Computer Networks 64 (2014) 369–389

Contents lists available at ScienceDirect

Computer Networks

journal homepage: www.elsevier.com/locate/comnet

CommonFinder: A decentralized and privacy-preserving common-friend measurement method for the distributed online social networks

Yongquan Fu [a], Yijie Wang [a,*], Wei Peng [b]

[a] Science and Technology on Parallel and Distributed Processing Laboratory, College of Computer, National University of Defense Technology, Hunan Province 410073, China

[b] College of Computer, National University of Defense Technology, Hunan Province 410073, China

ARTICLE INFO

Article history:
Received 19 November 2012
Received in revised form 9 October 2013
Accepted 3 February 2014
Available online 7 March 2014

Keywords:
Distributed online social network
Privacy
Common friends
Friend recommendation
Coordinate
Bloom filter

ABSTRACT

Distributed social networks have been proposed as alternatives for offering scalable and privacy-preserving online social communication. Recommending friends in the distributed social networks is an important topic. We propose CommonFinder, a distributed common-friend estimation scheme that estimates the numbers of common-friends between any pairs of users without disclosing the friends' information. CommonFinder uses privacy-preserving Bloom filters to collect a small number of common-friend samples, and proposes low-dimensional coordinates to estimate the numbers of common friends from each user to any other users. Simulation results on real-world social networks confirm that CommonFinder scales well, converges quickly and is resilient to incomplete measurements and measurement noises.

1. Introduction

Online social networks such as Facebook, YouTube, Flickr have become quite popular these days. However, there are also hot debates over the privacy protection of these centralized social services. For example, the service providers can peek at users' profiles at will for targeted advertisements or even sell these sensitive information to third parties for profits. As a result, improving the privacy protection of social networks becomes increasingly important.

Fortunately, borrowing the success of the P2P systems, the distributed online social networks (DOSN for short), e.g., Safebook [1], LotusNet [2], Cuckoo [3], diaspora [4], Peerson [5], Vis-a-Vis [6] have been proposed to better protect users' privacy by hosting the DOSN infrastructures based on decentralized end hosts. The key idea is to store personal data on decentralized nodes and to enforce strict access rules on who can visit a user's profile.

To increase the popularity of distributed online social networks, an open question is how to find potential friends for users, which is known as the **friend recommendation** problem. The centralized online social networks are able to compute possible friends by using the complete knowledge of users' profiles. Unfortunately, for DOSNs, it is commonly believed that end hosts are unwilling to publish their profiles to unknown entities because of the privacy leakage. As a result, it is highly desirable to develop scalable and privacy-preserving methods to quantify the possibility of being friends for DOSN users.

One of the most popular metric for friend recommendation is based on the **Number of Common Friends** (NCF for short), which has been shown to be very effective to recommend new friends or find old friends [7]. Measuring

* Corresponding author. Tel.: +86 13308491230 (Y. Wang).
E-mail addresses: yongquanf@nudt.edu.cn (Y. Fu), wangyijie@nudt.edu.cn (Y. Wang), wpeng@nudt.edu.cn (W. Peng).

http://dx.doi.org/10.1016/j.comnet.2014.02.020
1389-1286/

NCF values in the distributed online social networks is however, a difficult task. First, disclosing friend lists to non-friend users could severely leak users' privacy information [8,9]. Second, the NCF computation has to scale well since most end hosts have limited bandwidth capacity.

Talash [10] computes the common friends based on exchanging friend lists between pairs of users, which not only leaks personal information, but also does not scale well. The PSI approach [11–14] represents a list of friends with coefficients of a polynomial. It then exchanges Homomorphic encrypted coefficients via the communication links, and finally computes the sums of coefficients, which correspond to the set of common items in two friend lists. The PSI approach works well for semi-honest users, but increases the transmission bandwidth and the computation costs.

We propose a scalable and privacy-preserving distributed NCF estimation method called CommonFinder that estimates the NCF values between any pairs of users and is able to select top-k users that have the largest NCF values.

CommonFinder represents a user's friend list with the well-known Bloom filter. Our privacy analysis (Section 8) proves that the Bloom filter provides the differential privacy [15,16] for the friend list: Given a user i's Bloom filter, a curious entity is unable to determine who are user i's friends.

Unfortunately, exchanging the Bloom filters may need several KBytes bandwidth costs in order to control the false positives of the Bloom filter. To further increase the scalability of estimating the NCF values, we propose a distributed maximum margin matrix factorization method to estimate the NCF values by low-dimensional coordinates. As a result, *the transmission size is fixed to be the length of the coordinate that is independent of the length of the friend lists or Bloom filters*.

We model the NCF completion problem with the maximum margin matrix factorization method and provide a systematical design of a distributed NCF prediction method that significantly extends our prior work [17,18]. The MMMF method maps contiguous matrix factorization results to discrete NCF values with adaptive thresholds. These thresholds are learnt during the process of optimizing the matrix-factorization model. To adapt to the decentralization of users, we reformulate the MMMF method in separable objective functions that involve each user's coordinate and a small number of neighbors. For each user, we design a fully decentralized conjugate gradient optimization method to adjust the coordinates of each user that converges quickly.

Finally, we present extensive privacy analysis and simulation results over the real-world social network topologies. Our results show that CommonFinder not only protects users' friend lists, but also significantly improves the prediction accuracy and is more robust against incomplete or incorrect NCF measurements than previous methods.

The rest of the paper is organized as follows. Section 2 introduces the background information. Section 3 next defines the problem of predicting NCF values with preservation of users' privacy. Section 4 then analyzes typical characteristics of social-network data sets. Section 5 next presents an overview of our proposed NCF prediction methods. Section 6 presents the coordinate computation process. Section 7 next applies the decentralized coordinates to select top-k users for recommending friends. Section 8 then measures the degree of the privacy protection offered by CommonFinder. Section 9 next compares CommonFinder's performance with extensive simulation results. Section 10 next confirms CommonFinder is robust to Sybil accounts. Section 11 then summarizes related work. Section 12 concludes the paper. Table 1 shows key parameters.

2. Background

2.1. Distributed online social networks

We introduce the basic characteristics for existing DOSNs.

2.1.1. Social graph

Let a **user** be an online entity that participates in the DOSN. Each user A has a **friend list** S_A, where a **friend** B of a user A is a user B that establishes the **social link** or **friendship link** with user A on the DOSN. The **common friends** of two users A and B are represented by the subset of users that are both friends of user A and B at the same time.

Users and social links form a **social graph**. Each user and his/her friends are adjacent on the social graph, which correspond to one-hop **neighbors** to each other. The **degree** of each user amounts to the size of its neighbors on the social graph. The number of **hops** between two users amounts to the length of the shortest path for these two users on the social graph.

2.1.2. Privacy-preserving message routing

Users' data are usually stored into his/her own computer. For improving the availability of users' data when they are offline, some DOSNs like Safebook replicate each user's data on his/her friends' computers. These friends' replication will be used for offline users.

The social graph is maintained via some kind of Peer-to-Peer substrates. When a user joins the DOSN, each user's computer needs to be registered in the Peer-to-Peer substrate consisting of decentralized computers. Users' real names are decoupled with randomized keys called **identifiers** that are generated by cryptographic hash

Table 1
Notations and their meanings.

k_l	The number of hash functions in a Bloom filter
m_l	The length of a Bloom filter
N	The number of users on the DOSN
$\widehat{\mathbf{X}}$	The coordinate distance matrix
$\mathbf{Y}$	The pairwise NCF matrix between a set of users
L	The maximal NCF value
d	The coordinate dimension
θ	The NCF-mapping thresholds
k	The number of recommended users