# Graphlet decomposition of a weighted network

**Hossein Azari**
Harvard University

**Edoardo M. Airoldi**
Harvard University

## Abstract

We introduce the graphlet decomposition of a weighted network, which encodes a notion of social information based on social structure. We develop a scalable algorithm, which combines EM with Bron-Kerbosch in a novel fashion, for estimating the parameters of the model underlying graphlets using one network sample. We explore theoretical properties of graphlets, including computational complexity, redundancy and expected accuracy. We test graphlets on synthetic data, and we analyze messaging on Facebook and crime associations in the 19th century.

## 1 Introduction

In recent years there has been a surge of interest in collaborative projects similar to Wikipedia, social media platforms such as Facebook and Twitter, and services that rely on the social structure underlying these services, including cnn.com "popular on Facebook" and nytimes.com "most emailed". Efforts in the computational social sciences have begun studying patterns of behavior that result in organized social structure and interactions (e.g., see Lazer et al., 2009). Here, we develop a new tool to analyze data about social structure and interactions routinely collected in this context.

There is a rich literature of statistical models to analyze binary interactions or networks (Goldenberg et al., 2010). Arguably, however, only a few of them may be fit to weighted networks. Our approach is to decompose the graphon $\Lambda$, which defines an exchangeable model (Kallenberg, 2005) for integer-valued measurements of pairs of individuals $P(Y|\Lambda)$, in terms of a number of basis matrices that grows with the size of the network, $\Lambda = \sum_i \mu_i P_i$. The factorization of $\Lambda$ is related to models for binary networks based on the singular value decomposition and other factorizations (Hoff, 2009; Kim and Leskovec, 2010). We abandon the popular orthogonality constraint (Jiang et al., 2011) among the basis matrices $P_i$s to attain interpretability in terms of multi-scale social structure, and we chose not model zero edge weights to enable the estimation to scale linearly with the number of positive weights (Leskovec et al., 2010). These choices lead to a non-trivial inferential setting (Airoldi and Haas, 2011). Related methods utilize a network to encode inferred dependence among multivariate data (Coifman and Maggioni, 2006; Lee et al., 2008). We term our method "graphlet decomposition", as it is reminiscent of a wavelet decomposition of a graph. An unrelated literature uses the term graphlets to denote network motifs (Przulj et al., 2006; Kondor et al., 2009).

The key features of the graphlet decomposition we introduce in this paper are: the basis matrices $P_i$s are non-orthogonal, but capture overlapping communities at multiple scales; the information used for inference comes exclusively from positive weights; parameter estimation is linear in the number of positive weights. The basic idea is to posit a Poisson model for the edge weights $P(Y|\Lambda)$ and parametrize the rate matrix $\Lambda$ in terms of a binary factor matrix $B$. The binary factors can be interpreted as latent features that induce social structure through homophily. The factor matrix $B$ allows for an equivalent interpretation as basis matrices, which define overlapping cliques of different sizes. Inference is carried out in two stages: first we identify candidate basis matrices using Bron-Kerbosch, then we estimate coefficients using EM. The computational complexity of the inference is addressed in Section 3.2.

## 2 Graphlet decomposition and model

Consider observing an undirected weighted network, encoded by a symmetric adjacency matrix with integer entries and diagonal elements equal to zero. Intuitively, we want to posit a data generating process that can explain edge weights in terms of social information, quantified by community structure at multiple scales and possibly overlapping. We choose to represent com-

Appearing in Proceedings of the $15^{th}$ International Conference on Artificial Intelligence and Statistics (AISTATS) 2012, La Palma, Canary Islands. Volume XX of JMLR: W&CP XX. 

munities in terms of their constituent maximal cliques.

Below, we avoid notation whose sole purpose is to set to zero diagonal elements of the resulting matrices.

**Definition 1.** *The graphlet decomposition (GD) of a matrix $\Lambda$ with non-negative entries $\lambda_{ij}$ is defined as $\Lambda = B\,W\,B'$, where $B$ is an $N \times K$ binary matrix, $W$ is a $K \times K$ diagonal matrix. Explicitly, we have*

$$\begin{pmatrix} \lambda_{11} & \lambda_{12} & \dots & \lambda_{1N} \\ \lambda_{21} & \lambda_{22} & \dots & \lambda_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_{N1} & \lambda_{N2} & \dots & \lambda_{NN} \end{pmatrix} = B \begin{pmatrix} \mu_1 & 0 & \dots & 0 \\ 0 & \mu_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \mu_K \end{pmatrix} B'$$

*where $\lambda_{ii}$ is set to zero and $\mu_i$ is positive for each $i$.*

The basis matrix $B$ can be interpreted in terms overlapping communities at multiple scales. To see this, denote the $i$-th column of the matrix $B$ as $b_{\cdot i}$. We can re-write $\Lambda = \sum_{i=1}^{K} \mu_i\, P_i$, where each $P_i$ is an $N \times N$ matrix defined as $P_i = b_{\cdot i}\, b'_{\cdot i}$, in which the diagonal is set to zero. We refer to $P_i$ as basis elements. Think of the $i$-th column of $B$ as encoding a community of interest; then the vector $b_{\cdot i}$ specifies which individuals are members of that community, and the basis element $P_i$ is the binary graph that specifies the connections among the member of that community. The same individual may participate to multiple communities, and the communities may span different scales.

The model for an observed network $Y$ is then

$$Y \sim \text{Poisson}_+ \left(\textstyle\sum_{i=1}^{K} \mu_i\, P_i\right). \tag{1}$$

This is essentially a binary factor model with a truncated Poisson link and two important nuances. Standard factor models assume that entire rows of the matrix $Y$ are conditionally independent, even when the matrix is square and row and column $i$ refer to the same unit of analysis. In contrast, our model treats the individual random variables $y_{ij}$ as exchangeable, and imposes the restriction that positives entries in the factors map (in some way) to maximal cliques in the observed network. In addition, the matrix $Y$ is sparse; our model implies that zero edge weights carry no information about maximal cliques, and thus are not relevant for estimation. These nuances—map between cliques and factors, and information from positives entries only—produce a new and interesting inferential setting that scales to large weighted networks.

## 2.1 Inference

Computing the graphlet decomposition of a network $Y$ involves the estimation of a few critical parameters: the number of basis elements $K$, the basis matrix $B$ or equivalently the basis elements $P_{1:K}$, and the coefficients associated with these basis elements $\mu_{1:K}$.

