A FRAMEWORK FOR ATTRIBUTE-BASED COMMUNITY DETECTION WITH APPLICATIONS TO INTEGRATED FUNCTIONAL GENOMICS

HAN YU

Biostatistics, University at Buffalo,
Buffalo, NY 14220/EST, USA
E-mail: hyu9@buffalo.edu

RACHAEL HAGEMAN BLAIR

Biostatistics, University at Buffalo,
Buffalo, NY 14220/EST, USA
E-mail: hageman@buffalo.edu

Understanding community structure in networks has received considerable attention in recent years. Detecting and leveraging community structure holds promise for understanding and potentially intervening with the spread of influence. Network features of this type have important implications in a number of research areas, including, marketing, social networks, and biology. However, an overwhelming majority of traditional approaches to community detection cannot readily incorporate information of node attributes. Integrating structural and attribute information is a major challenge. We propose a flexible iterative method; inverse regularized Markov Clustering (irMCL), to network clustering via the manipulation of the transition probability matrix (aka stochastic flow) corresponding to a graph. Similar to traditional Markov Clustering, irMCL iterates between "expand" and "inflate" operations, which aim to strengthen the intra-cluster flow, while weakening the inter-cluster flow. Attribute information is directly incorporated into the iterative method through a sigmoid (logistic function) that naturally dampens attribute influence that is contradictory to the stochastic flow through the network. We demonstrate advantages and the flexibility of our approach using simulations and real data. We highlight an application that integrates breast cancer gene expression data set and a functional network defined via KEGG pathways reveal significant modules for survival.

Keywords: KEGG pathways, logistic regression, community detection, Markov clustering, omics, survival

1. Introduction

Community structure occurs when nodes exhibit a high-degree of connectivity to each other, and a lower degree of connectivity to other groups and nodes in the network.[1,2] The community detection problem has been studied extensively in Social Network Analysis (SNA). In the areas of bioinformatics and computational biology, the problem is also referred to as module detection or graph clustering.[3,4]

In a general sense, the community detection problem can be viewed as the clustering of a network. Classical graph clustering methods inlcude Kernighan-Lin algorithm,[5] hierarchical clustering methods,[6] spectral clustering,[7,8] Newman and Girvan algorithm,[9,10] and modularity-based algorithms comprise an important class of community detection methods.[11–13] Classical approaches to community detection cannot readily incorporate information of node attributes and rely solely on network structures. The simultaneous use of attribute and connectivity information can yield more accurate results and can be leveraged in downstream analysis

for prediction under attribute or network perturbations. Hanisch *et al.* derive the distance matrix by combining the structural and gene profiles distances, but require prior domain knowledge.[14] Zhou *et al.* represent attributes as additional nodes.[15] In this setting, attributes are restricted to discrete values, and consequently the size and complexity of the graph grows, and requires accounting of the different types nodes and edges.[16] Instead of graph partitioning, the algorithms of `CoPaM`[17] and `DME`[18] introduces a problem of identifying cohesive patterns or subnetworks satisfying a density threshold and cohesive constraints.

We have developed a novel community detection method that rely on stochastic flow in networks. Leveraging robust statistical classification methods, we bridge and simultaneously model the attribute and structural space. The methods that we propose are highly generalizable and flexible in their implementation. We showcase their flexibility through simulation and application that integrates breast cancer gene expression data set with KEGG ontologies and survival data.

2. Materials and Methods

Briefly, we begin by outlining *Markov CLustering (MCL)* and *regularized Markov CLustering (rMCL)* frameworks, which set the foundation of our approaches.[19,20] MCL is based on the notion that if a group of nodes belongs to the same community, then the stochastic flow from these nodes will be concentrated towards nodes in that community.[19] Performing random walks on a graph may reveal where *flows* gather, which suggests potential communities. In this setting, our focus is on undirected graphs, which have a symmetric adjacency matrix and have edge interpretations of association (not causation).

MCL algorithms depend on the iteration between two operators *expand* and *inflate*, until convergence, in order to identify communities in the network. Markov clustering utilizes a stochastic matrix that is initially derived from the adjacency matrix, $A_{adj} \in \mathbf{R}^{n \times n}$ of the graph. The stochastic matrix is defined as the matrix product, $M = A_0 \cdot D^{-1}$, where $A_0 = A_{adj} + I$, and $D \in \mathbf{R}^{n \times n}$ is the diagonal matrix containing the degree information for each node, $D(k,k) = \text{diag}\left(\sum_{i=1}^{n} A\left(i,k\right)\right)$. The operations in MCL and rMCL utilize the stochastic matrix, M, which has columns that can be interpreted as transition probabilities. In the classic MCL, the expand step at the $j+1^{th}$ iteration requires a matrix product $\tilde{M}_{j+1} = M_j \cdot M_j$.

