Community Detection on 2-Hop Topology

HU, Pili and LI, Yichao

December 13, 2011

Abstract

Community detection in social networks has drawn great interest in recent decades, especially inspired by the emerge of wide online social network sites.

This document serves as the report of CSCI5180 course project[3, 4]. For more information, codes, and data, please refer to our open source repository[1]. —**NEED-REVIEW-HERE**—

Contents

1 Introduction

Community detection has drawn great interest in the near decade. The explosion of online Social Network Sites make the study more practical but challenging. Traditional methodology aims at optimizing certain global metric. The resultant graph partition can thus be regarded as communities. The motivation of detecting communities on 2-hop topology comes in two folds:

1. **Data Availability**. The global topology of those SNS is kept confidential. However, users are allowed to view their buddies profile by default. This gives us the chance to construct a 2-hop subgraph from the whole graph. Take a step further, outsiders can always observe some forwarding sequences of tweet/status. This helps to reconstruct a subgraph more than 2-hop from the observer. The methodology developed here can thus be extended to such situations.

2. **Computation Availability**. Although those SNS providers have access to the global topology, it is computationally intractable to perform some well-developed global optimization. Local heuristics are widely used in many applications.

2 Data Preparation and Preprocessing

2.1 Data Source

The data source used in this paper is ”renren.com”, which is the most popular SNS in mainland China. At the mean while, ”renren.com” has the following good characteristics:

- Web UI of ”3g.renren.com” is very clean. Cawling through this interface doesn't require any API calls.

- Renren is a real name dense graph connected mainly through real communities. Intuitively speaking, people who are close to the observer should be observed within this 2-hop topology with high probability.

- Institution label for all nodes in the observed subgraph is available on ”renren.com”. It makes model validation easier and automated. If the model is proven effective, we can adapt it to other SNS's.

3 Statistics

4 Feature Selection

4.1 Random Walk Based Techniques

- PageRank with escape probability:

  $$\overrightarrow{v} = (1-\alpha)P^{\mathrm{T}}\overrightarrow{v} + \frac{\alpha}{n}\overrightarrow{1} \quad (1)$$

  Our graph is rooted and thus biased at the observer. So the traditional rank output by this algorithm can be a proxity measurement between other nodes and observer.

- Personalized PageRank:

  $$\overrightarrow{v} = (1-\alpha)P^{\mathrm{T}}\overrightarrow{v} + \alpha\frac{\overrightarrow{b}}{||\overrightarrow{b}||_1} \quad (2)$$

  With different escape vector $\overrightarrow{b}$, we can reach different goal:

  - Unsupervised learning. Denote the subscript of observer as o, then:

    $$b_i = \{ \begin{array}{cc} 1 & i = o \\ 0 & i \neq o \end{array} \quad (3)$$

  - Semi-supervised learning. Denote the labeled set as L and their label as l_i. The personalized escape vector can be:

    $$b_i = \{ \begin{array}{cc} l_i & i \in L \\ 0 & i \notin L \end{array} \quad (4)$$

4.2 Simple Proximity Measures

- Common Neighbours:

  $$Common(i,j) = |N(i) \cap N(j)| \quad (5)$$

- Adamic/Adar score:

  $$Adamic/Adar(i,j) = \sum_{k \in N(i) \cap N(j)} \frac{1}{\log|N(k)|} \quad (6)$$

  Adamic/Adar can be seen as an extension of simple common neighbours, which take the neighbours degree into consideration.

- Jaccards coefficient:

  $$J(i,j) = \frac{|N(i) \cap N(j)|}{|N(i) \cup N(j)|} \quad (7)$$

  This coefficient can be seen as another extension of common neighbours. It is effective when the graph is sparse.

5 Single Feature Evalutaion

5.1 Quality Functions

Evaluate some popular quality functions:

- Normalized cut:

$$Ncut(S) = \frac{\sum_{i \in S, j \in \overline{S}} A(i,j)}{\sum_{i \in S} d(i)} + \frac{\sum_{i \in S, j \in \overline{S}} A(i,j)}{\sum_{j \in \overline{S}} d(j)} \tag{8}$$

- Conductance:

$$Conductance(S) = \frac{\sum_{i \in S, j \in \overline{S}} A(i,j)}{\min\{\sum_{i \in S} d(i), \sum_{j \in \overline{S}} d(j)\}} \tag{9}$$

- Modularity:

$$Q = \sum_{i=1}^{k} \left[\frac{A(V_i, V_i)}{m} - \left(\frac{d(V_i)}{2m} \right)^2 \right] \tag{10}$$

Those quality functions are popular among different research groups. They also capture different characteristics of graphs. In this project, we'll check if these global metric can be used in our extreme local case.

5.2 Quality Functions

Draw the Receiver Operating Curve by varying proximity threshold. (Different TP and FP rate)

—NEED-REVIEW-HERE—

6 Model Training and Testing

7 Conclusion

8 Future Work

As is in our proposal [1], there are other metrics we want to try. Due to project time and report space, we only name a few in this section:

- Katz Score:

$$Katz(i,j) = \sum_{l=1}^{\infty} \beta^l A^l(i,j) \tag{11}$$

  The matrix series results in

$$Katz = (I - \beta A)^{-1} - I \tag{12}$$

  By tuning β, this score can measure number of paths between two nodes, with preferred length range.

- Simrank:
$$s(a,b) = \frac{\gamma}{|I(a)||I(b)|} \sum_{i \in I(a), j \in I(b)} s(i,j) \tag{13}$$

  This metric measures the expectation of γ^l, where l equals the time when two random walkers start from a and b meet. By proper matrix construction, Simrank can be solved as an eigenvalue problem.

For more information, interested readers are recommended to [2].

Acknowledgements

References

[1] Hu Pili and Li Yichao. Community detection on 2-hop topology. GitHub, https://github.com/Czlyc/2C-Web-Research, 12 2011. course project of CUHK/CSCI5180.

[2] C.C. Aggarwal. *Social network data analytics.* Springer-Verlag New York Inc, 2011.

[3] CSCI5180, Data Mining Lecture Notes, `http://www.cse.cuhk.edu.hk/~lwchan/teaching/csc5180.html`

[4] CSCI5180, Data Mining Tutorial Notes. (available in Moodle. Please contact the teacher/TA if you have interest)

Appendix

Declaration

Meta tools like PageRank, Adamic/Adar, etc, are learned from the book *Social Network Data Analytics.* For original sources of these algorithms, please refer to the bibliography pages of that book.

The analysis/adaptation/extension/application of these tools and the use in community detection on 2-hop topology mentioned in this document is originality. Please refer to our open source repository for citation issues.