We develop a two-stage estimation algorithm. First, we identify a candidate set of $K^c$ basis elements, using the Bron-Kerbosch algorithm, and obtain a candidate basis matrix $B^c$. Second, we develop an expectation-maximization (EM) algorithm to estimate the corresponding coefficients $\mu_{1:K^c}$; several of these coefficients will vanish, thus selecting a final set of basis elements. We study theoretical properties of this estimation algorithm in Section 3. Figure 1 illustrates the steps of the estimation process on a toy network.

### 2.1.1 Identifying candidate basis elements

The algorithm to identify the candidate basis elements proceeds by thresholding the observed network $Y$ at a number of levels, $t = \max(Y) \dots \min(Y)$, and by identifying all the maximal cliques in the corresponding sequence of binary networks, $Y^{(t)} = \mathbf{1}(Y \geq t)$, using the Bron-Kerbosch algorithm (Bron and Kerbosch, 1973). The resulting cumulative collection of $K^c$ maximal cliques found in the sequence of networks $Y^{(t)}$ is then turned into candidate basis elements, $P^c_i$, for $i = 1 \dots K^c$. Algorithm 1 details this procedure.

$B^c$ = empty basis set
**For** t = max($Y$) to min($Y$)
  $Y^{(t)} = \mathbf{1}(Y \geq t)$
  C = maximal cliques($Y^{(t)}$) using Bron-Kerbosch
  $B^c = B^c \cup C$
**Algorithm 1:** Estimate a basis matrix, $B^c$, encoding candidate elements, $P^c$, from a weighted network, $Y$.

Algorithm 1 allows us to avoid dealing with permutations of the input adjacency matrix $Y$ by inferring the basis elements $P_i$ independently of such permutation.

In Section 3.1 we show that, under certain conditions on the true underlying basis matrix $B$, Algorithm 1 finds a set of candidate basis elements that contains the true set of basis elements. In addition, in Section 3.2 we show that, under the same conditions on $B$ and additional realistic assumptions, we expect to have a number of candidate basis elements of the same order of magnitude of the true number of basis elements, $K^c = O(K)$, as the network size $N$ grows, given that the maximum weight is stationary, $\max(Y) = O(1)$.

### 2.1.2 Sparse Poisson deconvolution

Given the candidate set of basis elements encoded in $B^c$, we develop an EM algorithm to estimate the corresponding coefficients $\mu_{1:K^c}$. Under certain conditions on the true basis matrix $B$, this algorithm consistently estimates the coefficients by zeroing out the coefficients associated with the unnecessary basis elements.

Recall that we have $K^c$ candidate basis elements $P^c_{1:K^c}$. Let's define a set of statistics $T_k = \sum_{ij} P^c_{k,ij}$

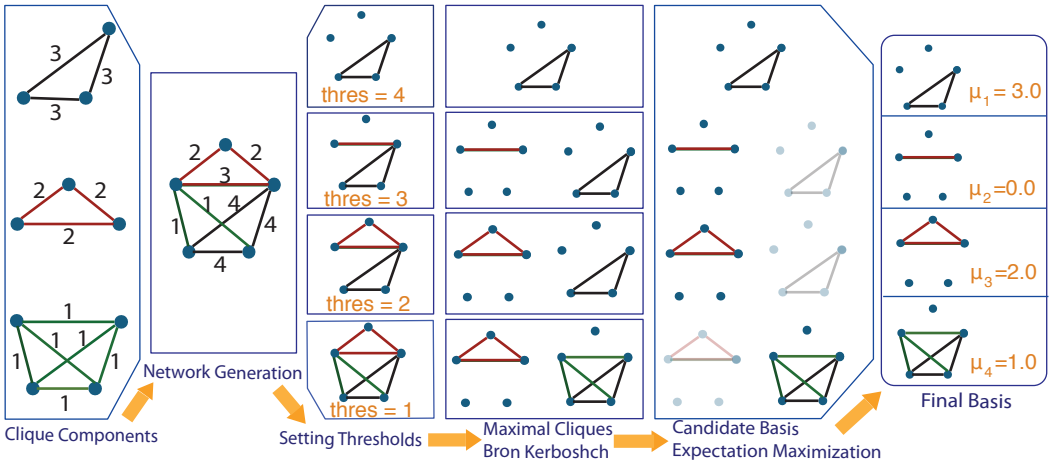


Figure 1: An illustration of the two-stage estimation algorithm for the graphlet decomposition.

for $k = 1 \dots K^c$, and let's introduce a set of latent $N \times N$ matrices $G_k$ with positive entries, subject to the only constrain that $\sum_k G_k = Y$. Algorithm 2 details the EM iterative procedure that will lead to maximum likelihood estimates for the coefficients $\mu_{1:K^c}$.

Initialize $\mu^{(0)}$
**While** $||\mu^{(t+1)} - \mu^{(t)}|| > \epsilon$
  **For** k=1 to K
    **E-step:** $G^{(t+1)}_{k,ij} = \mu^{(t)}_k \frac{Y_{ij} P^c_{k,ij}}{T_k \sum_m \mu^{(t)}_m P^c_{m,ij}}$
    **M-step:** $\mu^{(t+1)}_k = \sum_{i,j} G^{(t+1)}_{k,ij}$;
  t=t+1;

**Algorithm 2:** Estimate non-negative coefficients, $\mu$, for all candidate elements, $P^c$, in the basis matrix $B^c$.

This is the Richardson-Lucy algorithm for Poisson deconvolution (Richardson, 1972). The truncated Poisson likelihood can also be maximized directly, using a KL divergence argument (Csiszar and Shields, 2004) involving the discrete probability distribution obtained by normalizing the data matrix, $\bar{Y}$, and a linear combination of discrete probability distributions obtained by normalizing the candidate basis matrices, $\sum_k \omega_k \bar{P}_k$.

## 3 Theory

Here we develop some theory for graphlets when the graph $Y$ is generated by a non-expandable collection of basis matrices. The extent to which these results extend to the general case is unclear, as of this writing, however, the results help form some intuition about how graphlets encode information. See the Appendix for details of the proofs. We begin by establishing some notation.

