加权社会网络中的个性化隐私保护算法

陈春玲,熊 晶,陈 琳,余 瀚
(南京邮电大学 计算机学院 软件学院,江苏 南京 210003)

摘 要:针对加权社会网络中存在一部分用户不需要隐私保护或者需要某种特殊隐私保护的现象,提出了一种基于加权社会网络数据发布的个性化隐私保护方法。将社会网络中的隐私保护分为 3 个等级:不需保护 $L=0$、防止权重包攻击 $L=1$ 和防止敏感属性泄漏 $L=2$。对于 $L\neq0$ 的节点集,通过 k-度分组和修改权重包信息对节点进行匿名,使得每个分组满足权重包 k-匿名;在分组过程中,对于存在 $L=2$ 的分组要求其敏感属性满足 l-diversity。理论分析和实验表明:攻击者不能以大于 $1/k$ 的概率识别出某节点,且不能以大于 $1/l$ 的概率推断出节点的敏感信息。该方法能够满足社会网络中各用户对隐私保护的要求,同时降低了社会网络图的信息损失。

关键词:加权社会网络;隐私保护;个性化;权重包;敏感属性
中图分类号:TP301.6 **文献标识码**:A **文章编号**:1673-629X(2016)08-0088-05
doi:10.3969/j.issn.1673-629X.2016.08.019

Personalized Privacy Preservation Algorithm in Weighted Social Networks

CHEN Chun-ling, XIONG Jing, CHEN Lin, YU Han
(School of Computer Science & Technology, School of Software, Nanjing University of Posts and Telecommunications, Nanjing 210003, China)

Abstract: There is a phenomenon that some users do not need privacy protections or they need special privacy protections in social networks, so a personalized privacy protection method to meet these requirements is proposed based on weighted social network. The privacy protection is divided into three levels: without protection ($L=0$), preventing the weight bags attack ($L=1$), and preventing the sensitive attributes disclosure ($L=2$). For nodes with $L\neq0$, k-degree grouping and weight bag modifying is used to the anonymous nodes, which makes each group meets the k anonymity of weight bag. In the process of grouping, the group with $L=2$ has to ensure l-diversity for sensitive attributes. Theoretical analysis and experiments show that attackers can't identify a node with the probability over $1/k$ and infer node's sensitive attribute with the probability over $1/l$. The method satisfies the user's requirements in weighted social network, and the information loss is reduced.

Key words: weighted social network; privacy preservation; personalization; weight bag; sensitive attributes

1 概 述

前 Sun Microsystems 的 CEO, Scott McNealy 曾说过“反正你的隐私为零,那就克服它吧”。当时这句话震惊了很多人。然而,十几年过去了,事实证明他的说法是正确的。社会网络中的用户不断增多,使得社会网络在现实生活中已非常普遍,并深刻地影响着人们的日常生活。国外的 LinkedIn、Facebook、Twitter,国内的新浪微博、QQ、微信等社交平台都拥有了庞大的客户群。用户通过社会网络与他人分享个人信息,并接触到了更多的人,而用户之间的这种互动与交流必然会产生大量社会网络数据。分析数据有助于认识网络的拓扑结构和演化过程、用户的类别划分和行为倾向等,但是也会造成用户敏感信息泄漏,可能导致个人隐私受到威胁。近些年,对于社会网络的隐私保护研究已经引起了广泛的关注。

Samarati 等[1]提出 k-匿名隐私保护模型,但仅针对关系数据,并不适用于现在的社会网络;Terzi 等[2]将节点的度作为攻击者的背景知识,提出了 k-度匿名

收稿日期:2015-11-18 **修回日期**:2016-03-03 **网络出版时间**:2016-08-01
基金项目:国家自然科学基金资助项目(11501302)
作者简介:陈春玲(1961-),男,教授,硕士,CCF 会员,从事软件技术及其在通信中的应用、网络信息安全等方面的教学和科研工作;熊 晶(1991-),女,硕士研究生,研究方向为信息安全与隐私保护。
网络出版地址:http://www.cnki.net/kcms/detail/61.1450.TP.20160801.0842.018.html

