基金项目：国家重点研发计划项目(2017YFB0802300)
收稿日期：
*通讯联系人，E-mail：wanggy.cq@hotmail.com

一种基于用户结构和属性的无监督用户对齐方法

俞冬明[1,2]，李苑[1,2]，李智星[1,2]，王国胤[1,2*]

(1. 计算智能重庆市重点实验室，重庆，400065；

2.重庆邮电大学计算机科学与技术学院，重庆，400065)

摘　要：随着互联网应用的蓬勃发展，一个人在不同的社交网络平台上都拥有账户是很常见的。如何在多个社交网络上找到同一个人的账户，对于现在许多应用来说都是很重要的问题，这个问题也被称为用户对齐问题。在用户对齐问题上，目前有两个主要的挑战。首先，收集手工对齐的用户对作为训练数据代价非常大,但传统的有监督方法往往需要大量的标注数据才能获得较好的效果。其次，不同网络中的用户的结构和属性往往都不太相同，进一步增加了用户对齐的难度。提出一种无监督用户对齐方法 SPUAL（Soft Principle for User Alignment），设计了一种新颖的基于用户的属性与结构的软对齐一致性原则，通过无监督方法计算用户对是否服从此原则来推断用户对是否对齐。在几个公共数据集上进行的实验表明，该方法的性能比目前最先进的无监督方法都有明显提高。

关键词：用户对齐，社交网络，无监督，对齐原则

中图分类号：TP391　　**文献标识码：A**

An unsupervised user alignment method based on user structure and attribute

Yu Dongming[1,2], Li Yuan[1,2], Li Zhixing[1,2], Wang Guoyin[1,2*]

(1. Chongqing Key Lab of Computation Intelligence, Chongqing, 400065, China;

2. Chongqing University of Posts and Telecommunications, College of Computer Science and Technology, Chongqing, 400065, China)

Abstract: With the fast development of Internet applications, it's common for someone to have an account on a different social network platform. How to find out which accounts on multiple social networks are of the same person is an important issue for many applications today, which is also known as the user alignment problem. There are two major challenges when it comes to user alignment. First, it’s extremely expensive to collect manually aligned user pairs as training data, but traditional supervised methods often need a large amount of labeled data to achieve better results. Second, users on different networks often have different structures and attributes, which further increase the difficulty of user alignment. We propose an unsupervised user alignment method SPUAL (Soft Principle for User Alignment), design a novel soft alignment principle based on user attributes and structure, and then infer whether the user alignment is correct or not by calculating whether the users meet to the principles by unsupervised method. Experiments on several common datasets show that the performance of our method is much better than the most advanced unsupervised methods.

Key words: user alignment; social network; unsupervised; alignment principle

近年来，很多用户已经在不同的社交网络都拥有账户，如微博、Twitter、Instagram 或 LinkedIn。由于不同的社交网络平台的功能不同，它们利用各自的优势吸引用户，进行信息的寻找、分享和社交关系的维护等操作。为了更好地享受服务，用户往往会加入多个社交网络，如使用 Twitter 发布对突发热点事件的意见，使用 Instagram 分享自己的休闲活动。然而，不同的社交网络由不同的公司维护，彼此独立，所以无法判断不同的社交网络的不同用户是否属于同一个人。解决用户对齐的问题通常需要更好和更深入地了解个人用户，这通常会带来更好的商业机会。例如，用户对齐之后可以使社交网络的一部分用户的用户信息更加完整[1-2]，还可以了解社交网络之间的用户迁移模式[3]，或帮助社交网站更准确地推荐潜在的朋友等。

虽然有监督机器学习算法在用户对齐问题中得到了广泛的应用，但标注训练数据的工作量却不容小觑。首先，寻找已知对齐的用户对非常耗时，因为需要搜索整个网络并仔细评估大量的候选对；其次，还要求人工标记人员有广泛的专业知识，例如在处理有软件开发背景的用户之前，必须知道“SDE”是“软件开发工程师”的缩写。另外，很多社交网络的数据通常涉及个人隐私，尤其是企业内部的社交网络，不会轻易交给人工标记者来进行人工标记。

无监督学习的优势是可以从无标记的数据中根据特定的原则学习数据的模式，因此非常适合大规模无标记数据场景下的用户对齐。目前，大部分无监督方法都集中于基于社交网络的拓扑结构来推断用户对齐关系。例如，IsoRank[4]在社交网络中传播成对的拓扑结构相似性。NetAlign[5]利用基于网络拓扑结构的最大乘积信念进行传播。BigAlign 和 UniAlign[6]的方法是基于一个网络的邻接矩阵是另一个网络的噪声排列的假设来推断软对齐。这些方法背后的一个基本假设是拓扑一致性，也就是说，同一个用户在不同的社交网络中的社交关系具有一致性(例如，连接到相同或相似的邻居集)。然而，这种假设在某些社交网络中并不适用。例如，一个用户可能在一个社交网站(如 Facebook)上非常活跃，但在另一个网站(如 LinkedIn)上却表现得很安静[6]。在这些情况下，基于拓扑结构的方法可能会产生错误的对齐结果。此外,这些方法无法利用用户的属性信息，而这也损失了用户的信息。FINAL[7-8]提出用户对齐一致性原则，通过判断用户对是否满足对齐一致性原则来对判断用户对的对齐的一致性。虽然 FINAL 的思想很不错，但它要求用户对必须满足所有的对齐一致性原则，才保证他们的对齐性一致，这是严格的对齐过程。