**Definition 2.** *Given two adjacency matrices $P_1$ and $P_2$ both binary define $P_1 \subseteq P_2$ to mean $P_2(j,k) = 1$ whenever $P_1(j,k) = 1$; i.e. the induced graph by $P_1$ is a subgraph of $P_2$.*

**Definition 3.** *Define $P = \bigvee_{i \in I} P_i$ to mean $P(j,k) = 1$ whenever any $P_i(j,k) = 1$ and zero otherwise.*

**Definition 4.** *Given $\{P_k\}$, we say that $P'$ is an expansion of $P_k$ if both 2 and 3 are satisfied.*

**Definition 5** (non expandable basis)**.** *A collection of $\{P_k\}$ is non-expandable if for any expansion $P'$ of $P_k$ we can find $j$ such that $P' \subset P_j$.*

### 3.1 Identifiability

The first result asserts that the collection of Poisson rates $\Lambda$ can be uniquely decomposed into a collection of basis matrices $P_{1:K}$, here encoded by the equivalent binary matrix $B$, with weights $W = \text{diag}(\mu_{1:K})$.

**Theorem 1.** *Let $\Lambda$ be a symmetric $N \times N$ nonnegative matrix which is generated by a non-expandable basis. Then there exists a unique $N \times K$ matrix $B$ and $K \times K$ matrix $W$ such that $\Lambda = BWB'$.*

This implies that, in the absence of noise, the set of parameters $B, W, K$ are estimable from the data $Y$. The proof of Theorem 1 is constructive by means of Algorithm 3, which is fast but sensitive to noise. In practice, we propose a more robust algorithm that uses non-expandability to identify a collection of candidate basis elements $B^c$, which provably includes the unique non expandable basis $B$ that generated the graph.

**Theorem 2.** *Let $\Lambda = BWB'$, where $W$ is a diagonal matrix with nonnegative entries and $B$ encodes a non-expandable basis, both unknown. If $B^c$ is the output of Algorithm 1 with $\Lambda$ as input, then $B \subseteq B^c$.*

These two theorems, with $B_{ij}$'s IID Bernoulli random variables, may be used to generate random cliques.

### 3.2 Redundancy and complexity

Here, we derive an upper bound for the maximum number of basis elements encoded in $B^c$ in a network

with at most $K = c \log_2 N$ cliques. Networks with this many cliques include networks generated with most exchangeable graph models (Goldenberg et al., 2010), and are the largest networks with an implicit representation (Kannan et al., 1988).

**Theorem 3.** *Let the elements $B_{ik}$ of the basis matrix for a network $Y$ be IID Bernoulli random variables with parameter $p_N$. Then an asymptotic upper bound for the number of candidate basis elements, denoted $C_{N,p_N}$, identified by Algorithm 1 is*

$$\begin{aligned} C_{N,p_N} &\leq Q(2^{KH(p_N)} + K) \\ &= Q(N^{c_1 H(p_N)} + c \log_2 N), \end{aligned} \tag{2}$$

*where $Q$ is the number of thresholds in Algorithm 1.*

While Theorem 3 applies in general, a notion of redundancy of the candidate basis set is well defined only for networks generated by a non expandable basis, in which $B^c \subseteq B$. In this case, we define redundancy as the ratio $R_N \equiv C_{N,p_N}/K$. Theorem 2 states that number of candidate basis elements is never smaller than the true number of basis $K$. Thus Theorem 3 leads to the following upper bound on redundancy

$$R_{N,p_N} \leq Q\left(1 + \frac{N^{c_1 H(p_N)}}{c \log_2 N}\right). \tag{3}$$

In a realistic scenario $p_N = O(1/\log_2 N)$, the complexity of the candidate basis set is $O(Q \log_2 N)$, and the implied redundancy is at most $O(Q)$. Alternatively, if $p_N = O(1/K)$, the implied redundancy is at most $O(QK)$. These are regimes we often encounter in practice. These limiting behaviors of $p_N$ arise whenever the nodes in a network can be assumed to have a capacity that is essentially independent of the size of the network, e.g., individuals participate in the activities of a constant number of groups over time.

These calculations are only suggestive for networks generated by an expandable basis. However, non expandability tends to be a reasonable approximation in many situations, including whenever exchangeable graph models are good fit for the network. In the analysis of messaging patterns on Facebook in Section 4.3, for instance, we empirically observe $C_{N,p} \approx 2\hat{K}$.

### 3.3 Accuracy with less than $K$ basis elements

The main estimation Algorithm 2 recovers the correct number of basis elements $K$ and the corresponding coefficients $\mu$ and basis matrix $B$, whenever the true weighted network $Y$ is generated from a non expandable basis matrix. Here we quantify the expected loss in reconstruction accuracy if we were to use $\tilde{K} < K$ basis elements to reconstruct $Y$. To this end we introduce a norm for a network $Y$, and related metrics.

**Definition 6** ($\tau$-norm)**.** *Let $Y \sim Poisson(BWB')$, where $W = diag(\mu_1 \dots \mu_K)$, define the statistics $a_k \equiv \sum_{i=1}^{N} B_{ik}$ for $k = 1 \dots K$. The $\tau$-norm of $Y$ is defined as $\tau(Y) \equiv |\sum_{k=1}^{K} \mu_k a_k|$.*

Consider an approximation $\tilde{Y} = B\tilde{W}B'$ characterized by an index set $E \subset \{1 \dots K\}$, which specifies the basis elements to be excluded by setting the corresponding set of coefficients $\mu_E$ to zero. Its reconstruction error is $\tau(Y - \tilde{Y}) = |\sum_{k \notin E} \mu_k a_k|$, and its reconstruction accuracy is $\tau(\tilde{Y})/\tau(Y)$. Thus, given a network matrix $Y$ representable exactly with $K$ basis elements, the best approximation with $\tilde{K}$ basis elements is obtained by zeroing out the lowest $K - \tilde{K}$ coefficients $\mu$.

We posit the following theoretical model,

$$\mu_k \cdot a_k/N \quad \sim \quad Gamma(\alpha, 1), \tag{4}$$

for $k = 1 \dots \tilde{K}$. This model may be used to compute the expected accuracy of an approximate reconstruction based on $\tilde{K} < K$ basis elements, since the magnitude of the ordered $\mu_k$ coefficients that are zeroed out are order statistics of a sample of Gamma variates (Shawky and Bakoban, 2009). Given $K$ and $\alpha$, we can compute the expected coefficient magnitudes and the overall expected accuracy $\tau_0$.

