Detecting Multiple Stochastic Network Motifs in Network Data

Kai Liu, William K. Cheung, and Jiming Liu

Department of Computer Science
Hong Kong Baptist University
224 Waterloo Road, Kowloon Tong, Hong Kong
{kliu,william,jiming}@comp.hkbu.edu.hk

Abstract. Network motif detection methods are known to be important for studying the structural properties embedded in network data. Extending them to stochastic ones help capture the interaction uncertainties in stochastic networks. In this paper, we propose a finite mixture model to detect multiple stochastic motifs in network data with the conjecture that interactions to be modeled in the motifs are of stochastic nature. Component-wise Expectation Maximization algorithm is employed so that both the optimal number of motifs and the parameters of their corresponding probabilistic models can be estimated. For evaluating the effectiveness of the algorithm, we applied the stochastic motif detection algorithm to both synthetic and benchmark datasets. Also, we discuss how the obtained stochastic motifs could help the domain experts to gain better insights on the over-represented patterns in the network data.

Keywords: Stochastic motifs, finite mixture models, expectation maximization algorithm, social networks.

1 Introduction

Network motifs, also known as simple building blocks of complex networks, are defined as patterns of interactions that appear in different parts of a network more frequently than those found in randomized networks. With the network represented as a graph, network motifs can be interpreted as the over-represented subgraph patterns embedded in the graph. Since the pioneering work by Shen-Orr *et. al* [1], there have been a lot of research works on detecting network motifs in biological networks [4,5,6] with the objective to gain insights on the relationship between the network structural properties and the functions they possess. Milo *et al.* [2,3] generalized the idea to characterize a broad range of networks, including ecosystem food webs, neuronal networks, World Wide Web, etc. Recently, network motif detection has also been applied to social network analysis. For example, an email based social network can be well characterized by the Z-score distribution of embedded 3-node subgraph patterns [8,9].

Most of the existing works on network motif detection assume that the network motif is deterministic, which means that the corresponding subgraph

P.-N. Tan et al. (Eds.): PAKDD 2012, Part II, LNAI 7302, pp. 205–217, 2012.

patterns either appear completely or are missing totally. Deterministic network motif detection methods could give inaccurate results if the motifs exhibit stochastic properties. The corresponding stochastic network can be modeled as a mixture of a background random ensemble and families of mutually similar but not necessarily identical interconnection patterns represented by a stochastic network motif [6] (which is also called probabilistic motif in [5]). Stochastic motif detection can then be casted as a missing-value inference and parameter estimation problem under a Bayesian framework. Expectation-Maximization(EM) algorithm and Gibbs sampling can readily be adopted [6,10].

Recently, Liu *et al.* [11] applied the finite mixture model to analyze social media but with the assumption that there is only one stochastic motif. This paper generalizes this work to model stochastic network as a finite mixture model with k components ($k \geq 1$) and adopt the Bayesian approach for detecting the optimal set of multiple stochastic motifs. The paper is organized as follow. Section 2 presents the problem formulation. Evaluation results obtained via experiments performed based on both synthesis and benchmark datasets are reported in Section 3. Section 4 concludes the paper with future research directions.

2 Network Motif Analysis in Social Media

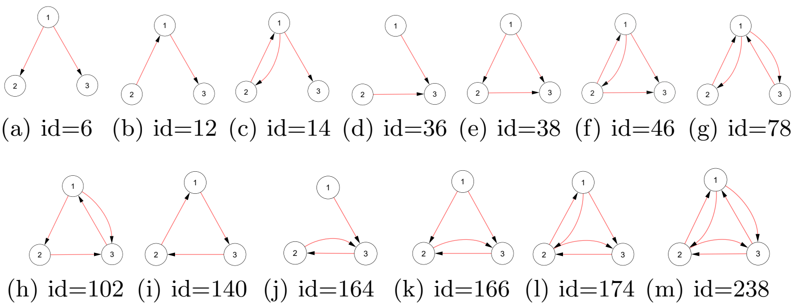
Analyzing triads embedded in networks have long been found important in conventional social network analysis. However, local interaction patterns (or termed as “ties” in social network analysis community) which are salient for characterizing the overall structure of the networks could appear with stochastic variations. It makes conventional motif detection methods problematic as demonstrated in [11]. For large online networks which contain interactions of millions of different individual entities, the incorporation of stochastic models becomes especially essential for more robust motif detection. This is analogous to the need of hidden Markov Model (HMM) for more robust speech recognition and that of conditional random field (CRF) for information extraction. For stochastic motif detection, the target of detection is the embedded network motifs (foreground) and the other links are modeled as the random background.

Relationships or interactions among N elementary units in a population could be represented as a graph G with N nodes and a set of edges denoted by an adjacency matrix $\boldsymbol{A} = (a_{ij})_{N \times N}$. For directed graphs, $a_{ij} = 1$ if there is a directed edge pointing from node i to node j, and 0 otherwise. For undirected graphs, $a_{ij} = 1$ if node i and node j are connected, and 0 otherwise. Subsets of nodes in G with only the local connectivity considered define subgraphs of G. A subgraph S with n nodes can be described by an adjacency matrix $X_S = (x_{ij})_{n \times n}$, where x_{ij} is either 0 or 1 to indicate its connectivity. By sampling subgraphs of a relatively small size (say, triads) from G, the frequency distribution of their appearance can characterize the local structural properties of the graph. To extend from this, a set of subgraphs with “structurally similar” adjacency matrices defines a stochastic network subgraph pattern which if over-represented defines a stochastic network motif M.