由于社交网络复杂度高、噪声大的特点，在不同的社交网络中对齐的用户的属性和拓扑结构可能会不同，又因为不同的社交网络注重性不同，如果对用户进行严格的对齐，会导致很多的错误对齐。针对这个问题，本文提出一种基于用户结构和属性的无监督用户对齐方法 SPUAL(Soft Principle for User Alignment)，主要贡献在于：

(1)提出一种无监督用户对齐方法 SPUAL，利用基于用户的属性与结构建立新的用户的软对齐一致性原则来推断用户对是否对齐；

(2)将用户对齐的问题转化成一个二次优化问题，并且转换成矩阵形式，从而可以更有效率地进行求解；

(3)进行了大量的实验，将 SPUAL 与最先进的无监督方法进行比较，证明了方法 SPUAL 的有效性。

1 相关工作

用户对齐问题近些年来已经引起了广泛的研究兴趣，已有大量的相关文献。目前的一些主流方法主

要分为监督的方法和无监督的方法。

有监督方法通过从用户的属性中提取特征，并使用监督分类器模型来预测用户对是否对。Goga et al[9]基于属性的距离相似性特征训练逻辑斯蒂回归分类器，对候选对进行二分类。Liu et al[10]使用基于词袋和距离的特征，将用户对齐问题作为多目标优化问题进行求解。Zhang et al[1]采用基于词袋的特征和基本关系特征，并结合局部和全局一致性提出基于能量的模型。最近，许多基于嵌入的方法也被提了出来[11-13]。然而，监督的方法都需要昂贵的带标签的训练数据，因此应用场景受限.

IsoRank[4]是一种经典的无监督用户对齐方法，它是受到了 PageRank[14]的启发。IsoRankN[15]扩展了原有的 IsoRank 算法，并使用类似于 PageRank-Nibble[16]的方法对多个网络进行对齐。Bayati et al[17]提出一种利用最大乘积信念传播进行网络对齐的最大权值匹配算法。NetAlign[5]将用户对齐问题转化为一个整形二次规划问题，以最大限度地增加网络的平方数。Zhang and Philip[18]分两步解决了多个匿名网络用户对齐问题，即无监督用户对齐推理和传递多网络匹配。这些方法都假设同一个用户在不同的社交网络的社交关系具有一致性，但这个假设在某些社交网络中并不适用，并且无法利用用户的属性信息。

FINAL[7]提出用户对齐一致性原则，通过判断用户对是否满足所有的对齐一致性原则来判断用户对的对齐一致性。然而现在很多用户注重隐私保护，很多对齐的用户对不一定满足所有的用户对齐一致性原则。其次，FINAL 把每一个对齐一致性原则都等同对待，这在某些社交网络中并不适用，因为不同的社交网络注重性不同，有的社交网络注重用户的个人信息，这个时候属性一致性就比拓扑一致性重要很多。而有的社交网络注重社交关系，比如 twitter、微博，此时拓扑一致性就显得格外重要。针对这个问题，本文提出一种无监督用户对齐方法 SPUAL，基于用户的属性与结构建立新的用户软对齐一致性原则来推断用户是否对齐。

2 问题定义

用户对齐的问题可以简单描述为在多个输入的社交网络之间找到对应的用户，这些在不同社交网络中的对应用户属于真实世界中的同一个人。表 1 总结了这篇文章使用的主要的符号。本文用黑体的大写字母来表示矩阵（比如$\boldsymbol{A}$），黑体的小写字母来表示向量(比如$\boldsymbol{s}$)，小写字符来表示标量（比如a）。用$A(i,j)$来表示矩阵$\boldsymbol{A}$第i行和第j列的元素，用$A(i:)$和$A(:j)$分别来表示矩阵$\boldsymbol{A}$的第i行和第j列，用$\boldsymbol{A}^T$来表示矩阵$\boldsymbol{A}$的转置，$\boldsymbol{D}$是矩阵的度矩阵。一个矩阵的矢量化用$vec(:)$表示，并且结果向量用对应的黑体小写字母表示，例如$\boldsymbol{a} = vec(\boldsymbol{A})$。