**Theorem 4.** *The theoretical accuracy of the best approximate Graphlet decomposition with $\tilde{K}$ out of $K$ basis elements is: $\tau_0(\tilde{K}, K, \alpha) = \sum_{j=1}^{\tilde{K}} \frac{f(j,K,\alpha)}{\alpha K}$; where $f(j, K, \alpha) = \binom{K}{j} \sum_{q=0}^{j-1} (-1)^q \binom{j-1}{q} \frac{f(1,K-j+q+1,\alpha)}{K-j+q+1}$, and $f(1, K, \alpha) = \frac{K}{\Gamma(\alpha)} \sum_{m=0}^{(\alpha-1)(K-1)} c_m(\alpha, K-1) \frac{\Gamma(\alpha+m)}{K^{\alpha+m}}$, in which the coefficients $c_m$ are defined by the recursion $c_m(\alpha, q) = \sum_{i=0}^{i=\alpha-1} \frac{1}{i!} c_{m-i}(\alpha, q-1)$ with boundary conditions $c_m(\alpha, 1) = \frac{1}{i!}$ for $i = 1 \dots \alpha$.*

Figure 2 illustrates this result on simulated networks. The solid (red) line is the theoretical accuracy computed for $K = 30$ and $\alpha = 1$, the relevant parameters used to to simulate the sample of 100 weighted networks. The (blue) boxplots summarize the empirical accuracy at a number of distinct values of $\tilde{K}/K$.

## 4 Results

We evaluate graphlets on real and simulated data. Simulation results on weighted networks in Section 4.1 show that the estimation problem is well-posed and both the binary matrix $B$ and the coefficients $\mu$ are estimable without bias, while results on binary networks in Section 4.2 enable the analysis of the comparative performance with existing methods for binary networks. We also develop an analysis of messaging patterns on Facebook for a number of US colleges and an analysis of historical crime data in Sections 4.3–4.4.

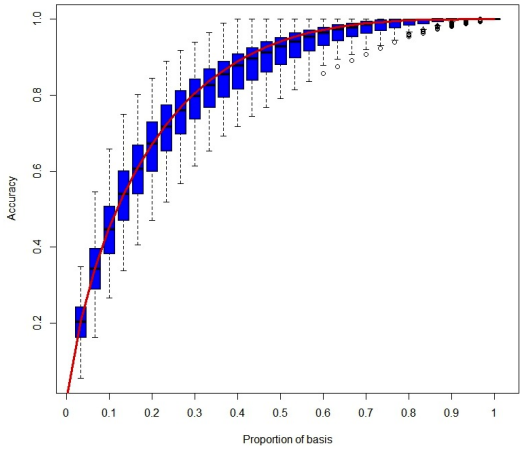


Figure 2: Theoretical and empirical accuracy for different fractions of basis elements $\tilde{K}/K$ with $\alpha = 0.1$. The ratio $\tilde{K}/K$ also provides a measure of sparsity.

Overall, these results suggest that graphlets encode a new quantification of social information; we provide a concrete illustration of this idea in Section 5.1.

### 4.1 Parameter estimation

We set out to evaluate whether the parameters underlying graphlets—the entries of binary matrix $B$, its size $K$, and the non-negative coefficients $\mu$—are estimable from one weighted network sample $Y$.

We simulated $M = 100$ networks from the model, each with 50 nodes, using random values for the parameters $\mu, B$ and $K$. For the $i$-th network, the number of basis elements $K_i$ was sampled from a Poisson distribution with rate $\lambda = 30$. The coefficients $\mu_{ji}$ were sampled from a Gamma distribution with parameters $\alpha = 1$ and $\beta = 10$. The entries of the basis matrix $B$ were sampled from a Bernoulli distribution with probability of success $p = 0.04$. The index $i = 1 \dots M$ runs over the networks and the index $j = 1 \dots K_i$ runs over the basis elements for the $i$-th network.

Algorithms 1–2 were used to estimate the parameters $B, \mu$ and $K$, for each synthetic network. The estimation error of $\hat{K}$ is $|\hat{K} - K|$. Since the estimated matrix $\hat{B}$ is unique only up to a permutation of its columns, the Procrustes transform (Kendall, 1989) was used to identify the optimal rotation of the matrix $\hat{B}$ to be compared to the true matrix $B$. The estimation error of $\hat{B}$ is computed as the normalized $L_2$ distance between $B$ and $\hat{B}$ after Procrustes. The estimation error of $\hat{\mu}$ is computed as the normalized $L_2$ distance between $\mu$ and $\hat{\mu}$. The average error in reconstructing the simulated networks is computed as the $L_1$ distance between $Y$ and $\hat{Y}$ normalized to the total number of counts in the network, averaged over 100 simulations, denoted $|\tilde{Y}|_1$. This metric gives more weight to larger cliques, since the clique size enters as a quadratic term. We also provide the average error in reconstructing the simulated networks based on the $\tau$-norm, which is less sensitive to incorrectly estimating larger cliques. In addition, we compute the average error in reconstructing whether or not edges have positive weights, rather than their magnitude, denoted error in $\mathbf{1}(\tilde{Y} > 0)$.

Table 1 summarizes the simulation results for weighted network reconstructions obtained by setting a target accuracy $\tau_0$, ranging from 0.20 to 1.00—no error. The last row of Table 1 supports the claim that the estimation problem is well-posed; the estimation algorithm stably recovers the true parameter values. The only sources of non-zero error are the estimation of the number of basis elements and the estimation of the basis elements themselves, as encoded by the matrix $B$. However, these errors are negligible and do not seem to impact the estimation of the coefficients $\mu$, nor to propagate to the network reconstruction independently of the metric we use to quantify it.

The average bias in estimating the optimal number of basis elements $K$ is 0.07, with a standard deviation of 0.256. Column $\tilde{K}/\hat{K}$ provides the fraction of the estimated optimal number of basis elements $\hat{K}$ that is necessary to achieve the desired target accuracy, in terms of the $\tau$-norm, using Theorem 4. Column two provides the empirical $\tau(\tilde{Y})$ error, defined as one minus the accuracy, corresponding to the reported fraction of basis elements. The empirical accuracy matches the expected accuracy given by Theorem 4 well. For instance, the Theorem 4 implies that we need 45% of the basis elements in order to achieve an accuracy of 0.85, and the empirical accuracy is slightly in excess of 0.85, on average. Figure 2 shows similar results, for a larger set of 29 values for the target accuracy $\tau_0$.

The first two columns of Table 1 report the average $L_1$ error and the average $\tau$-based error, properly normalized to have range in zero to one. Cliques in the simulated networks range between 2 and 5 nodes, for most networks. The differences in the reconstruction error are due to the weight of these larger cliques. For instance, the empirical $L_1$ accuracy obtained by using 45% of the basis elements is around 0.65, as opposed to a $\tau$ accuracy of around 0.85, on average.