2.1 Canonical Forms of Subgraphs for Modeling Stochastic Motifs

Enumerating or sampling subgraphs from a network is the pre-processing step needed before related stochastic models can be applied. A well-known problem is the handling of subgraph isomorphism. Intuitive speaking, structurally equivalent subgraph instances could have their nodes labeled in different orders, making those equivalent subgraphs associated with very different adjacency matrices. Identifying subgraph isomorphism itself is NP-complete in general [13]. Some heuristic computational tricks could be applied to reduce the computational complexity issue on average. In this work, an efficient graph/subgraph isomorphism testing algorithm Nauty [12] is used to check for structurelly equivalent subgraphs and relabel them based on a canonical one so that their appearances can be well aggregated and the stochastic model learning can be accurate.

The remaining question is the choice of the canonical forms. Existing methods for detecting deterministic motifs assume that the motifs are independent and the choices of the canonical forms for the isomorphically equivalent groups of subgraph instances can just be independently considered. However for stochastic motifs, one should expect a stochastic motif model which gives a high probability value to the canonical form of a subgraph pattern A should give also a relatively high value to that of a subgraph pattern B which is a subgraph of A. In other words, the canonical forms for the different isomorphically equivalent subgraph patterns should be chosen in such a way that one being the subset of another should be "aligned" as reflected in their node labeling orders. With this considered, we carefully derived the set of canonical subgraph patterns for subgraphs with 3 nodes as shown in Figure 1. For subgraphs with more than 3 nodes, we are currently studying the possibility of building the corresponding canonical forms efficiently by joining and/or extending the canonical adjacency matrices of 3-node subgraphs [14].



(a) id=6 (b) id=12 (c) id=14 (d) id=36 (e) id=38 (f) id=46 (g) id=78

(h) id=102 (i) id=140 (j) id=164 (k) id=166 (l) id=174 (m) id=238

Fig. 1. All possible subgraphs of canonical form with 3 nodes

2.2 Finite Mixture Model

With the assumption that a stochastic network can be modeled as a mixture of families of independent foreground stochastic motifs embedded in a background random ensemble, each subgrah in the stochastic network can be regarded as either generated from the background or from one of the foreground motifs. In this paper, we extend from the mixture model in [6,11] that multiple stochastic motifs can be detected and the number of motifs required can be estimated.

Assuming there exist k stochastic motifs $\boldsymbol{M}_f = \{M_1, \cdots, M_k\}$ which are represented as a set of probability matrices $\boldsymbol{\Theta}_f = \{\Theta_1, \cdots, \Theta_k\}$, with $\Theta_h = (\theta^h_{ij})_{n\times n}, 0 \leq \theta_{ij} \leq 1, 1 \leq h \leq k$. θ^h_{ij} denotes the probability that there is an edge from node i to j in the h-th motif. The background ensemble M_0 is characterized by a family of randomized networks generated from a given stochastic network which contain the same number of nodes and edges, and the same statistics for the nodes' in/out degrees.

Moreover, let $\{S_1, \cdots, S_W\}$ denote a set of subgraph instances sampled from a given network, $\boldsymbol{X} = \{\boldsymbol{X}^1, \cdots, \boldsymbol{X}^W\}$ denote the adjacency matrices corresponding to the subgraph instances (observed data) where $\boldsymbol{X}^w = (x^w_{ij})_{n\times n}$ and $x^w_{ij} = \{0, 1\}$, Z^w_h denotes an indicator variable taking the value of 1 if subgraph instance S_w comes from the model M_h or 0 otherwise, and thus $\boldsymbol{Z} = (\boldsymbol{Z}^1, \cdots, \boldsymbol{Z}^W)^T$ form the missing data of the problem, where $\boldsymbol{Z}^w = (Z^w_0, \cdots, Z^w_k)^T$. The probability that $\boldsymbol{X}^w$ comes from M_h is given as

$$p(\boldsymbol{X}^w|\Theta_h) = \prod_{i=1}^{n}\prod_{j=1}^{n}(\theta^h_{ij})^{x^w_{ij}}(1 - \theta^h_{ij})^{1-x^w_{ij}}. \tag{1}$$

Also, let $\boldsymbol{\lambda} = (\lambda_0, \cdots, \lambda_k)$ be the mixing portion of the mixture model which also denotes the prior probabilities of $\Pr(Z^w = 1), w = \{1, \cdots, W\}$.