表 1 符号和意义

Table 1 Symbols and notation

符号	意义
$G = \{\boldsymbol{A}, \boldsymbol{N}\}$	一个社交网络
$\boldsymbol{A}$ $\boldsymbol{N}$	网络的邻接矩阵 网络用户的属性矩阵
n_1,n_2	网络G_s和网络G_t的用户数
K	用户节点的属性个数
a, b	网络G_s的用户索引

x, y	网络G_s的用户索引
v, w	向量化对齐的节点对索引
$\boldsymbol{I}, \boldsymbol{l}$	分别是单位矩阵和值全为 1 的向量
$\boldsymbol{H}$,$\boldsymbol{S}$	对齐前的偏好和对齐矩阵
α, β, m	对齐原则的权重以及正则化参数
$\boldsymbol{a} = vec(\boldsymbol{A})$ $\boldsymbol{D} = diag(\boldsymbol{d})$ $\otimes$	将矩阵$\boldsymbol{A}$以列的顺序向量化 对角化向量$\boldsymbol{d}$ 克罗内克积

本文用一个二元组来表示一个网络：$G = \{\boldsymbol{A} \in R^{n \times n}, \boldsymbol{N} = \{\boldsymbol{N}^1, \boldsymbol{N}^2, \ldots, \boldsymbol{N}^k\}\}$，$\boldsymbol{A}$是社交网络图的邻接矩阵，$\boldsymbol{N}$是社交网络用户的属性矩阵集合，集合的每一个元素$\boldsymbol{N}^k \in R^{n \times n}$都是一个对角矩阵，用来表示用户节点是否具有属性$k$。如果用户节点$a$具有属性$k$，那么$\boldsymbol{N}^k(a, a) = 1$，否则$\boldsymbol{N}^k(a, a) = 0$。其中$n$为网络中用户的数量。所以，正式地将带属性的用户对齐问题的定义如下：

问题：基于用户属性与结构的用户对齐

输入：一个源社交网络$G_s\{\boldsymbol{A}_s, \boldsymbol{N}_s\}$，一个目标社交网络$G_t\{\boldsymbol{A}_t, \boldsymbol{N}_t\}$，一个可选的对齐偏好矩阵$\boldsymbol{H}$。

输出：输出一个$n_2 \times n_1$的对齐矩阵$\boldsymbol{S}$，其中$\boldsymbol{S}(x, a)$表示了源网络用户x和目标网络用户a的相似度。

上述定义中有一个可选的$n_2 \times n_1$的输入矩阵$\boldsymbol{H}$，$\boldsymbol{H}$中的每个元素都反映了两个输入网络中两个用户节点对齐的可能性。它蕴含了社交网络中的先验知识，如果缺少这种先验知识，就将$\boldsymbol{H}$的每个元素都设为相等。

3 SPUAL

这一节先简单的描述 FINAL[7]，然后在此基础上提出一种新的无监督用户对齐方法 SPUAL。

3.1 FINAL

FINAL[6]的核心是对齐一致性原则，主要思想是如果这两对用户节点本身满足对齐一致性原则，那么两个输入网络的两对节点之间的对齐应该是一致的。用图 1 的例子来详细说明：图 1 中的两对节点是a和x以及b和y，如果这两对节点对满足对齐一致性原则，那么两对节点的对齐性应该是一致的。其中，对齐一致性原则包括：

(1)拓扑一致性：a和b在网络G_s是邻居关系，x和y在网络G_t也是邻居关系。

(2)属性一致性：a和x拥有相同的用户属性，b和y也拥有相同的用户属性。

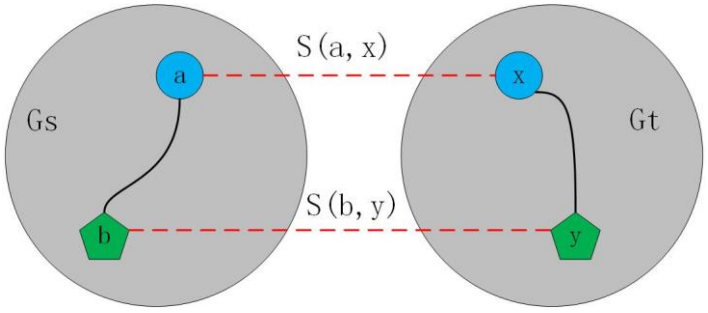

图 1 对齐一致性的说明

Fig. 1 An illustration of alignment consistency

假设如果用户a和用户x已经对齐了，如果它们的邻居b和y的用户属性相同，则b和y就有很大的机会对齐。基于这个假设，FINAL 提出以下的目标函数，希望最小化目标函数(式(1))来得到对齐矩阵$\boldsymbol{S}$：

$$J_1(\boldsymbol{S}) = \sum_{a.b.x.y}[\frac{\boldsymbol{S}(x,a)}{\sqrt{f(x,a)}} - \frac{\boldsymbol{S}(y,b)}{\sqrt{f(y,b)}}]^2 \times [\boldsymbol{A}_1(a,b)\boldsymbol{A}_2(x,y) \times l(\boldsymbol{N}_1(a,a) = \boldsymbol{N}_2(x,x))l(\boldsymbol{N}_1(b,b) = \boldsymbol{N}_2(y,y))] \quad (1)$$

$$f(x,a) = \begin{cases} \sum_{b,y} \boldsymbol{A}_1(a,b)\boldsymbol{A}_2(x,y)\, l(\boldsymbol{N}_1(b,b) = \boldsymbol{N}_2(y,y)) & \text{如果} \boldsymbol{N}_1(a,a) = \boldsymbol{N}_2(x,x) \\ 1 & \text{其他} \end{cases} \quad (2)$$