The accuracy in reconstructing the presence or absence of edges in the network, rather than the magnitude of the corresponding weights, is in between the $L_1$ accuracy and the $\tau$ accuracy, for any fraction $\tilde{K}/\hat{K}$.

### 4.2 Link reconstruction

Graphlet is a method for decomposing a weighted network. However, here we evaluate how graphlet fares

Table 1: Performance of the estimation algorithm for a target accuracy $\tau_0$, on synthetic data.

| $\lvert\tilde{Y}\rvert_1$ error | $\tau(\tilde{Y})$ error | error in $\mathbf{1}(\tilde{Y} > 0)$ | error in $B$ | error in $\mu$ | $\tilde{K}/\hat{K}$ | $\tau_0$ |
|---|---|---|---|---|---|---|
| $0.97 \pm 0.012$ | $0.65 \pm 0.074$ | $0.76 \pm 0.062$ | $50.75 \pm 10.893$ | $0.55 \pm 0.082$ | $0.08 \pm 0.035$ | 0.20 |
| $0.92 \pm 0.016$ | $0.46 \pm 0.026$ | $0.61 \pm 0.053$ | $42.40 \pm 8.134$ | $0.32 \pm 0.041$ | $0.16 \pm 0.039$ | 0.50 |
| $0.79 \pm 0.036$ | $0.23 \pm 0.012$ | $0.40 \pm 0.057$ | $29.14 \pm 5.536$ | $0.10 \pm 0.015$ | $0.33 \pm 0.050$ | 0.75 |
| $0.56 \pm 0.063$ | $0.09 \pm 0.006$ | $0.24 \pm 0.048$ | $17.04 \pm 3.757$ | $0.02 \pm 0.004$ | $0.54 \pm 0.062$ | 0.90 |
| $0.00 \pm 0.000$ | $0.00 \pm 0.000$ | $0.00 \pm 0.000$ | $0.67 \pm 0.618$ | $0.00 \pm 0.000$ | $1.01 \pm 0.033$ | 1.00 |

in terms of the computational complexity versus accuracy trade-off when compared to the performance current methods for binary networks can achieve.

For this purpose, we generated 100 synthetic networks from the model, each with 100 nodes, choosing random values for the parameters as we did in Section 4.1. The number of basis was sampled from a Poisson with rate $\lambda = 12$. The coefficients were sampled from a Gamma with parameters $\alpha = 2$ and $\beta = 10$. The entries of the basis matrix were sampled from a Bernoulli with probability of success $p = 0.1$. We then obtained the corresponding binary networks by resetting positive edge weights to one, $\mathbf{1}(Y > 0)$.

We compute comparative performance for a few popular models of networks that can be used for link prediction. These include the Erdös-Rényi-Gilbert random graph model (Erdös and Rényi, 1959; Gilbert, 1959), denoted ERG, a variant of the degree distribution model fitted with importance sampling (Blitzstein and Diaconis, 2010), denoted DDS, the latent space cluster model (Handcock et al., 2007), denoted LSCM, and the stochastic blockmodel with mixed membership (Airoldi et al., 2008), denoted MMSB. A comparison with singular value decomposition in terms binary link prediction is problematic, since using a few singular vectors would not lead to binary edge weights.

Table 2 summarizes the comparative performance results. The accuracy of graphlet is reported for different fractions of the estimated optimal number of basis elements $\hat{K}$, ranging from 25% to 100%—no error. The accuracy is consistently high, while the running time is a only a fraction of a second, independent of the desired reconstruction accuracy. This is because the complexity of computing the *entire* graphlet decomposition is linear in the number of edges present in the network. Thus specifying a smaller fraction of the number of basis elements is a choice based on storage considerations, rather than on runtime. The comparative performance results suggest that graphlet decomposition is very competitive in predicting links.

Table 2: Link reconstruction accuracy and runtime for a number of competing methods, on simulated data. The graphlet decomposition is denoted Glet, with the fraction of basis elements used quoted in brackets.

| Method | Runtime (sec) | Accuracy |
|---|---|---|
| Glet (25%) | 0.0636 | $92.7 \pm 0.30$ |
| Glet (50%) | 0.0636 | $94.7 \pm 0.15$ |
| Glet (75%) | 0.0636 | $97.0 \pm 0.08$ |
| Glet (90%) | 0.0636 | $98.9 \pm 0.05$ |
| Glet (100%) | 0.0636 | $100.0 \pm 0.00$ |
| ERG | 0.0127 | $86.0 \pm 1.20$ |
| DDS | 11.1156 | $89.0 \pm 0.85$ |
| LSCM | 6.3491 | $90.0 \pm 0.70$ |
| MMSB | 2.8555 | $93.0 \pm 0.40$ |

This binary link prediction simulation study is one way to compare graphlet to interesting existing methods, which have been developed for binary graphs.

### 4.3 Analysis of messaging on Facebook

Here we illustrate graphlet with an application to messaging patterns on Facebook. We analyzed the number of public wall-post on Facebook, over a three month period, among students of a number of US colleges. While our data is new, the US colleges we selected have been previously analyzed (Traud et al., 2011).

Table 3 provides a summary of the weighted network data and of the results of the graphlet decomposition. Salient statistics for each college include the number of nodes and edges. The table reports the number of estimated basis elements $\hat{K}$ for each network and the runtime, in seconds. The $\tau(\tilde{Y})$ error incurred by using a graphlet decomposition is reported for different fractions of the estimated optimal number of basis elements $\hat{K}$, ranging from 10% to 100%—no error.

Overall, these results suggest that the compression of wall-posts achievable on collegiate network is substantial. A graphlet decomposition with about 10% of the optimal number of basis elements already leads to a reconstruction error of 10% or less, with a few exceptions. Using 25% of the basis elements further reduces the reconstruction error to below 5%.