方法来防止节点再识别攻击;Zhou 等[3]提出的匿名化方法使得对任意节点的邻域子图,至少有 $k-1$ 个节点的邻域子图与其同构;Jiao 等[4]提出了个性化的社会网络数据保护方法,允许用户根据自身要求设置隐私保护程度,同时为了保护节点的敏感属性(例如疾病、收入)不被泄漏,引入了 l -diversity 思想[5-7],避免同质性攻击。上述研究成果都是基于不带权值的社会网络,在现实生活中,对边的权值分析也是具有重要意义的,它可以表示社会网络中各节点之间的连接关系,如朋友网络(朋友圈)中的边权值表示联系次数,商业交易网络中的边权值表示交易次数等。Maria 等[8]提出了基于加权社会网络的 k -匿名保护方法,使得攻击者不能以大于 $1/k$ 的概率识别出目标节点;Tsai 等[9-10]通过修改权值方法,使得任意两节点之间的最短路径条数满足 $k_1 \leqslant k \leqslant k_2$,避免任意两节点之间的最短路径泄漏问题;Chen 等[11]提出了 k -histogram-inverse- l -diversity 匿名方法,通过修改节点的权重包信息,使得同一等价类中节点的权重包信息相同;陈可等[12]提出了 k -可能路径匿名模型来保护加权社会网络中的最短路径信息泄漏问题;兰丽辉等[13-14]采用向量来描述加权社会网络,通过随机分割和聚类分割两种方式将加权社会网络表示为若干个子图,用向量表示每个子图,将所有子图的向量构成的集合作为加权社会网络的发布模型,但目前还不能很好地应用于社会网络中。

现实生活中存在一种现象:只有很少一部分人需要较高级别的隐私保护;对于度相对较大的节点集,只有其中很少一部分节点需要进行较高级别的隐私保护[4]。例如,将生活中的公众人物抽象为一个节点,该节点的邻居节点会比普通节点多,聚类系数更大,但是相对来说,他们的隐私保护要求更低。在加权社会网络的隐私保护研究中,每个人对隐私保护的要求不尽相同,而目前的研究并没有考虑到用户的这一需求,而是将所有用户节点都进行同一隐私保护,导致对某些节点的隐私保护过度,同时增加了隐私保护的代价,降低了社会网络信息的有效性。因此,文中提出一种基于加权社会网络数据发布的个性化隐私保护算法。

2　相关概念

2.1　加权社会网络

为了形象地描述社会网络结构,文中将其抽象成加权图模型 $G=(V,E,W,S,L)$。其中,V 为节点集,表示社会网络中的用户;E 为边集,表示社会网络中的节点关系;W 为边权值集,表示节点关系的紧密程度;S 为敏感标签集,表示社会网络中的用户敏感属性;L 为隐私级别集,表示社会网络中的用户隐私要求。

图 1 为某公司员工交流中心的模型,边权值表示两节点一周内联系的次数,如果两节点没有连接,则表示他们本周没有联系过。节点的上半部分标识表示员工的薪资,属于敏感属性,下半部分标识表示隐私级别,可由节点自身进行设置。例如节点 a 和 b 本周联系了 2 次,a 和 d 在本周没有联系过,且 a 的薪资为 7.5 k,隐私级别为 1。