The stochastic motif detection problem can thus be casted as a maximum likelihood estimation problem for $\boldsymbol{\Theta} = \{\boldsymbol{\Theta}_f, \boldsymbol{\lambda}\}$ where the log-likelihood function for the complete data is given as

$$l(\boldsymbol{\Theta}) = \log p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\Theta}) \quad = \sum_{w=1}^{W}\sum_{h=0}^{k} Z^w_h \log \lambda_h + \sum_{w=1}^{W}\sum_{h=0}^{k} Z^w_h \log p(\boldsymbol{X}^w|\Theta_h). \tag{2}$$

The EM algorithms for estimating $\boldsymbol{\Theta}_f$ and $\boldsymbol{\lambda}$ will be presented in the next section.

For the background model, we are interested in the probability of observing the subgraph instance S_w in the background model $p(\boldsymbol{X}^w|\Theta_0)$ instead of Θ_f. As in [6,11], the background model is estimated by counting the subgraph instances in randomized networks. We first generate a set of randomized networks. For each randomized network described by an adjacency matrix $\boldsymbol{A} = (a_{ij})_{N\times N}$, we randomly choose pairs of connections and repeatedly swap the target of them until the network is well randomized, while keeping the incoming and outgoing degrees of each node remain unchanged, i.e., keeping the summation of each row and each column in the adjacency matrix unchanged. Subgraphs are then sampled from the randomized networks. $p(\boldsymbol{X}^w|\Theta_0)$ is estimated as N_w/N_{total}, where N_w is the number of the subgraph S_w sampled from the ensemble of the randomized networks and N is the total number of subgraphs sampled with the same size with S_w.

2.3 Basic EM Algorithm

For learning probabilistic models with missing data (unknown motifs for our case), the Expectation-Maximization (EM) algorithm [15] is typically used for obtaining the Maximum Likelihood (ML) estimates of the model parameters.

The EM algorithm produces a sequence of estimates by alternatingly applying the E-step and M-step until it converges.

- E-step: Compute the complete data expectation of log-likelihood $E[l(\boldsymbol{\Theta})]$ given the observed data $\boldsymbol{X}$ and the current estimates of model parameters $\hat{\boldsymbol{\Theta}}$. We have
$$E[l(\boldsymbol{\Theta})] = \sum_{w=1}^{W}\sum_{h=0}^{k} E[Z_h^w]\log\hat{\lambda}_h + \sum_{w=1}^{W}\sum_{h=0}^{k} E[Z_h^w]\log p(\boldsymbol{X}^w|\hat{\Theta}_h), \quad (3)$$
where $$E[Z_h^w] = E[Z_h^w|\boldsymbol{X},\hat{\boldsymbol{\Theta}}] = \frac{p(\boldsymbol{X}^w|\hat{\Theta}_h)\hat{\lambda}_h}{\sum_{j=0}^{k} p(\boldsymbol{X}^w|\hat{\Theta}_j)\hat{\lambda}_j} \quad (4)$$
- M-step: The model parameters are estimated by maximizing the expectation of the log-likelihood, given as
$$(\boldsymbol{\lambda}^*, \boldsymbol{\Theta}_f^*) = \arg\max_{\boldsymbol{\lambda},\boldsymbol{\Theta}_f} E[l(\boldsymbol{\Theta})]. \quad (5)$$
And the updating rules for $\boldsymbol{\lambda}$ and Θ_h are given as
$$\lambda_h^* = \frac{1}{W}\sum_{w=1}^{W} E[Z_h^w] \qquad \lambda_0^* = 1 - \sum_{h=1}^{k}\lambda_h^* \quad (6)$$
$$(\theta_{ij}^h)^* = \frac{\hat{\alpha}_{ij}^h}{\hat{\alpha}_{ij}^h + \hat{\beta}_{ij}^h}, \quad \hat{\alpha}_{ij}^h = \sum_{w=1}^{W} E[Z_h^w]x_{ij}^w, \quad \hat{\beta}_{ij}^h = \sum_{w=1}^{W} E[Z_h^w](1 - x_{ij}^w). \quad (7)$$

Note that $p(\boldsymbol{X}^w|\Theta_0)$ is estimated based on the subgraph statistics in the ensemble of randomized networks as explained in the previous section.

2.4 Learning the Optimal Number of Motifs

To determine the optimal number of stochastic motifs automatically, we adopt the *Component-wise EM for Mixture*(CEM^2) which was proposed to integrate both the model parameter estimation and model selection steps into one single EM algorithm [16]. The general idea of CEM^2 is to update the parameters of each component one by one so that the component with very low support by the data can be pruned. CEM^2 starts from all possible k-component mixtures and prunes the died components($\lambda_h = 0$) sequentially at each EM iteration.

CEM^2 implements the *minimum message length*(MML) criterion [17] to select the number of components. The best parameter estimate for the mixture model is the one minimizing the message length $L[\boldsymbol{\Theta}, \boldsymbol{X}]$, which is given by

$$L[\boldsymbol{\Theta}, \boldsymbol{X}] = L[\boldsymbol{\Theta}] + L[\boldsymbol{X}|\boldsymbol{\Theta}], \quad (8)$$

where $L[\boldsymbol{\Theta}]$ is the minimum message length for prior information, and $L[\boldsymbol{X}|\boldsymbol{\Theta}]$ is the minimum message length for data which can be estimated as $-\log p(\boldsymbol{X}|\boldsymbol{\Theta})$. As in [16], the final cost function (message length) $L[\boldsymbol{\Theta}, \boldsymbol{X}]$ is given by

$$L[\boldsymbol{\Theta}, \boldsymbol{X}] = \frac{N}{2}\sum_{m:\lambda_h>0}\log(\frac{W\lambda_h}{12}) + \frac{k_{nz}}{2}\log\frac{W}{12} + \frac{k_{nz}(N+1)}{2} - \log p(\boldsymbol{X}|\boldsymbol{\Theta}), \quad (9)$$

where k_{nz} is the number of components with non-zero probability, and N is the number of parameters specifying each component. The detailed steps of CEM^2 to motif detection can be found in Algorithm 1.

