

Community Discovery in Social Networks via Heterogeneous Link Association and Fusion

Lei Meng *

Ah-Hwee Tan *

Abstract

Discovering social communities of web users through clustering analysis of heterogeneous link associations has drawn much attention. However, existing approaches typically require the number of clusters a priori, do not address the weighting problem for fusing heterogeneous types of links and have a heavy computational cost. In this paper, we explore the feasibility of a newly proposed heterogeneous data clustering algorithm, called Generalized Heterogeneous Fusion Adaptive Resonance Theory (GHF-ART), for discovering user communities in social networks. Different from existing algorithms, GHF-ART performs real-time matching of patterns and one-pass learning which guarantee its low computational cost. With a vigilance parameter to restrain the intra-cluster similarity, GHF-ART does not need the number of clusters a priori. To achieve a better fusion of multiple types of links, GHF-ART employs a weighting function to incrementally assess the importance of all the feature channels. Extensive experiments have been conducted to analyze the performance of GHF-ART on two heterogeneous social network data sets. The promising results comparing with existing methods demonstrate the effectiveness and efficiency of GHF-ART.

Keywords— User community discovery, Heterogeneous social networks, heterogeneous data clustering, multimodal feature weighting

1 Introduction

Clustering [4] for discovering communities of users in social networks [1] has been an important task for the understanding of collective social behavior [24] and associative mining such as social link prediction and recommendation [2, 3]. However, with the popularity of social websites such as Facebook, users may communicate and interact with each other easily and diversely, such as posting blogs and tagging documents. The availability of those social media data, on one hand, enables the extraction of rich link information among users for further analysis. On the other hand, new challenges have arisen for traditional clustering techniques to perform commu-

nity discovery from heterogeneous social networks, such as the scalability to large social networks, techniques for link representation and methods for fusing heterogeneous types of links.

In the recent years, many works have been done on the clustering of heterogeneous data. Existing methods may be considered in four categories: multi-view clustering approach [5, 12, 13, 17], spectral clustering approach [11, 14, 21, 23], matrix factorization approach [6, 15] and aggregation approach [8, 9]. However, they have several limitations for clustering heterogeneous social network data in practice. Firstly, existing algorithms typically involve iterative optimization which does not scale well to big data sets. Secondly, most of them need the number of clusters a priori, which is hard to decide in practice. Thirdly, most of those algorithms do not consider the weighting problem when fusing multiple types of links. Since different types of links have their own meanings and levels of feature values, equal or empirical weights for them may bias their importance in similarity measure and may not yield satisfactory performance.

In this paper, we explore the feasibility of Generalized Heterogeneous Fusion Adaptive Resonance Theory (GHF-ART) for identifying user groups in the heterogeneous social networks. GHF-ART [10], extended from Fusion ART [16], has been proposed for clustering web multimedia data through the fusion of an arbitrary rich level of heterogeneous data resources such as images, articles and surrounding text. For clustering data patterns of social networks, we develop a set of specific feature representation and learning rules for GHF-ART to handle various heterogeneous types of social links, including relational links, textual links in articles and textual links in short text.

GHF-ART has several key properties different from existing approaches. Firstly, GHF-ART performs online and one-pass learning so that the clustering process can be done in just a single round of pattern presentation. Secondly, GHF-ART does not need the number of clusters a priori. Thirdly, GHF-ART employs a weighting function, termed *Robustness Measure (RM)*, which adaptively tunes the weights for different feature chan-

*School of Computer Engineering, Nanyang Technological University, Singapore, Email: {meng0027, asahtan}@ntu.edu.sg

nels according to their importance in pattern representation in order to achieve a satisfactory level of overall similarity across all the feature channels. Besides, GHF-ART not only globally considers the overall similarity across all the feature channels, but also locally evaluates the similarity obtained from each channel. This helps to handle cases when users share some common interests but behave differently in some other aspects.

We analyze the performance of GHF-ART on two public social network data sets, namely the YouTube data set [8] and the BlogCatalog data set [7], through parameter sensitivity analysis, clustering performance comparison, effectiveness evaluation of *Robustness Measure* and time cost. The experimental results show that GHF-ART outperforms and is much faster than many existing heterogeneous data clustering algorithms.

The remainder of this paper is summarized as follows. Section 2 reviews existing works on the problem of heterogeneous data clustering. Section 3 formulates the problem of community discovery in the heterogeneous social networks. The technical details of GHF-ART are described in Section 4. Section 5 presents the analysis of experimental results. The final section concludes and highlights future work.

2 Related Work

The task of identifying social groups of users via heterogeneous social links is related to the problem of heterogeneous data clustering. Considering different model formulation, existing approaches can be categorized into four categories: 1) **The multi-view clustering approach** [5, 12, 13, 17] considers to use two clustering models for two types of independent features. Subsequently, the learnt parameters of them are further refined by learning from each other iteratively. However, this approach is restricted to two types of links. 2) **The spectral clustering approach** [11, 14, 21, 23] typically models each feature modality as a graph and uses different unified objective function to identify an overall best cut of the graphs, which is typically an embedding vector and needs traditional clustering algorithms to obtain the final results. However, it typically requires heavy computation; 3) **The Matrix factorization approach** [6, 15] factorizes a similarity matrix into two or three matrices by optimizing a unified objective which considers all types of features. The cluster membership of patterns are identified by finding the cluster indicator matrix that contains the projection values of each data pattern to a pre-defined number of clusters. 4) **The aggregation approach** [8, 9] follows the idea of first obtaining the relational vectors [8] or similarities [9] between patterns for each type of features and then integrating them to produce the final results.