The inflate operator, $M_{j+1}^{inf} = Inflate(\tilde{M}_{j+1}, r)$, can be understood as the component-wise exponentiation $\tilde{m}(i,j)^r$, $\forall\, i,j = 1,\ldots,n$, where the inflation operator, r, is a constant. Following inflation, M_{j+1}^{inf} is converted to a stochastic matrix, M_{j+1}, and a new iteration is started. Importantly, the expand operator alone would give rise to a Markov Chain via a random walk on the graph. However, due to the inflation operator the process cannot be regarded as a Markov Chain. Inflation is critical to accentuate strong ties and paths, and deemphasize weak ones. The inflation constant, r, controls the degree at which this strengthening and weakening is enforced, and has a direct impact on the cluster formation. Upon convergence of MCL to steady-state, the stochastic matrix can be understood in terms of *attractors*. The matrix is sparse, and the *attractors* have at least one positive value in their row. The indices of these positive values, together with the attractor, form the community.

A regularized version of Markov Clustering, rMCL, was proposed and has been shown

to overcome some fragmentation issues in the communities. The rMCL algorithm follows the same iterative approach, with an *expand* step that is replaced by a *regularization operation*, $M_{j+1} = M_j \cdot M_0$, where M_0 is the initial stochastic matrix formed from the network adjacency matrix.[20] The regularize step ensures that the original structural information is still utilized for the graph clustering process after the first iteration. Unfortunately, the regularized MCL does not naturally converge to a steady state with the same desirable interpretations in terms of community membership. In order to achieve this, at each iteration, a *prune* step is added that forces some smaller entries of the stochastic matrix to zero using a heuristic threshold. The pruning aims to eliminate entries that are small relative to other entries in the matrix.[20]

2.1. *inverse regularized Markov Clustering (irMCL)*

We propose a flexible method, **i**nverse **r**egularized **M**arkov **CL**ustering (irMCL), which utilizes the *expand* and *inflate* operators, but relies on an alternative concept of community that emphasizes the spreading of *influence* or *information* in a non-exclusive manner. Our approach relies on the following modeling assumptions:

> **(A1)** Spreading of information/influence from Node i to Node j will not affect that from Node i to other nodes, $k \neq j$.
> **(A2)** Nodes in the same community are influenced or share information from similar group of nodes.
> **(A3)** Nodes with larger degrees tend to be more influential.
> **(A4)** If an individual is highly influenced by a group of nodes, such influence tends to be self-amplified.
> **(A5)** Spread of information between nodes with similar attributes is easier, and thus should be a function of the attributes similarity measures between nodes.

In this model, the community membership of a node is measured by *information* that flows *into* the nodes, as opposed to MCL and rMCL, where a feature is the stochastic flow that *exits* this node. Accordingly, we term this procedure "inverse regularized Markov Clustering" (irMCL). These assumptions naturally give higher weights to nodes in the network with high degrees and naturally incorporate attribute information in a flexible manner. Similar to MCL, we denote $A_{adj} \in \mathbf{R}^{n \times n}$ as the adjacency matrix of graph $\mathcal{G}$. We define a symmetric *spread matrix* as: $A = A_{adj} + I$, which defines the graph with the addition of self loops.

Algorithm 1 shows the full details of the irMCL approach. At each iteration, the initial spread matrix used to regularize. Repeated use of the spread matrix naturally puts more weight on the high degree nodes in the network (A3), and is unique to our approach. The same inflation operator as in MCL is used according to assumption (A4). Convergence is tracked empirically by examining the mean squared difference as the difference between M_j and M_{j-1}, defined as $\sum_{i=1}^{n}\sum_{k=1}^{n}\left(m_{ik}^{(j)} - m_{ik}^{(j-1)}\right)^2/n$, where $m_{ik}^{(j)}$ is the entry of M_j.

The output of this iterative method is a stochastic matrix, where the rows with high similarity are likely to belong to the same community. In our applications, we utilize complete linkage, and estimate the similarity using a euclidean distance. Silhouette plots are utilized for the determination of the number of clusters via average silhouette width.[21]

Algorithm 2.1 Feature derivation for inverse Regularized Markov Clustering (iRMCL)

Initialize:
$A_{\text{adj}} \in \mathbf{R}^{n \times n}$ Adjacency Matrix
$A_0 = A_{\text{adj}} + I$
for $k = 1$ to n **do**
  $D_0(k,k) = diag\left(\sum_{i=1}^{n} A_0(i,k)\right)$
end for
set: $r > 1$

Repeat until stopping criteria is met
for $j = 1$ to m **do**
  $M_j \leftarrow M_{j-1} \cdot A_0$
  $M_j^{\text{infl}} = \text{Inflate}(\tilde{M}_j, r)$
  for $k = 1$ to n **do**
    $D_j(k,k) = diag\left(\sum_{i=1}^{n} M_j^{\text{infl}}(i,k)\right)$
  end for
  $M_{\text{j}} = M_j^{\text{infl}} \cdot D_j^{-1}$
end for

Output: M_{j} for row clustering

2.2. *attribute inverse regularized Markov Clustering (airMCL)*

The irMCL algorithm is based solely on network connectivity. We propose a natural extension for clustering of networks that contain nodes with heterogenous attributes. In this setting, we use the term *attribute* to loosely to define features of the nodes. In the biological context, this could include, for example, a measurement of a phenotype, gene expression, or demographic information. The term *heterogenous* is used to describe the set of attributes defined on the network, which can be continuous or categorical. We call this method **a**ttribute **i**nverse **r**egularized **M**arkov **CL**ustering (airMCL), because it connects the inverse regularized Markov Clustering (irMCL) approach with statistical classification methods, for the purpose of community detection in attributed networks.

