

Inferring Social Ties across Heterogeneous Networks

Jie Tang
Department of Computer
Science
Tsinghua University
Beijing 100084, China
jietang@tsinghua.edu.cn

Tiancheng Lou
Institute for Interdisciplinary
Information Sciences
Tsinghua University
Beijing 100084, China
ltc08@tsinghua.edu.cn

Jon Kleinberg
Department of Computer
Science
Cornell University
Ithaca NY 14853
kleinber@cs.cornell.edu

ABSTRACT

It is well known that different types of social ties have essentially different influence on people. However, users in online social networks rarely categorize their contacts into “family”, “colleagues”, or “classmates”. While a bulk of research has focused on inferring particular types of relationships in a specific social network, few publications systematically study the generalization of the problem of inferring social ties over multiple heterogeneous networks. In this work, we develop a framework for classifying the type of social relationships by learning across heterogeneous networks. The framework incorporates social theories into a factor graph model, which effectively improves the accuracy of inferring the type of social relationships in a target network by borrowing knowledge from a different source network. Our empirical study on five different genres of networks validates the effectiveness of the proposed framework. For example, by leveraging information from a coauthor network with labeled advisor-advisee relationships, the proposed framework is able to obtain an F1-score of 90% (8-28% improvements over alternative methods) for inferring manager-subordinate relationships in an enterprise email network.

Categories and Subject Descriptors

J.4 [Social and Behavioral Sciences]: Miscellaneous; H.2.8 [Database Management]: Data Mining; H.3.3 [Information Search and Retrieval]: Text Mining

General Terms

Algorithms, Experimentation

Keywords

Inferring social ties, Social network, Predictive model, Social influence analysis

1. INTRODUCTION

Our social networks are complex and consist of many overlapping parts. Nobody exists merely in one social network. People are connected via different types of social ties in different networks. For example, in an enterprise email network, where people

are connected by sending/receiving emails to/from others, relationships between people can be categorized into manager-subordinate, colleague, etc.; in a mobile communication network, the relationship types could include family, colleagues, and friends. It is well known that the different types of social ties have essentially different influence on people. A graduate’s research topic may be mainly influenced by his or her advisor, while other parts of his everyday life will be more influenced by his close friends. Awareness of these different types of social relationships can benefit many applications. For example, if we could have extracted friendships between users from a mobile communication network, we can leverage the friendships for a “word-of-mouth” promotion of a new product.

However, in most online networks (e.g., Facebook, Twitter, LinkedIn, YouTube, and Slashdot), such information (relationship type) is usually unavailable. Users may easily add links to others by clicking “friend request”, “follow” or “agree”, but do not often take the time to create labels and maintain their friend list. Indeed, one survey of mobile phone users in Europe shows that only 16% of users have created contact groups on their mobile phones [10, 26]; our preliminary statistics on the LinkedIn data also shows that more than 70% of the connections have not been well labeled. In addition, the availabilities of labeled relationships in different networks are very unbalanced. In some networks, such as Slashdot, it might be easy to collect the labeled relationships (e.g., trust/distrust relationships between users). However, in most other networks, it may be infeasible to obtain the labeled information. A challenging question is: can we leverage the labeled relationships from one network to infer the type of relationships in another totally different network?

Motivating Examples To clearly illustrate the problem, Figure 1 gives an example of inferring social ties across a product-reviewer network and a mobile communication network. In Figure 1, the left sub-figure is the input to our problem: a reviewer network, which consists of reviewers and relationships between reviewers; and a mobile network, which consists of mobile users and their communication relationships (via calling or texting message). The right sub-figure shows the output of our problem: the inferred social ties in the two networks. In the reviewer network, we infer the trust/distrust relationships and in the communication network, we identify friendships, colleagues, and families. The middle of Figure 1 is the component of knowledge transfer for inferring social ties in different networks. This is the key objective of this work. The fundamental challenge is how to bridge the available knowledge from different networks to help infer the different types of social relationships.

The problem is non-trivial and poses a set of unique challenges. First, what are the fundamental factors that form the structure of different networks? Second, how to design a generalized frame-

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

WSDM’12, February 8–12, 2012, Seattle, Washington, USA.
Copyright 2012 ACM 978-1-4503-0747-5/12/02 ...\$10.00.

