Community Detection based on Structural and Attribute Similarities

The Anh Dang, Emmanuel Viennet*

*L2TI - Institut Galilée - Université Paris-Nord
99, avenue Jean-Baptiste Clément
93430 Villetaneuse - France
theanh.dang@univ-paris13.fr
emmanuel.viennet@univ-paris13.fr

Résumé. La recherche de communautés est importante pour de nombreux problèmes d'analyse des réseaux sociaux. Une communauté est définie comme un ensemble de nœuds qui interagissent d'avantage entre eux qu'avec le reste du réseau. La connaissance de ces communautés peut faciliter de nombreuses tâches telles que la recommandation d'amis, la visualisation, etc. Dans les réseaux réels, en plus de la structure topologique (i.e, liens), les attributs des nœuds sont souvent disponible. Les méthodes existantes de détection de communautés sont généralement basées sur la structure du graphe et ne prennent pas en compte les attributs des nœuds. Dans cet article, nous proposons deux algorithmes qui intègrent la structure et les attributs du graphe pour extraire les communautés. Ils détectent des communautés telles que les nœuds dans la même communauté soient densément connectés et portent des attributs proches. Les résultats démontrent que nos méthodes donnent des communautés significatives que nous comparons à celles trouvées par les méthodes de l'etat de l'art .

1 Introduction

Social networks of various kinds demonstrate a feature called community structure. Individuals in a network tend to form closely-knit groups. The groups are called communities or clusters in different context. Community detection is the task of detecting these cohesive groups in a social network (Tang et Liu (2010), Fortunato (2009)). In many real-world networks, in addition to topological structure, content information is also available. Data is associated to the nodes and in the form of text, images, etc. When content data is available, it might be relevant to extract groups of nodes that are not only connected in the social graph but also share similar attributes.

Many existing community detection techniques only focus on the topological structure of the graph. On the other hand, data clustering has been studied for a long time but most of the efficient algorithms (e.g., k-means, EM) do not deal with relational data. The work of incorporating structural and attribute data has not been throughly studied yet in the context of large social graphs. In this paper, we study the relationship between semantic similarity of users and the topology of social networks (homophily concept). We propose two approaches

to extract communities on several real-world datasets. Based on our evaluations, we conclude that our methods are able to discover more relevant communities.

2 Community detection algorithms

In this section, we present two methods to discover communities in an attributed graph, given a similarity measure.

2.1 Algorithm SAC1

Our first approach is based on the modification of Newman's well-known modularity function. Given a graph of n nodes and m edges, $G_{i,j}$ represents the link (i, j), d_i is the degree of node i. If a graph is partitioned into K clusters, Newman's modularity (Clauset et al. (2004)) is defined as

$$Q_{Newman} = \sum_{l=1}^{K} \sum_{i \in C_l, j \in C_l} \frac{1}{2m} (G_{ij} - \frac{d_i \cdot d_j}{2m}) \tag{1}$$

In this case, the link strength $S(i, j)$ between two nodes i and j is measured by comparing the true network interaction G_{ij} with the expected number of connections $(d_i \cdot d_j)/2m$

$$S(i,j) = \frac{1}{2m} \cdot \left(G_{i,j} - \frac{d_i \cdot d_j}{2m} \right)$$

Newman's modularity does not include the attribute similarity between nodes. We define the "modularity attribute" Q_{Attr} of a partition as

$$Q_{Attr} = \sum_{C} \sum_{i,j \in C} simA(i,j) \tag{2}$$

where $simA$ is the attribute similarity function.
Next, we introduce a composite modularity as a weighted combination of modularity structure (1) and modularity attribute (2)

$$Q = \sum_{C} \sum_{i,j \in C} \alpha \cdot S(i,j) + (1-\alpha) \cdot simA(i,j) \tag{3}$$

α is the weighting factor, $0 \leq \alpha \leq 1$.

The next step is to find an approximate optimization of Q (direct optimization is a NP-hard problem Brandes et al. (2006)). We follow an approach directly inspired by the Louvain algorithm Blondel et al. (2008). The algorithm starts with each node belonging to a separated community. A node is then chosen randomly. The algorithm tries to move this node from its current community. If a positive gain is found, the node is then placed to the community with the maximum gain. Otherwise, it stays in its original community. This step is applied repeatedly until no more improvement is achieved.
The first phase is completed when there is no more positive gain by moving of nodes. Following Louvain, we can reapply this phase by grouping the nodes in the same communities to a new

community-node. The weights between new nodes are given by the sum of the weight of the links between nodes in the corresponding communities Blondel et al. (2008). To determine the attribute similarity between two communities, we propose two approaches. The first is to sum up the similarity of their members.The second way is to set to the similarity of their prototypes.