其中，$a,b = 1,\dots,n_2$，$x,y = 1,\dots,n_1$；$l(\cdot)$是一个函数，如果里面的等式成立，这个函数等于 1 否则为 0；$f(\cdot)$是一个用户节点对规范化函数，它的计算如式(2)。函数$f(x,a)$是为了计算有多少对用户是用户x和a的邻居并且它们的用户属性一致(比如b,y)。

3.2 SPUAL

FINAL 要求用户对必须满足所有的用户对齐一致性原则，才保证他们的对齐保持一致，这其实是一种严格对齐的过程。SPUAL 的目标是实现软对齐,即当用户对不满足某些对齐一致性原则时，仍然可能保持对齐的一致。SPUAL 提出了软对齐一致性原则：

(1)拓扑一致性：a和b在网络G_s是邻居关系，x和y在网络G_t也是邻居关系。

(2)属性一致性：a和x拥有相同的用户属性，b和y也拥有相同的用户属性。

(3)拓扑属性一致性：a和b在网络G_s是邻居关系，x和y在网络G_t也是邻居关系，并且a和x拥有相同的用户属性，b和y也拥有相同的用户属性。

为了实现软对齐，SPUAL 根据新的软对齐一致性原则提出了以下目标函数(式(3))，希望最小化$J_2(\boldsymbol{S})$来得到对齐矩阵$\boldsymbol{S}$：

$$\begin{aligned} J_2(\boldsymbol{S}) = \sum_{a.b.x.y} & [\frac{\boldsymbol{S}(x,a)}{\sqrt{f(x,a)}} - \frac{\boldsymbol{S}(y,b)}{\sqrt{f(y,b)}}]^2 \times \alpha \cdot \boldsymbol{A}_1(a,b)\boldsymbol{A}_2(x,y) \\ & + \beta l\big(\mathbf{N}_1(a,a) = \boldsymbol{N}_2(x,x)\big) l\big(\boldsymbol{N}_1(b,b) = \boldsymbol{N}_2(y,y)\big) \\ & + \lambda \cdot \mathbf{A}_1(a,b)\boldsymbol{A}_2(x,y) \times l(\boldsymbol{N}_1(a,a) = \boldsymbol{N}_2(x,x))l(\boldsymbol{N}_1(b,b)\boldsymbol{N}_2(y,y)) \end{aligned} \quad (3)$$

可以看出，在式(3)的右端，SPUAL 没有将各项对齐一致性原则相乘，而是相加，这样可以有效地避免对齐的用户对可能结构和属性不一致的情况，并且对不同的对齐一致性原则赋予不同的权重，这也解决了不同的社交网络侧重点不同的问题。

由于现在社交网络的用户属性不够完善，大部分用户的属性，如年龄、性别等，可能都是相同的，直接将属性一致性和其他对齐一致性原则相加会带来很大的噪声。所以，提出新的目标函数，如式(4)所示：

$$\begin{aligned} J_2(\boldsymbol{S}) = \sum_{a.b.x.y} & [\frac{\boldsymbol{S}(x,a)}{\sqrt{f(x,a)}} - \frac{\boldsymbol{S}(y,b)}{\sqrt{f(y,b)}}]^2 \times \alpha \cdot \boldsymbol{A}_1(a,b)\boldsymbol{A}_2(x,y) \\ & + \beta \cdot \mathbf{A}_1(a,b)\boldsymbol{A}_2(x,y) \times l(\boldsymbol{N}_1(a,a) = \boldsymbol{N}_2(x,x))l(\boldsymbol{N}_1(b,b) = \boldsymbol{N}_2(y,y)) \end{aligned} \quad (4)$$

其中，$a,b = 1,\dots,n_2$，$x,y = 1,\dots,n_1$；$l(\cdot)$的计算方式和式(1)相同，$0 < \alpha,\beta < 1$；$f(\cdot)$同样也是一个用户节点对规范化函数，如式(5)所示：

$$f(x,a) = \sum_{b,y} \alpha \cdot \boldsymbol{A}_1(a,b)\boldsymbol{A}_2(x,y) + \beta \cdot \boldsymbol{A}_1(a,b)\boldsymbol{A}_2(x,y) \times l(\boldsymbol{N}_1(b,b) = \boldsymbol{N}_2(y,y)) \quad (5)$$

