Detection Overlapping Communities in Social System

PINKI
Research Scholar, Department of Computer Science
Ganga Institute of Technology, Kablana
E-mail ID: PINKISOLANKI17@gmail.com

Abstract

Some of the most popular sites in the Web today are social tagging systems or folksonomies
(e.g. Delicious, Flickr, LastFm etc.) where users share resources and collaboratively annotate
resources with tags which help in the search, personalized recommendation and organization of
the resources.
Folksonomies are modelled as tripartite (user-resource-tag) hypergraphs in order to study their
network properties, and detecting communities of similar nodes from such networks is a challenging
and well-studied problem. However, most existing algorithm for community detection in
folksonomies assign unique communities to nodes, whereas in reality, nodes in folksonomies are
associated with multiple overlapping communities – users have multiple topical interests, and
the same resource is often tagged with semantically different tags. In this work, we propose the first algorithm to detect overlapping communities in folksonomies
using the complete hypergraph structure. Our algorithm converts a hypergraph into its corresponding
weighted line-graph, using measures of hyperedge similarity, whereby any community
detection algorithm on unipartite graphs can be used to produce overlapping communities in the
folksonomy.

***Index Terms*—Introduction, Proposed algorithm, experiment on synthetic hypergraph, experiments on real folksonomies, conclusion, References.**

Introduction

A number of the most popular sites in the Web today are online social systems where users form social relationships with one another and generate and share various forms of contents. Among these social systems, some are specifically designed for content sharing. This type of websites are known as Social Tagging Systems. Here, users share contents or resources in these sites, and collaboratively annotate resources with descriptive keywords (known as 'tags') in order to facilitate search and retrieval of interesting resources. Examples of such websites include Delicious.. Thomas Vander Wal coined a term Folksonomy1 to describe social tagging systems. The word 'Folksonomy' is a combination of two words – 'folk' and 'taxonomy'. In such systems, ways for classification and categorization evolve through the practice of collaboratively creating and managing tags. For this reason, folksonomies are also known as Collaborative Tagging Systems. In this work, we use the terms 'Social Tagging System' and 'Folksonomy' interchangeably.

.

1.1 Folksonomy as Hypergraph

Hypergraph model of folksonomies includes user, resource and tag nodes, where an hyperedge
(u, t, r) indicates that user u has assigned tag t to resource r. Figure 1.1 shows a toy example of
tripartite hypergrap A toy example of Tripartite Hypergraph. Three types of nodes aregraphically
represented as Blue Circles, Orange Rectangles and Black Diamonds respectively. Each trianglecreated by connecting these three type of nodes is a hyperedge.Detecting communities from such hypergraphs is a challenging problem – this not only helps inefficient search and recommendation of resources or friends to users, but also in the organizationof the vast amount of resources present in folksonomies into semantic categories.

1.2 Existence of Overlapping Communities

Several algorithms have been proposed for detecting communities in hypergraph But, almost all of the prior approaches do not consider an importantaspect of the problem – they assign a single community to each node, whereas in reality, nodesin folksonomies frequently belong to *multiple overlapping communities*. For instance, users havemultiple topic On the
other hand, an algorithm detecting multiple overlapping communities would place the photo in
both communities related to flowers and the colour 'yellow', and thus raise the chances that this
popular photo is recommended to the above mentioned us

1.3 Identifying Overlapping Communities

To the best of our knowledge, only two studies have addressed the problem of identifying overlapping communities in folksonomies.
1. Wang et al. [10] proposed considering only the user-tag relationships (i.e. the user-tag bipartite projection of the hypergraph),
2. Papadopoulos et al. [5] detected overlapping *tag* communities by taking a projection of the
hypergraph onto the set of tags.

1.4 Link Clustering

Though a node in a network can be associated to multiple semantic topics, a *link* (or edge,
the terms are used interchangeably) is usually associated with only one semantic
instance, a user can have multiple topical interests, but each link created by the user is likely to
be associated with exactly one of his interests.Link clustering algorithms utilize this notion to detect overlapping communities, by clustering *links* instead of the more conventional approach of clustering nodes..