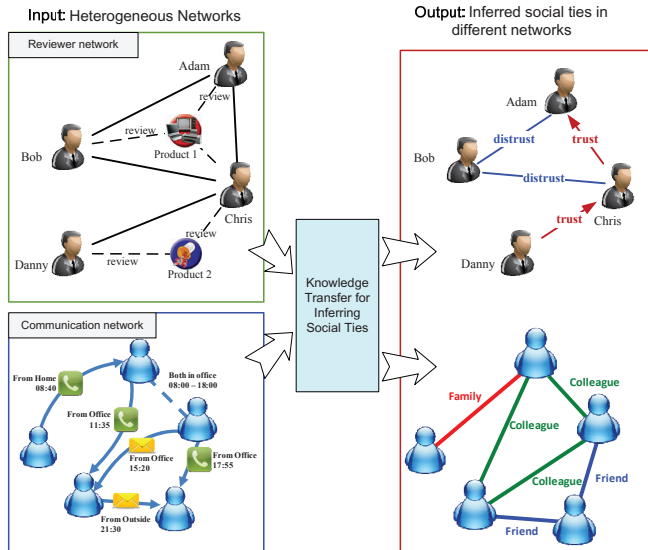


Figure 1: Example of inferring social ties across two heterogeneous networks: a product-reviewer network and a mobile communication network.

work to formalize the problem in a unified way? Third, as real social networks are getting larger with hundreds of millions of nodes, how to scale up the model learning algorithm to adapt to the growth of large real networks?

Results In this work, we aim to conduct a systematic investigation of the problem of inferring social ties across heterogeneous networks. We precisely define the problem and propose a transfer-based factor graph (TranFG) model. The model incorporates social theories into a semi-supervised learning framework, which can be used to transfer supervised information from a source network to help infer social ties in a target network.

We evaluate the proposed model on five different genres of networks: Epinions, Slashdot, Mobile, Coauthor, and Enron. We show that the proposed model can significantly improve the performance (averagely +15% in terms of F1-Measure) for inferring social ties across different networks comparing with several alternative methods. Our study also reveals several interesting phenomena for social science:

- Social balance is satisfied well on friendship (or trust) networks; but not (< 20% with a large variance) on user communication networks (e.g., mobile communication network).
- Users are more likely (up to +152% higher than chance) to have the same type of relationship with a user who spans a structural hole. Disconnected users have an even higher likelihood.
- It was validated that social status is satisfied in many networks. We further discover that several frequent forms of triads have a similar distribution in different networks (Coauthor and Enron).
- Opinion leaders are more likely (+71%+84%) to have a higher social status than ordinary users.

Organization Section 2 formulates the problem; Section 3 introduces the data set and our observations over different networks. Section 4 explains the proposed model and describes the algorithm for learning the model; Section 5 gives the experimental setup and

Section 6 presents the results; finally, Section 7 discusses related work and Section 8 concludes.

2. PROBLEM DEFINITION

In this section, we first give several necessary definitions and then present the problem formulation. To simplify the explanation, we frame the problem with two social networks: a source network and a target network, although the generalization of this framework to multiple-network setting is straightforward.

Let $G = (V, E^L, E^U, \mathbf{X})$ denote a partially labeled social network, where E^L is a set of labeled relationships and E^U is a set of unlabeled relationships with $E^L \cup E^U = E$; \mathbf{X} is an $|E| \times d$ attribute matrix associated with edges in E with each row corresponding to an edge, each column an attribute, and an element x_{ij} denoting the value of the j^{th} attribute of edge e_i . The label of edge e_i is denoted as $y_i \in \mathcal{Y}$, where \mathcal{Y} is the possible space of the labels (e.g., family, colleague, classmate).

Input: The input to our problem consists of two partially labeled networks G_S (source network) and G_T (target network) with $|E_S^L| \gg |E_T^L|$ (with an extreme case of $|E_T^L| = 0$). Please note that the two networks might be totally different (with different sets of vertices, i.e., $V_S \cap V_T = \emptyset$, and different attributes defined on edges).

In real social networks, the relationship could be undirected (e.g., friendships in a mobile network) or directed (e.g., manager-subordinate relationships in an enterprise email network). To keep things consistent, we will concentrate on the undirected network. In addition, the label of a relationship may be static (e.g., the family-member relationship) or change over time (e.g., the manager-subordinate relationship). In this work, we focus on static relationships.

Learning Task: Given a source network G_S with abundantly labeled relationships and a target network G_T with a limited number of labeled relationships, the goal is to learn a predictive function $f : (G_T|G_S) \rightarrow Y_T$ for inferring the type of relationships in the target network by leveraging the supervised information (labeled relationships) from the source network.

Without loss of generality, we assume that for each possible type y_i of relationship e_i , the predictive function will output a probability $p(y_i|e_i)$; thus our task can be viewed as obtaining a triple $(e_i, y_i, p(y_i|e_i))$ to characterize each link e_i in the social network. There are several key issues that make our problem formulation different from existing works on social relationship mining [4, 6, 27, 29, 30]. First, the source network and the target network may be very different, e.g., a coauthor network and an email network. What are the fundamental factors that form the structure of the networks? Second, the label of relationships in the target network and that of the source network could be different. How reliably can we infer the labels of relationships in the target network by using the information provided by the source network? Third, as both the source and the target networks are partially labeled, the learning framework should consider not only the labeled information but also the unlabeled information.

3. DATA AND OBSERVATIONS

3.1 Data Collection

We try to find a number of different types of networks to investigate the problem of inferring social ties across heterogeneous networks. In this study, we consider five different types of networks: Epinions, Slashdot, Mobile, Coauthor, and Enron. Table 1

