Network Marker Selection for Untargeted LC–MS Metabolomics Data

Qingpo Cai, Emory University
Jessica A. Alvarez, Emory University
Jian Kang, Emory University
Tianwei Yu, Emory University

Journal Title: Journal of Proteome Research
Volume: Volume 16, Number 3
Publisher: American Chemical Society | 2017-03-01, Pages 1261-1269
Type of Work: Article | Post-print: After Peer Review
Publisher DOI: 10.1021/acs.jproteome.6b00861
Permanent URL: https://pid.emory.edu/ark:/25593/s8k4n

Final published version: http://dx.doi.org/10.1021/acs.jproteome.6b00861

Copyright information:
© 2017 American Chemical Society.

Accessed November 30, 2019 7:38 PM EST
Network Marker Selection for Untargeted LC–MS Metabolomics Data

Qingpo Cai†, Jessica A. Alvarez‡, Jian Kang*,¶, and Tianwei Yu*,†
†Department of Biostatistics and Bioinformatics, Emory University, Atlanta, Georgia 30322, United States
‡Department of Medicine, Emory University, Atlanta, Georgia 30322, United States
¶Department of Biostatistics, University of Michigan, Ann Arbor, Michigan 48109, United States

Abstract

Untargeted metabolomics using high-resolution liquid chromatography–mass spectrometry (LC–MS) is becoming one of the major areas of high-throughput biology. Functional analysis, that is, analyzing the data based on metabolic pathways or the genome-scale metabolic network, is critical in feature selection and interpretation of metabolomics data. One of the main challenges in the functional analyses is the lack of the feature identity in the LC–MS data itself. By matching mass-to-charge ratio (m/z) values of the features to theoretical values derived from known metabolites, some features can be matched to one or more known metabolites. When multiple matchings occur, in most cases only one of the matchings can be true. At the same time, some known metabolites are missing in the measurements. Current network/pathway analysis methods ignore the uncertainty in metabolite identification and the missing observations, which could lead to errors in the selection of significant subnetworks/pathways. In this paper, we propose a flexible network feature selection framework that combines metabolomics data with the genome-scale metabolic network. The method adopts a sequential feature screening procedure and machine learning-based criteria to select important subnetworks and identify the optimal feature matching simultaneously. Simulation studies show that the proposed method has a much higher sensitivity than the commonly used maximal matching approach. For demonstration, we apply the method on a cohort of healthy subjects to detect subnetworks associated with the body mass index (BMI). The method identifies several subnetworks that are supported by the current literature, as well as detects some subnetworks with plausible new functional implications. The R code is available at http://web1.sph.emory.edu/users/tyu8/MSS.

Graphical Abstract

*Corresponding Authors: jiankang@umich.edu. Phone: +1 (734)763-1607. tianwei.yu@emory.edu, +1 (404) 727-7671.

ORCID
Qingpo Cai: 0000-0001-7603-4096

Notes
The authors declare no competing financial interest.

Supporting Information
The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jproteome.6b00861.
Selection of the cutoff value to remove highly connected nodes (PDF)
Plots for top 30 ego-networks in our application study (PDF)
Keywords
metabolic network data; optimal matching; feature selection

INTRODUCTION

Metabolomics is the comprehensive analysis of metabolites, that is, low molecular weight components, in a biological system.\textsuperscript{1} In recent years, metabolomics has become one of the major areas of interest in high-throughput biology.\textsuperscript{2,3} There are two general categories of metabolomics: targeted and untargeted. While targeted metabolomics seeks to accurately quantify a limited number of metabolites, untargeted metabolomics seeks to profile the entire metabolome in an unbiased manner.\textsuperscript{4,5} It helps to discover biomarkers, unravel disease etiology, evaluate systematic response to drugs, and detect environmental chemicals in humans.\textsuperscript{2,6}

Untargeted metabolomics is largely made possible by the advances of high-resolution mass spectrometry platforms, which generate highly accurate mass-to-charge ratio (\(m/z\)) measurements, greatly facilitating metabolite identification.\textsuperscript{7} Complex preprocessing routines are necessary to ensure high-quality peak detection, quantification, and alignment across profiles.\textsuperscript{2,8} After alignment, an aligned peak across the LC–MS profiles is referred to as a feature. In the downstream data analysis, a major aspect is functional analysis, that is, finding pathways or subnetworks that are associated with the clinical outcome.\textsuperscript{3}