The link between irMCL and is achieved through use of multiple logistic regression, in which the attribute information is regressed on the vectorized structure of the network.[22] This approach gives rise to probabilistic estimate of association between network structure and attributes directly, which is embedded into the *weights* for edges in the spread matrix for Algorithm 1. Specifically, airMCL relies on vectorized versions of distance matrices, which reflect the similarity (or lack thereof) between individuals for an attribute or set of attributes. The distance matrix, $D \in R^{n \times n}$ is symmetric, and the entries $d(i,j) = d(j,i)$ convey the similarity between nodes i and j for a given set of attributes. Consequently, vectorizing the strict upper triangular portion (not including the diagonal) of these matrices maps the pairwise information between nodes and attributes into a vectorized space. This set of vectors forms the set of predictors for the logistic regression modeling.

More formally, let Z_k be the vectorized strict upper triangular regions D_k, in the same way as the vectorization of A_{adj}. The logistic model is defined as:

$$\log\left(\frac{Pr(Y=1|Z)}{1-Pr(Y=1|Z)}\right) = \beta_0 + \sum_{k=1}^{p} \beta_k Z_k, \tag{1}$$

where β_0 is an intercept term, and $\beta_1 \dots \beta_p$ are the regression coefficients for the vectorized attributes. The left hand side of Equation 1 is the log-odds ratio. We can directly estimate the odds ratio using the estimated coefficients $\hat{\beta}$ for each pairwise-relationship: $w = \exp\left(\sum_{k=1}^{p} \hat{\beta}_k Z_k\right)$, which is embedded into the *weights* for edges in the spread matrix for Algorithm 1.

Implementations rMCL and airMCL are performed in the R programming language (https://www.r-project.org/). A library `airMCL` that implements these algorithms will be made available in the CRAN repository upon publication.

2.3. *Simulations*

We examine the performance irMCL and airMCL using a variety of network simulations following the general framework proposed by Girvan and Newman.[9] In our simulations, we consider networks containing 128 nodes that are divided into four communities of 32 nodes each. Vertices are connected independently and randomly with a probability P_{in} for those within the same community, and P_{out} for vertices in different communities ($P_{out} < P_{in}$). The probabilities are selected such that the average degree of a vertex is 16. The expected number of links to a vertex in a different community is defined as z_{out}, while the expected number of links to a vertex in the same community is defined as z_{in}. Note that the community structure is less defined (weak) when z_{out} is larger.

Within simulations of different connectivity patterns, we examined single continuous and categorical attributes, as well as their combination. Categorical attributes in the ith group were generated from a multinomial distribution:

$$p(X = x) = \begin{cases} p, & x = i \\ \dfrac{1-p}{3}, & x \in \{1,2,3,4\}/i \end{cases}$$

The values of p were set to 0.9, 0.6, 0.3 to mimic strong, moderate, and weak associations to the network structure, respectively. Note that when p takes large value (0.9), the attribute X is highly homogeneous within communities. When p is small, however, it implies X has high variability within each group, and will be less informative for the purpose of community detection.

A normal distribution, $N(\mu_i, 1)$, was used for continuous attributes of group i. The difference between means of consecutive groups $\Delta\mu = \mu_{i+1} - \mu_i$ was set at 4, 2, or 0.5, to convey strong, moderate, and weak levels of association, respectively, between structural and attribute information. Within the simulation framework, we also set out to determine how sensitive our methods are to noise in network in the form of missing links. For each scenario, we performed community detection on the full network, and networks with up to 30% of their links missing at random. We compared our methods, airMCL and irMCL , with rMCL and a fast-greedy method.[11] We also examined an irMCL-adhoc method, which can be only applied to networks with single categorical attribute. In this setting, irMCL-adhoc assigns a fixed weight of 0.5 when the two nodes have different attribute values, regardless of the structural relevance.

Mixed attributes were also explored for different combinations of continuous and categorical levels of association. The mixed attribute simulations described previously were also

carried out to explore performance for networks varying from well defined communities (small z_{out}) to poorly defined communities (large z_{out}). The clustering by attribute information alone is also performed. For continuous attributes, Euclidean distance and hierarchical clustering with complete linkage is used. For categorical attribute, the attribute value is directly used as cluster label. For combination of two heterogeneous attributes, the larger average performance between continuous and categorical is used, because they cannot be combined for clustering.