Algorithm 1. CEM^2 Algorithm

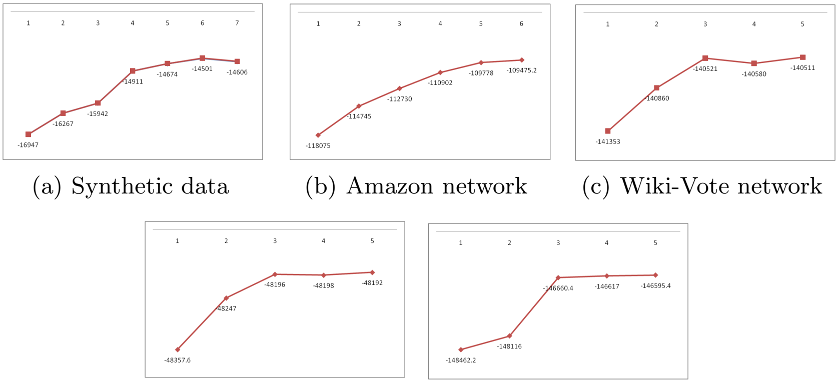
```
Input: Subgraphs $\boldsymbol{X} = \{\boldsymbol{X}_1, \cdots, \boldsymbol{X}_W\}$, $\epsilon$, $k_{min}$, $k_{max}$, initial parameters $\hat{\boldsymbol{\Theta}}(0) = \{\hat{\Theta}_1, \cdots, \hat{\Theta}_{k_{max}}; \hat{\lambda}_1, \cdots, \hat{\lambda}_{k_{max}}\}$
Output: Mixture model with optimal $\boldsymbol{\Theta}^*$
$t \leftarrow 0, k_{nz} \leftarrow k_{max}, L_{min} \leftarrow +\infty$
$u_h^w \leftarrow p(\boldsymbol{X}^w|\hat{\Theta}_h), c_h \leftarrow \max\{0, (\sum_{w=1}^{W} E[Z_h^w]) - \frac{N}{2}\}$, for $h = 1, \cdots, k_{max}$, and $w = 1, \cdots, W$
while $k_{nz} \geq k_{min}$ do
  repeat
    $t \leftarrow t + 1$
    for $h = 1$ to $k_{max}$ do
      E-step: $E[Z_h^w] = u_h^w \hat{\lambda}_h (\sum_{j=0}^{k_{max}} u_j^w \hat{\lambda}_j)^{-1}$, $\lambda_h \leftarrow c_h (\sum_{j=0}^{k_{max}} c_j)^{-1}$
      M-step: $\{\hat{\lambda}_1, \cdots, \hat{\lambda}_{k_{max}}\} \leftarrow \{\hat{\lambda}_1, \cdots, \hat{\lambda}_{k_{max}}\}(\sum_{h=0}^{k_{max}} \hat{\lambda}_h)^{-1}$
        $\hat{\lambda}_0 = 1 - \sum_{h=1}^{k_{max}} \hat{\lambda}_h$
        if $\hat{\lambda}_h > 0$ then
          update $\boldsymbol{\Theta}_f$ according to Eq.(7), and $u_h^w \leftarrow p(\boldsymbol{X}^w|\hat{\Theta}_h)$
        else
          $k_{nz} \leftarrow k_{nz} - 1$
        end if
    end for
    $\hat{\boldsymbol{\Theta}}(t) = \{\hat{\Theta}_1, \cdots, \hat{\Theta}_{k_{max}}; \hat{\lambda}_0, \cdots, \hat{\lambda}_{k_{max}}\}$
    calculate $L[\hat{\boldsymbol{\Theta}}(t), \boldsymbol{X}]$ according to Eq.(9)
  until $L[\hat{\boldsymbol{\Theta}}(t-1), \boldsymbol{X}] - L[\hat{\boldsymbol{\Theta}}(t), \boldsymbol{X}] < \epsilon|L[\hat{\boldsymbol{\Theta}}(t-1), \boldsymbol{X}]|$
  if $L[\hat{\boldsymbol{\Theta}}(t-1), \boldsymbol{X}] \leq L_{min}$ then
    $L_{min} \leftarrow L[\hat{\boldsymbol{\Theta}}(t-1), \boldsymbol{X}]$
    $\boldsymbol{\Theta}^* \leftarrow \hat{\boldsymbol{\Theta}}(t)$
  end if
  $h^* \leftarrow \arg\min_h\{\hat{\lambda}_h > 0\}$, $\hat{\lambda}_h \leftarrow 0$, $k_{nz} \leftarrow k_{nz} - 1$
end while
```

3 Experimental Results

In this section, we present experimental results to demonstrate first the correctness of the detected stochastic motifs using synthetic datasets. Then, we further present the results of applying the stochastic motif detection algorithm to some real datasets and provide interpretation of the results obtained.