1.5 Organization of the Thesis

Chapter 2 gives a summary about prior works in community detection in graphs as well as in
hypergraphs. Our link-clustering based algorithm is detailed in Chapter 3. We compare the
performance of the proposed algorithm with the existing algorithms by Papadopoulos et al. [5]
and Wang et al. [10]. Extensive experiments on synthetically generated hypergraphs show that
our proposed algorithm out-performs both these algorithms (Chapter 4). Further, using data
from three popular real folksonomies – Delicious

2. PROPOSED ALGORITHM – OHC

In this section, we present the proposed link-clustering algorithm for detecting overlapping communities in tripartite hypergraphs, which we name as 'Overlapping Hypergraph Clustering' algorithm (abbreviated to 'OHC'). As discussed earlier, a folksonomy is modelled as a tripartite hypergraph G = (V,E) where the vertex set V consists of 3 partite sets V_X, V_Y and V_Z. Each hyperedge in hyperedge set E connects a triple of nodes (a, b, c) where a 2 V_X, b 2 V_Y and c 2 V_Z.
For a given hypergraph G, we compute the weighted line graph G′ which is a unipartite graph in which the hyperedges in G are nodes, and two nodes e_1 and e_2 in G′ are
connected by an edge if e_1 and e_2 are adjacent in G (i.e. the two hyperedges have at least one common node in G). The weight of the edge (e_1, e_2) in G′ represents the similarity _ between the two hyperedges e_1 and e_2 in the hypergraph G, which is computed as follows.
Let $N_X(i)$, $N_Y(i)$ and $N_Z(i)$ denote the set of neighbours of node i of type V_X, V_Y and V_Z respectively (if i 2 V_X, then $N_X(i)$ = _ since nodes in the same partite set are not linked). Similarity between two adjacent hyperedges e_1 = (a, b, c) and e_2 = (p, q, r) (where a, p 2 V_X; b, q 2 V_Y; c, r 2 V_Z and assumed a = p) is measured by the relative overlap among the neighbours of the non-common nodes of the same type:

$$\_(e_1, e_2) = \frac{|S \cap S'|}{|S \cup S'| + |N_Y(c) \cap N_Y(r)| + |N_Z(b) \cap N_Z(q)|}$$

where $S = N_X(b) \cap N_X(c)$ and $S' = N_X(q) \cap N_X(r)$. Non-adjacent hyperedges are considered to have zero similarity. It can be noted that the similarity for hyperedges can be computed in various other ways like expressing hyperedges as feature vectors and measuring cosine similarity or Pear-

Initially, the nodes in each partite set are evenly distributed among each community under consideration (e.g. $|V_X|/C$ nodes in set V_X are assigned to each of the C communities). Subsequently, fraction of nodes are selected at random from each of V_X, V_Y and V_Z, and each selected node is assigned to some randomly chosen communities apart from the one it already has been assigned to. Nodes assigned to the same community are then randomly selected, one from each partite set, and interconnected with hyperedges. The number of hyperedges is decided based on the specified density _.
The above assignment of communities to nodes constitutes the 'ground truth'. After a hypergraph is generated, information about the communities is hidden, and then communities are detected from the hypergraph by different community detection algorithms. The community structure detected by each algorithm is compared with the ground truth using the metric Normalized Mutual Information.
Normalized Mutual Information (NMI): NMI is an information-theoretic measure of similarity between two partitioning of a set of elements, which can be used to compare two community structures for the same graph (as identified by different algorithms). The traditional definition of NMI does not consider the case of a node being present in multiple communities; hence we use an alternative definition of NMI considering overlapping communities, as proposed in [8]. The NMI value is in the range [0, 1]; higher the NMI value, the more similar are the two community structures (refer to [8] for details).

3. EXPERIMENTS ON SYNTHETIC HYPERGRAPHS

In this section, we evaluate the performance of our proposed OHC algorithm by comparing with the existing algorithms by Wang et al. [15] and Papadopoulos et al. [13], which are henceforth referred to as 'CL' (abbreviation of 'Correlational Learning') and 'HGC' (as referred by the respective authors) respectively 1.
Since evaluation of clustering is difficult without the knowledge of 'ground truth' regarding the community memberships of nodes, we have used synthetically generated hypergraphs with a known community structure for evaluation of the algorithms. We discuss the generation of synthetic hypergraphs and the metric used to evaluate the algorithms, followed by the results of experiments on synthetic hypergraphs.

3.1 Generation of Synthetic Hypergraphs

Synthetic hypergraphs are generated using a modified version of the method used in [15]. The generator algorithm

takes the following as input: (i) Number of nodes in a partite set (all 3 partite sets V_X, V_Y and V_Z are assumed to contain equal number of nodes), (ii) Number of communities C, (iii) Fraction of nodes which belong to multiple communities and (iv) Hyperedge density _ (i.e. fraction of total number of hyperedges possible in the hypergraph).

1We acknowledge the authors of [13,15] for providing us with the implementations of their algorithms.