2.2 Algorithm SAC2

Our first algorithm SAC1 repetitively checks all nodes, leading to $O(n^2)$ complexity. To reduce the computational cost, we propose another approach that only makes use of a node's nearest neighbors. Given an attributed graph : $G = (V, E, X)$, we define a k-nearest neighbor graph (k-NN) $G_k = (V, E_k)$ as a directed graph in which each node has exactly k edges, connecting to its k most similar neighbors in G. The similarity measure between 2 nodes i and j is defined as

$$S(i, j) = \alpha \cdot G_{i,j} + (1 - \alpha) \cdot simA(i, j)$$

where $simA(i, j)$ is the attribute similarity function, $G_{i,j}$ represents the link (i, j). Note that we can replace $G_{i,j}$ by other similarity measurements such as Jaccard similarity, cosine similarity, etc. Zhou et al. (2009) discussed several similarity metrics based on local information. Similar to the previous algorithm, we use α as a weighting factor.

We apply the measurement S in the first place to construct the nearest neighbor graph. In G_k, a structural edge represents the similarity between nodes (in terms of structure and attribute) in the original graph G.

The naive approach to build k-NN graph uses $O(n^2)$ time and $O(nk)$ space. However substantial effort has been devoted to speed up the process, such as parallel algorithms (Connor et Kumar (2009), Lieberman et al. (2008)), approximation algorithms (Chen et al. (2009), Arya et al. (1994)). In most recent work, Dong et al. (2011) introduced $NN - Descent$, an algorithm for approximate k-NN construction with an arbitrary similarity measure. The method is scalable with the empirical cost $O(n^{1.14})$.

We propose a simple algorithm with two phases : the first phase is constructing a k-NN graph G_k and the second phase is finding structural communities in G_k to obtain the final clustering. In phase 2, various methods can be employed to find communities. In our experiments, we choose Louvain as the detection method because of its scalability. We set k equal to the average degree of the nodes in the graph.

3 Results

We perform experiments on several real social networks : Political Blogs Dataset (Adamic et Glance (2005)), Facebook Friendship Datasets(Traud et al. (2010)). We extract the communities using different methods : Louvain algorithm on unweighted graph, attribute-based clustering and our proposed algorithms. We compare the number of communities, size of communities, modularity structure, modularity attribute, density and entropy of the clustering. Experimental results show that in terms of community size, our methods bring in a more balanced distribution, i.e, able to merge tiny communities to bigger ones. In terms of homogeneity, our methods result in communities with lower entropy (higher attribute similarities) than Louvain's. In term of density, our methods achieve higher density than attribute-based clustering.

Références

Adamic, L. A. et N. Glance (2005). The political blogosphere and the 2004 u.s. election : divided they blog. In *Proceedings of the 3rd international workshop on Link discovery*, LinkKDD '05, New York, NY, USA, pp. 36–43. ACM.

Arya, S., D. M. Mount, N. S. Netanyahu, R. Silverman, et A. Y. Wu (1994). An optimal algorithm for approximate nearest neighbor searching in fixed dimensions. In *ACM-SIAM SYMPOSIUM ON DISCRETE ALGORITHMS*, pp. 573–582.

Blondel, V. D., J.-L. Guillaume, R. Lambiotte, et E. Lefebvre (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics : Theory and Experiment 2008*(10), P10008 (12pp).

Brandes, U., D. Delling, M. Gaertler, R. Goerke, M. Hoefer, Z. Nikoloski, et D. Wagner (2006). Maximizing modularity is hard.

Chen, J., H. Fang, et Y. Saad (2009). Fast approximate knn graph construction for high dimensional data via recursive lanczos bisection. *Journal of Machine Learning Research 10*(2009), 1989–2012.

Clauset, A., M. E. J. Newman, et C. Moore (2004). Finding community structure in very large networks. *Physical Review E 70*, 066111.

Connor, M. et P. Kumar (2009). Fast construction of k-nearest neighbor graphs for point clouds. *IEEE Transactions on Visualization and Computer Graphics 16*(4), 599–608.

Dong, W., C. Moses, et K. Li (2011). Efficient k-nearest neighbor graph construction for generic similarity measures. In *Proceedings of the 20th international conference on World wide web*, WWW '11, New York, NY, USA, pp. 577–586. ACM.

Fortunato, S. (2009). Community detection in graphs. *Physics Reports 486, 75-174 (2010)*.

Lieberman, M. D., J. Sankaranarayanan, et H. Samet (2008). A fast similarity join algorithm using graphics processing units. *2008 IEEE 24th International Conference on Data Engineering 25*(April), 1111–1120.

Tang, L. et H. Liu (2010). *Community Detection and Mining in Social Media (Synthesis Lectures on Data Mining and Knowledge Discovery)*, Chapter 3. Morgan-Claypool.

Traud, A. L., E. D. Kelsic, P. J. Mucha, et M. A. Porter (2010). Comparing community structure to characteristics in online collegiate social networks. SIAM Review, in press (arXiv :0809.0960).

Zhou, T., L. Lu, et Y.-C. Zhang (2009). Predicting missing links via local information. *European Physical Journal B 71*(4), 623–630.

Summary

We propose two methods that couple topological structure as well as attribute information in the detection process. Experimental results in real social networks demonstrated that our methods achieve a flexibility in combining structural and attribute similarities, hence able to bring in more meaningful communities.