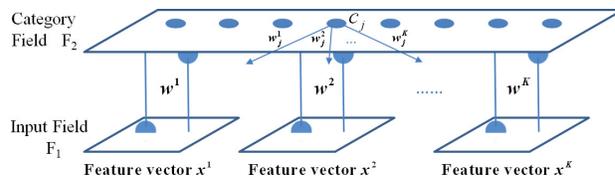


Figure 1: The architecture of GHF-ART for integrating K types of feature vectors.

3 Problem Statement

The community discovery problem in heterogeneous social networks is to identify a set of social user groups by evaluating different types of links between users such that members in a group interact with each other more frequently and share more common interests than those outside the group.

Consider a set of users $\mathcal{U} = \{u_1, \dots, u_N\}$ and their associated multiple types of links $\mathcal{L} = \{l_1, \dots, l_K\}$, such as contact links and subscription links. Each user u_n can be represented by a multi-channel input pattern $\mathcal{I} = \{\mathbf{x}^1, \dots, \mathbf{x}^K\}$, where \mathbf{x}^k is a feature vector extracted from the k -th link.

Consequently, the community discovery task is to identify a set of clusters $\mathcal{C} = \{c_1, \dots, c_J\}$ according to the similarities among the user patterns evaluated within and across different types of links. As a result, given a user $u_N \in c_J$ and two users $u_p \in c_J$ and $u_q \notin c_J$, for $\forall p, q$ such that $u_p, u_q \in \mathcal{U}$, we have $S_{u_N, u_p} > S_{u_N, u_q}$, where S_{u_N, u_p} denotes the overall similarity between u_N and u_p . Namely, users in a cluster may consistently have a higher degree of similarity in terms of all types of links than those belonging to the other clusters.

4 GHF-ART for Clustering Heterogeneous Social Links

GHF-ART [10] is designed for clustering composite patterns which are represented by multiple types of features. As shown in Fig. 1, GHF-ART consists of K independent feature channels in the input field which may handle an arbitrarily rich level of heterogeneous links and a category field consisting of clusters. GHF-ART processes input patterns one at a time during which each of them is either identified as a novel template/exemplar which incurs the generation of a new cluster or categorized into an existing cluster of similar patterns.

In the following subsections, we illustrate the key steps in GHF-ART in terms of representation of commonly used social links, heterogeneous link fusion for pattern similarity measure, learning strategies for cluster template generalization and weighting algorithm for heterogeneous links. The complete algorithm of GHF-ART is shown at the end of this section.

4.1 Heterogeneous Link Representation In GHF-ART, each social user with multi-modal links is represented by a multi-channel input pattern $\mathcal{I} = \{\mathbf{x}^k\}_{k=1}^K$, where \mathbf{x}^k is the feature vector for the k -th feature channel. When presenting to GHF-ART, the input patterns undergo two normalization procedures. Firstly, *min-max normalization* is employed to guarantee that the input values are in the interval $[0, 1]$. Secondly, *complement coding* [18] normalizes the input feature vector by concatenating \mathbf{x}^k with its complement vector $\bar{\mathbf{x}}^k$ such that $\bar{\mathbf{x}}_i^k = 1 - \mathbf{x}_i^k$.

To fit GHF-ART with the social network data, we categorize commonly used social links into three categories and develop the respective representation methods accordingly, as discussed below.

4.1.1 Density-based Features for Relational Links

Relational links, such as contact and co-subscription links, use the number of interactions as the strength of connection between users. Considering a set of users $\mathcal{U} = \{u_1, \dots, u_N\}$, the density-based feature vector of the n -th user u_n is represented by $[f_{n,1}, \dots, f_{n,N}]$, wherein $f_{n,i}$ reflects the density of interactions between the user u_n and the i -th user u_N .

4.1.2 Text-similarity Features for Articles

Text-similarity features are used to represent the articles of users with long paragraphs such as blogs. Considering a set of users $\mathcal{U} = \{u_1, \dots, u_N\}$ and the word list $\mathcal{G} = \{g_1, \dots, g_M\}$ of all the M distinct keywords from their articles, the text-similarity feature vector of the n -th user u_n is represented by $[f_{n,1}, \dots, f_{n,M}]$, where $f_{n,i}$ indicates the importance of keyword g_i to represent the user u_n , which can be computed by term frequency-inverse document frequency (tf-idf).

4.1.3 Tag-similarity Features for Short Text

Tag-similarity features are used to represent short text, such as tags and comments. The key difference of short text from article is that short text consists of few but meaningful words. Given a set of user $\mathcal{U} = \{u_1, \dots, u_N\}$ and the corresponding word list $\mathcal{G} = \{g_1, \dots, g_H\}$ of all the H distinct words, the tag-similarity feature vector of the n -th user u_n is expressed by $[f_{n,1}, \dots, f_{n,H}]$. Following the representation method for meta-information in [10], given that \mathcal{G}_n is the word list of u_n , $f_{n,i}$ ($i = 1, \dots, H$) is given by

$$(4.1) \quad f_{n,i} = \begin{cases} 1, & \text{if } g_i \in \mathcal{G}_n \\ 0, & \text{otherwise} \end{cases}.$$

4.2 Heterogeneous Link Fusion for Pattern Similarity Measure

GHF-ART performs the selection of best-matching cluster to the input pattern and evaluates the fitness between them through a two-way

similarity measures: a bottom-up measure to select a winner cluster by globally considering the overall similarity across all the feature channels; and a top-down measure to locally evaluate if the similarity for each feature channel meets the vigilance criteria.