Performance is assessed using the Adjusted Rand Index (ARI) as a measure of agreement between two data clusterings.[23,24] Let S be a set of n elements and consider two partitions of S to compare, $X = \{X_1, \ldots, X_r\} \in S$ and $Y = \{Y_1, \ldots, Y_s\} \in S$. The ARI assumes the generalized hypergeometric distribution as the model of randomness, where the two partitions are picked at random such that the number of classes and clusters are fixed.[24] Specifically, letting n_{ij} denote the number of objects in common between X_i and Y_j and $a_i = \sum_j n_{ij}$, and $b_j = \sum_i n_{ij}$, the ARI is defined as:[24]

$$\text{ARI} = \frac{\sum_{ij}\binom{n_{ij}}{2} - [\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}]/\binom{n}{2}}{\frac{1}{2}[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}] - [\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}]/\binom{n}{2}}.$$

For each parameter setting, 100 simulated networks are tested and the standard error is calculated.

2.4. *Application to functional genomics*

We applied the airMCL method to a breast cancer microarray dataset by Van Der Vijver *et al*.[25] The data was obtained from the package `seventyGenesData` available in Bioconductor (https://www.bioconductor.org/). Our objective was to infer communities using airMCL and identify those which relate to survival. Briefly, the data consists of 295 tumor samples from a 295 women with breast cancer. Survival data was also made available for all each patient in this population. The duration for survival analysis in this study is Time To Metastasis (TTM). In this study, 101 metastasis events occurred and 194 censored data points.

The *input* to airMCL requires specification of an adjacency matrix for a corresponding network and a set of attributes. In our application, we define the network using the `KEGG` database.[26] The 24,496 transcripts in the dataset were mapped to KEGG pathways using Entrez gene identifiers with the Bioconductor annotation package `KEGG.db`. In order to obtain a 1: 1 mapping, when several transcripts mapped to a gene, the one with the most variation across the sample was retained for the modeling. After mapping, the data set consisted of 295 samples and 4,715 genes that represent nodes in the network. Transcript abundance was represented by the log10 of the ratio between each sample and the reference RNA.[25] The adjacency matrix (input) was determined through an pathway-based gene network that was formed by placing links between genes when they are present in the same KEGG pathway. The functional network consists of 4,715 nodes (genes) and 883,557 edges.

Node attributes for the airMCL are defined through a measure of dissimilarity of the gene expression data. Several dissimilarity options are feasible and we expand on this point in the discussion. The dissimilarity measure is defined as $d_{i,j} = 1 - |r_{i,j}|$, where $r_{i,j}$ is the Pearson correlation coefficient between the ith and jth genes. Logistic regression models are fit using the vectorized pairwise dissimilarity on edges (1 linked, 0 for unlinked pairs) as the predictor,

and the vectorized adjacency matrix as the response variable. However, the gene network has $4,715$ nodes, implying more than 11 million observations in the regression. Moreover, the sparsity of the network gives rise a severe class imbalance. To alleviate the computational complexity and address imbalance, we randomly selected the unlinked node pairs so as to have the same number as that of the edges.

Survival analysis is performed on TTM using a Cox proportional hazard model.[27] Benjamini and Hochberg method was used to control the false discovery rate.[28] A threshold of P-value< 0.05 was used to identify modules whose overall expression levels are significantly associated with the time to metastasis. Kaplan-Meier estimates were calculated for each significant module based on stratification of the 295 patients into two groups, using the median overall expression levels of the module. Specifically, $w_{kl} = \frac{1}{m_l} \sum_{i \in c_l}^{m_l} z_{ik}$, where w_{kl} is the average expression level of lth module for kth patient, c_l is the set of node index of lth module, and m_l is the number of nodes in this module.

3. Results

Each simulation was run to convergence. Some general trends persisted for the different parameter and attribute simulations (Figure 1). The overall performance of rMCL was poor, but relatively stable across missing links and different levels of association between structure and attribute. This was the case for categorical, continuous, and mixed attribute settings. When the attribute associations are moderate and weak, fast-greedy shows advantages over the other methods when the missing links is larger (Figure 1B-C,E-F).

When a categorical attribute is highly relevant to true groups ($p = 0.9$), the inclusion of attribute information significantly improved the performance (Figure 1A). In this case, the airMCL and post-hoc weighting were both useful in boosting performance. The performance for post-hoc weighting degrades as the attribute association weakens (Figures 1B-C). For continuous attributes, the airMCL is superior for strong associations across all levels of missing links (Figure 1D), and is the top-performer for moderate association with fewer missing links (Figure 1E). When the associations are weak for continuous attributes, airMCL is competitive with irMCL for scenarios with few missing links (Figure 1F). In simulations with multiple heterogeneous attributes (Figure 2G-I), the airMCL successfully extracts the structurally relevant information and improves the performance over clustering using structural information only (irMCL).