214

Initially, the nodes in each partite set are evenly distributed among each community under consideration (e.g. $|V_X|/C$ nodes in set V_X are assigned to each of the C communities). Subsequently, fraction of nodes are selected at random from each of V_X, V_Y and V_Z, and each selected node is assigned to some randomly chosen communities apart from the one it already has been assigned to. Nodes assigned to the same community are then randomly selected, one from each partite set, and interconnected with hyperedges. The number of hyperedges is decided based on the specified density _.

The above assignment of communities to nodes constitutes the 'ground truth'. After a hypergraph is generated, information about the communities is hidden, and then communities are detected from the hypergraph by different community detection algorithms. The community structure detected by each algorithm is compared with the ground truth using the metric Normalized Mutual Information.

Normalized Mutual Information (NMI): NMI is an information-theoretic measure of similarity between two partitioning of a set of elements, which can be used to compare two community structures for the same graph (as identified by different algorithms). The traditional definition of NMI does not consider the case of a node being present in multiple communities; hence we use an alternative definition of NMI considering overlapping communities, as proposed in [8]. The NMI value is in the range [0, 1]; higher the NMI value, the more similar are the two community structures (refer to [8] for details).

3.2 Results of Experiments

The CL and HGC algorithms produce only user and tag communities respectively. Hence, while calculating the NMI value for these algorithms, we have used the community memberships of only the user (respectively, tag) nodes according to the ground truth. On the other hand, the proposed OHC algorithm gives composite communities containing all three types of nodes. Hence, while evaluate OHC, we considered the community memberships of all three types of nodes.

For all the following experiments, $|V_X| = |V_Y| = |V_Z| =$ 200 and number of communities C = 20. For each result, random hypergraphs were generated 50 times using the same set of parameter values and the average performances over all 50 runs are reported.

Performance w.r.t. number of hyperedges: To study how the number of hyperedges affects the performance of the clustering algorithms, we generated synthetic hypergraphs having various hyperedge densities _ = 0.1, 0.2, . . ., 1.0. In each of these hypergraphs, 10% of nodes in each partite set belonged to multiple communities (i.e., = 0.1). The NMI values for the three algorithms are shown in Figure 1(a).

It can be clearly seen that, across all hyperedge densities, OHC performs significantly better than HGC and CL algorithms. A possible explanation for this is that, as stated earlier, the proposed OHC algorithm utilizes the complete tripartite structure of the hypergraph, whereas both CL and HGC algorithms work on unweighted projections which is known to result in significant loss of information [6].

Also note that even for very low hyperedge densities, when detecting community structures is difficult, the Comparison of proposed OHC algorithm with CL and HGC algorithms – variation of NMI values (a) with varying hyperedge density when 10% nodes belong to multiple communities and (b) with varying fraction of nodes in multiple communities, keeping hyperedge density constant at 0.2 proposed OHC algorithm performs very well resulting in NMI scores above 0.8. This makes OHC suitable for real world folksonomies where hyperedge density is typically low.

Performance w.r.t. fraction of nodes in multiple communities: A node belonging to multiple communities creates hyperedges to nodes in all those communities; hence, from the perspective of a particular community, the hyperedges created by this member node to nodes in other communities reduces the exclusivity of this particular community. As the number of nodes in multiple overlapping communities increases, the fraction of inter-community hyperedges increases making the community structure more difficult to identify.

We generated synthetic hypergraphs by varying the fraction of nodes in multiple communities () while keeping hyperedge density (_) constant at 0.2. This low value of hyperedge density was chosen to measure the effectiveness of the algorithms in sparse environment (as in real-world foksnomies). Figure 1(b) shows that OHC performs consistently better than HGC and CL algorithms in this case as well. Further, as the community structure becomes more and more complex, the information loss as a result of projections becomes increasingly more crucial, hence the performance of the HGC and CL algorithms degrade sharply with increase in . On the other hand, the performance of our OHC algorithm shows relatively much greater stability.

The above experiments clearly validate our motivation and show that considering the complete tripartite structure of hypergraphs can result in better identification of community structure, as compared to considering

4.EXPERIMENTS ON REAL FOLKSONOMIES

Now we apply the proposed OHC algorithm to gain insights into the community structures prevalent in real folksonomies. For this, we use the publicly available datasets [3] of the folksonomies Delicious (`www.delicious.com`), LastFm (`www.last.fm`) and MovieLens (`movielens.umn.edu`). The statistics of these data sets are summarized in Table 1.