改进之后的函数$f(x,a)$计算的是用户x和a的用户属性一致邻居对的个数加上用户x和a的邻居对个数。注意，函数$l(\cdot)$计算的是两个输入用户节点有多少属性值相同，可将其分解如式(6)：

$$l(\boldsymbol{N}_1(a,a) = \boldsymbol{N}_2(a,a)) = \sum_{k=1}^{K} \boldsymbol{N}_1^k(a,a)\, \boldsymbol{N}_2^k(x,x) \quad (6)$$

将式(6)代入带式(4)和式(5)，可以得到式(7)和式(8)：

$$J_2(\boldsymbol{S}) = \sum_{a.b.x.y} [\frac{\boldsymbol{S}(x,a)}{\sqrt{f(x,a)}} - \frac{\boldsymbol{S}(y,b)}{\sqrt{f(y,b)}}]^2 \times \alpha \cdot \boldsymbol{A}_1(a,b)\boldsymbol{A}_2(x,y) +$$

$$\beta \cdot \sum_{k.k'=1}^{K} \boldsymbol{A}_1(a,b)\boldsymbol{A}_2(x,y)\, \boldsymbol{N}_1^k(a,a)\boldsymbol{N}_2^k(x,x)\boldsymbol{N}_1^{k'}(b,b)\boldsymbol{N}_2^{k'}(y,y) \quad (7)$$

$$f(x,a) = \sum_{b,y}[\alpha \cdot \boldsymbol{A}_1(a,b)\boldsymbol{A}_2(x,y) + \beta \cdot \sum_{k=1}^{K} \boldsymbol{N}_1^k(b,b)\, \boldsymbol{N}_2^k(y,y)\boldsymbol{A}_1\,(a,b)\boldsymbol{A}_2(x,y) \quad (8)$$

为了更方便地优化目标函数，将目标函数改写为矩阵的形式。通过将对齐矩阵矢量化(比如$\boldsymbol{s} = vec(\boldsymbol{S})$)，根据矩阵点乘和克罗内克积的定义，式(7)可以被改写为式(9)：

$$J_2(\boldsymbol{s}) = \sum_{v,w}[\frac{s(v)}{\sqrt{\boldsymbol{D}(v,v)}} - \frac{s(w)}{\sqrt{\boldsymbol{D}(w,w)}}]^2\, \boldsymbol{W}(v,w) = \boldsymbol{s}^T(\boldsymbol{I} - \boldsymbol{W}')\boldsymbol{s} \quad (9)$$

其中：

$$v = n_2(a-1) + x$$

$$w = n_2(b-1) + y$$

$$\boldsymbol{W} = \alpha \cdot \boldsymbol{A}_1 \otimes \boldsymbol{A}_2 + \beta \cdot \boldsymbol{N}(\boldsymbol{A}_1 \otimes \boldsymbol{A}_2)\boldsymbol{N}$$

$$\boldsymbol{W}' = \boldsymbol{D}^{-1/2}\boldsymbol{W}\boldsymbol{D}^{1/2}$$

$\boldsymbol{W}'$是$\boldsymbol{W}$的对称归一化矩阵。$\boldsymbol{W}$的对角度矩阵$\boldsymbol{D} = \alpha\boldsymbol{D}_1 \otimes \boldsymbol{D}_2 + \beta\boldsymbol{D}_N$，其中$\boldsymbol{D}_1$和$\boldsymbol{D}_2$分别是$\boldsymbol{A}_1$和$\boldsymbol{A}_2$的度矩阵，$\boldsymbol{D}_N$的定义如式(10)所示：

$$\boldsymbol{D}_N = diag(\sum_{k.k'=1}^{K}(\boldsymbol{N}_1^k\boldsymbol{A}_1\,\boldsymbol{N}_1^{k'}\boldsymbol{l}) \otimes (\boldsymbol{N}_2^k\boldsymbol{A}_2\boldsymbol{N}_2^{k'}\boldsymbol{l}) \quad (10)$$

其中，$\boldsymbol{D}$的一些元素可能等于 0，令那些值的负二分之一次方等于 0，即$\boldsymbol{D}(v,v)^{-1/2} = 0$。通过上述描述，可将用户对齐问题的优化描述为式(11)：

$$argmin_{\boldsymbol{s}} J(\boldsymbol{s}) = m\boldsymbol{s}^T(\boldsymbol{I} - \boldsymbol{W}')\boldsymbol{s} + (1-m)||\boldsymbol{s} - \boldsymbol{h}||_F^2 \quad (11)$$

其中，$||\ ||_F^2$是弗罗贝尼乌斯范数，m是正则化参数，$\boldsymbol{h} = vec(\boldsymbol{H})$。与式(9)比较，式(11)多了一个正则化项，目的是使计算出的对齐矩阵不会与先验对齐偏好相差过大，同时也防止计算出来的对齐矩阵的元素全为 0。当没有这样的先验信息时，则令$\boldsymbol{h}$为一个均匀列向量。