### 4.4 Analysis of historical crime data

More recently, we have used the graphlet decomposition to analyze crime associations records from the 19th century. These records are available at the

Table 3: $\tau(\tilde{Y})$ error for different fractions of the estimated optimal number of basis elements $\hat{K}$.

| college | nodes | edges | $\hat{K}$ | sec | 10% | 25% | 50% | 75% | 90% | 100% |
|---|---|---|---|---|---|---|---|---|---|---|
| Amherst | 2235 | 181907 | 10151 | 124 | 6.5 | 2.3 | .84 | .33 | .14 | 0 |
| Caltech | 769 | 33311 | 3735 | 51 | 5.7 | 1.8 | .65 | .27 | .11 | 0 |
| CMU | 6637 | 499933 | 11828 | 169 | 14.9 | 5.2 | 1.89 | .73 | .34 | 0 |
| MIT | 6440 | 502503 | 13145 | 191 | 11.5 | 4.7 | 1.68 | .65 | .30 | 0 |
| Rice | 4087 | 369655 | 12848 | 155 | 8.7 | 3.2 | 1.25 | .51 | .22 | 0 |
| Swarthmore | 1659 | 122099 | 7856 | 96 | 6.3 | 2.6 | 1.06 | .44 | .20 | 0 |

South Carolina Department of Archives and History (Columbia, SC), Record Group 44, Series L 44158, South Carolina, Court of General Sessions (Union County) Indictments 1800-1913. Nodes in the network are 6221 residents of Union County, during the years 1850 to 1880. Edge weights indicate the number of crimes involving pairs of residents. Details on the structure of this data set and its historical relevance may be found in Frantz-Parsons (2005, 2011).

A first step in the analysis was to explore the degree to which there is an optimal level of granularity, for the observed associations, at which a non-trivial criminal structure emerges. We pursued this analysis by fitting graphlets to the raw criminal association network raised to different powers, $Y^k$ for $k = 1 \dots 5$. We found that $Y^2$ contained a substantial degree of criminal social structure. At this resolution, $k = 2$, two residents are associated with crimes in which they are either directly involved, or in which they are involved indirectly through their direct criminal associates. The results suggest that there are several tight pockets of criminality scattered throughout Union County.

We were interested in developing an interpretation for these pockets of criminality. One compelling hypothesis is that each pocket captures a gang or a clan. If this holds, we might expect the variation in the gangs' relations to be explained by the spatial organization of the townships they operate in, to a large degree.

To test this hypothesis, we built a meta network of criminal associations, in which each gang identified in the first step of the analysis is represented by a meta node. Two or more gangs were connected in the meta network whenever a criminal record involved at least one of their members. A graphlet decomposition of the meta network identified pockets criminal activity at the gang level. Plotting these gang level criminal associations reveals the townships' spatial organization, to a large degree. Figure 3 shows the pockets of criminality identified by the graphlet analysis of the original criminal associations, $Y^2$, as green nodes associated with numbers, as well as the pockets of criminality identified by the graphlet analysis of the meta network, as purple nodes associated with letters.

In summary, we were abel to successfully use graphlets as an exploratory tool to capture and distill an interesting, multi-scale organization among criminals in 19th century Union County, from historical records.

## 5 Discussion

Taken together, our results suggest that the graphlet decomposition implies a new notion of social information, quantified in terms of multi-scale social structure. We explored this idea with a simulation study.

### 5.1 A new notion of "social information"

Graphlet quantifies social information in terms of community structure (i.e., maximal cliques) at multiple scales and possibly overlapping. To illustrate this notion of social information, we simulated 200 networks: 100 Erdös-Rényi-Gilbert random graphs with Poisson

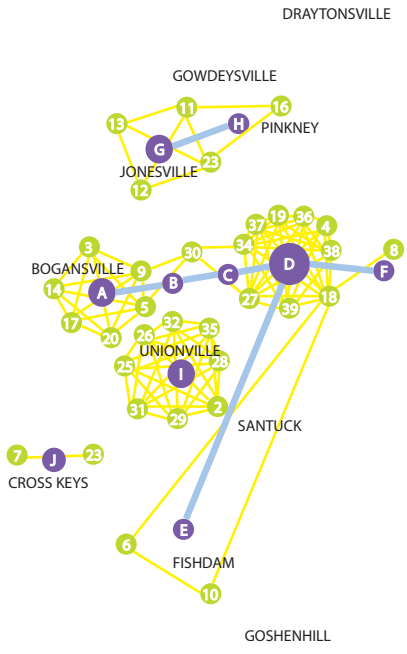


Figure 3: A two-step analysis of the Union County criminal association network using graphlets reveals tight pockets of criminality (the numbered nodes) and local criminal communities (the lettered nodes).

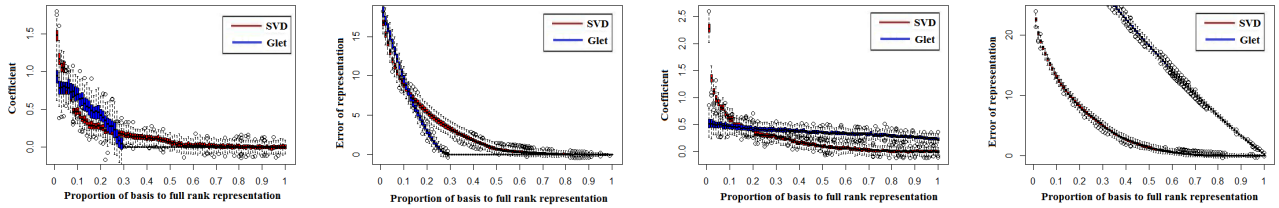


Figure 4: Comparison between graphlet and SVD decompositions. Panels 1–2 report results for on the networks simulated from the model underlying graphlets. Panels 3–4 report results on the Poisson random graphs.

weights and 100 networks from the model described in Section 2. While, arguably, the Poisson random graphs do not contain any such information, the data generating process underlying graphlets was devised to translate such information into edge weights.

As a baseline for comparison, we consider the singular value decomposition (Searle, 2006), applied to the symmetric adjacency matrices with integer entries encoding the weighted networks. The SVD summarizes edge weights using orthogonal eigenvectors $v_i$ as basis elements and the associated eigenvalues $\lambda_i$ as weights, $Y_{N\times N} = \sum_{i=1}^{N} \lambda_i \, v_i \, v_i'$. However, SVD basis elements are not interpretable in terms of community structure, thus SVD should not be able to capture the notion of social information we are interested in quantifying.

We applied graphlets and SVD to the two sets of networks we simulated. Figure 4 provides an overview of the results. Panels 1–2 report results for on the networks simulated from the model underlying graphlets. Panels 3–4 report results on the Poisson random graphs. Panels 1 and 3 show the box plots of the coefficients associated with each of the basis elements, for graphlets in blue and for SVD in red. Panels 2 and 4 show the box plots of the cumulative representation error as a function of the number of basis elements utilized. Graphlets coefficients decay more slowly than SVD coefficients on Poisson graphs (panel 3). Because of this, the error in reconstructing Poisson graphs achievable with graphlets is consistently worse then the error achievable with SVD (panel 4). In contrast, graphlets coefficients decay rapidly to zero on networks with social structure, much sharply then the SVD coefficients (panel 1). Thus, the reconstruction error achievable with graphlets is consistently better then the error achievable with SVD (panel 2).