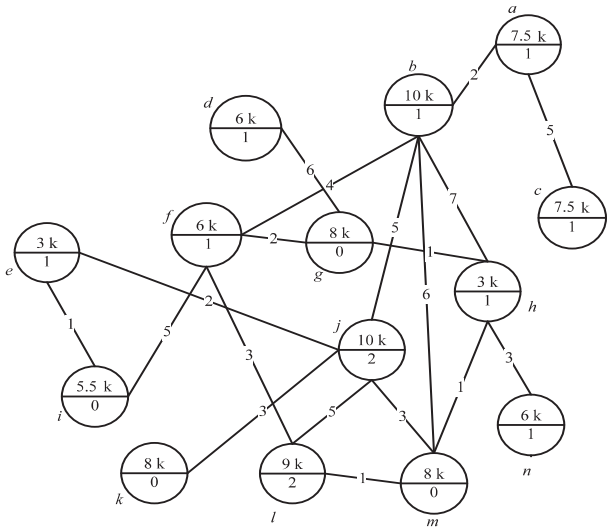

图 1　某公司员工交流中心网络图

2.2　权重包匿名

权重包:节点的权重包表示与该节点关联的边的权值序列,文中将此序列按照逆序排列。图 1 中 a 的权重包为<5,2>,b 的权重包为<7,6,5,4,2>。

权重包 k -匿名:假设攻击者已经掌握了某个节点的权重包信息,很容易识别出目标节点。为了避免节点受到此类攻击,文中采用了权重包 k -匿名的思想。通过修改权重包信息,使得每个分组中节点的权重包均一致,这样攻击者就不能以大于 $1/k$ 的概率识别出目标节点[11]。

2.3　敏感属性匿名

在对节点进行分组时,可能会出现某个分组中的敏感属性值都一样的情形。而对于 $L=2$ 的节点要求进行敏感属性匿名,使得攻击者不能获取敏感属性。为了满足用户的隐私保护要求,文中采用 l -diversity 思想[11],使得每个分组中的敏感属性值不同的节点数不小于 l,因此攻击者获取目标节点的敏感属性值的概率不会大于 $1/l$。

2.4　隐私级别

用户根据自身需求设置相应的隐私保护级别,文中将用户隐私级别 L 分三个等级:

(1) $L=0$,表示用户对隐私保护没有要求;

(2) $L=1$,表示用户希望攻击者不能通过其权重包识别出自己;

(3) $L=2$,表示用户希望攻击者不能通过其权重

包识别出自己,且不能识别出自己的敏感属性。

2.5 信息损失度 loss

社会网络的信息损失由三部分组成:边的添加与删除、节点的添加与删除以及权值的修改。式(1)表示信息损失度 loss 的计算方法。

$$\text{loss} = \alpha \frac{|\Delta\text{edge}|}{|E|} + \beta \frac{|\Delta\text{node}|}{|V|} + \gamma \frac{|\Delta\text{weight}|}{|E|} \tag{1}$$

其中,$\alpha + \beta + \gamma = 1$;$|\Delta\text{edge}|$ 为添加与删除的边总数;$|\Delta\text{node}|$ 为添加与删除的节点总数;$|\Delta\text{weight}|$ 为权值修改的边总数。

3 加权网络的个性化隐私保护算法

3.1 分组算法

图 1 表示一个原始的社会网络图 G,现在需要对 G 进行隐私保护,假设 $k=3$, $l=2$。匿名算法如下:

(1)将节点(隐私级别 $L>0$)的度按逆序排列,节点 v 记录为(v, $v.\text{degree}$, S, L),得到度序列:

$P = \{(b,5,10\text{k},1),(j,5,10\text{k},2),(f,4,6\text{k},1),(h,4,3\text{k},1),(l,3,9\text{k},2),(a,2,7.5\text{k},1),(e,2,3\text{k},1),(c,1,7.5\text{k},1),(d,1,6\text{k},1),(n,1,6\text{k},1)\}$

(2)将节点分组,选取前 k 个节点为当前组,如果在该组中有 $L=2$ 的节点,且不满足 l-diversity,则将下一节点加入到当前组中,直到满足 l-diversity,得到分组 $C_1=(b,j,f)$。