Tuning the parameter z_{out} in the simulations enables us to test the performance of our approaches in scenarios where the communities are not well defined. The performance of irMCL is comparable to fast greedy algorithm, and actually slightly outperforms fast-greedy under z_{out} ranges from 1 to 6 (Figure 2A-C). In our simulations, large z_{out} represents networks in which there is poor community structure. The airMCL's use of attributes offsets this poor structure and is the top-performing method in these extreme scenarios.

We applied the airMCL method to a breast cancer dataset using a KEGG pathway-based network and gene expression attributes.[25] A correlation-based similarity was utilized for the attributes, and the estimated coefficient for the logistic regression was -0.7624 and significant. Convergence was observed 15 iterations. The clustering of the rows of the stochastic matrix was

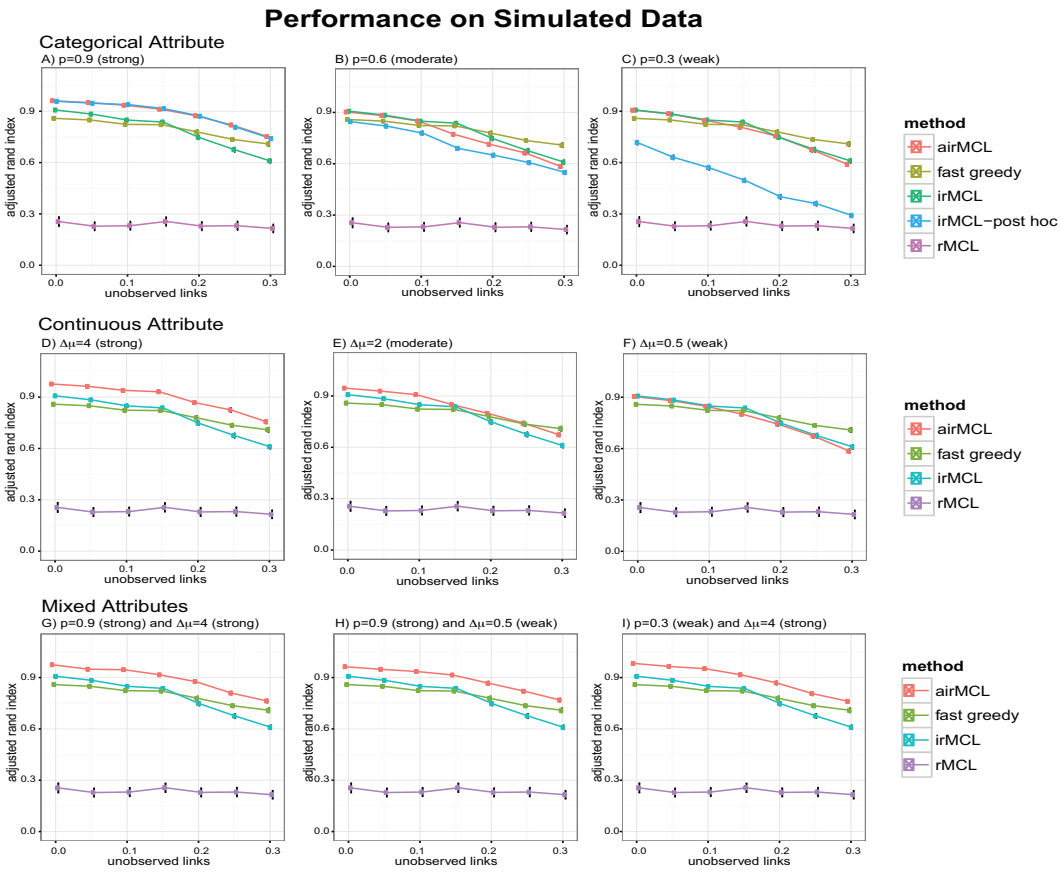

Fig. 1. Simulation results for community detection for a categorical attribute (top row), continuous attribute (second row), and a mixture of a continuous and categorical attributes (third row). Relationships between categorical attributes and community structure were simulated to be **(A)** strong, **(B)** moderate, and **(C)** weak, respectively. Likewise, for continuous attributes **(D-F)**. For the mixed attribute simulation the categorical/continuous relationships between attribute and structure considered were **(G)** strong/strong, **(H)** strong/weak, and **(I)** weak/strong.

determined using the maximum average silhouette, which was 0.85, and yielded 434 clusters. Note that the *rule of thumb* for strong structure is an average silhouette between $0.71 - 1$.[21]

Only modules with size ≥ 8 were selected for survival analysis, and the overall activation status of each module was used for the covariate (see M&M) for predicting TTM. Cox proportional hazard model was used and a multiple testing adjustment was made. A threshold criteria of P-value< 0.05, both methods yields six modules whose overall expression levels are significantly associated with the time to metastasis. Table 1 shows the summary of modules detected and a full listing of module members is available in the Supplement (posted on https://sphhp.buffalo.edu/biostatistics/news-events/workshops/). The adjusted p-values in Table 1 are from Cox regression.