4.2.1 Bottom-Up Similarity Measure for Category Choice In the first step, a choice function is employed to evaluate the overall similarity between the input pattern and the template weight of each cluster in the category field, which is defined by

$$(4.2) \quad T(c_j, \mathcal{I}) = \sum_{k=1}^K \gamma^k \frac{|\mathbf{x}^k \wedge \mathbf{w}_j^k|}{\alpha + |\mathbf{w}_j^k|},$$

where \mathbf{w}_j^k denotes the weight vector for the k -th feature channel of the j -th cluster, contribution parameter $\gamma^k \in [0, 1]$ is the weight for the k -th feature channel, choice parameter $\alpha \approx 0$ is a positive real value to balance the denominator, the operation \wedge is defined by $(\mathbf{p} \wedge \mathbf{q})_i \equiv \min(\mathbf{p}_i, \mathbf{q}_i)$, and $|\cdot|$ is the ℓ_1 norm. The choice function evaluates the proportion of intersection between the feature vectors of the input pattern and the prototypes of the winner across all the feature channels so that the winner cluster with the best matching feature distribution in the category field is identified.

4.2.2 Top-Down Similarity Measure for Template Matching

After identifying the winner cluster c_{j^*} , a match function is used to evaluate if the selected winner matches the input pattern in each feature channel. For the k -th feature channel, the match function is defined by

$$(4.3) \quad M(c_{j^*}, \mathbf{x}^k) = \frac{|\mathbf{x}^k \wedge \mathbf{w}_{j^*}^k|}{|\mathbf{x}^k|}.$$

If the match function value for each of the K feature channels satisfies the respective vigilance criterion defined by $M(c_{j^*}, \mathbf{x}^k) > \rho$ for $k = 1, \dots, K$, where $\rho \in [0, 1]$ is the vigilance parameter, a resonance occurs so that the input pattern is categorized into the winner cluster. Otherwise, a reset occurs to select a new winner from the rest of the clusters in the category field.

4.3 Learning from Heterogeneous Links

4.3.1 Learning from Density-based and Text-similarity Features

The density-based features and textual features for articles use a distribution to represent the characteristics of a user. Therefore, GHF-ART should be able to learn the generalized distribution of similar patterns in the same cluster so that the users with similar feature distribution can be identified.

To this end, we use the learning function of Fuzzy ART [18]. Assuming the k -th feature channel is for

density-based features, the corresponding learning function of the winner cluster c_{j^*} is therefore defined by

$$(4.4) \quad \hat{\mathbf{w}}_{j^*}^k = \beta(\mathbf{x}^k \wedge \mathbf{w}_{j^*}^k) + (1 - \beta)\mathbf{w}_{j^*}^k,$$

where $\beta \in [0, 1]$ is the learning parameter. We observe that the updated weight values will not be larger than the old ones so that this learning function may incrementally identify the key features by preserving the key features which have stably high values while depressing the features which are unstable in values.

4.3.2 Learning from Tag-similarity Features

We use the learning function for meta-information in GHF-ART [10] to model the cluster prototypes for tag-similarity features. Assuming the k -th feature channel is for tag-similarity features of short text, given the k -th feature vector $\mathbf{x}^k = [x_1^k, \dots, x_H^k]$ of the input pattern \mathcal{I} , the winner cluster c_{j^*} with L users and the corresponding weight vector $\mathbf{w}_{j^*}^k = [w_{j^*,1}^k, \dots, w_{j^*,H}^k]$ of c_{j^*} for the k -th feature channel, the learning function for $w_{j^*,h}^k$ is defined by

$$(4.5) \quad \hat{w}_{j^*,h}^k = \begin{cases} \eta w_{j^*,h}^k & \text{if } x_h^k = 0 \\ \eta(w_{j^*,h}^k + \frac{1}{L}) & \text{otherwise} \end{cases},$$

where $\eta = \frac{L}{L+1}$. (4.5) models the cluster prototype for the tag-similarity features by the probabilistic distribution of tag occurrences. Thus, the similarity between tag-similarity features can be considered as the number of common words. During each round of learning, the keywords with high frequency to occur in the cluster are given high weights while those of the noisy words are incrementally decreased.

4.4 Adaptive Weighting of Heterogeneous Links GHF-ART employs the *Robustness Measure* (R - M) to adaptively tune γ for different feature channels, which evaluates the importance of different feature channels by considering the intra-cluster scatters.

Considering a cluster c_j with L users, each of which is denoted by $\mathcal{I}_l = \{\mathbf{x}_l^1, \dots, \mathbf{x}_l^K\}$ for $l = 1, \dots, L$, and the corresponding weight vectors for the K feature channels denoted by $\mathcal{W}_j = \{\mathbf{w}_j^1, \dots, \mathbf{w}_j^K\}$, the *Difference* for the k -th feature channel of c_j is defined by

$$(4.6) \quad D_j^k = \frac{\frac{1}{L} \sum_l |\mathbf{w}_j^k - \mathbf{x}_l^k|}{|\mathbf{w}_j^k|}.$$

Considering all the clusters, the *Robustness* of the k -th feature channel can be measured by

$$(4.7) \quad R^k = \exp(-\frac{1}{J} \sum_j D_j^k).$$

As the weights for the respective feature channels, the contribution parameter for the k -th feature channel γ^k is defined by

Algorithm 1 GHF-ART

Input: Input patterns $\mathcal{I}_n = \{\mathbf{x}^k |_{k=1}^K\}$, α , β and ρ .