Unlike other omics technologies, metabolomic profiling by LC–MS does not directly provide a critical piece of information—the molecular identities of the features. A single metabolite can produce one or more ion species, due to the presence of various adduct ions, multiple charge states, and isotopic peaks.\textsuperscript{9} One way to tackle this issue is to first reduce the data by grouping and annotating features derived from the same metabolite.\textsuperscript{10,11} However, this can be difficult for metabolites that exist in low abundance in the biological sample. Hence a common practice in biomarker selection and functional analysis is to match features individually to known metabolites based on mass-to-charge ratio (\(m/z\)).\textsuperscript{12} Retention time can also be used to improve the matching when such information is available on known metabolites, and it can help to determine whether two features are likely to be derived from the same metabolite.
Often a feature can be matched to more than one known metabolite, and a metabolite can be matched to multiple features. This is due to several reasons including (1) some metabolites have the same molecular weight, (2) some features have extremely close m/z values, and (3) various adduct ions and isotopic peaks are possible. At the same time, due to the sensitivity limits of the technology, some metabolites are not detected in the data. Figure 1 displays a subnetwork with potentially matched features, which is derived from real data. Several features have been matched to multiple metabolites, and some metabolites have been matched to multiple features.

So far, metabolic pathway analyses are conducted without addressing the matching uncertainty problem, that is, one feature can be matched to more than one metabolite, with only one of the matchings being true. In addition, analyzing the biological network directly, without dissecting the overall network artificially into pathways, has been shown to be a very promising approach in other areas of omics. Currently for metabolic network analysis, there are few dedicated network analysis methods available. Although methods can be borrowed from the gene expression field, such methods are not designed to take into account the matching uncertainty issue.

To fill the gap, we propose a unified framework for network feature selection from the metabolic network along with optimal matching detection. First, we adopt the ego-network concept for easy delineation of subnetworks from a large-scale biological network; Next, we develop a sequential optimizing procedure which conducts feature selection of subnetworks based on their predictive power of the clinical outcome, and detects the optimal matching between features and metabolites. To the best of our knowledge, we are the first to address the matching uncertainty issue in metabolomic network analysis. Our proposed framework provides a very flexible sequential optimization procedure that can incorporate various machine learning algorithms to identify the most important subnetworks while finding optimal matching, including the Naive Bayes method which is in concept close to the common enrichment-based methods.

The method involves matching of features to adduct ions of metabolites. In actual application the user can choose what adduct ions and isotope peaks should be allowed. There is clearly a trade-off. The more adduct ions and isotope peaks allowed, the more potential matching between features and metabolites. However, at the same time, more false matchings are included in the computation, because there are features derived from pure noise in untargeted metabolomics data. In this study we choose to use a conservative approach, allowing only four common adduct ions and the most abundant isotopes: [M + H]⁺, [M + Na]⁺, [M + K]⁺, and [M + NH4]⁺. We evaluate the performance of our proposed method using simulation studies, and illustrate the proposed framework on a metabolome-wide association study (MWAS) of body mass index (BMI) in a healthy cohort.

**METHOD**

**Problem Setup**

Suppose the data set contains n samples with p features. For i = 1,⋯, n and j = 1,⋯, p, denoted by x_ij observation i for feature j and by y_i an outcome variable, which could be
continuous or categorical. Write \( x_j = (x_{1j}, \ldots, x_{nj})^T \), \( X = (x_1, \ldots, x_p) \) and \( y = (y_1, \ldots, y_n)^T \). A network of \( q \) metabolites is also given. Let \( D = \{ D_{kl} \}_{q \times q} \) denote the distance matrix of the metabolic network, where \( D_{kl} \in \{ 1, 2, \ldots \} \) indicates the distance between metabolite \( k \) and \( l \). The distance here is defined as the shortest path between two nodes in a graph. In the rest of paper we will not distinguish metabolite and node. Given the allowed adduct ions and m/z difference tolerance level, let \( m_k \) denote the number of features that could possibly match to metabolite \( k \) in the network and \( f_k = \{ f_{k1}, \ldots, f_{km_k} \} \) be the collection of those features, where \( f_{kh} \in \{ 1, \ldots, p \} \), \( h \in \{ 1, \ldots, m_k \} \). The uniqueness of this problem lies in the fact that \( f_k \cap f_l \neq \emptyset \), for some \( k \neq l \). This indicates that one feature may be matched to multiple metabolites.