4.1 Results

For all three datasets, OHC algorithm successfully groups semantically related resources and tags and the users tag-

215

Dataset	users	resources	tags	hyperedges
Delicious	1,867	69,226	53,388	437,593
LastFm	1,892	17,632	11,946	186,479
MovieLens	2,113	10,197	13,222	47,957

Table 1: Statistics of real folksonomy datasets

ging these resources. As an illustration, Table 2 shows the resources and tags placed in some example communities for each of the three datasets. It is evident that the resources and tags that are placed in the same community are often related to a common semantic theme. A closer look at Table 2 reveals that the algorithm also correctly identifies nodes that are related to multiple overlapping communities (themes). For instance, the band 'Van Halen' is placed in two different communities detected from LastFm. The Wikipedia article about 'Van Halen' justifies this placement, stating their

genre as both 'Hard Rock' and 'Heavy Metal'.
There are substantial amounts of overlap detected by OHC algorithm in all three datasets. Figure 2 shows the cumulative distribution of the fraction of communities which overlap with a given number of other communities, for LastFm and MovieLens. A similar pattern was detected in Delicious, which we omit due to lack of space.

4.2 Evaluation of Communities Detected

The principal difficulty in evaluating the communities detected in real folksonomies is the absence of 'ground truth' regarding the community memberships of nodes, since their huge size makes it impossible for human experts to evaluate the quality of identified communities. Hence, we use the following two methods for evaluation.
First, we use the graph-based metric Conductance, which has been shown to correctly conform with the intuitive notion of communities and is extensively used for evaluating quality of communities in OSNs (see [9] for details). As conductance is defined only for unipartite networks, we compare tag communities detected by HGC with the tag nodes in the communities identified by our OHC algorithm.
Second, in case of the folksonomies which allow users to form a social network among themselves, we can assume that users having similar interests are likely to be linked in the social network, or to have a common social neighbourhood (a property known as homophily). We utilize this notion to evaluate the user communities detected by CL algorithm and the user nodes in the communities identified
Figure 3: Distribution of conductance values of tag communities identified by OHC and HGC algorithms, for LastFm (main plot), Delicious and MovieLens (both inset)
Comparison of conductance values: The conductance value ranges from 0 to 1 where a lower value signifies better community structure [9]. Figure 3 shows the cumulative distribution of conductance values of detected tag communities by the two algorithms. Across all three datasets, OHC produces more communities having lower conductance values, which implies that OHC can find communities of better quality than obtained by HGC algorithm. The reason for this superior performance is that OHC groups semantically related nodes into relatively smaller cohesive communities instead of creating a few number of generalized large communities. For examples of semantically related communities, refer to Table 2.
Comparing detected user communities with social network: In case of folksonomies which allow users to form a social network, there can be two types of relationships among users – explicit social connections in the social network, and implicit connections through their tagging behaviour (e.g. tagging the same resource) in the hypergraph. A community detection algorithm for hypergraphs utilizes the implicit relationships to identify the community structure, and we propose to evaluate the detected community structure using the explicit connections that the users themselves create (in the social network). For instance, if a large fraction of the users who are socially linked (or share a common social neighbourhood in the social network) are placed in the same community (by the algorithm), the detected community structure can be said to group together users having common interests.
Hence, to compare the community structure identified by two algorithms, we consider the user-pairs who are within a certain distance from each other in the social network (where distance 1 implies friends, i.e. two users who are directly linked), and compute the fraction of such user-pairs who have been placed in a common community by the algorithm. Figure 4(a) shows the results for the proposed OHC algorithm and the CL algorithm, for the LastFm dataset. Across all distances, OHC places a larger number of user-pairs who share a common social neighborhood, in a common community than the CL algorithm. Also, as the distance between two users in the social network increases, both algorithms put a smaller fraction of such user-pairs in the same community.
We can also investigate the reverse question – among the users who are placed in a common community (by a community detection algorithm), what fraction of these users are
216
Community Theme Examples of member nodes
LastFm Artists Hard Rock Van Halen, Deep Purple, Aerosmith, Alice Cooper, Guns N' Roses, Scorpions, Bon Jovi
(resources) Heavy Metal Van Halen, Deep Purple, Aerosmith, Iron Maiden, Motorhead, Metallica
LastFm Tags Metal blues rock, psychedelic rock, rap metal, nu metal, metal, symphonic metal, doom metal
Rock blues rock, psychedelic rock, rap metal, nu metal, progressive rock, art rock, soft rock
MovieLens Movies Superhero The Incredibles, Shrek, Shrek 2, Incredible Hulk, Batman Begins, Spider-Man, Superman
(resources) Animation The Incredibles, Shrek, Shrek 2, Incredible Hulk, Kung fu Panda, Toy Story
MovieLens Tags Criticism violent, brutal, waste of celluloid, disturbing, junk, tragically stupid, lousy script
Violence violent, brutal, murder, fatality, civil war, great villain, dark, serial killer, war film
Delicious Tags Web 2.0 socialnetworking, socialmedia, php, drupal, xml, webdesign, twitter, skype, ruby
Table 2: Examples of communities detected by OHC algorithm. Nodes related to a common theme are

successfully clustered. Nodes related to multiple themes (italicized) are placed in overlapping communities.