3.1 Results on Synthetic Networks

We generated a set of synthesized networks for correctness evaluation. Each network is generated by 1) creating a group of subgraphs coming from a known set of reference stochastic motifs as foreground models and 2) adding random links among the subgraphs to generate random background. In particular, we chose the subgraphs commonly found in many real networks, e.g., id is 38, 46, 166, 174, and 238 (see Fig. 1) as the reference motifs. We then applied our method to the synthetic networks we generated. In order to avoid the EM algorithm being



(a) Synthetic data (b) Amazon network (c) Wiki-Vote network

(d) Slashdot network (e) Epinions network

Fig. 2. The plot of the expected log likelihood under different numbers of motifs

trapped into local optima, we ran the EM algorithm several times with different initializations to report the best $\boldsymbol{\Theta}$ in terms of the likelihood value.

Figs. 3(a) - 3(e) show the stochastic motifs obtained by the multiple motif detection method in the synthetic networks. According to Fig. 2(a), the value of $E[l(\boldsymbol{\Theta})]$ increases a lot when the motif number varies from 1 to 5. There is a sharp drop in the increasing rate for the value of E$[l(\boldsymbol{\Theta})]$ when $k = 5, 6, 7$. This is consistent to the fact that there are 5 reference motifs used for generating the synthetic networks. A similar conclusion can be drawn by referring to Table 1.

Table 1. λs when the number of motifs is 5, 6 and 7 ($\times 10^{-2}$)

	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7
$k = 5$	7.1	9.8	8.6	7.7	**2**	-	-
$k = 6$	3.8	7.9	8.5	10.1	**13.5**	0.2	-
$k = 7$	9.1	9.5	9.1	6.0	**9.6**	0.1	0.1

Table 2. Dataset statistics ($\times 10^3$)

	Amazon	Wiki	Slash	Epinions
# nodes	262	8	77	76
# edges	1,235	104	828	509
# subgraphs	7,685	13,329	67,361	70,911

3.2 Results on Benchmark Datasets

We have also applied the stochastic motif detection algorithm to large-scale social network datasets named "Amazon", "Wiki-Vote", "Slashdot" and "Epinions" which are obtained as described in [18,19]. The dataset "Amazon" considers the Customers Who Bought This Item Also Bought feature of the Amazon website. If a product A is frequently co-purchased with product B, the graph contains an directed edge from node A to node B. "Wiki-Vote" is a network consisting of voting interaction for Wikipedia admin candidates. The link refers to a vote from a user to an admin candidate represented a user agree or disagree the promotion of the admin candidate. "Slashdot" is a social network of technology blog. The links in this network are the designations of "friends" or "foes". "Epinions" is a trust network, where we can know the trust or distrust relations of the users from the directed links between each other. Table 2 lists the statistics of these four datasets. These networks have order of tens to hundreds of thousands of nodes and hundreds of thousands to millions of edges. In each network, we know the directions of all the edges.

Figs. 2(b) - 2(e) show the expected maximum log likelihood values E$[l(\boldsymbol{\Theta})]$ of the mixture models with different component numbers in the four datasets. Here we can determine the best number of motifs by visual inspection to identify the points where the increase of E$[l(\boldsymbol{\Theta})]$ starts to slow down. Also, by referring to tables 3 - 6, the values of λ with different motif numbers in these four datasets also hint us the best optimal number to choose from (as marked with bold face in the tables). For instance, for Amazon network, when the number of motifs k is set to 6, the value of λ_6 is very small. This hints that the 6-th motif is only "supported" by a very limited number of subgraph instances and thus 5 stochastic motifs could be enough. Similar results were obtained for Wiki-Vote, Slashdot and Epinions networks.

Table 3. λs for Amazon ($\times 10^{-2}$)

	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6
$k = 4$	2.3	3.5	2.1	1.5	-	-
$k = 5$	1.7	1.6	2.3	2.3	**1.6**	-
$k = 6$	2.2	2.7	1.4	1.3	**1.8**	0.1

Table 4. λs for Wiki-Vote ($\times 10^{-2}$)

	λ_1	λ_2	λ_3	λ_4
$k = 2$	1.0	1.2	-	-
$k = 3$	2.0	1.1	**1.1**	-
$k = 4$	1.2	3.9	**4.7**	0.2

Table 5. λs for Slashdot ($\times 10^{-3}$)