(3)将分组中所有节点度的平均值(四舍五入取整)作为该分组的标准度 $\text{avg}(C_i)$,例如 $\text{avg}(C_1)=5$。对下一节点进行分组,采用局部最优化思想,分别计算此节点的 C_{new} 和 C_{merge} 值。如果 $C_{\text{merge}} < C_{\text{new}}$,则将 h 加入到前一分组中,否则创建新的分组。C_{new} 表示创建下一个新的分组各节点的 $|\Delta d|$ 之和,$|\Delta d| = \text{avg}(C_i) - v.\text{degree}, v \in C_i$。$C_{\text{merge}}$ 表示将节点加入到前面的分组时的 $|\Delta d^*|$ 与创建下一个新的分组各节点的 $|\Delta d|$ 之和,$\Delta d^* = \text{avg}(C_{i-1}) - v.\text{degree}, v \in C_{i-1}$。例如节点 h, $C_{\text{new}}=2$, $C_{\text{merge}}=l+1$,故创建新的分组 $C_2=(h,l,a)$[4]。

(4)重复步骤 3,直到待分组的节点数小于等于 k,得到分组 $C_3=(e,c,d)$, $C_4=(n)$。

(5)分组结束后,判断最后一个分组是否满足分组要求。如果不满足,则与前一分组进行合并,将 C_3 与 C_4 合并,得到 $C_3=(e,c,d,n)$。

故图 1 的分组为 $C_1C_2C_3$。

3.2 图匿名算法

对节点进行分组后,需要对图进行匿名,使得攻击者不能识别出目标节点。为了防止攻击者通过权重包识别出目标节点,要求每个分组满足权重包 k-匿名特性。

(1)分组后,每个分组首先要求节点度相同,这样才可以达到权重包相同。故对于 $\Delta d \neq 0$ 的节点需要进行 k-度匿名,使得每一组的节点度都相同。对分组中的各节点进行 k-度匿名会出现两种情况:

①对于 $\Delta d>0$ 的节点集 N_1,优先选择在 N_1 中的两个节点之间添加噪声边(不存在边关联),权值为 0。如果 $|N_1|=1$ 或者 N_1 节点集中的任意两节点都存在边,则添加噪声节点,此时噪声节点的敏感属性值与该节点相同,隐私级别 $L=0$,噪声边权重为 0。

②对于 $\Delta d<0$ 的节点集 N_2,优先选择删除 N_2 中的两个节点之间的关联边。如果依然存在 $\Delta d<0$ 的节点,则删除其与隐私级别 $L=0$ 的节点之间的关联边。如果不存在这样的边,则删除其中一条边,并将该边关联的另一节点连接到噪声节点上,权值不变。

(2)权重包的匿名:对于每个分组,根据各边权重出现的频率,选取前 n 个值作为该分组的权重包标准,$n=\text{avg}(C_i)$。修改每个节点的权重包,使得分组中的各节点权重包一致,尽量不改变原来的权值。C_1 的权重包为<6,5,4,3,2>, C_2 的权重包为<5,4,3,2>, C_3 的权重包为<5>。由于每条边关联了两个节点,导致修改某个边的权值后,在匿名其关联的另外一个节点权重包时,需要再次修改边的权值。为了解决这个问题,采取泛化的方法,保存多次修改后的边权值。例如,匿名节点 b 的权重时,将边 e_{ab} 的权值改为 2,而在匿名节点 a 的权值时,需要将 e_{ab} 的权值改为 3,故将 e_{ab} 的权值修改为<2,3>,即 e_{ab} 的权值可能为 2,也可能为 3。

