Identifying Overlapping Communities in Folksonomies or Tripartite Hypergraphs

Saptarshi Ghosh Pushkar Kane Niloy Ganguly
Department of CSE, Indian Institute of Technology Kharagpur, India
{saptarshi, kane, niloy}@cse.iitkgp.ernet.in

ABSTRACT

Online folksonomies are modeled as tripartite hypergraphs, and detecting communities from such networks is a challenging and well-studied problem. However, almost every existing algorithm known to us for community detection in hypergraphs assign unique communities to nodes, whereas in reality, nodes in folksonomies belong to multiple overlapping communities e.g. users have multiple topical interests, and the same resource is often tagged with semantically different tags. In this paper, we propose an algorithm to detect overlapping communities in folksonomies by customizing a recently proposed edge-clustering algorithm (that is originally for traditional graphs) for use on hypergraphs.

Categories and Subject Descriptors

E.1 [**Data Structures**]: Graphs and Networks; G.2.2 [**Graph Theory**]: Hypergraphs

General Terms

Algorithms, Measurement

Keywords

Folksonomy, tripartite hypergraph, overlapping community

1. INTRODUCTION

Social tagging systems or folksonomies (e.g. Delicious, Flickr) can be modeled as tripartite hypergraphs having user, resource and tag nodes. Detecting communities from hypergraphs is practically important to identify users having similar topical interests as well as similar resources and tags; this helps in classification of resources into semantic categories and recommendation of potential friends and resources of matching interest to users of the folksonomy.

Though several algorithms for community detection in hypergraphs have been proposed (e.g. [2]), one important aspect of the problem that has seldom been considered is that nodes in folksonomies frequently belong to *multiple overlapping communities* (rather than a single community). Most users have multiple topics of interest, and thus link to resources and tags of many different semantic categories. Similarly, the same resource (e.g. photo, web-page) is frequently associated with semantically different tags by users who appreciate different properties of the resource.

The only work known to us on detecting overlapping communities in folksonomies is [3] which considers communities of tags only. However, detecting overlapping communities of users and resources in folksonomies in equally necessary for personalized recommendation and categorization of resources and tags. As a motivating example, consider a popular photo of a daffodil in Flickr. Since many users are likely to tag the photo with 'flower' (or 'daffodil'), as compared to few users using the tag 'yellow', algorithms assigning single communities to nodes would place this photo in the community related to flowers (or daffodils). Community-based recommendation schemes, which recommend resources to users based on common-memberships in communities, would thus overlook the fact that this photo is an excellent candidate for recommendation to a user who favours tagging objects that are yellow-coloured (e.g. photos of yellow cars, sunset, etc). On the other hand, an algorithm detecting multiple overlapping communities would place the photo in both communities related to flowers and the color 'yellow', and thus raise the chances that this popular photo is recommended to the said user.

Out of the few algorithms for detecting overlapping communities of nodes in traditional graphs (but not for hypergraphs), a recently proposed one identifies communities as a set of closely inter-related *edges*, hence different edges created by a node make the node a part of multiple overlapping communities [1]. In this paper, we identify overlapping communities in folksonomies by customizing the algorithm in [1] for use on hypergraphs.

2. PROPOSED ALGORITHM

A tripartite hypergraph is denoted as $G = (V, E)$ where the set of nodes V is composed of three partite sets (types) V^X, V^Y and V^Z, and E is the set of hyperedges; each hyperedge connects triples of nodes (a, b, c) where $a \in V^X$, $b \in V^Y$, $c \in V^Z$. Further, let the notations $N^X(i)$, $N^Y(i)$ and $N^Z(i)$ denote the set of neighbours of node i of type V^X, V^Y and V^Z respectively.

The proposed algorithm performs an agglomerative hierarchical clustering of hyperedges using single-linkage similarity among clusters of hyperedges. Algorithm 1 gives our

WWW '11, Hyderabad, India

Algorithm 1 Compute similarity of two hyperedges

Input: hyperedges $e_1 = (a, b, c)$ and $e_2 = (p, q, r)$
Output: sim, similarity between e_1 and e_2

```
if a ≠ p and b ≠ q and c ≠ r then
    sim ← 0    /* hyperedges are non-adjacent */
else
    /* without loss of generality, let a = p; either of the
    other two pairs may be common as well */
    S1 ← N^X(b) ∪ N^X(c) ; S2 ← N^Y(c) ; S3 ← N^Z(b) ;
    S1' ← N^X(q) ∪ N^X(r) ; S2' ← N^Y(r) ; S3' ← N^Z(q) ;
    sim ← (|S1 ∩ S1'| + |S2 ∩ S2'| + |S3 ∩ S3'|) / (|S1 ∪ S1'| + |S2 ∪ S2'| + |S3 ∪ S3'|)
end if
```

customized measure for the similarity of hyperedges - the similarity between two adjacent hyperedges (i.e. having at least one node in common) is measured by the relative overlap among the neighbours of the non-common nodes of the same type, whereas non-adjacent hyperedges are assumed to have zero similarity. The hierarchical clustering, continued until all hyperedges belong to a single cluster, builds a dendrogram, and cutting this dendrogram at some suitable level gives communities of hyperedges. The optimal level for the cut, on which the quality of the obtained communities depend, is decided based on the *partition density* (p.d.) metric [1] as follows.