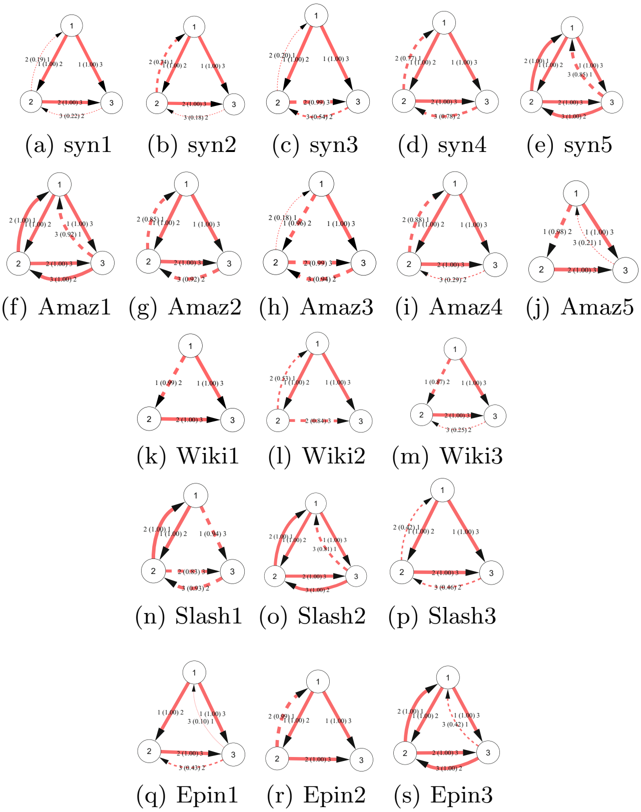
	λ_1	λ_2	λ_3	λ_4	λ_5
$k = 3$	0.8	1.3	**2.9**	-	-
$k = 4$	1.5	1.8	**1.6**	0.42	-
$k = 5$	1.1	2.2	**1.4**	0.32	0.39

Table 6. λs for Epinions ($\times 10^{-3}$)

	λ_1	λ_2	λ_3	λ_4	λ_5
$k = 3$	8.4	4.7	**8.0**	-	-
$k = 4$	8.1	5.7	**2.3**	0.45	-
$k = 5$	9.2	4.8	**5.8**	0.54	0.27

Fig. 3 shows the stochastic motifs detected in the datasets we used. Similar to [11], one can make interpretations on the networks of study based on the motifs extracted, which can in turn be validated by related domain experts. For instance, we made the following observations which seems revealing some local structural properties of the networks:

– By referring to the results obtained based on the Amazon dataset (Figs. 3(f) - 3(j)), we observed the following patterns: i) a 3-node pattern (Fig. 3(f)) where three products are always co-purchased bidirectionally; ii) a 3-node pattern (Fig. 3(g)) where only two pairs of products are co-purchased bidirectionally but not the third pair; iii) some other 3-node patterns where only one pair of products are co-purchased bidirectionally but not the other two (Figs. 3(h),3(i)); and iv) a 3-node pattern where the co-purchasing is never done directionally for the related products. It could be interesting to further analyze whether the four patterns are corresponding to different product characteristics, which could in turn result in some more context specific product recommendation methodologies.



(a) syn1 (b) syn2 (c) syn3 (d) syn4 (e) syn5

(f) Amaz1 (g) Amaz2 (h) Amaz3 (i) Amaz4 (j) Amaz5

(k) Wiki1 (l) Wiki2 (m) Wiki3

(n) Slash1 (o) Slash2 (p) Slash3

(q) Epin1 (r) Epin2 (s) Epin3

Fig. 3. Stochastic motifs detected in different datasets with the edge width showing the corresponding probability of edge appearance, and the edge label gives the actual probabilities. E.g., 1(0.92)2 means the occurrence probability of edge x_{12} is 0.92. A dashed edge here implies a probability value less than 1.

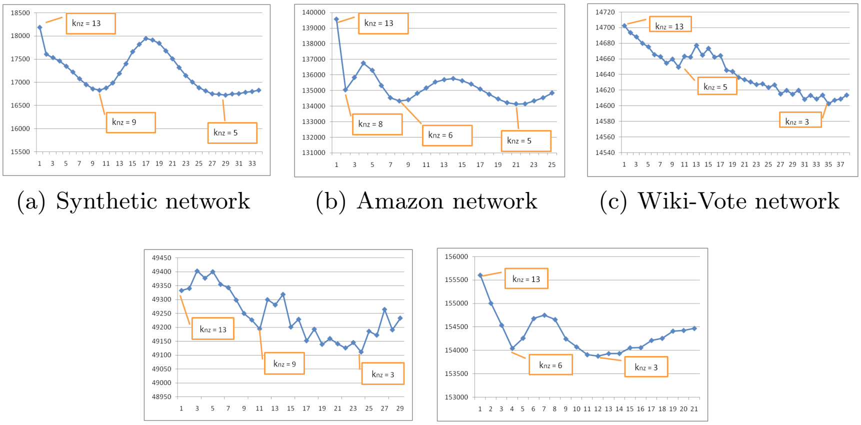
- From the results obtained based on the Wiki-Vote dataset (Figs. 3(k) - 3(m)), it is also interesting to observe the following patterns: i) a 3-node pattern (Fig. 3(k)) where co-voting never occurs; and ii) some 3-node patterns where co-voting only occasionally occurs for one pair of voters but not the other pairs (Figs. 3(l) and 3(m)). In general, co-voting activities within a triad are not commonly observed. We believe that this could be related to the user psychology behind the voting process, requiring again further investigation effort with respect to the corresponding application context.
- All the motifs detected in these social networks consist of a basic feed-forward loop structure (structure of Fig. 1(e)) with some additional edges. The feed-forward loop structure is the most popular deterministic motif found in biological and social networks, which follows the status theory in social networks proposed in [18]. E.g., if A regards B as having higher status (a link from A to B), and B regards C as having higher status (a link from B to C), so A should regard C as having higher status and hence be inclined to link

from A to C. So, the feed-forward loop structure is often over-represented while the feedback loop (structure of Fig. 1(i) having link from C to A) is under-represented instead.