图 2 为 G 的发布图 G^*。粗实线表示修改了权值的边或添加的噪声边,虚线为匿名过程中已删除的边,细实线表示原始边,节点 o 为添加的噪声节点。

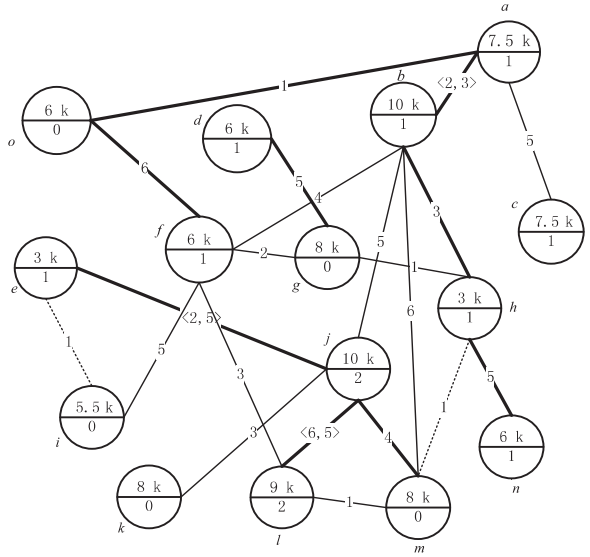

图 2 G 的发布图 G^*

3.3　算法分析

假设攻击者掌握了目标节点的权重包信息,在发布的社会网络图中,存在两种情况:第一,在权重包匿名过程中,权重包的信息被修改过,使得在发布的社会网络图中不存在这样的权重包,这样攻击者就识别不出目标节点,例如图 2 中的节点 b;第二,在权重包匿名过程中,要求同一个分组中每个节点的权重包一致,这样攻击者不能以大于 $1/k$ 的概率识别出目标节点。在分组过程中,如果存在某个节点的隐私级别 $L=2$,则要求该组中敏感属性值不同的节点个数大于等于 l,因此,攻击者也不能以大于 $1/l$ 的概率推测出目标节点的敏感属性值。综上所述,该算法是有效的。

由于在生成节点分组的过程中需要遍历每个节点,故时间复杂度为 $O(n)$。分组结束后,对于 $L\neq 0$ 的节点,需要进行 k-度匿名。在对 $\Delta d\neq 0$ 的节点集匿名过程中,需要删除和增加边,而找到最优解的策略则需要遍历该节点集,因此进行节点 k-度匿名的时间复杂度为 $O(n^2)$。例如现在需要使 $\Delta d>0$ 的节点集满足 k-度匿名,则需要遍历 $\Delta d>0$ 节点集,判断是否存在这样的两个节点,不存在边的连接。在对权重包的匿名过程中,需要将节点的权重包修改为所在分组的标准权重包,而选择权重包的标准需要的时间复杂度为 $O(n)$。因此,总的时间复杂度为 $O(n^2)$。

4　实　验

4.1　实验环境

实验在 Windows 7 操作系统上进行,CPU2.0 GHz,内存 2 GB,编程工具使用 Visual C++ 6.0。实验数据集来自微信(http://weixin.qq.com/),包含 500 个节点的网络图,1 482 条边,节点平均度为 5.928,各节点的隐私保护级别(0~2)由程序随机产生,各节点的敏感属性值由程序随机产生。

4.2　实验结果分析

算法考虑了现实情况,避免对不需要进行隐私保护的节点进行匿名,在满足用户的隐私要求下,减小了开销,降低了数据的损失。分组是关键,在进行分组时考虑节点满足 k-匿名的同时,判断当前分组是否存在要求敏感属性值满足 l-多样性的节点。k 值越大,用户的隐私保护程度就越高,造成的信息损失也会越大,执行时间越长;l 值越大,攻击者获取目标节点的敏感属性值概率越小;社会网络中节点的隐私级别分布比例,也会对社会网络的信息损失和执行时间产生重要影响。通过修改参数 k 和 l,以及社会网络的隐私级别分布,从时间和信息损失两个方面对实验结果进行分析。

图 3 和图 4 分别表示了隐私级别分布比例,以及 k 值对算法的执行时间和信息损失度产生的影响,$l=3$,$X:Y:Z$ 表示隐私级别为 $L=0$、$L=1$ 和 $L=2$ 的节点分布比例。