- 1: Present $\mathcal{I}_1 = \{\mathbf{x}^k |_{k=1}^K\}$ to the input field.
- 2: Set $J = 1$. Create a node c_J such that $\mathbf{w}_J^k = \mathbf{x}^k$ for $k = 1, \dots, K$.
- 3: set $n = 2$.
- 4: **repeat**
- 5: Present \mathcal{I}_n to the input field.
- 6: For $\forall c_j$ ($j = 1, \dots, J$), calculate the choice function $T(c_j, \mathcal{I}_n)$ according to (4.2).
- 7: Identify the winner cluster c_{j^*} so that $j^* = \arg \max_{j: c_j \in F_2} T(c_j, \mathcal{I}_n)$. If $T(c_{j^*}, \mathcal{I}_n) = 0$, go to 10.
- 8: Calculate the match function $M(c_{j^*}, \mathbf{x}^k)$ for $k = 1, \dots, K$ according to (4.3).
- 9: If $\exists k$ such that $M(c_{j^*}, \mathbf{x}^k) < \rho^k$, set $T(c_{j^*}, \mathcal{I}_n) = 0$, $j^* = 0$, go to 7.
- 10: If $j^* \neq 0$, Update $\mathbf{w}_{j^*}^k$ for $k = 1, \dots, K$ according to (4.4) and (4.5) respectively and update γ according to (4.7)-(4.10).
- 11: If $j^* = 0$, set $J = J + 1$, create a new node c_J such that $\mathbf{w}_{J+1}^k = \mathbf{x}^k$ for $k = 1, \dots, K$, update γ according to (4.11)
- 12: $n = n + 1$.
- 13: **until** All the input patterns are presented.

Output: Cluster Assignment Array $\{A_n |_{n=1}^N\}$.

$$(4.8) \quad \gamma^k = \frac{R^k}{\sum_{k=1}^K R^k}.$$

The respective incremental update equations for the contribution parameters are further derived for the following two cases:

- **Resonance in existing cluster:** Assume that the input pattern $\mathcal{I}_{L+1} = \{\mathbf{x}_{L+1}^1, \dots, \mathbf{x}_{L+1}^K\}$ is assigned to an existing cluster c_j . For the k -th feature channel, the corresponding update equations for the density-based and text-similarity features and tag-similarity features are defined by (4.9) and (4.10) respectively:

$$(4.9) \quad \hat{D}_j^k = \frac{\eta}{|\hat{\mathbf{w}}_j^k|} (|\mathbf{w}_j^k| D_j^k + |\mathbf{w}_j^k - \hat{\mathbf{w}}_j^k| + \frac{1}{L} |\hat{\mathbf{w}}_j^k - \mathbf{x}_{L+1}^k|)$$

$$(4.10) \quad \hat{D}_j^k = \frac{\eta}{|\hat{\mathbf{w}}_j^k|} (\eta D_j^k + |\hat{\mathbf{w}}_j^k - \eta \mathbf{w}_j^k| + \frac{1}{L} |\hat{\mathbf{w}}_j^k - \mathbf{x}_{L+1}^k|).$$

After the update for all feature channels, the updated contribution parameter can then be obtained by calculating (4.7)-(4.8).

- **Generation of new cluster:** When generating a new cluster, the differences of other clusters re-

main unchanged. Therefore, it just introduces a proportionally change of *Difference*. Considering the *robustness* R^k ($k = 1, \dots, K$) for all of the feature channels, the update contribution parameter for the k -th feature channel is derived as:

$$(4.11) \quad \hat{\gamma}^k = \frac{(R^k)^\eta}{\sum_{k=1}^K (R^k)^\eta}.$$

4.5 Time Complexity Comparison The time complexity of GHF-ART with *Robustness Measure* has been demonstrated to be $O(n_i n_c n_f)$ in [10], where n_i is the number of input patterns, n_c is the number of clusters and n_f is the total number of features.

In comparison with existing heterogeneous data clustering algorithms, the time complexity of LMF [6] is $O(t n_i n_c (n_c + n_f))$, PMM [8] is $O(n_i^3 + t n_c n_i n_f)$, SRC [14] is $O(t n_i^3 + n_c n_i n_f)$ and NMF [15] is $O(t n_c n_i n_f)$, where t is the number of iteration. We observe that GHF-ART has a much lower time complexity.

5 Experiments

5.1 YouTube Data Set

5.1.1 Data Description The YouTube data set ¹ is a heterogeneous social network data set, which is originally used to study the community detection problem via heterogeneous interactions of users. This data set contains 15,088 users from YouTube website and involves five types of relational links, including contact network, co-contact network, co-subscription network, co-subscribed network and favorite network.

5.1.2 Evaluation Measure Since there is no ground truth labels of users in this data set, we adopt the following five evaluation measures: 1) *Cross-Dimension Network Validation (CDNV)* [8], which evaluates how well the cluster structure learnt from one or more types of links fits the network of the other type of links. A larger value indicates a better performance; 2) *Average Density (AD)* measures the average probability of two users in the same cluster having connection, defined by $AD = \frac{1}{J} \frac{1}{K} \sum_j \sum_k \frac{2e_j^k}{n_j(n_j-1)}$, where e_j^k is the number of edges of the k -th link in cluster c_j and n_j is the number of patterns in c_j ; 3) *Intra-cluster sum-of-squared error (Intra-SSE)* measures the weighted average of *SSE* within clusters across feature modalities, defined by $Intra-SSE = \sum_j \sum_{\mathbf{x}_i^k \in c_j} \sum_k \frac{n_j}{\sum_j n_j} (\mathbf{x}_i^k - \bar{\mathbf{x}}_j^k)^2$, where \mathbf{x}_i^k is the feature vector of the i -th pattern for the k -th link and $\bar{\mathbf{x}}_j^k$ is the mean value of all the $\mathbf{x}_i^k \in c_j$; 4) *Between-cluster SSE (Between-SSE)* measures the average distance between two cluster centers to evaluate

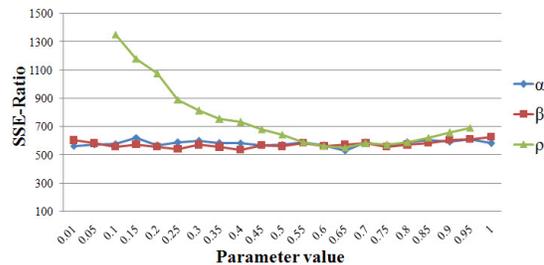


Figure 2: The clustering performance of GHF-ART on the YouTube data set in terms of *SSE-Ratio* by varying the values of α , β and ρ respectively.