5.Conclusion

In this paper, we proposed the first algorithm to detect overlapping communities considering the full tripartite hypergraph structure of folksonomies. Through extensive experiments on synthetic as well as real folksonomy networks, we showed that the proposed algorithm out-performs existing algorithms that consider projections of hypergaphs. The proposed algorithm can be effectively used in recommending interesting resources and friends to users. Our future work will be to build such a recommendation system utilizing the proposed algorithm.

6.Acknowledgements

The authors wish to thank Melanie Aurnhammer, Andreas Hotho and Gerd
Stumme for very interesting discussions. This research has been partly sup-
ported by the TAGora project funded by the Future and Emerging Tech-
nologies program (IST-FET) of the European Commission under the contract
IST-34721. The information provided is the sole responsibility of the authors
Emergent Community Structure in Social Tagging Systems 13
and does not reect the Commission's opinion. The Commission is not responsible for any use that may be made of data appearing in this publication.

7.References

1. A. Mates, Folksonomies - Cooperative Classi_cation and Communication
Through Shared Metadata, Computer Mediated Communication, LIS590CMC,
Graduate School of Library and Information Science, University of Illinois
Urbana-Champaign (2004).
2. T. Hammond, T. Hannay, B. Lund and J. Scott, Social Bookmarking Tools (I):
A General Review, D-Lib Magazine 11(4), (2005).
3. S. Golder and B. A. Huberman, Usage patterns of collaborative tagging systems,
Journal of Information Science 32, 198 (2006).
4. C. Cattuto, V. Loreto, and L. Pietronero, Semiotic Dynamics and Collaborative
Tagging, Proc. Natl. Acad. Sci. USA 104, 1461 (2007).
5. T. Vander Wal, Explaning and Showing Broad and Narrow Folksonomies,
http://www.personalinfocloud.com/2005/02/explaining and .html,
(2005).
6. A. Hotho, R. J□aschke, C. Schmitz and G. Stumme, Emergent Semantics in
BibSonomy, Proc. Workshop on Applications of Semantic Technologies. In: eds.
C. Hochberger and R. Liskowsky, Informatik f□ur Menschen. Band 2, p94 (2006).
7. G. Salton and M.J. McGill, Introduction to modern information retrieval
(McGraw-Hill, 1983).
8. A. Barrat, M. Barthelemy, R. Pastor-Satorras and A. Vespignani, The archi-
tecture of complex weighted networks, Proc. Natl. Acad. Sci. USA 101, 3747
(2004).
9. A. Capocci, V.D.P. Servedio, G. Caldarelli and F. Colaiori, Physica A 352, 669
(2005).
10. M.E.J. Newman, Phys. Rev. E 74, 036104 (2006).
11. L. Steels, Semiotic Dynamics for Embodied Agents, IEEE Intelligent Systems,
21, 32 (2006).
12. Ciro Cattuto, Christoph Schmitz, Andrea Baldassarri, Vito D. P. Servedio, Vittorio
Loreto, Andreas Hotho, Miranda Grahl and Gerd Stumme, Network Prop-
erties of Folksonomies, AI Communications Journal, Special Issue on "Network
Analysis in Natural Sciences and Engineering", S. Hoche and A. N□urnberger
and J□urgen Flach Eds., IOS PRESS, 20, n.4, 245{262 (2007).
13. Georgia Koutrika, Frans Adjie E_endi, Zolt_an Gy□ongyi, Paul Heymann and
Hector Garcia-Molina, Combating spam in tagging systems, AIRWeb '07: Proceedings
of the 3rd international workshop on Adversarial information retrieval
on the web, p. 57{64, ACM Press (2007).
14. Paul Heymann and Georgia Koutrika and Hector Garcia-Molina, Fighting Spam
on Social Web Sites: A Survey of Approaches and Future Challenges, IEEE
Internet Computing, 11, 35{46 (2007)
15. Ana G. Maguitman, Filippo Menczer, Heather Roinestad and Alessandro
Vespignani, Algorithmic detection of semantic similarity WWW '05: Proceed-
ings of the 14th international conference on World Wide Web 107{116 (2005).
16. Ana Maguitman, Filippo Menczer, Fulya Erdinc, Heather Roinestad and
Alessandro Vespignani,