As inspired by [18], we plan to make reference to different social psychology theories developed in social science to validate and gain further insights and thus explanation on the underlying social behaviors embedded in the social media.

3.3 Effectiveness of CEM^2 in Estimating Optimal Number of Motifs

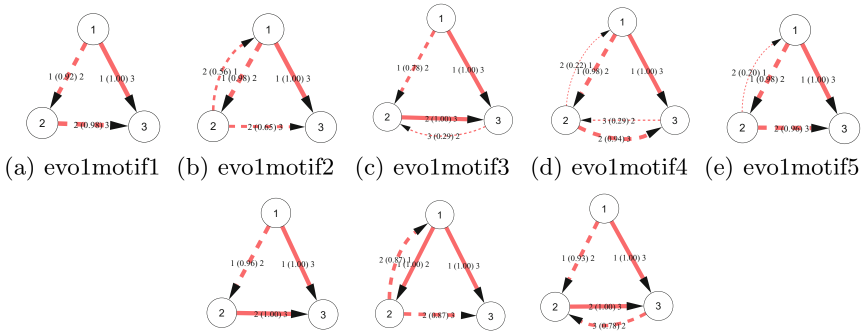
Fig. 4 shows how the cost functions $L(\boldsymbol{\Theta}, \boldsymbol{X})$ evolve throughout the CEM^2 iterations. Starting from the maximum possible number of motifs ($k_{nz} = 13$ for motif size is 3), the cost function decreases as the CEM^2 iterations proceed. When some components are pruned as described in the algorithm, the value of the cost function would increase to some extent. After some iterations, the remaining motifs will then be learned to better fit to the data, and thus the cost function decreases again. The number of motifs is automatically estimated by choosing with the one which gives the lowest cost function value. For synthetic data, the mixture model with 5 motifs gives the lowest cost function value, which is consistent to that estimated using the basic EM. In the four social networks we used, the cost functions have the lowest values when the motif numbers become 3, 3, 3 and 5 respectively, which are also consistent to the results observed in Section 3.2, but here we need to run CEM^2 only once. Fig. 5 shows the evolution of motifs annihilation by taking the Wiki-Vote network as an example. Figs. 5(a) - 5(e) are the motifs when the number of motifs is 5. With the iteration continues until convergence, the number of motifs becomes to 3, the corresponding motifs are listed in Figs. 5(f) - 5(h).



(a) Synthetic network (b) Amazon network (c) Wiki-Vote network

(d) Slashdot network (e) Epinions network

Fig. 4. The evolution of cost functions $L(\boldsymbol{\Theta}, \boldsymbol{X})$ until convergence in different datasets, the x-axis gives iteration times, k_{nz} means the number of none-zero components



(a) evo1motif1 (b) evo1motif2 (c) evo1motif3 (d) evo1motif4 (e) evo1motif5

(f) evo2motif1 (g) evo2motif2 (h) evo2motif3

Fig. 5. The evolution of motif pruning by taking Wiki-Vote network as an example

3.4 Computational Complexity

The overall complexity include those for subgraph sampling, generating of random networks, and the parameter estimation via the EM algorithms.

The complexity of sampling subgraphs of n nodes in a network is $R_S = O(N_s K^{n-1} n^{n+1})$, where K is a small constant value corresponding to the average node degree in the network, and N_s is the number of subgraphs sampled. The background model is simulated by randomized networks which generated by the switch method as in [2,11], where many rounds of "switchings" of two edges randomly selected from the real network are conducted while keeping the in/out degree of each node fixed. In so doing, the complexity of generating a random network is $O(T_s N_e)$ (the number of switches), where T_s is the switch times per edge (a random number in the range of $100 - 200$) and N_e is the number of edges in the real network. Overall time complexity for pre-processing is $O(N_s K^{n-1} n^{n+1}(1+N_r)+N_r T_s N_e)$, where N_r is the number of random networks.

For the basic EM algorithm, the complexity of each iteration is $O(n^2 N_s)$. So, the total complexity of EM algorithm together with pre-processing is $O(N_s \times K^{n-1} n^{n+1}(1+N_r)+N_r \times T_s \times N_e + k \times I \times n^2 N_s)$, where I is the iteration times of EM algorithm and k is optimal number of motifs. For CEM^2, it is only slightly computationally heavier than the basic EM algorithm due to the multiple E-steps to recompute $E[Z_h^w]$ [16]. As updating $E[Z_h^w]$ needs only full computation of Eq.(4) for $j = h$. For $j \neq h$, the terms $(X^w|\theta_h)$, which could contribute a lot to the computational cost of E-step, remain unchanged and thus only need to be computed once per sweep, like in the basic EM. However, the basic EM should be run several times with different motif numbers. CEM^2 is needed to run only once. So, the overall time complexity of CEM^2 is lighter than basic EM.