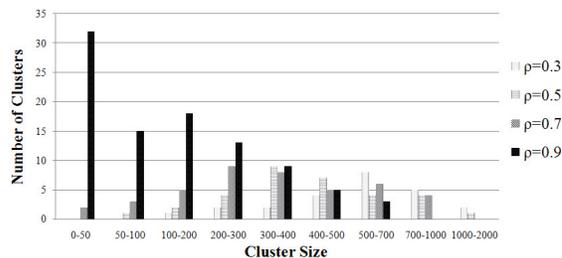


Figure 3: The cluster structures generated by GHF-ART on the Youtube data set in terms of different values of vigilance parameter ρ .

how well-separated the clusters are from each other, defined by $Between-SSE = \sum_j \sum_i \sum_k \frac{1}{J(J-1)} (\bar{\mathbf{x}}_j^k - \bar{\mathbf{x}}_i^k)^2$; and 5) $SSE-Ratio = Intra-SSE / Between-SSE$ gives an overall performance.

5.1.3 Parameter Selection Analysis We initialized $\alpha = 0.01$, $\beta = 0.6$ and $\rho = 0.6$ and studied the change in performance of GHF-ART in terms of *SSE-Ratio* by varying one of them while fixing others, as shown in Fig. 2. We observe that despite some small fluctuations, the performance of GHF-ART is roughly robust to the change in the values of α and β . Regarding the vigilance parameter ρ , we find the performance is improved when ρ increases up to 0.65 and degrades when $\rho > 0.85$. We further analyzed the cluster structures generated under different values ρ , which is shown in Fig. 3. We observe that the increase of ρ leads to the generation of more clusters, which may contribute to the compactness of clusters. At $\rho = 0.9$, a significant number of small clusters are generated, which degrades the performance in terms of recall.

To study the selection of ρ , we analyzed the cluster structure at $\rho = 0.5$ and 0.7 at which the best performance is obtained. We observe that when ρ increases from 0.5 to 0.7, the number of small clusters, which contain less than 100 patterns, increases. Therefore, we assume that when a suitable ρ is reached, the number of small clusters starts to increase. If this idea works, an interesting empirical way to select a reasonable value of ρ is to tune the value of ρ until a small number of small

¹<http://socialcomputing.asu.edu/datasets/YouTube>

Table 1: The clustering performance of GHF-ART, K-means, SRC, LMF, NMF and PMM under the best setting of pre-defined number of clusters (“ k ”) ($\rho = 0.6$ and 0.65 when $k = 35$ and 37 respectively for GHF-ART) in terms of *CDNV*, *Average Density (AD)*, *Intra-SSE*, *Between-SSE* and *SSE-Ratio* on the YouTube data set.

	<i>CDNV</i>		<i>AD</i>		<i>Intra-SSE</i>		<i>Between-SSE</i>		<i>SSE-Ratio</i>	
	value	k	value	k	value	k	value	k	value	k
K-means	0.2446	43	0.0572	40	7372.4	41	9.366	40	774.14	41
SRC	0.2613	37	0.0691	35	6593.6	36	10.249	35	652.34	36
LMF	0.2467	39	0.0584	38	6821.3	41	9.874	37	694.72	40
NMF	0.2741	36	0.0766	35	6249.5	36	10.746	34	591.57	35
PMM	0.2536	36	0.0628	37	6625.8	37	9.627	34	702.25	35
GHF-ART	0.2852	37	0.0834	37	5788.6	37	10.579	35	563.18	37

clusters, less than 10% of the total number of clusters, are identified.

5.1.4 Clustering Performance Comparison We compared the performance of GHF-ART with four existing heterogeneous data clustering algorithms, namely the Spectral Relational Clustering (SRC) [14], Linked Matrix Factorization (LMF) [6], Non-negative Matrix Factorization (NMF) [15] and Principal Modularity Maximization (PMM) [8]. Since SRC and PMM need K-means to obtain the final clusters, we also employed K-means with Euclidean distance as a baseline.

To make a fair comparison, since GHF-ART needs to perform min-max normalization, we applied the normalized data as input to the other algorithms. For GHF-ART, we fixed $\alpha = 0.01$ and $\beta = 0.6$. For K-means, we concatenated the feature vectors of the five types of links. For SRC, we use the same weight values as GHF-ART. The number of iteration for K-means, SRC, LMF, NMF and PMM was set to 50.

We obtained the clustering results of GHF-ART with different values of ρ ranging from 0.3 to 0.9 and those of K-means, SRC, LMF, NMF and PMM with different pre-defined numbers of clusters ranging from 20 to 100. The best performance of each algorithm for each evaluation measure is reported in Table 1. We observe that the best performance of each algorithm is typically achieved with 34 – 41 clusters. GHF-ART usually achieves the best performance with $\rho = 0.65$ which is more consistent than other algorithms. GHF-ART outperforms other algorithms in terms of all the evaluation measures except *between-SSE*, but the result of GHF-ART is still competitive to the best one.

5.1.5 Correlation Analysis of Heterogeneous Networks We first ran GHF-ART under $\alpha = 0.01$, $\beta = 0.6$ and $\rho = 0.65$ and showed the trace of contribution parameters for each type of links during clustering in Fig. 4. We observe that the weights for all types of features begin with 0.2. The initial fluctuation at $n = 1500$ is due to the incremental generation of new

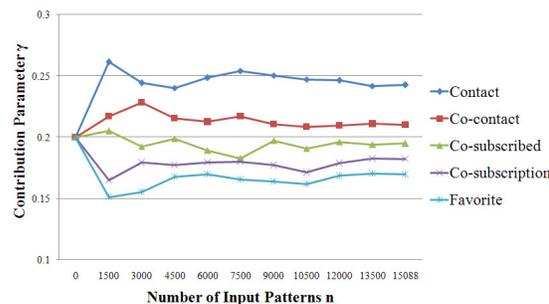


Figure 4: Trace of contribution parameters for five types of links during clustering with the increase in the number of input patterns.