In order to utilize the Kaplan-Meier product limit estimator, for each of the six modules, the 295 patients were split into two groups (low-expression and high-expression) using the median of overall expression levels as cut-off. The survival curves are shown in Figure 3. Log-rank tests were used to test the difference between survival curves of high- and low-expression

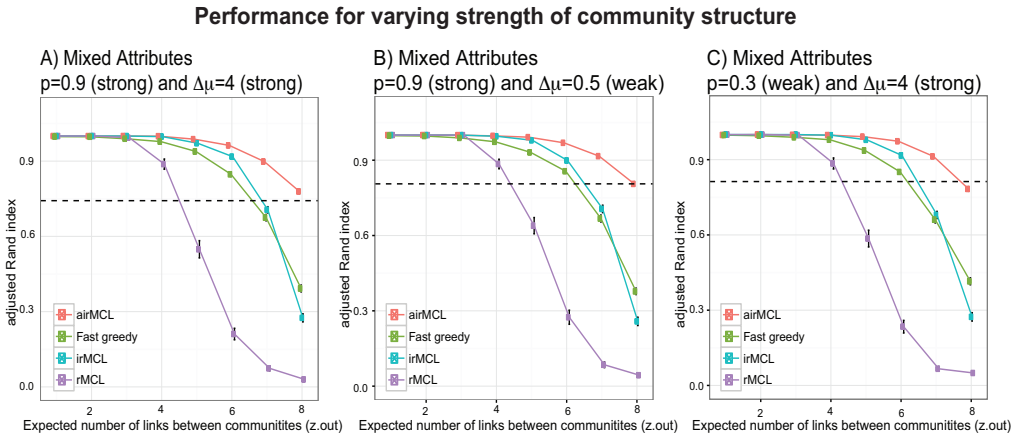

Fig. 2. Comparison of the performance of airMCL/irMCL (with/without attributes) with rMCL and fast greedy method in synthetic networks using adjusted Rand index against z_{out}. The attributes are mixed, which include attributes with **(A)** high-relevance categorical ($p = 0.9$) and high-relevance continuous ($\Delta\mu = 4$), **(B)** high-relevance categorical ($p = 0.9$) and weak-relevance continuous ($\Delta\mu = 0.5$), and **(C)** weak-relevance categorical ($p = 0.3$) and high-relevance continuous ($\Delta\mu = 4$). The horizontal black dashed line indicating the average ARI using attribute information alone.

Table 1: Breast Cancer Module Summarization

Module	Size	Pathways represented	P-value
1	8	Hedgehog signaling pathway (hsa04340)	0.02195
2	27	Pathway in cancers (hsa05200)	0.02195
		MAPK signaling pathway (hsa04010)	
		Adherens junction (hsa04520)	
		Regulation of actin cytoskeleton (hsa04810)	
		Melanoma (hsa05218)	
		Prostate cancer(hsa05215)	
		Oocyte meiosis (hsa04114)	
3	82	Ribosome pathway (hsa03010)	0.02195
4	25	Cell cycle pathway (hsa04110)	0.02195
		Non-homologous end-joining (hsa03450)	
5	19	Pathway in cancers (hsa05200)	0.03541
		Mismatch repair (hsa03430)	
		Colorectal cancer (hsa05210)	
		Small cell lung cancer (hsa05222)	
		Pancreatic cancer (hsa05212)	
		Thyroid cancer (hsa05216)	
6	35	Proteosome pathway (hsa03050)	0.03614

groups. The unadjusted p-values of log-rank tests are shown in Figure 3.

4. Discussion

The design of airMCL is such that the impact of the attributes on community formation depends on the strength of the association between attributes and network structure. Consequently, those weak associations are naturally dampened. Our approach is similar to spirit to the weighting that is done in neural network via an activation function (usually a sigmoid),