Let \( t_j \) denote the number of possible metabolites that could match with feature \( j \) in the network and \( u_j = \{ u_{j1}, \ldots, u_{jt} \} \) be the collection of those metabolites, where \( u_{jg} \in \{ 1, \ldots, q \} \), for \( g = 1, \ldots, t_j \). In addition to the metabolomics data, suppose \( r \) demographic covariates are collected, denoted by \( z_i = (z_{i1}, \ldots, z_{ir})^T \). Write \( Z = (z_1, \ldots, z_n)^T \). The goal of this paper is to develop a framework to simultaneously select important network markers and identify the optimal matching of features to the metabolites on the network, while adjusting for demographic covariates.

### Metabolic Ego Networks

In this study, we adopt the ego-network approach to delineate the subnetwork structure, which is a well-defined notation in social network studies,\(^{19}\) and previously applied in the genomics setting.\(^{18}\) An ego-network consists of a centroid node, referred as ego-node, and its neighborhood defined as a set of nodes within certain distance to the ego-node over the network. We refer to this distance as the ego-radius. An ego-network can be grown by increasing the corresponding ego-radius and including more nodes. Let \( \mathcal{R} \) be the upper bound of the ego-radius for all the possible ego-networks in the network. We fix \( \mathcal{R} = 2 \) in the following content. Given an ego-radius, we can obtain all the nodes and potentially matched features of the ego network. Furthermore, we can evaluate the performance of the ego network based on a criterion, that is, capability of the matched features to predict the clinical outcome in cross-validation, based on which we can rank all the ego-networks. Our framework is general such that any machine learning or statistical predictive model can be used, as long as they are capable of variable selection.

Table 1 provides a summary of all the notations and their definitions used in the general workflow (Figure 2). Specifically, let \( v^s_k \) be a set of metabolites with distance smaller than or equal to \( s \) to metabolite \( k \). Denote by \( F^s_k \) the matched feature set of metabolites \( v^s_k \) and by \( e^s_k \) the predictive error of ego-network for metabolite \( k \) with ego-radius \( s \), which is calculated using some machine learning algorithms based on outcome \( y \) and covariates \( \{ Z, F^s_k \} \). The left box of Figure 2 is the proposed algorithm to determine the ego-network for each node in the network without considering the multiple matching issue.

### Optimal Matching

Next we address the multiple matching issue for network feature selection. The proposed method uses a sequential feature screening procedure to select important subnetworks and identifies the optimal matching. Let \( E_k \) denote the ego-network for node \( k \), and \( e_k \) is the
predictive error of \( E_k \). In addition, \( F_k \) and \( N_k \) are the set of features and nodes of \( E_k \). \( \tilde{F}_k \) is the set of features selected from \( \{ F_k, Z \} \) of optimal matching, besides demographic variables. Let \( T (T \leq q) \) be the number of top ego-networks that we select and conduct the optimal matching. \( T \) is prespecified and fixed in the algorithm. The algorithm for ego-network selection and optimal matching is described in Figure 2.

Our main idea for developing this algorithm lies in that, for features that can match to multiple metabolites, the true matching more likely corresponds to the one where it yields the lowest predictive error, together with neighboring metabolites. In each iteration, we first select the ego-network with the lowest predictive error and conduct a statistical feature selection procedure within the ego network. On the basis of the feature selection results, we assign each selected feature to the ego-network, and most likely to a specific metabolite when no two metabolites share a molecular weight in the ego network. This can bring changes to the matching between features and metabolites in some other ego-networks. Then we keep the selected ego-network fixed and refit predictive models for all other ego-networks affected by the change of matching. We repeat the procedures for ego-network selection and feature/matching selection until enumerating all ego-networks or a predefined number of iterations is met.

In many applications, it is desirable to consider demographic variables, such as age, gender, ethnicity, etc. Our framework easily accommodates this need by forcing all the demographic variables to be used in the predictive model fitting, regardless of the feature selection results. In rare occasions, features with the same \( m/z \) value but different retention time are matched to the same metabolite. However, only one matching can be true. In this case, the matching with the lowest predictive error is retained.

**SIMULATION**

Simulation studies are conducted to evaluate the performance of our proposed method. Here we directly adopt the KEGG human metabolic network\(^{20}\) as well as a real metabolomics data set. The KEGG network was downloaded and extracted using R packages KEGREST\(^{21}\) and igraph.\(^{22}\) For the real data, demographic covariates are omitted in simulation. In total, 1074 features are matched to 944 unique metabolites in the network, with another 1306 metabolites not matched by any feature. Of all these features, 685 have only one matched metabolite and 389 have been matched to multiple metabolites. For the purpose of simulation, a random matching for these 389 features with multiple matched nodes is set as the true matching.