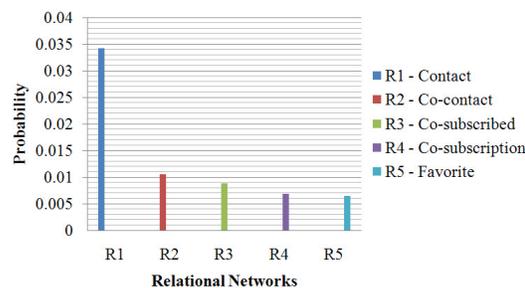


Figure 5: The probability that pairs of patterns falling into the same cluster are connected in each of the five relational networks.

clusters. After $n = 12000$, the weight values of all types of features become stable.

We further analyzed the probability of pairs of connected patterns falling into the same cluster to study how each type of relational networks affects the clustering results, as shown in Fig. 5. We observe that the order of relational networks is consistent with the results shown in Fig. 4. This demonstrates the validity of *Robustness Measure*. Among all types of links, the contact network achieves a much higher probability than other relational networks. This may be due to the fact that the contact network is much sparser than the other four networks. As such, we may expect that the links of contact network are more representative.

Table 2: The clustering performance of GHF-ART, K-means, SRC, LMF, NMF and PMM under the best setting of pre-defined number of clusters (“ k ”) ($\rho = 0.15, 0.2$ and 0.25 when $k = 158, 166$ and 174 respectively for GHF-ART) on the BlogCatalog data set in terms of *Average Precision (AP)*, *Cluster Entropy ($H_{cluster}$)*, *Class Entropy (H_{class})*, *Purity* and *Rand Index (RI)*.

	AP		$H_{cluster}$		H_{class}		$Purity$		RI	
	value	k								
K-means	0.6492	185	0.5892	185	0.5815	165	0.6582	185	0.5662	170
SRC	0.7062	175	0.5163	175	0.4974	160	0.7167	175	0.6481	170
LMF	0.6626	175	0.5492	175	0.5517	155	0.6682	175	0.6038	165
NMF	0.7429	175	0.4836	175	0.4883	155	0.7791	175	0.6759	165
PMM	0.6951	170	0.5247	170	0.5169	165	0.6974	170	0.6103	165
GHF-ART	0.7884	174	0.4695	174	0.4865	158	0.8136	174	0.6867	166

5.2 BlogCatalog Data Set

5.2.1 Data Description The BlogCatalog data set² is crawled in [7] and used for discovering the overlapping social groups of users. It consists of the raw data of 88,784 users, each of which involves the friendship to other users and the published blogs. Each blog of a user is described by several pre-defined categories, user-generated tags and six snippets of blog content.

We extracted three types of links, including a friendship network and two textual similarity networks in terms of blog content and tags. By filtering infrequent words from tags and blogs, we obtained 66,418 users, 6,666 tags and 17,824 words from blogs. As suggested in [7], we used the most frequent category in the blogs of a user as the class label and obtained 147 class labels.

5.2.2 Evaluation Measure With the ground truth labels, we used *Average Precision (AP)*, *Cluster Entropy* and *Class Entropy* [22], *Purity* [19] and *Rand Index* [20] as the clustering evaluation measures. *Average Precision*, *Cluster Entropy* and *Purity* evaluate the intra-cluster compactness. *Class Entropy* evaluates how well the classes are represented by the minimum number of clusters. *Rand Index* considers both cases.

5.2.3 Parameter Selection Analysis We studied the influence of parameters to the performance of GHF-ART on the BlogCatalog data set with the initial setting of $\alpha = 0.01$, $\beta = 0.6$ and $\rho = 0.2$, as shown in Fig. 6. We observe that, consistent with those in Fig. 2, the performance of GHF-ART is robust to the change in the choice and learning parameters. As expected, the performance of GHF-ART varies a lot due to the change in ρ . This curve may also be explained by the same reason for that in Fig. 2.

To validate our findings to select a suitable ρ in section 5.1.3, we analyzed the cluster structures corresponding to the four key points of ρ , as shown

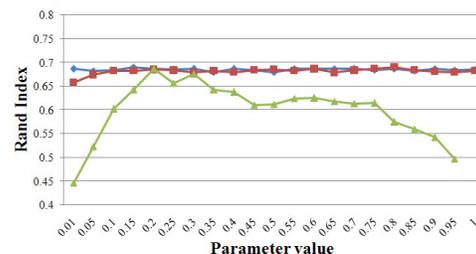


Figure 6: The clustering performance of GHF-ART on the BlogCatalog data set in terms of Rand Index by varying the values of α , β and ρ respectively.

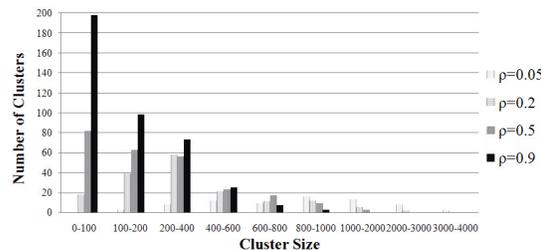


Figure 7: The cluster structures generated by GHF-ART on the BlogCatalog data set in terms of different values of vigilance parameter ρ .

in Fig. 7. We observe that, at $\rho = 0.2$, nearly 20 small clusters with less than 100 patterns are generated. Interestingly, we find that the number of small clusters is also around 10% of the total number of clusters, which fits the findings that we observe on the YouTube data set. This demonstrates the feasibility of the proposed empirical way to select a suitable value of ρ .

