Finding Similar Questions with Community Detection

Jake Carlson

March 14, 2019

Abstract

A report on implementing algorithms to find community structure in large graphs. We will explore several different methods for finding communities, mainly modularity, Walktrap, and Spin-Glass. I implement each of these methods in Python and show how they perform on Stack Overflow questions and answers.

Contents

1 Introduction 3

2 Background 3

3 Algorithms 3

3.1 Modularity . . . 3

3.2 Walktrap . . . 3

3.3 Spin-Glass . . . 3

4 Evaluation 3

5 Conclusion 3

1 Introduction

The widespread usage of social networks has made large, graph-based data sets widely available. It is possible to find communities within this graph data by defining what a community is, and then by finding a way to extract that information from the graph data set. Community structure is defined as densely connected groups of verticies, joined by sparser connections between groups [1].

2 Background

3 Algorithms

3.1 Modularity

3.2 Walktrap

3.3 Spin-Glass

4 Evaluation

To examine the usefulness of these algorithms, I will use them to detect communities in Stack Overflow data. The Stack Overflow data is contains a table of questions, a table of answers to those questions, and a list of tags associated with each question. In the questions and answers tables, there are also fields indicating the user who posted the question. In this data, community structure exists in the form of questions that are about related topics. For example, we would expect a user who is active in the C++ community to answer many questions about C++. It is also possible for users to be active in multiple different communities.

Initially I thought it would be useful to represent both users and questions in the graph we are mining for communities, however, this proved to not be useful. Representing both users and questions forms a bipartite graph where all edges go between user and question nodes. Although this better represents the original data, this makes it much harder to do community detection. For example, if a user is active in multiple communities, there will be more connections joining these communities and it will be harder to find the community structure.

To account for this, I manipulated the graph to better encode the tag information. To do this I removed the user nodes, and created edges between questions that share a tag. It can be seen that this will significantly increase the connectedness of the graph, but this connectedness is useful for algorithms such as Walktrap.

5 Conclusion

References

[1] M. E. J. Newman, Detecting community structure in networks. Eur. Phys. J. B 38, 321–330 (2004).