Given some features have zero readings from some samples, which can be caused by either true nonpresence or nondetection due to low signal strength, we use features with less than 20% zeros to generate the response variable and use features with less than 50% zeros to calculate the predictive error of an ego-network. For each simulation, we randomly choose a subnetwork as the ground truth ego-network and randomly sample more than two features with less than 20% zeros from the selected subnetwork to generate a response variable using a logistic regression model; that is, for \( i = 1, \ldots, n \)
where the response variable \( y_i \in \{0,1\} \). The sample size is 499, and the number of true predictors is between 2 and 11.

We generate 100 data sets in this simulation study. The mean and median number of features in the true ego-network are 3.61 and 3, respectively. In this simulation, we only make changes to the top 20 ego-networks ranked by classification accuracy (\( T = 20 \)). Four methods are considered here for comparisons: logistic regression (LR), naive Bayes classifier (NBC), random forest (RF), and support vector machine (SVM). We calculate the predictive error using a 5-fold cross validation. The Recursive Feature Elimination (RFE) procedure implemented in \( \text{R} \) package “caret” is adopted for feature selection. For comparison, we also conduct the simulation without considering the multiple matching issue—every ego-network uses all the features possibly matched to the ego-network, which we refer to as “maximum matching” in the following discussion. Note that all currently existing methods implicitly use the maximum matching method as they do not consider the multiple matching issue. Also, the naive Bayes (NB) approach paired with the maximum match is in essence similar to the predominant enrichment-based analysis, such as Mummichog etc.

Figure 3a presents a summary of the network marker selection accuracy for our proposed optimal matching method and the maximum matching method. We first compare the sensitivity and specificity of selecting the correct features from all features. Because most features are in the negative class, we use the precision-recall curve to summarize the results from each simulated data set, and compute the area under curve (AUC). We then use boxplots to compare the AUC values of the methods (Figure 3a). It is clear that the optimal matching produced better AUC than maximum matching. Among the four prediction methods, logistic regression performed the best, which is no surprise given that the data is simulated from a logistic regression model. Among the more flexible machine learning methods, Random Forest achieved the best performance.

We then focus on the true ego networks that are used to generate the \( y \) value, and compare the selection accuracy of the features in the true ego network. Figure 3 panels b–e display the precision (true positive divided by all selected) of the features, where the optimal matching approach produces a better precision (red) compared to the maximum matching approach (blue) in most cases.

Figure 3f compares the recovery of the predictive features in the true ego network. Four categories for the selection results are considered here: features in selected ego-network are exactly the same as all the true features (Same); features in selected ego-network contain all the true features and some false positive features (Large); features in selected ego-network are a subset of the true features (Small); features in selected ego-network and true features partially overlap (Mixed). Computational time for different methods are also compared. Computation time is the average CPU time in seconds per simulation across 100
simulations. All the simulations are executed on a desktop computer with 3.40 GHz i7 CPU and 16 GB memory.

In addition to the frequency in all categories, the size ratios are also calculated (Figure 3g) except for the “Same” category. In a “Large” or “Small” selected ego-network, the size ratio is defined as the ratio of the number of selected features over the number of true features. For a “Mixed” selected ego-network, the size ratio is defined as the ratio of the number of the shared features over the number of true features. Of note, a “Large” selected ego-network with a smaller size ratio has a more accurate selection. On the other hand, a “Small” selected ego-network or a “Mixed” selected ego-network with a larger size ratio indicates a better selection result. Overall, the optimal matching approach (Figure 3b, orange bars) produces better size ratios.

APPLICATION

Data Set

We test our network marker selection framework in the Emory-Georgia Tech Predictive Health Initiative Cohort of the Center for Health Discovery and Well Being. This is an ongoing, cohort of generally healthy university employees, ages 18 and older, recruited between January 2008 and February 2013 (http://predictivehealth.emory.edu). All participants are free of any acute illness, uncontrolled or unstable chronic disease, hospitalizations within the year prior to study entry, substance or drug abuse within the past year, or active malignant neoplasm or history of malignancy other than basal cell skin cancer within the previous 5 years. Subjects undergo an extensive medical and metabolic assessment annually. The study is approved by the Emory Institutional Review Board, and all participants provide informed consent prior to any testing. For this study, only subjects with available high-resolution plasma metabolomics data are assessed (N = 371). For the metabolic network, we used the KEGG human metabolic network, and removed all nodes with degrees of 20 or higher. Such highly connected nodes are involved in too many reactions for their concentration level to be informative. In addition, the subnetwork surrounding such a node may be too diverse to carry a clear biological theme. From a network analysis point of view, the presence of such nodes makes the distance between most node pairs very small, making it difficult to select meaningful subnetworks. We conducted a systematic study of network characteristics versus the cutoff value, and determine 20 is a good cutoff value (Supplementary File 1).