5.2.4 Clustering Performance Comparison We compared the performance of GHF-ART with the same set of algorithms compared in the YouTube data set under the same parameter settings as mentioned in section 5.1.4, except the number of clusters. We varied the value of ρ from 0.1 to 0.4 with an interval of 0.05 and the number of clusters from 150-200 with an interval of 5. The best performance for each algorithm with the number of clusters is shown in Table 2. We observe that

²<http://dmml.asu.edu/users/xuwei/datasets.html#Blogcatalog>

Table 3: The five biggest clusters identified by GHF-ART with class labels, top tags, cluster size and *Precision*.

Cluster Rank	Class Label	Top Tags	Cluster Size	<i>Precision</i>
1	Personal	music, life, art, movies, Culture	2692	0.7442
2	Blogging	news, blog, blogging, SEO, Marketing	2064	0.8166
3	Health	health, food, beauty, weight, diet	1428	0.7693
4	Personal	life, love, travel, family, friends	1253	0.6871
5	Entertainment	music, movies, news, celebrity, funny	1165	0.6528

GHF-ART obtained much better performance (at least 4% improvement) than the other algorithms in terms of *Average Precision*, *Cluster Entropy* and *Purity*. This indicates that GHF-ART may well identify similar patterns and produce more compact clusters. Competitive performance is obtained by SRC and NMF in terms of *Class Entropy*. Considering the number of clusters under the best settings, we find that GHF-ART identifies a similar number of clusters to other algorithms, which demonstrates the effectiveness of GHF-ART.

5.2.5 Case Study We further studied the identified communities by GHF-ART. First, we listed the five biggest clusters discovered, as shown in Table 3. We observe that those clusters are well formed to reveal the user communities since more than 1000 patterns are grouped with a reasonable level of precision. We also observe that most of the top tags discovered by the cluster weight values are semantically related to their corresponding classes. Interestingly, the clusters ranked 1 and 4 belong to the class “Personal”. This may be because, according to our organized statistics, “Personal” is much larger than other classes. However, in the top 5 tags, only “life” is shared by them. To have an insight of the relation between these two clusters, we plot the tag clouds for them. As shown in Fig. 8, we observe that the two clusters share many key tags such as “love”, “travel”, “personal” and “film”. Furthermore, when looking into the large number of smaller tags in the clouds, we find that such tags in Fig. 8(a) are more related to “music” and enjoying “life”, such as “game”, “rap” and “sport”, while those in Fig. 8(b) are more related to “family” life, such as “kids”, “parenting” and “wedding”. Therefore, although the shared key tags indicate their strong relations to the same class “Personal”, they are separated into two communities due to the differences in the sub-key tags.

5.2.6 Time Cost Analysis To evaluate the efficiency of GHF-ART on big data, we further analyzed the time cost of GHF-ART, K-means, SRC, LMF, NMF and PMM with the increase in the number of input patterns. To make a fair comparison, we set the number of clusters $k = 166$ for K-means, SRC, LMF, NMF and PMM and set $\rho = 0.2$ for GHF-ART so that the num-

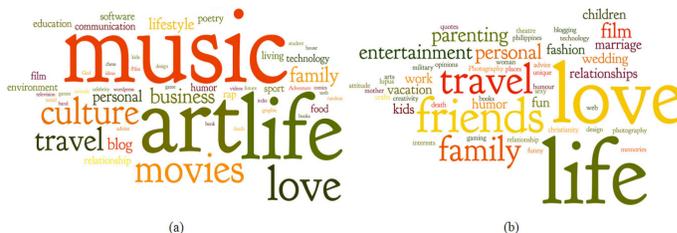
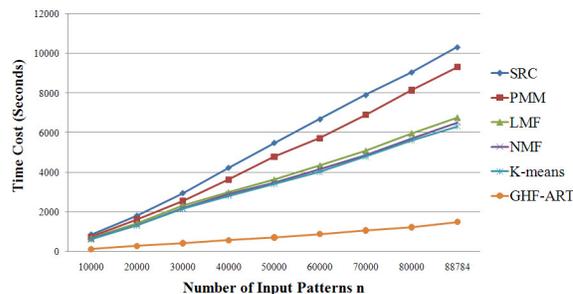
Figure 8: The tag clouds generated for the (a) 1st and (b) 4th biggest clusters. A larger font of tag indicates a higher weight in the cluster.

Figure 9: Time cost of GHF-ART, K-means, SRC, LMF, NMF and PMM on the BlogCatalog Data set with the increase in the number of input patterns.

bers of the generated clusters for all the algorithms are the same. In Fig. 9, we observe that GHF-ART runs much faster than the other algorithms. Whereas the other algorithms incur a great increase of time cost with the increase in the number of input patterns, GHF-ART maintains a relatively small increase. This demonstrates the scalability of GHF-ART to big data.

6 Conclusion

In this paper, we have explored the feasibility of GHF-ART for the community discovery problem in the heterogeneous social networks. Comparing with existing heterogeneous data clustering algorithms [6, 8, 14, 15] for clustering heterogeneous social networks, GHF-ART has several advantages including: 1) Scalability to big data: GHF-ART performs real-time matching of patterns and one-pass learning which guarantee low computational cost; 2) Doing away with the number of clusters a priori: GHF-ART employs a vigilance parameter to restrain the intra-cluster similarity so that clusters may be incrementally identified; 3) Considering hetero-

geneity of links: GHF-ART considers different representation of learning functions for heterogeneous types of links, which is flexible and may produce better representation for heterogeneous links; 4) Incorporating global and local similarity evaluation in pattern similarity measure; and 5) Incorporating a weighting algorithm for heterogeneous link fusion.