For further speedup, as the data are assumed independent and identically distributed ($i.i.d$) and thus can be partitioned into multiple subsets, our method can also take the advantage of parallel computing on GPUs [20] so as to be more scalable to large-scale datasets.

4 Conclusion and Future Works

Motif detection provides an important tool to assist the study of structural properties in network data for domains like bioinformatics and on-line social media. We proposed the use of the finite mixture model to detect multiple stochastic motifs in network data and a related CEM algorithm for automatically determining the optimal number of motifs embedded and the model parameters of the motifs. We applied the method to both synthetic and several benchmark datasets and discussed how the obtained motifs could be used to gain an in-depth understanding of the underlying stochastic local interaction patterns.

Our method works well for analyzing the network structural properties based on small motifs (i.e., 3 or 4 nodes). For future work, more scalable (possibly parallel) implementation will be needed if the analysis is to be carried out for motifs of various sizes. From the perspective of further improving modeling and thus the analysis power, related research directions include: 1) extending the method to take into consideration the sign of the edges, and 2) incorporating the timing information on edges to detect temporal motifs as a family of similar interaction patterns which over-represented throughout the period of time.

Acknowledgments. This work was supported by the General Research Fund (HKBU210410) from the Research Grant Council of the Hong Kong Special Administrative Region, China.

References

1. Shen-Orr, S., Milo, R., Mangan, S., Alon, U.: Network motifs in the transcriptional regulation network of Escherichia coli. Nature Genetics 31(1), 64–68 (2002)
2. Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovski, D., Alon, U.: Network motifs: Simple building blocks of complex networks. Science 298(5594), 824–827 (2002)
3. Milo, R., Itzkovitz, S., Kashtan, N., Levitt, R., Shen-Orr, S., Ayzenshtat, I., Sheffer, M., Alon, U.: Superfamilies of evolved and designed networks. Science 303(5663), 1538–1541 (2004)
4. Mangan, S., Alon, U.: Structure and function of the feedforward loop network motif. PNAS USA 100(21), 11980–11985 (2003)
5. Berg, J., Michael, L.: Local graph alignment and motif search in biological networks. PNAS USA 101(41), 14689–14694 (2004)
6. Jiang, R., Tu, Z., Chen, T., Sun, F.: Network motif identification in stochastic networks. PNAS USA 103(25), 9404–9409 (2006)
7. Kashtan, N., Itzkovitz, S., Milo, R., Alon, U.: Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs. Bioinformatics 20(11), 1746–1758 (2004)
8. Juszczyszyn, K., Kazienko, P., Musiał, K.: Local Topology of Social Network Based on Motif Analysis. In: Lovrek, I., Howlett, R.J., Jain, L.C. (eds.) KES 2008, Part II. LNCS (LNAI), vol. 5178, pp. 97–105. Springer, Heidelberg (2008)
9. Musial, K., Juszczyszyn, K.: Motif-based analysis of social position influence on interconnection patterns in complex social network. In: Proceedings of First Asian Conference on Intelligent Information and Database Systems, pp. 34–39 (2009)

10. Jiang, R., Chen, T., Sun, F.: Bayesian models and Gibbs sampling strategies for local graph alignment and motif identification in stochastic biological networks. Communications in Information & Systems 9(4), 347–370 (2009)
11. Liu, K., Cheung, W.K., Liu, J.: Stochastic network motif detection in social media. In: Proceedings of 2011 ICDM Workshop on Data Mining in Networks (2011)
12. McKay, B.: Nauty user's guide (version 2.4). Australian National University (2007)
13. Garey, M., Johnson, D.: Computers and intractability: A guide to the theory of np-completeness. Freeman San Francisco (1979)
14. Huan, J., Wang, W., Prins, J.: Efficient mining of frequent subgraphs in the presence of isomorphism. In: Proceedings of 2003 IEEE ICDM, pp. 549–552 (2003)
15. Dempster, A., Laird, N., Rubin, D.: Maximum likelihood from incomplete data via the EM algorithm. J. of the Royal Statistical Society. Series B 39(1), 1–38 (1977)
16. Figueiredo, M., Jain, A.: Unsupervised learning of finite mixture models. IEEE Transactions on PAMI 24(3), 381–396 (2002)
17. Wallace, C., Dowe, D.: Minimum message length and kolmogorov complexity. The Computer Journal 42(4), 270–283 (1999)
18. Leskovec, J., Huttenlocher, D., Kleinberg, J.: Signed networks in social media. In: Proceedings of the 28th International Conference on Human Factors in Computing Systems, pp. 1361–1370 (2010)
19. Leskovec, J., Adamic, L., Huberman, B.: The dynamics of viral marketing. ACM Transactions on the Web 1(1), 5–44 (2007)
20. Kumar, N., Satoor, S., Buck, I.: Fast parallel expectation maximization for gaussian mixture models on gpus using cuda. In: 11th International Conference on High Performance Computing and Communications, pp. 103–109 (2009)