Results

We choose to use BMI as the outcome variable to assess our proposed framework because of the vast literature, including metabolomics studies linking BMI, obesity, and adiposity to major metabolic pathways. This would allow us to evaluate the biological plausibility of our models. We compute the classification accuracy based on the four methods: logistic regression (LR), naive Bayes classifier (NBC), random forest (RF), and support vector machine (SVM). All methods result in some degree of biological plausibility with regard to ego-network links to BMI. However, the random forest method, in addition to being one of the best-performing methods in the simulation study (above), provides the most consistent
ego-networks in terms of the resultant ego-nodes and selected metabolites fitting within a specific metabolic pathway or common unifying metabolite. Given the nature of ego-networks, some of the selected ego-networks are partially overlapping, as their ego nodes are neighbors in the KEGG network. Some metabolic pathways are represented by several of the selected ego-networks. We select the top 30 ego-networks generated using the random forest method, excluding those supported by a single feature. They are metabolically connected to several pathways or specific metabolites, all of which have been biologically linked to BMI.

Several of the ego-networks are related to the tricarboxylic acid (TCA) cycle. The TCA cycle is an essential mitochondrial component of the metabolism of carbohydrates, fats, and proteins for the production of energy. Impaired mitochondrial activity has long been implicated in the development in obesity and its metabolic sequelae given the role of the mitochondria in energy expenditure and lipid storage and mobilization. It is, therefore, expected that BMI would be linked to such a key pathway, in addition to several metabolites that function as substrates for the TCA cycle (pyruvate, lactate, alanine, cysteine, glutamate, phenylalanine, and tryptophan). Specific intermediates of the TCA cycle identified as either ego-nodes or metabolites within the ego-networks include citrate, oxalosuccinate, and cis-aconitate (Figure 4a). Obesity has been shown to impair the rate-limiting enzyme of the TCA cycle, citrate synthase, which catalyzes the production of citrate.

Our data are also supported by a recent metabolomics study showing a relationship between BMI and several intermediates of the TCA cycle, including cis-aconitate (Figure 4a), as well as lactate and several amino acids and their intermediates, including alanine (Figure 4b), tryptophan (Figure 4b, nonego node), cysteine (Figure 4c), and phenylalanine (a nonego node among the top 30 ego networks, Supplementary File 1). Interestingly, the tryptophan intermediates within our ego-networks, anthranilate, 3-hydroxyanthranilate, and 3-hydroxy-L-kynurenine are consistent with studies indicating that obesity induces the increase of indoleamine 2,3-dioxygenase through a pro-inflammatory pathway. Additional metabolomics and amino acid studies confirm the relationship between BMI or other indicators of adiposity and the circulating amino acids related to our selected ego-networks.

Intermediates in the metabolism of sulfur-containing amino acids, cysteine and glutathione, are prominent among our BMI-associated ego-networks (Figure 4c). Cysteine has been implicated in the promotion of obesity through various epidemiological and experimental studies. Glutathione, the major intracellular antioxidant, is decreased in circulation in obesity, consistent with the oxidative environment associated with excess adiposity. Hydrogen sulfide is identified as an ego-node in our study (Supplementary File 1). This metabolite has been shown to suppress oxidative stress by promoting the transport of cysteine toward glutathione production, and circulating hydrogen sulfide is inversely associated with obesity.

Prostaglandin H2 is an ego-node that is associated with BMI (Figure 4d). Prostaglandins are eicosanoids derived from arachidonic acid that play important roles in pro-inflammatory responses. Prostaglandins, in various biological sources, correlate positively with BMI and/or obesity. Metabolites within this ego-network included arachidonate and its other
derivative lipid mediators, thromboxane A2 and leukotriene A4, as well as related intermediates. Arachidonic acid, thromboxanes, and leukotrienes also correlate with adiposity. The link between BMI and the arachidonic acid pathway may reflect dietary differences in polyunsaturated fatty acid intakes.