We have empirically analyzed the performance of GHF-ART on the YouTube and the BlogCatalog heterogeneous social network data sets, in terms of parameter selection, clustering performance comparison, time cost and two case studies to analyze the effectiveness of *Robustness Measure* and the discovered communities.

Although our work has so far obtained encouraging experimental results, there are several directions for further investigation. Firstly, the length of the feature vectors used to represent relational networks of users in GHF-ART equals to the number of users, which results in a high space complexity. Therefore, feature reduction techniques or hashing methods are preferred to reduce memory consumption. Secondly, visual data such as images and videos are becoming more important in our social life and should also be considered as an important social link between users. Thus, identifying a social network data set with visual links and studying the feasibility of GHF-ART for effective fusion of visual links will be an interesting extension work.

Acknowledgment

This research is supported in part by Interactive and Digital Media Programme Office (IDMPO), National Research Foundation (NRF) hosted at Media Development Authority (MDA) under Grant No.: MDA/IDM/2012/8/8-2 VOL 01.

References

- [1] J. Yang and J. Leskovec, *Defining and Evaluating Network Communities based on Ground-truth*, In ICDM, pp. 745–754, 2012.
- [2] Y. Dong, J. Tang, S. Wu, J. Tian, N. V. Chawla, J. Rao and H. Cao, *Link Prediction and Recommendation across Heterogeneous Social Networks*, In ICDM, pp. 181–190, 2012.
- [3] Y. Yang, N. Chawla, Y. Sun and J. Han, *Predicting Links in Multi-Relational and Heterogeneous Networks*, In ICDM, pp. 755–764, 2012.
- [4] J. J. Whang, X. Sui, Y. Sun and I. S. Dhillon, *Scalable and Memory-Efficient Clustering of Large-Scale Social Networks*, In ICDM, pp. 705–714, 2012.
- [5] I. Drost, S. Bickel and T. Scheffer, *Discovering Communities in Linked Data by Multi-View Clustering*, From Data and Information Analysis to Knowledge Engineering, pp. 342–349, 2006.
- [6] W. Tang, Z. Lu and I. S. Dhillon, *Clustering with Multiple Graphs*, In ICDM, pp. 1016–1021, 2009.
- [7] X. Wang, L. Tang, H. Gao and H. Liu, *Discovering Overlapping Groups in Social Media*, In ICDM, pp. 569–578, 2010.
- [8] L. Tang, X. Wang and H. Liu, *Uncovering Groups via Heterogeneous Interaction Analysis*, In ICDM, pp. 503–512, 2009.
- [9] G. Bisson and C. Grimal, *Co-clustering of Multi-View Datasets: a Parallelizable Approach*, In ICDM, pp. 828–833, 2012.
- [10] L. Meng, A.-H. Tan and D. Xu, *Semi-supervised Heterogeneous Fusion for Multimedia Data Co-clustering*, IEEE Transactions on Knowledge and Data Engineering, 2013.
- [11] D. Zhou and C. J. C. Burges, *Spectral Clustering and Transductive Learning with Multiple Views*, In ICML, pp. 1159–1166, 2007.
- [12] K. Chaudhuri, S. M. Kakade, K. Livescu and K. Sridharan, *Multi-View Clustering via Canonical Correlation Analysis*, In ICML, pp. 129–136, 2009.
- [13] A. Kumar and H. Daume III, *A Co-training Approach for Multi-View Spectral Clustering*, In ICML, pp. 393–400, 2011.
- [14] B. Long, X. Wu, Z. Zhang and P. Yu, *Spectral Clustering for Multi-Type Relational Data*, In ICML, pp. 585–592, 2006.
- [15] Y. Chen, L. Wang and M. Dong, *Non-negative Matrix Factorization for Semisupervised Heterogeneous Data Coclustering*, IEEE Transactions on Knowledge and Data Engineering, vol. 22, no. 10, pp. 1459–1474, 2010.
- [16] A.-H. Tan, G. A. Carpenter and S. Grossberg, *Intelligence through Interaction: Towards a Unified Theory for Learning*, In LNCS, vol. 4491, pp. 1094–1103, 2007.
- [17] S. Bickel and T. Scheffer, *Multi-View Clustering*, In ICDM, pp. 19–26, 2004.
- [18] G. A. Carpenter, S. Grossberg and D. B. Rosen, *Fuzzy ART: Fast Stable Learning and Categorization of Analog Patterns by an Adaptive Resonance System*, Neural Networks, vol. 4, no. 6, pp. 759–771, 1991.
- [19] Y. Zhao and G. Karypis, *Criterion functions for document clustering: experiments and analysis*, Technical Report, 2002.
- [20] R. Xu and D. C. Wunsch II, *BARTMAP: A Viable Structure for Biclustering*, Neural Networks, vol. 24, no. 7, pp. 709–716, 2011.
- [21] M. Rege, M. Dong and J. Hua, *Graph Theoretical Framework for Simultaneously Integrating Visual and Textual Features for Efficient Web Image Clustering*, In WWW, pp. 317–326, 2008.
- [22] J. He, A.-H. Tan, C.-L. Tan and S.-Y. Sung, *On Quantitative Evaluation of Clustering Systems*, In Clustering and Information Retrieval, Kluwer Academic Publishers, pp. 105–133, 2003.
- [23] X. Wang, B. Qian, J. Ye and I. Davidson, *Multi-Objective Multi-View Spectral Clustering via Pareto Optimization*, In SDM, pp. 234–242, 2013.
- [24] K. Zhang, D. Lo, E.-P. Lim, and P. K. Prasetyo, *Mining indirect antagonistic communities from social interactions*, vol. 35, no. 3, pp. 553–583, 2013.