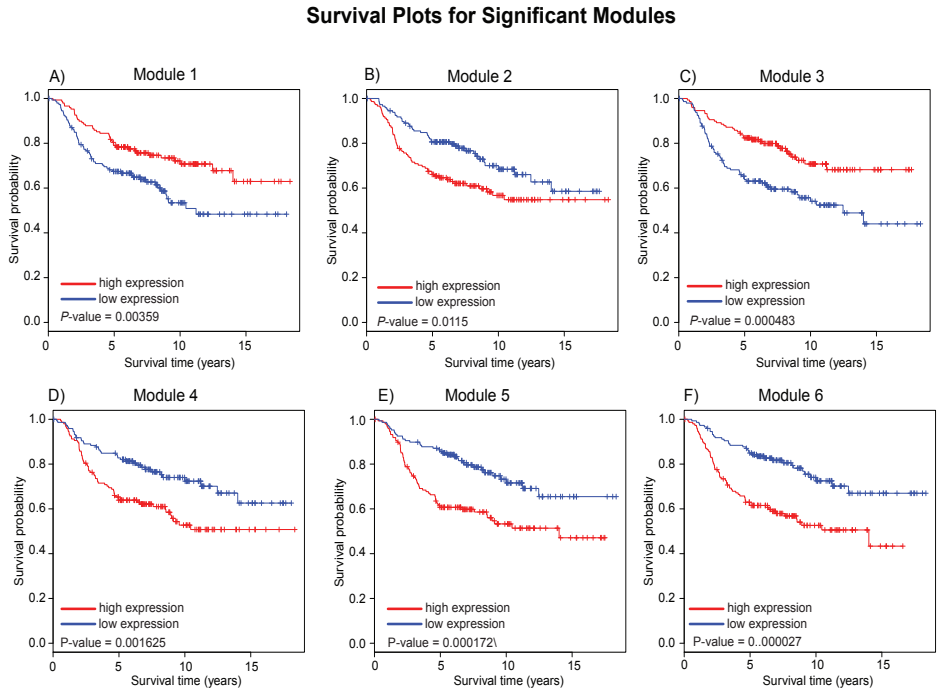

Fig. 3. **(A-F)** Kaplan-Meier survival plots for modules $1-6$. Estimate is based on the partition of the sample into two groups using median values of overall expression for each module (see methods). Red indicates higher expression, blue is for lower expression, and the unadjusted P-values for the log-rank tests are shown.

which weights the features in the input layer. In severely weak settings, the airMCL operates like the irMCL. A challenge attribute information may be irrelevant, or even contradict, the structure of the network. In our simulations, bringing in attribute with weak signals did not derail performance (Figure 1C,F,G-I). This is important as it is not up to the user to specify what attributes are important by weighting, or even eliminating them. In contrast, in the categorical case, we observed with the ad-hoc weighting can derail performance, especially in light of weak attribute associations (Figure 1C).

The fit of the logistic model itself reveals the strength of the relationship between attribute similarity network structure. Examining the regression coefficients (Equation 1) of the model can guide in model development, e.g., choice of similarity, subsets of features. For example, hypothesis testing on the coefficients (e.g., $H_0 : \beta_j = 0$) can reveal the significance of the attribute similarity as a predictor of structure. We have found this useful as a way of selecting a similarity measure for the attributes.

An important feature of the airMCL approach is that the derived inputs for the logistic regression can be handled in a flexible manner. If the set of attributes is heterogenous, one can partition the attributes into multiple subsets, and estimate distance matrices over these subsets independently. This approach enables a unique choice of similarity measure most appropriate for the given attribute or set of attributes. Differences in scales, even within variables of the

same type, can also be managed by subsetting attributes. Collectively, the vectorization of the different distances would give rise to multiple predictors for the logistic regression.

In the breast cancer application, some of the identified pathways are consistent with that reported by Van't Veer *et al.*,[29] such as pathways in cell cycle regulation (Module 4) and signal transduction (Module 2). In addition, we also found that ribosome pathway is associated with breast cancer metastasis. This is consistent with the results reported by Belin *et al.*, that dysregulation of ribosome biogenesis is related to enhanced tumor aggressivity.[30] Activation of hedgehog pathway is also reported in tumors including breast cancers,[31,32] and is related to cancer metastasis.[33] Figure 3 shows that module over-expression (red) is often associated with higher hazards of metastasis. The up-regulation of Module 1 (hedgehog signaling pathway) is unexpectedly associated with better prognosis. This can be explained by the fact that up-regulated genes in this module encode inhibitors in this pathway ($GAS1$, $RAB23$, and $CK1$), which is biologically plausible.

In our simulations, we have simulated balanced communities of moderate size. However, we have also observed good performance, in terms of computational time and accuracy, in the simulation of balanced larger communities. In the case of unbalanced communities, we have achieved good performance in moderate sized simulation networks and real social networks. However, a limitation of our approach is applications to large (1000+ nodes) unbalanced networks. Addressing this form of scalability will be a direction of future research.

We have focussed on a specific application to gene expression cancer data to showcase our method. However, the `airMCL` is generalizable in the sense that it can be used in connection with data that contains a network structure and a set of attributes. The term *attribute* can be loosely defined to encompass demographic information, clinical data, omics data, and combinations of different types of data. The combination of multiple sources of data is known to be a major challenge, and our approach directly integrates them into the community detection. Framing the problem of relating the attributes to the structure via classification has several advantages. Arguably the most important of these advantages is the ability to monitor and quantify loss. Framing the connection between structure and attributes as a supervised learning problem enables the use of statistical classification methods. In this work, we outlined the framework in terms of the classic multiple logistic regression model.[22] However, several classification methods may be more or less suitable depending on the dimension of the graph and attributes, and also the correlation of predictors. Within the classification methods framework are opportunities to utilize the bias-variance tradeoff for model and feature selection. This is a direction of future research, which we anticipate will guide in elimination of extraneous attributes (and potentially nodes), and protect against overfitting.