These results support our claim that graphlets is able to distill and quantify a notion of social information in terms of social structure.

### 5.2 Concluding remarks

The graphlet decomposition of a weighted network is an exploratory tool, which is based on a simple but compelling statistical model, it is amenable to theoretical analysis, and it scales to large weighted networks.

SVD is the most parsimonious orthogonal decomposition of a data matrix $Y$ (Hastie et al., 2001). Graphlets abandon the orthogonality constraint to attain interpretability of the basis matrix in terms of social information. In the presence of social structure, graphlet is a better summary of the variability in the observed connectivity. SVD always achieves a lower reconstruction error, however, when a few basis elements are considered (see Figure 4) suggesting that orthogonality induces a more useful set of constraints whenever a very low dimensional representation is of interest.

Graphlets provide a parsimonious sumary of complex social structure, in practice. The simulation studies in Sections 4.1 and 5.1 suggest that graphlets are leveraging the social structure underlying messaging patterns on Facebook to deliver the high compression/high accuracy ratios we empirically observe in Table 3.

The information graphlets utilize for inference comes exclusively from positive weights in the network. This feature is the source of desirable properties, however, because of this, graphlets are also sensitive to missing data; zero edge weights cannot be imputed. More research is needed to address this issue properly, while maintaining the properties outlined in Section 3. Empirical solutions include considering powers of the original network, as in Section 4.4, or cumulating weights over a long observation window, as in Section 4.3. Intuitively, cumulating messages over long time windows will reveal a large set of basis elements. Algorithm 2 can then be used to project message counts over a short observation window onto the larger set of basis elements to estimate which elements were active in the short window and how much. Recent empirical studies suggest three months as a useful cumulation period (Cortes et al., 2001; Kossinets and Watts, 2006).

### Acknowledgments

This work was partially supported by National Science Foundation grants DMS-1106980, IIS-1017967, CAREER award IIS-1149662, and by Army Research

Office grant 58153-MA-MUR, all to Harvard University. EMA is an Alfred P. Sloan Research Fellow.

## References


E. M. Airoldi and B. Haas. Polytope samplers for inference in ill-posed inverse problems. In *Journal of Machine Learning Research*, volume 15 of *W&CS (AiStat)*, 2011.

E. M. Airoldi, D. M. Blei, S. E. Fienberg, and E. P. Xing. Mixed membership stochastic blockmodels. *Journal of Machine Learning Research*, 9:1981–2014, 2008.

J. K. Blitzstein and P. Diaconis. A sequential importance sampling algorithm for generating random graphs with prescribed degrees. *Internet Mathematics*, 6(4):489–522, 2010.

C. Bron and J. Kerbosch. Algorithm 457: Finding all cliques of an undirected graph. *Communications of the ACM*, pages 575–577, 1973.

R. R. Coifman and M. Maggioni. Diffusion wavelets. *Applied and Computational Harmonic Analysis*, 21 (1):53–94, 2006.

Corinna Cortes, Daryl Pregibon, and Chris Volinsky. Communities of interest. *Proceedings of Intelligent Data Analysis*, pages 105–114, 2001.

Imre Csiszar and Paul C. Shields. *Information Theory and Statistics: A Tutorial*. Foundations and Trendsô in Communications and Information Theory, 2004.

P. Erdös and A. Rényi. On random graphs. *Publicationes Mathematicae Debrecen*, 5:290–297, 1959.

E. Frantz-Parsons. Midnight rangers: Costume and performance in the reconstruction-era ku klux klan. *Journal of American History*, 92:3:811–836, 2005.

E. Frantz-Parsons. Klan skepticism and denial in reconstruction-era public discourse. *Journal of Southern History*, 67:1:53–90, 2011.

E. N. Gilbert. Random graphs. *Annals of Mathematical Statistics*, 30:1141–1144, 1959.

A. Goldenberg, A. X. Zheng, S. E. Fienberg, and E. M. Airoldi. Statistical models of networks. *Foundations and Trends in Machine Learning*, 2(2):1–117, 2010.

M. Handcock, A. E. Raftery, and J. Tantrum. Model-based clustering for social networks (with discussion). *Journal of the Royal Statistical Society, Series A*, 170:301–354, 2007.

T. Hastie, R. Tibshirani, and J. H. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer-Verlag, 2001.

P. D. Hoff. Multiplicative latent factor models for description and prediction of social networks. *Computational & Mathematical Organization Theory*, 15 (4):261–272, 2009.

X. Jiang, Y. Yao, H. Liu, and L. Guibas. Compressive network analysis. Manuscript, April 2011. URL http://arxiv.org/abs/1104.4605.

O. Kallenberg. Probabilistic symmetries and invariance principles. In *Probability and its Applications*. Springer, New York, 2005.

S. Kannan, M. Naor, and S. Rudich. Implicit representation of graphs. *Proceedings of 20th Annual ACM Symposium on Theory of Computing*, 1988.

D. G. Kendall. A survey of the statistical theory of shape. *Statistical Science.*, 4:87–99, 1989.

M. Kim and J. Leskovec. Multiplicative attribute graph model of real-world networks. In *7th Workshop on Algorithms and Models for the Web Graph (WAW)*, 2010.

R. Kondor, N. Shervashidze, and K. M. Borgwardt. The graphlet spectrum. In *International Conference on Machine Learning*, volume 26, 2009.

G. Kossinets and D. J. Watts. Empirical analysis of an evolving social network. *Science*, 311:88–90, 2006.

D. Lazer, et al. Computational social science. *Science*, 323(5915):721–723, 2009.

A. B. Lee, B. Nadler, and L. Wasserman. Treelets: An adaptive multiscale basis for sparse unordered data. *Annals of Applied Statistics*, 2(2):435–471, 2008.

J. Leskovec, D. Chakrabarti, J. Kleinberg, C. Faloutsos, and Z. Ghahramani. Kronecker graphs: An approach to modeling networks. *Journal of Machine Learning Research*, 11:985–1042, 2010.

N. Przulj, D. G. Corneil, and I. Jurisica. Efficient estimation of graphlet frequency distributions in protein–protein interaction networks. *Bioinformatics*, 22(8):974–980, 2006.

W. H. Richardson. Bayesian-based iterative method of image restoration. *Journal of the Optical Society of America*, 62(1):55–59, 1972.

