(a_{ij} a_{i\ell} a_{j\ell} a_{rs} a_{rt} a_{st}). \tag{13}$$

The proof is essentially a counting exercise. Consider the index sets (i,j,ℓ) and (r,s,t) for $i < j < \ell$ and $r < s < t$, and let $\mathbf{R}$ denote a $C(n) \times C(n)$ matrix of cross-products of these index sets with elements $\{(i,j,\ell)(r,s,t)\}$. For example, if $n = 4$, then the cross-products matrix $\mathbf{R}$ is given by

$$\mathbf{R} = \begin{bmatrix} (1,2,3)(1,2,3) & (1,2,4)(1,2,3) & (1,3,4)(1,2,3) & (2,3,4)(1,2,3) \\ (1,2,3)(1,2,4) & (1,2,4)(1,2,4) & (1,3,4)(1,2,4) & (2,3,4)(1,2,4) \\ (1,2,3)(1,3,4) & (1,2,4)(1,3,4) & (1,3,4)(1,3,4) & (2,3,4)(1,3,4) \\ (1,2,3)(2,3,4) & (1,2,4)(2,3,4) & (1,3,4)(2,3,4) & (2,3,4)(2,3,4) \end{bmatrix}. \tag{14}$$

In general, there are $C(n)$ elements of $\mathbf{R}$ that satisfy $|(i,j,\ell) \cap (r,s,t)| = 3$ (these are the diagonal elements), and since each triangle shares an edge with $3n-9$ other triangles, there are $C(n)(3n-9)$ elements that satisfy $|(i,j,\ell) \cap (r,s,t)| = 2$. Lastly, there are $C(n)\left[C(n) - 3n + 8\right]$ elements that satisfy $|(i,j,\ell) \cap (r,s,t)| = 0$ or 1 (these are triangle pairs that do not share an edge), where the notation $|\cdot|$ denotes set cardinality. Also, note that

$$\mathrm{E}(a_{ij} a_{i\ell} a_{j\ell} a_{rs} a_{rt} a_{st}) = \begin{cases} p^6, & |(i,j,\ell) \cap (r,s,t)| = 0 \text{ or } 1 \\ p^5, & |(i,j,\ell) \cap (r,s,t)| = 2, \\ p^3, & |(i,j,\ell) \cap (r,s,t)| = 3, \end{cases}$$

so that equation (13) can be re-written as

$$E(T^2) = C(n)\left(\left[C(n) - 3n + 8\right]p^6 + (3n-9)p^5 + p^3\right) \tag{15}$$

and after subtracting $E(T)^2$ from equation (15) and simplifying we obtain the expression in equation (12). ■

B: Means, Variances and Covariances of Triangle Counts for Partitioned Bernoulli Graph

Suppose the vertex set is partitioned into k mutually exclusive groups, then the number of observed triangles in group m, of size n_m, is denoted by T_m and

$$E(T_m) = C(n_m)p^3 \tag{16}$$

and

$$Var(T_m) = C(n_m)[(3n_m - 9)p^5 - (3n_m - 8)p^6 + p^3]. \tag{17}$$

Let $T_b = T - \sum_{m=1}^{k} T_m$ denote the number of between group triangles, i.e., the number of triangles observed *between* the k groups. Then

$$E(T_b) = \left[C(n) - \sum_{m=1}^{k} C(n_m)\right] p^3 \tag{18}$$

where $n = n_1 + \cdots + n_k$ and

$$Var(T_b) = Var(T) + \sum_{m=1}^{k} Var(T_m) + 2 \sum_{m \neq m'} Cov(T_m, T_{m'}) - 2 \sum_{m=1}^{k} Cov(T, T_m) \tag{19}$$

where $Cov(T_m, T_{m'}) = 0$ since triangles in group m do not share edges with triangles in group m' ($m \neq m'$), $Var(T)$ and $Var(T_m)$ ($m = 1, ..., k$) are given in equations (12) and (17), respectively, and $Cov(T, T_m) = E(TT_m) - E(T)E(T_m)$ with

$$E(TT_m) = C(n_m)\left[C(n) - 3n + 8\right] p^6 + C(n_m)(3n - 9)p^5 + C(n_m)p^3. \tag{20}$$

Using equation (20) one can easily show

$$Cov(T_m, T_b) = Cov(T, T_m) - Var(T_m) = 3C(n_m)\left[(n - n_m)p^5(1 - p)\right]. \tag{21}$$

Proof. The proof to equation (20) is similar to the proof of $E(T^2)$. The only difference being that the matrix of cross products $\mathbf{R}$ has dimensions $C(n) \times C(n_m)$, where $n_m < n$. Notice that there are $C(n_m)$ possible triangles in group m, and each triangle shares an edge with $3n - 9$ other triangles. As a result, $C(n_m)$ elements of $\mathbf{R}$ satisfy $|(i, j, \ell) \cap (r, s, t)| = 3$, $C(n_m)(3n - 9)$ elements satisfy $|(i, j, \ell) \cap (r, s, t)| = 2$, and the remaining elements, $C(n_m)\left[C(n) - 3n + 8\right]$, satisfy $|(i, j, \ell) \cap (r, s, t)| = 0$ *or* 1. Thus $E(TT_m)$ can be written as given in equation (20). ∎