3.3 算法优化

式(11)中的目标函数本质上是一个二次函数。可以将其导数设为零来求解：

$$\frac{\partial J_2(\boldsymbol{s})}{\partial \boldsymbol{s}} = 2((\boldsymbol{I} - m\boldsymbol{W}')\boldsymbol{s} + 2(1-m)\boldsymbol{h} = 0$$

则可以推导出式(12)：

$$\boldsymbol{s} = m\boldsymbol{W}'\boldsymbol{s} + (1-m)\boldsymbol{h} = m\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{W}\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{s} + (1-m)\boldsymbol{h} =$$

$$m\boldsymbol{D}^{-1/2}[\alpha \cdot \boldsymbol{A}_1 \otimes \boldsymbol{A}_2 + \beta \cdot \boldsymbol{N}(\boldsymbol{A}_1 \otimes \boldsymbol{A}_2)\boldsymbol{N}]\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{s} + (1-m)\boldsymbol{h} =$$

$$m\alpha\boldsymbol{D}^{-1/2}\boldsymbol{A}_1 \otimes \boldsymbol{A}_2)\boldsymbol{D}^{-1/2}\boldsymbol{s} + m\beta\boldsymbol{D}^{-1/2}\boldsymbol{N}(\boldsymbol{A}_1 \otimes \boldsymbol{A}_2)\boldsymbol{N}\boldsymbol{D}^{-1/2}\boldsymbol{s} + (1-m)\boldsymbol{h} \quad (12)$$

根据式(12)，可以利用迭代算法来计算$\boldsymbol{s}$，但是在迭代过程需要计算$\boldsymbol{A}_1$和$\boldsymbol{A}_2$的克罗内克积，时间复杂度为$O(M^2)$，M为社交网络关系的个数。为了降低时间复杂度，利用克罗内克积的性质$(vec(\boldsymbol{ABC}) = (\boldsymbol{C}^T \otimes \boldsymbol{A})vec(\boldsymbol{B}))$，将式(12)改写为式(13)：

$$\boldsymbol{s} = m\alpha\boldsymbol{D}^{-1/2}vec(\boldsymbol{A}_2\boldsymbol{Q}_1{\boldsymbol{A}_1}^T) + m\beta\boldsymbol{D}^{-1/2}\boldsymbol{N}vec(\boldsymbol{A}_2\boldsymbol{Q}_2{\boldsymbol{A}_1}^T) + (1-m)\boldsymbol{h} \quad (13)$$

其中，$\boldsymbol{Q}_1$和$\boldsymbol{Q}_2$是矩阵$\boldsymbol{q}_1 = \boldsymbol{D}^{-1/2}\boldsymbol{s}$和$\boldsymbol{q}_2 = \boldsymbol{N}\boldsymbol{D}^{-1/2}\boldsymbol{s}$按列顺序重新排列的$n_2 \times n_1$的矩阵。根据上述推导，SPUAL 的具体细节总结如下：

算法　基于用户结构和属性的无监督用户对齐算法 SPUAL
输入： (1)网络$G_s\{\boldsymbol{A}_s,\boldsymbol{N}_s\}$和$G_t\{\boldsymbol{A}_t,\boldsymbol{N}_t\}$； (2)可选的对齐前的偏好矩阵$\boldsymbol{H}$； (3)对齐一致性原则的权重和正则化参数$\alpha,\beta,m$； (4)迭代次数$t_{\max}$
输出：对齐矩阵$\boldsymbol{S}$ ① 计算度矩阵$\boldsymbol{D}$和用户属性矩阵$\boldsymbol{N}$； ② 初始化$\boldsymbol{s}=\boldsymbol{h}=vec(\boldsymbol{H})$ ③ while $t\le t_{\max}$ do ④　　计算矩阵$\boldsymbol{Q}_1$和$\boldsymbol{Q}_2$ ⑤　　更新式(12) ⑥　　令$t=t+1$ ⑦ 将向量$\boldsymbol{s}$按列顺序重新排列为$n_2\times n_1$的矩阵$\boldsymbol{S}$

4　实验与结果

4.1　实验设置

实验中采用了三种真实世界的具有用户属性的数据集 Flickr-Lastfm，Flickr-Myspace，Douban Online-Douban Offline。

Flickr-Lastfm：根据部分已知的对齐用户，从中抽取两个子网络，这两个子网络分别拥有 12974 个用户和 15436 的用户。将用户的性别作为一个用户属性，根据用户节点的 pagerank 得分对用户节点进行排序[1]，并将前 1%的用户节点标记为“意见领袖”，接下来 10%的用户节点标记为“中产阶级”，其余节点标记为“普通用户”。使用用户名相似度(编辑距离)来表示$\boldsymbol{H}$。