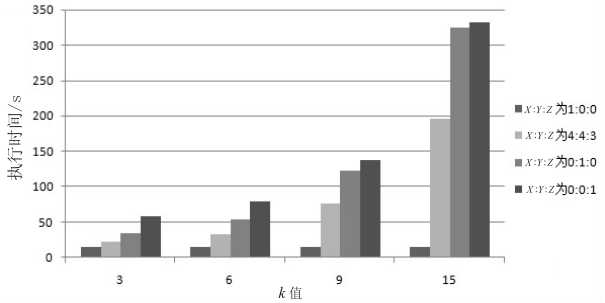

图 3　执行时间比较

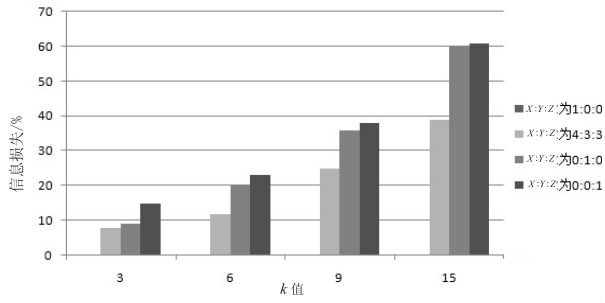

图 4　信息损失度比较

当分布比例为 1∶0∶0 时,所有的节点均不要求进行隐私保护,因此 k 值的改变对执行时间和信息损失没有影响,且信息损失度为 0。随着 k 值的增大,当分布比例为 4∶3∶3,0∶1∶0 以及 0∶0∶1 时,执行时间和信息损失不断增大。当 k 远大于 l 时,分布比例为 0∶1∶0 和 0∶0∶1 的执行时间和信息损失趋于相等。一般 k 值越大,分布比例为 4∶4∶3 的优势越能得到体现。

图 5 表示的是算法执行时间随 l 值的变化规律,其中 $k=6$,隐私级别分布比例为 4∶3∶3。随着 l 值的增大,执行时间不断增长,且增长比率也随之增大。当 l 远小于 k 时,l 值对执行时间的影响不明显;当 l 接近 k 时,l 对执行时间的影响则比较明显。

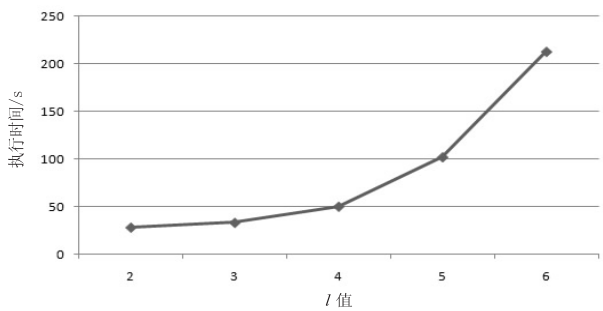

图 5　执行时间

5　结束语

对于加权社会网络的隐私保护方法还处在不断研究中。文中提出了一种基于加权社会网络的个性化隐

私保护方法,先除去不需要进行保护的节点,再将社会网络按照节点度分组,要求每个分组中的节点满足 k -度匿名;对于存在 $L=2$ 的节点所在分组,要求此分组满足 l -多样性;分组结束后,对每个分组根据各边权重出现的频率,确定该分组的权重包标准;通过修改边的权值实现节点的权重包 k -匿名,而对于需要重复修改权值的边,通过泛化方法保留每次修改后的权值。实验结果表明,算法在满足用户的隐私保护要求的前提下,降低了对社会网络结构信息的破坏。但在处理敏感属性满足 l -多样性的过程中只考虑单一敏感属性,并且对引入的噪声节点未进行匿名处理,以后还需针对上述问题不断完善。

参考文献:

[1] Samarati P, Sweeney L. Generalizing data to provide anonymity when disclosing information[C]//Proceedings of the seventeenth ACM SIGACT-SIGMOD-SIGART symposium on principles of database systems. [s. l.]: ACM, 1998.

[2] Liu K, Terzi E. Towards identity anonymization on graphs [C]//Proceedings of ACM SIGMOD international conference on management of data. Vancouver: ACM, 2008: 93-106.

