Social Network Friend Reccomendations

Social and Economic Networks - Literature Review

Guillem Bagaria Portet

June 23, 2018

Contents

Chapter 1

Introduction

1.1 The data

The goal of this project is to implement and test the performance of a next-connection prediction algorithm in a social network with communities, that is a friend suggester algorithm. For this project I will be using a ready-made dataset provided by Stanford department that developed and maintains SNAP (Stanford Network Analysis Project) named *email-EuAll*.

The network was generated using email data from a large European research institution. The information about all incoming and outgoing email between members of the research institution has been anonymized. There is an edge (u, v) in the network if person u sent person v at least one email. The e-mails only represent communication between institution members (the core), and the dataset does not contain incoming messages from or outgoing messages to the rest of the world.

The dataset also contains "ground-truth" community memberships of the nodes. Each individual belongs to exactly one of 42 departments at the research institute.

This network represents the "core" of the *email-EuAll* network, which also contains links between members of the institution and people outside of the institution (although the node IDs are not the same).

Table 1.1: Dataset Statistics

Nodes	1005
Edges	25571
Nodes in largest WCC	986 (0.981)
Edges in largest WCC	25552 (0.999)
Nodes in largest SCC	803 (0.799)
Edges in largest SCC	24729 (0.967)
Average clustering coefficient	0.3994
Number of triangles	105461
Fraction of closed triangles	0.1085
Diameter (longest shortest path)	7
90-percentile effective diameter	2.9

The dataset can be found in the SNAP email-Eu-core network Dataset information page at `https://snap.stanford.edu/data/email-Eu-core.html` [June 2018].

1.2 Literature

1.2.1 Paper *The Link-Prediction Problem for Social Networks* by David Liben-Nowell and Jon Kleinberg

This is a very solid and exhaustive source for this project. It explores different methods to calculate similarity exclusively by topological means, establishes ways to evaluate each method and ranks them in an extensive comparison section.

"Experiments on large co-authorship networks suggest that information about future interactions can be extracted from network topology alone, and that fairly subtle measures for detecting node proximity can outperform more direct measures."

"There are many reasons exogenous to the network why two scientists who have never written an article together will do so in the next few years: For example, they may happen to become geographically close when one of them changes institutions. Such collaborations can be hard to predict. But one also senses that a large number of new collaborations are hinted at by the topology of the network: Two scientists who are "close" in the network will have colleagues in common and will travel in similar circles; this social proximity suggests that they themselves are more likely to collaborate in the near future."

Methods for Link Prediction

- Methods Based on Node Neighborhoods
- Common neighbors
- Jaccard's coefficient and Adamic/Adar
- Preferential attachment
- Methods Based on the Ensemble of All Paths
- Hitting time, PageRank, and variants.
- Katz (1953), Katz-Bonacich
- SimRank (Jeh & Widom, 2002)

1.2.2 Paper *SimRank: A Measure of Structural-Context Similarity* by Glen Jeh and Jennifer Widom

This paper is referenced in the previous one as one as the methods explored. The method proposed follows the idea that "two objects are similar if they are referenced by similar objects".

1.2.3 Paper *Social Friend Recommendation Based on Multiple Network Correlation* by Shangrong Huang, Jian Zhang, Lei Wang and Xian-Sheng Hua

In this paper a network correlation (NC) method is used to improve friend recommendations in social networks. The use of both a projection to a lower dimensional space as well as using tags (meta data) other than network topology will provide some insight into further work for a working algorithm.

1.2.4 Book *Networks, Crowds, and Markets: Reasoning about a Highly Connected World* by David Easley

Many fundamentals will be used for this project found in this book, especially sections regarding:

- **Network Homophily** refers to the theory in network science which states that, based on node attributes, similar nodes may be more likely to attach to each other than dissimilar ones

- **Similarity** of two nodes in a network occurs when they fall in the same equivalence class. There are three fundamental measures of network similarity: structural equivalence, automorphic equivalence, and regular equivalence.

- **Betweenness Measures and Graph Partitioning** as centrality in a graph based on shortest paths in different communities.

Other items might be added.