**DISCUSSION AND CONCLUSION**

Functional analysis, including network analysis and pathway analysis, is important for data interpretation and feature selection in metabolomics data. For untargeted metabolomics, the issue of multiple matching has existed for a long time and has been overlooked, which can lead to erroneous results. In this paper, we propose a flexible sequential optimizing procedure that can incorporate various machine learning algorithms to address this multiple matching issue in metabolomics data, along with identifying subnetworks which are highly relevant to the clinical outcome. The method ranks ego networks. The number of top ego-networks to study is a user-defined parameter. In practice, one can also choose the parameter based on predictive accuracy, that is, stopping the program when the prediction accuracy is smaller than a threshold.

Simulation studies show that our method greatly improves the selection accuracy compared with the existing maximum matching approach. Application to a real data set also proves that our method can detect important subnetworks associated with the outcome variable. The same idea can be easily adapted to pathway analysis, where predetermined pathways, rather than ego-networks, are used. We note that the method is based on matching of m/z values to theoretical values of known metabolites, which can only indicate, but not confirm the identities of features. Experimental approaches, such as chemical spike-in and LC–MS/MS, should be used to confirm the identities of features found to be relevant.

Metabolomics data are very complex, which brings many interesting and difficult statistical issues. As we mentioned above, many features have lots of zeros, either because the metabolite is truly nonpresent in the samples, or because the low peaks cannot be differentiated from noise using current technology. Imputing these missing data in features could improve the power for statistical inference in analyzing metabolomics data. This could be seen as a possible extension of our paper. Another possible extension is to develop a systematic Bayesian modeling framework for feature selection over the network while addressing the multiple matching issue.

**Supplementary Material**

Refer to Web version on PubMed Central for supplementary material.

**Acknowledgments**

**Funding**

This work has been partially supported by NIH Grants 1R01MH105561-02, P20HL113451, K01DK102851, and RF1AG051633.
References


Figure 1.
An example subnetwork presented as a bipartite network, with potentially matched LC–MS features linked to the metabolites (dotted lines). The red nodes represent metabolites, the green nodes represent reaction, and blue nodes represent LC–MS data features. Four adduct ions are considered: $[M + H]^+$, $[M + Na]^+$, $[M + K]^+$, and $[M + NH_4]^+$, and the m/z tolerance is 10 ppm.
Figure 2.
General workflow of the method.
Figure 3.
A comparison between the proposed optimal matching and the maximum matching method.
(a) Boxplots of area under the curve (AUC) of feature-level precision-recall (PR) curve. (b–e) Comparison of precision in the true ego network for individual simulated data sets. Red, optimal matching; blue, maximum matching. (f) Feature selection accuracy for the true ego network. Ratios of the four categories are followed by 95% confidence interval. Computing time (last column) is followed by standard deviation. (g) Selection ratios of the methods.
Figure 4.
Some example ego-networks selected by Random Forest. Red dotted line means the match between feature and node is eliminated by our algorithm.
<table>
<thead>
<tr>
<th>notation</th>
<th>definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Z$</td>
<td>demographic covariates</td>
</tr>
<tr>
<td>$\mathcal{G}$</td>
<td>largest ego-radius</td>
</tr>
<tr>
<td>$V^s_k$</td>
<td>nodes with distance $s$ to node $k$</td>
</tr>
<tr>
<td>$F^s_k$</td>
<td>feature set of nodes $V^s_k$</td>
</tr>
<tr>
<td>$\epsilon^s_k$</td>
<td>predictive error for $F^s_k$ and $Z$</td>
</tr>
<tr>
<td>$E_k$</td>
<td>Ego-network for node $k$</td>
</tr>
<tr>
<td>$T$</td>
<td>number of iterations for optimal matching</td>
</tr>
<tr>
<td>$F_k$</td>
<td>feature sets in $E_k$</td>
</tr>
<tr>
<td>$#(F_k)$</td>
<td>number of unique features in $F_k$</td>
</tr>
<tr>
<td>$N_k$</td>
<td>node sets in $E_k$</td>
</tr>
<tr>
<td>$\epsilon_k$</td>
<td>predictive error of $E_k$</td>
</tr>
<tr>
<td>$M$</td>
<td>predefined largest predictive error</td>
</tr>
<tr>
<td>$F^*$</td>
<td>selected feature from ${F_k, Z}$</td>
</tr>
<tr>
<td>$Q$</td>
<td>set for ego node id</td>
</tr>
</tbody>
</table>