5. Acknowledgements

HY and RHB were supported through NSF DMS 1312250 and NSF DMS 1557593.

References

1. L. Danon, A. Diaz-Guilera, J. Duch and A. Arenas, *Journal of Statistical Mechanics: Theory and Experiment* **2005**, p. P09008 (2005).

2. M. E. Newman, *The European Physical Journal B-Condensed Matter and Complex Systems* **38**, 321 (2004).
3. S. E. Schaeffer, *Computer Science Review* **1**, 27 (2007).
4. S. Horvath, *Weighted Network Analysis: Applications in Genomics and Systems Biology* (Springer Science & Business Media, 2011).
5. B. W. Kernighan and S. Lin, *Bell system technical journal* **49**, 291 (1970).
6. S. C. Johnson, *Psychometrika* **32**, 241 (1967).
7. M. Fiedler, *Czechoslovak Mathematical Journal* **23**, 298 (1973).
8. W. E. Donath and A. J. Hoffman, *IBM Journal of Research and Development* **17**, 420 (1973).
9. M. Girvan and M. E. Newman, *Proceedings of the National Academy of Sciences* **99**, 7821 (2002).
10. M. E. Newman and M. Girvan, *Physical review E* **69**, p. 026113 (2004).
11. A. Clauset, M. E. Newman and C. Moore, *Physical review E* **70**, p. 066111 (2004).
12. M. E. Newman, *Physical review E* **69**, p. 066133 (2004).
13. M. E. Newman, *Proceedings of the National Academy of Sciences* **103**, 8577 (2006).
14. D. Hanisch, A. Zien, R. Zimmer and T. Lengauer, *Bioinformatics* **18**, S145 (2002).
15. Y. Zhou, H. Cheng and J. X. Yu, *Proceedings of the VLDB Endowment* **2**, 718 (2009).
16. L. Akoglu, H. Tong, B. Meeder and C. Faloutsos, Pics: Parameter-free identification of cohesive subgroups in large attributed graphs., in *SDM*, 2012.
17. F. Moser, R. Colak, A. Rafiey and M. Ester, Mining cohesive patterns from graphs with feature vectors., in *SDM*, 2009.
18. E. Georgii, S. Dietmann, T. Uno, P. Pagel and K. Tsuda, *Bioinformatics* **25**, 933 (2009).
19. S. Van Dongen, *SIAM Journal on Matrix Analysis and Applications* **30**, 121 (2008).
20. V. Satuluri and S. Parthasarathy, Scalable graph clustering using stochastic flows: applications to community discovery, in *Proceedings of the 15th ACM SIGKDD International conference on Knowledge Discovery and Data Mining*, 2009.
21. P. J. Rousseeuw, *Journal of computational and applied mathematics* **20**, 53 (1987).
22. D. W. Hosmer Jr and S. Lemeshow, *Applied logistic regression* (John Wiley & Sons, 2004).
23. W. M. Rand, *Journal of the American Statistical Association* **66**, 846 (1971).
24. L. Hubert and P. Arabie, *Journal of classification* **2**, 193 (1985).
25. M. J. Van De Vijver, Y. D. He, L. J. van't Veer, H. Dai, A. A. Hart, D. W. Voskuil, G. J. Schreiber, J. L. Peterse, C. Roberts, M. J. Marton *et al.*, *New England Journal of Medicine* **347**, 1999 (2002).
26. M. Kanehisa, S. Goto, Y. Sato, M. Kawashima, M. Furumichi and M. Tanabe, *Nucleic acids research* **42**, D199 (2014).
27. D. R. Cox and D. Oakes, *Analysis of survival data* (CRC Press, 1984).
28. Y. Benjamini and Y. Hochberg, *Journal of the Royal Statistical Society. Series B (Methodological)* , 289 (1995).
29. L. J. Van't Veer, H. Dai, M. J. Van De Vijver, Y. D. He, A. A. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen *et al.*, *Nature* **415**, 530 (2002).
30. S. Belin, A. Beghin, E. Solano-Gonzàlez, L. Bezin, S. Brunet-Manquat, J. Textoris, A.-C. Prats, H. C. Mertani, C. Dumontet and J.-J. Diaz, *PloS one* **4**, p. e7147 (2009).
31. M. Kubo, M. Nakamura, A. Tasaki, N. Yamanaka, H. Nakashima, M. Nomura, S. Kuroki and M. Katano, *Cancer research* **64**, 6071 (2004).
32. J. Taipale and P. A. Beachy, *nature* **411**, 349 (2001).
33. J. M. Bailey, P. K. Singh and M. A. Hollingsworth, *Journal of cellular biochemistry* **102**, 829 (2007).