The p.d. of a community C of edges (or hyperedges, in case of hypergraphs) is the number of edges in C, normalized by the minimum and maximum number of edges possible among the induced nodes (i.e. nodes that are touched by the edges in C). The global p.d. for a given partitioning of the edges (hyperedges) is the average p.d. of all communities weighted by the fraction of edges present in each community. We customize the p.d. metric for use on hypergraphs, whose details are omitted for lack of space. Similar to [1], the dendrogram is cut at that level at which the global p.d. is maximum. Thus each hyperedge is placed into a single community, and a node inherits membership of all the communities into which its edges are placed.

3. EXPERIMENTS

Experiments are performed using synthetic hypergraphs generated by a modified version of the method used in [2]. For each generated hypergraph, $V^X = V^Y = V^Z = 10$, while the number of communities considered is set to 4. Each node is initially assigned to a random community; subsequently, α fraction of nodes are selected at random from each of V^X, V^Y and V^Z, and is assigned to an arbitrary number of additional randomly-selected communities. We consider different values of $\alpha = 0.1$, 0.2, 0.5, 0.8 and 1.0, with increasing values implying more complex community structure. Nodes of the same community are then randomly selected, one from each partite-set, and interconnected with hyperedges.

Users in real-world folksonomies often tag a few resources related to topics that are different from their topics of primary interest, according to their transient interests at different times. Such taggings are known to adversely affect the performance of algorithms that assign a single community to nodes. To test whether the proposed algorithm can identify both the primary and transient interests of users, a second set of hypergraphs were generated, where 10% of the hyperedges interconnect randomly-selected nodes from *different* communities; we denote these as 'scattered' hyperedges.

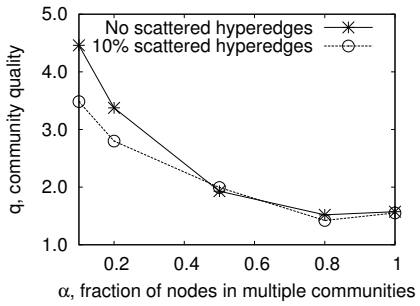

Figure 1: Quality of detected communities

The above assignment of communities to nodes constitutes the 'ground truth'. After a hypergraph is generated, information about the communities is hidden, and overlapping communities are detected from the hypergraph by the proposed algorithm.

Evaluation of Performance: The goodness of the communities detected is evaluated using the 'community quality' metric (q) defined in [1], which measures the true (i.e. according to ground truth) average similarity between pairs of nodes that are assigned to the same community by the algorithm, divided by the true average similarity between all pairs of nodes (null model). Values for q greater than 1.0 indicate that the detected communities contain more similar nodes compared to the null model.

Results: Fig. 1 shows the community quality obtained for different values of α. Each data-point in the figure is the average of 10 individual experiments. The community quality for all cases is higher than 1.0, implying that meaningful communities of similar nodes are detected by the proposed algorithm. For higher values of α, i.e. as the complexity of the community structure increases, there is an expected fall in q; however, the comparable q-values for the two sets of hypergraphs (with and without scattered hyperedges) signify that when most of the nodes belong to several communities, the presence of few hyperedges due to transient interests of users do *not* adversely affect the performance of the algorithm; this makes the algorithm suitable for use on real-world folksonomy data.

4. CONCLUSION

The algorithm proposed in this paper is one of the first steps towards detecting overlapping communities in hypergraphs. In future, we plan to use the algorithm on data from real-world folksonomies to explore the scope of improvement in tasks like recommendation of resources and tags.

5. REFERENCES

[1] Y.-Y. Ahn, J. P. Bagrow, and S. Lehmann. Link communities reveal multiscale complexity in networks. *Nature*, 466(7307):761–764, August 2010.

[2] T. Murata. Modularity for heterogeneous networks. In *ACM Hypertext*, pages 129–134, June 2010.

[3] S. Papadopoulos, Y. Kompatsiaris, and A. Vakali. Leveraging collective intelligence through community detection in tag networks. In *CKCaR*, September 2009.