Flickr-Myspace：和 Flickr-Lastfm 相同，根据已知对齐的用户，从中抽取两个子网络，其中 Flickr 子网络有 6714 个用户，Myspace 子网络有 10733 个用户。对于用户节点属性和先验对齐偏好$\boldsymbol{H}$，使用和 Flickr-Lastfm 相同的方法。

Douban Online-Douban Offline：首先为豆瓣数据集构建一个对齐场景[19]，根据用户在社交聚会中的共现情况来构建 offline 网络，共有 1118 个用户。从包含所有离线用户的 Online 网络中提取一个包含 3906 个用户节点的子网络。将用户的位置作为节点属性，根据用户之间的度相似度来计算$\boldsymbol{H}$。

ACM-DBLP：根据部分已知的对齐用户，从中抽取两个子网络，这两个子网络分别拥有 9872 个用户和 9916 的用户。选择作者最活跃的会议作为用户属性，并且根据用户之间的度相似度来计算$\boldsymbol{H}$。

表 2 显示了以上四个数据集的基本统计数据。

表 2 数据集信息

Table 2 Data set information

Data Set	Flickr-Lastfm	Flickr-Myspace	Douban	ACM-DBLP
Number	12974-15436	6714-10733	3906-1118	9872-9916
Attributes	Sex, Pagerank	Sex, Pagerank	Location	Conference
$\boldsymbol{H}$	Username Similarity	Username Similarity	Degree Similarity	Degree Similarity

为了证明 SPUAL 的有效性，将 SPUAL 与六种设计良好或最先进的方法进行比较，包括 Regal[20]，FINAL[7]，IsoRank[4]，NetAlign[5]，UniAlign[6]，Klau's Algorithm[21]。

4.2 有效性分析

和 FINAL 一样，本文采用启发式贪婪匹配算法[22-23]求出两个输入社交网络之间的用户一对一的对齐，并根据已知的对齐计算对齐准确率。实验结果如表 3 所示。

表 3 不同数据集上不同算法对齐准确率的对比

Table 3 Accuracy on different datasets

Methods	Flickr-Lastfm	Flickr-Myspace	Douban	ACM-DBLP
Regal	0.01	0.01	0	0.003
IsoRank	0.4	0.36	0.07	0.21
NetAlign	0.43	0.45	0.01	0.03
UniAlign	0.1	0.03	0.01	0.01
Klau's Algorithm	0.38	0.4	0.07	0.12
FINAL	0.665	0.640	0.239	0.210
SPUAL	0.677	0.663	0.249	0.241

数据集 Flickr-Lastfm 和 Flickr-Myspace 的参数为$m = 0.3, t_{\max} = 30, \alpha = 0.7, \beta = 0.3$；数据集 Douban Online-Douban Offline 的参数为$m = 0.82, t_{\max} = 30, \alpha = 0.9, \beta = 0.1$。从表 3 可以看出，SPUAL 都优于其他的方法。首先，SPUAL 的准确率远高于 IsoRank[3]，NetAlign[4]，UniAlign[5]和 Klau's Algorithm[19]，这是因为仅仅利用用户的拓扑结构或者属性不能很好地对用户进行对齐。而 Regal 的准确率低是因为 Regal 在进行用户嵌入的时候，是将两个网络放在一起进行嵌入的，当两个网络的结构差距过大时这种方法的局限性很大，并且在降低时间复杂度时是用一种近似的方法来求解用户的相似性矩阵，这导致最终的结果也会有很大的误差。同时，在新的软对齐一致性原则下，进行软对齐的 SPUAL 的效果要优于 FINAL，这也验证了 SPUAL 的合理性和正确性。

4.3 参数分析

为了了解 SPUAL 的参数是如何影响性能的，本研究在数据集 Flickr-Lastfm 和 Flickr-Myspace 中通过改变参数$\alpha - \beta$来分析不同的参数对精度的影响，实验结果如表 4 所示。可以看出，在不同的参数下，SPUAL 的效果都是要优于其他方法的，这也再一次验证了 SPUAL 的合理性和正确性。从表中还可以看出结构一致性占的比重比属性一致性更大，这也符合现实中社交网络数据的真实情况。

表 4 SPAUL 在两个数据集中使用不同参数的对齐准确率的对比

Table 4 Accuracy on different parameters

Parameter($\alpha - \beta$)	Flickr-Lastfm	Flickr-Myspace
0.7-0.3	0.677	0.663
0.8-0.2	0.675	0.655
0.9-0.1	0.670	0.659

5 总结

为了解决社交网络用户对齐的问题，本文结合用户结构和属性，通过建立新的用户软对齐一致性原则，提出一种无监督方法 SPUAL 来推断用户对是否对齐，并对其进行了优化。在细节上，SPUAL 通过给不同的软对齐一致性原则不同的权重，再采用叠加的方式来判断用户对是否满足软对齐一致性原则，实现了软对齐的目标。在三个真实世界的数据集上进行实验来评估 SPUAL 算法，结果证明该算法的性能明显优于现有的方法。但目前 SPUAL 仅能处理静态网络的对齐问题，而在实际应用中用户网络往往是快速变化的，未来将基于 SPUAL 研究动态网络的用户对齐问题。

参考文献

[1] Zhang Y, Tang J, Yang Z, et al. Cosnet: Connecting heterogeneous social networks with local and global consistency. Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Sydney, NSW, Australia: ACM, 2015: 1485-1494.