[3] Zhou B, Pei J. Preserving privacy in social networks against neighborhood attacks[C]//Proc of IEEE international conference on data engineering. [s. l.]: IEEE Computer Society, 2008: 506-515.

[4] Jiao J, Liu P, Li X. A personalized privacy preserving method for publishing social network data[M]//Theory and applications of models of computation. [s. l.]: Springer International Publishing, 2014: 141-157.

[5] Zhou B, Pei J. The k-anonymity and l-diversity approaches for privacy preservation in social networks against neighborhood attacks[J]. Knowledge & Information Systems, 2011, 28(1): 47-77.

[6] Tripathy B K, Sishodia M S, Jain S, et al. Privacy and anonymization in social networks[M]//Social networking. [s. l.]: Springer International Publishing, 2014.

[7] Song Y, Karras P, Xiao Q, et al. Sensitive label privacy protection on social network data[M]//Scientific and statistical database management. Berlin: Springer, 2012: 562-571.

[8] Skarkala M E, Maragoudakis M, Gritzalis S, et al. Privacy preservation by k-anonymization of weighted social networks [C]//Proceedings of the 2012 international conference on advances in social networks analysis and mining. [s. l.]: IEEE Computer Society, 2012: 423-428.

[9] Tsai Y C, Wang S L, Hong T P, et al. [K1, K2]-anonymization of shortest paths[C]//Proc of international conference on advances in mobile computing & multimedia. [s. l.]: ACM, 2014: 317-321.

[10] Tsai Y C, Wang S L, Hong T P, et al. Extending [K1, K2] anonymization of shortest paths for social networks[M]//Multidisciplinary social networks research. Berlin: Springer, 2015: 187-199.

[11] Chen Ke, Zhang Hongyi, Wang Bin, et al. Protecting sensitive labels in weighted social networks[C]//Proc of web information system and application conference. [s. l.]: IEEE, 2013: 221-226.

[12] 陈 可,刘向宇,王 斌,等. 防止路径攻击的加权社会网络匿名化技术[J]. 计算机科学与探索,2013,7(11):961-972.

[13] 兰丽辉,鞠时光. 基于向量相似的权重社会网络隐私保护[J]. 电子学报,2015,43(8):1568-1574.

[14] 兰丽辉,鞠时光. 基于差分隐私的权重社会网络隐私保护[J]. 通信学报,2015,36(9):145-159.

2016 全国第十四届嵌入式系统学术会议征文

2016 年全国嵌入式系统学术会议(ESTC 2016)将于 2016 年 10 月 29 日 ~30 日在上海举办。ESTC 2016 以“安全可信嵌入式系统设计、验证与应用”为主题。会议论文范围包括但不限于: 1. 可信嵌入式计算; 2. 高性能嵌入式计算; 3. 移动计算与情境感知计算; 4. 智能制造与工业控制; 5. 物联网技术; 6. 信息物理融合系统; 7. 嵌入式系统结构; 8. 单片机与智能硬件; 9. 嵌入式操作系统与中间件; 10. 微处理器与微系统技术; 11. 软硬件协同设计与验证; 12. 嵌入式系统安全与可靠性技术; 13. 形式化验证与可信评估; 14. 可靠性模型与预测; 15. 嵌入式系统课程建设与教育; 16. 嵌入式系统应用技术。

投稿地址:https://easychair.org/conferences/?conf=estc2016。

投稿截止时间:英文: 2016 年 7 月 15 日,中文:2016 年 8 月 10 日,Poster: 2016 年 9 月 10 日。

论文发表:录用的英文论文将在 IEEE CPS(EI 检索)或 Springer CCIS(EI 检索)上刊载;录用的中文论文将全部推荐至《计算机学报》、《软件学报》、《通信学报》、《计算机研究与发展》、《计算机应用与软件》或《计算机技术与发展》、《计算机教育》等期刊;录用的 Poster 将全部在会上展出并在专委会网站发布。