S. R. Searle. *Matrix Algebra Useful for Statistics*. Wiley-Interscience, 2006.

A. I. Shawky and R. A. Bakoban. Order statistics from exponentiated gamma distribution and associated inference. *Int. J. Contemp. Math. Sciences,*, 4 (2):71–91, 2009.

A. L. Traud, E. D. Kelsic, P. J. Mucha, and M. A. Porter. Community structure in online collegiate social networks. *SIAM Review*, 53(3):526–543, 2011.

# A Proofs of theorems and lemmas

In the following, $\Lambda \circ C$ denotes the element-wise (Hadamard) product of two matrices $\Lambda$ and $C$.

## A.1 Proof of Theorem 1

*Proof.* We will show $B$ can be recovered from $\Lambda$ using the deterministic algorithm 3.

Set $\Lambda' = \Lambda$
**While**(There is at least one non-zero element in $\Lambda'$)
$\mu_c = \min_{i,j}(\Lambda \circ C)(i,j)$
$ij_c$=$argmin_{ij}(\Lambda \circ C)(i,j)$
C=largest clique in $\Lambda'$ that includes the edge $ij_c$
$\Lambda'$=$\Lambda' - \mu_c C$
Add $(C, \mu_c)$ to $(B, W)$ respectively.

**Algorithm 3:** Algorithm in theorem 1 for uniqueness of non expandable basis

In each step, due to non-expandability, the maximal clique of $\Lambda'$, $C$, will not be a subset of any other clique and will contain an edge which only belongs to $C$, using Lemma 1 below. The weight of this edge in $\Lambda'$ will be $\mu_c$ the coefficient of clique $C$ in the composition. Also, edges of clique $C$ in the network will not have weights less than $\mu_c$ because all the coefficients in composition($\mu$s) are positive. So, in each step, the weight corresponding to the edge(s) unique to $C$ will be correctly detected and subtracted from the network.

Hence, using algorithm 3 we can find basis $B$ and coefficients $W$ of matrix in $\Lambda' = BWB^T$. Since the algorithm is deterministic given $\Lambda$, it will provide us only one set of unique NEB. □

**Lemma 1.** *If $B$ generates a nonexpandable basis set $S$, any clique $C \in S$ will contain an edge that does not belong to any other clique in $S$.*

*Proof.* Proof by contradiction: If every edge in $C$ overlaps with another clique in $S$, then $C$ can be decomposed to other cliques, which contradicts non-expandability of $S$. □

## A.2 Proof of Theorem 2

*Proof.* Recall that $P_k = b_{.k}b'_{.k}$ and $\Lambda = \sum_l \mu_l P_l$. For any $k \in \{1, 2, .., K\}$, setting a threshold on the edge weights of the network $\Lambda$, such that edges with weights above the threshold are selected, obtains a binary network. Define $G_k$ to be those indices $\ell$ for which $B_\ell$ covers $B_k$. If the above mentioned threshold is set to $th_k = \sum_{l \in G_k} \mu_l$, then a binary network $\Lambda_{th=th_k}$ will be revealed. Below we will prove that $P_k$ is a maximal clique for $\Lambda_{th=th_k}$ using Lemma 2.

To show that $P_k$ is a maximal cliques of $\Lambda_{th=th_k}$ we need to prove the following two statements:

(1) $P_k \subseteq \Lambda_{th=th_k}$ which means clique $P_k$ should be included in $\Lambda_{th=th_k}$. This statement is true because the weights of the edges in $\Lambda$ that correspond to $P_k$, are greater than $th_k$ since $\mu$s are positive.

(2) $P_k$ can not be part of a larger clique in $\Lambda_{th=th_k}$. By contradiction, assume such a larger clique exists and name it $P'$ then $P_k \subset P' \subseteq \Lambda_{th=th_k}$. Considering that the basis is non-expandable, given lemma (2) and $P' \subseteq \Lambda_{th=th_k} \subseteq \bigvee_{i \in T \backslash G} P_{l_i}$ then we can find $j \in T \backslash G$ that $P' \subseteq P_j$. Hence, we have $P_k \subset P' \subseteq P_j$ which leads to $P_k \subset P_j$ for a $j \in T \backslash G$. This is clearly a contradiction since such $j$'s are excluded in $G$.

With $P_k$'s being one of the maximal cliques of thresholded binary networks at $th_k$ and considering the fact that the threshold level will be equal to $\sum_{i \in I} \mu_i$ for any $I \in \{G_1, G_2, ..., G_K\}$ (because $\min_{ij} \Lambda_{ij} \leq \sum_{l \in I} \mu_l \leq \max_{ij} \Lambda_{ij}$ ), it is clear that $B \subseteq B_c$. This ends the proof of theorem (2). □

**Lemma 2.** $\Lambda_{th=th_k} \subseteq \bigvee_{i \in T \backslash G_k} P_i$, *for* $T = 1 \dots K$.

*Proof.* We will prove this by contradiction. Assume the above claim is not true. In that case, an edge in $\Lambda_{th=th_k}$ can be found which belongs to complementary set: $\bigvee_{i \in G_k} P_i \backslash \bigvee_{i \in T \backslash G_k} P_i$. However, the weight for this edge is more than $th_k = \sum_{l \in G_k} \mu_l$ which is not possible for edges included in $\bigvee_{i \in G_k} P_i \backslash \bigvee_{i \in T \backslash G_k} P_i$ because the maximum weight for edges obtained from this group of cliques is at most $th_k = \sum_{l \in G_k} \mu_l$. End of proof of the lemma (2). □

## A.3 Proof of Theorem 3

*Proof.* Using the definition of Graphlet, $P_1, .., P_K$ are cliques generated from matrix $B$. Considering that $Y = \sum_l \mu_l P_l$, any clique $P'$ detected at different thresholds will be in the form of $P' = \bigwedge_{l \in I} P_l$ for an $I \subseteq \{1, 2, .., K\}$. Each clique will be detected at most $Q$ times. Then the total number of detected cliques will not be more than $Q$ times $2^K$, the total number of possible $I$s. However, based on the distribution of elements of matrix $B$ we know the typical set for $I$s have size of $2^{KH(p_N)}$, asymptotically for large $n$, where $H(.)$ is entropy function. This follows the fact that the bit strings for nodes will be in a typical set where each bit-string has $Kp_N$ bits of one. Hence, the number of nonempty common intersections among a nontypical set of cliques will converge to zero and we will have $C_{N,p} = Q2^{KH(p_N)} + QK$. □