[2] Chen Z, Yu X, Song B, et al. Community-based network alignment for large attributed network. Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. Singapore: ACM, 2017: 587-596.

[3] Manners H N, Elmsallati A, Guzzi P H, et al. Performing local network alignment by ensembling global aligners. 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). Kansas City, MO, USA: IEEE, 2017: 1316-1323.

[4] Smalter A, Huan J, Lushington G. Gpm: A graph pattern matching kernel with diffusion for chemical compound classification. 2008 8th IEEE International Conference on BioInformatics and BioEngineering. Athens, Greece: IEEE, 2008: 1-6.

[5] Bayati M, Gerritsen M, Gleich D F, et al. Algorithms for large, sparse network alignment problems. 2009 Ninth IEEE International Conference on Data Mining. Miami, Florida, USAI: IEEE, 2009: 705-710.

[6] Koutra D, Tong H, Lubensky D. Big-Align: Fast bipartite graph alignment. 2013 IEEE International Conference on Data Mining (ICDM). Dallas, TX, USA: IEEE, 2013.

[7] Zhang S, Tong H. Final: Fast attributed network alignment. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, CA, USA: ACM, 2016: 1345-1354.

[8] Zhang S, Tong H, Tang J, et al. ineat: Incomplete network alignment. 2017 IEEE International

Conference on Data Mining. New Orleans, LA, USA : IEEE, 2017: 1189-1194.

[9] Goga O, Perito D, Lei H, et al. Large-scale correlation of accounts across social networks. University of California at Berkeley, Berkeley, California, Tech. Rep. TR-13-002, 2013.

[10] Liu S, Wang S, Zhu F, et al. HYDRA: large-scale social identity linkage via heterogeneous behavior modeling. Acm Sigmod International Conference on Management of Data. Snowbird, UT, USA: ACM, 2014.

[11] Zhou X, Liang X, Du X, et al. Structure based user identification across social networks[J]. IEEE Transactions on Knowledge and Data Engineering, 2017, 30(6): 1178-1191.

[12] Liu L, Cheung W K, Li X, et al. Aligning Users across Social Networks Using Network Embedding. Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-16). New York, NY, USA: IJCAI/AAAI Press, 2016.

[13] Zhang J, Chen B, Wang X, et al. Mego2vec: embedding matched ego networks for user alignment across social networks. Proceedings of the 27th ACM International Conference on Information and Knowledge Management. Torino, Italy: ACM, 2018: 327-336.

[14] Page L, Brin S, Motwani R, et al. The PageRank citation ranking: Bringing order to the web. Technical Report. Stanford InfoLab, 1999.

[15] Liao C S, Lu K, Baym M, et al. IsoRankN: spectral methods for global alignment of multiple protein networks[J]. Bioinformatics, 2009, 25(12):i253-i258.

[16] Andersen R , Chung F R K , Lang K J . Local Graph Partitioning using PageRank Vectors.[J]. Foundations of Computer Science, 2006.

[17] Bayati M, Shah D, Sharma M. Maximum Weight Matching via Max-Product Belief Propagation[J]. IEEE Transactions on Information Theory, 2005(3).

[18] Zhang J, Philip S Y. Multiple anonymized social networks alignment. 2015 IEEE International Conference on Data Mining. Atlantic City, NJ, USA: IEEE, 2015: 599-608.

[19] Zhong E, Fan W, Wang J, et al. ComSoc: adaptive transfer of user behaviors over composite social network. Acm Sigkdd International Conference on Knowledge Discovery & Data Mining. Beijing, China: ACM, 2012.

[20] Heimann M, Shen H, Safavi T, et al. Regal: Representation learning-based graph alignment. Proceedings of the 27th ACM International Conference on Information and Knowledge Management. Torino, Italy: ACM, 2018: 117-126.

[21] Klau G W. A new graph-based method for pairwise global network alignment[J]. BMC Bioinformatics, 2009, 10(1 Supplement).

[22] Kollias G, Mohammadi S, Grama A. Network Similarity Decomposition (NSD): A Fast and Scalable Approach to Network Alignment[J]. IEEE Transactions on Knowledge & Data Engineering, 2012, 24(12):22.

[23] Zheng Z, Zheng L, Yang Y. Pedestrian alignment network for large-scale person re-identification[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2018.