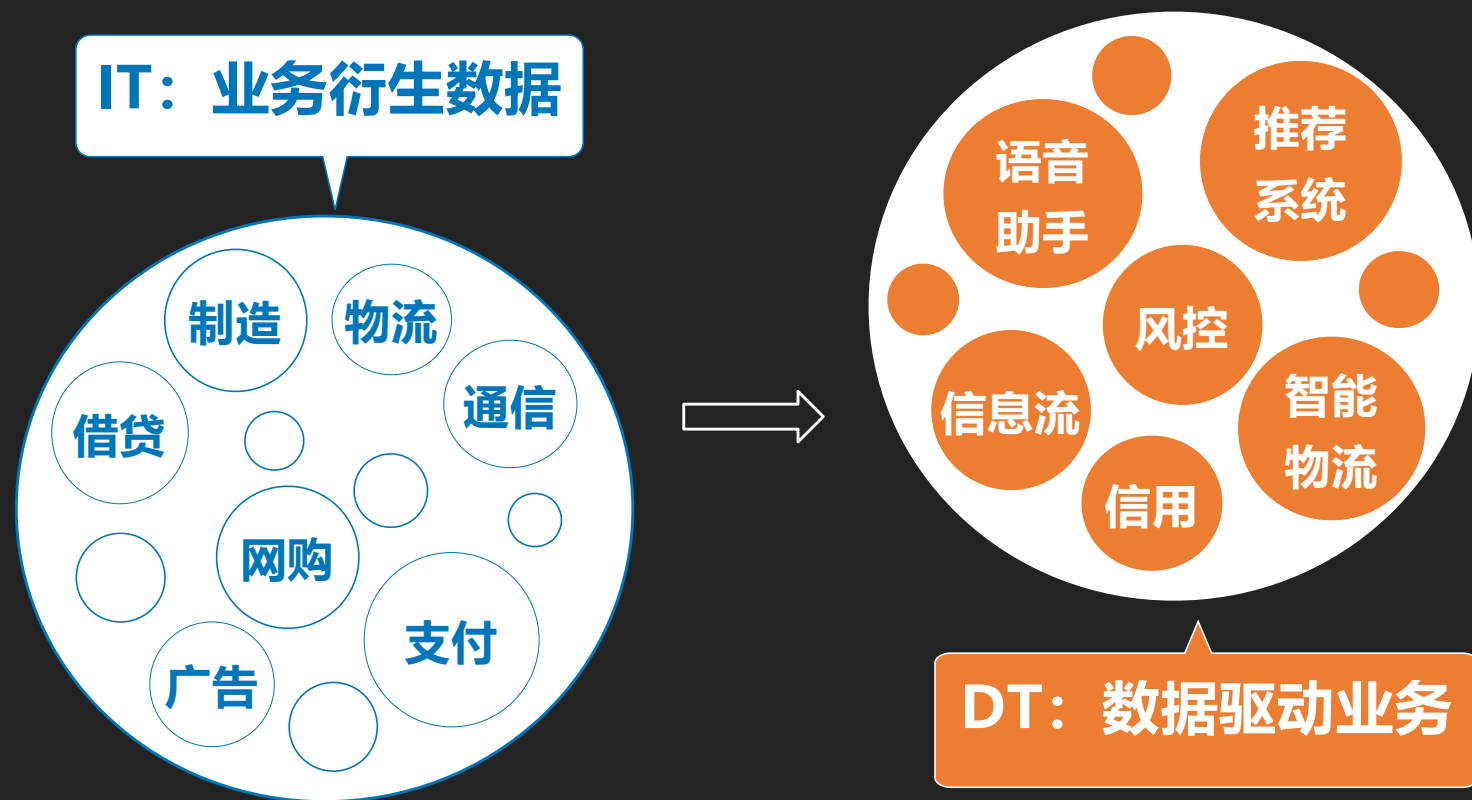


安全多方计算技术 及其在阿里大数据中的应用

阿里安全双子座实验室 黄智聪



共享经济可以最大化的利用资源，创造更多的价值

数据是新能源，需要流动起来，“数据孤岛”是难以产生价值的。



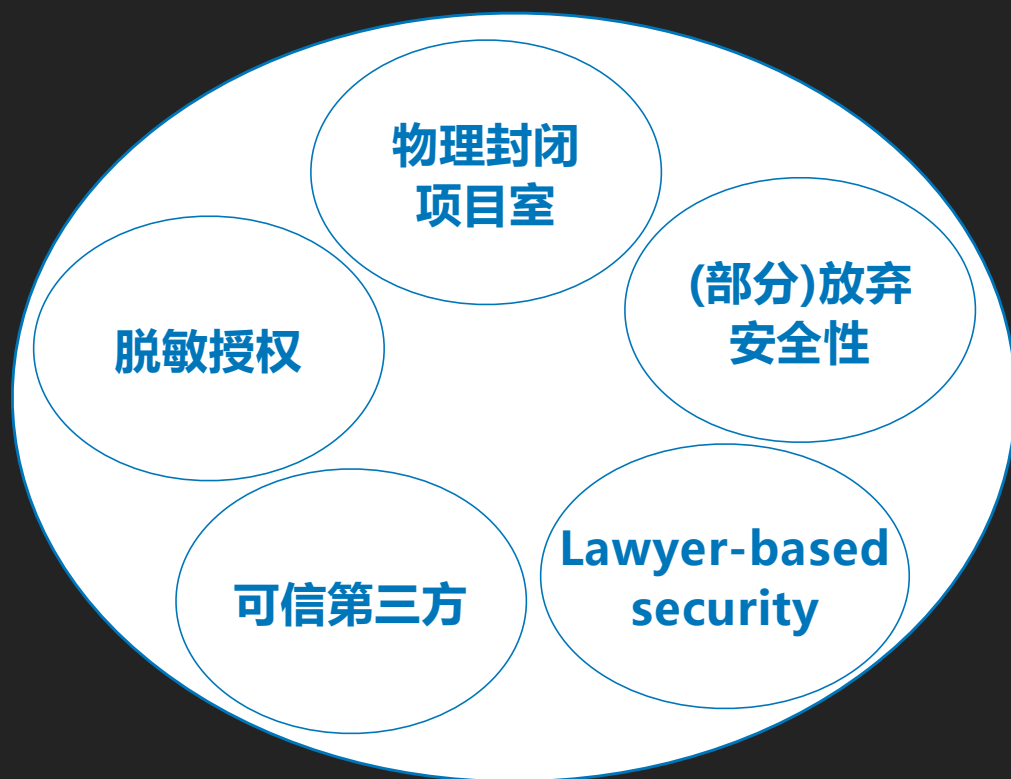
矛盾

数据与一般资源不一样，是可拷贝的，所有权和使用权很难分离

只要授权他人访问数据，对方方便具有了数据的传播能力



现有的“使用权-所有权”问题解决方案



- 整体安全强度不够
- 数据可用性差
- 迁就数据所有者, 不考虑数据使用者
 - 算法、模型参数也需要保护
 - 数据使用者也提供数据进行联合计算

思考

- 数据共享需求是为了获取数据价值
 - 例：合作医疗、合作风控、合作科研...

数据共享可以产生价值 ✓
获取数据价值 = 获取原始数据 ✕

- 能否做到数据可用但不可见？

- 数据可用不可见-技术背景
 - 安全多方计算(Secure Multiparty Computation, MPC)
- MPC在阿里大数据中的应用
 - 私有交集 (Private set intersection, PSI)
 - 机器学习 (Machine learning)

安全多方计算-引子



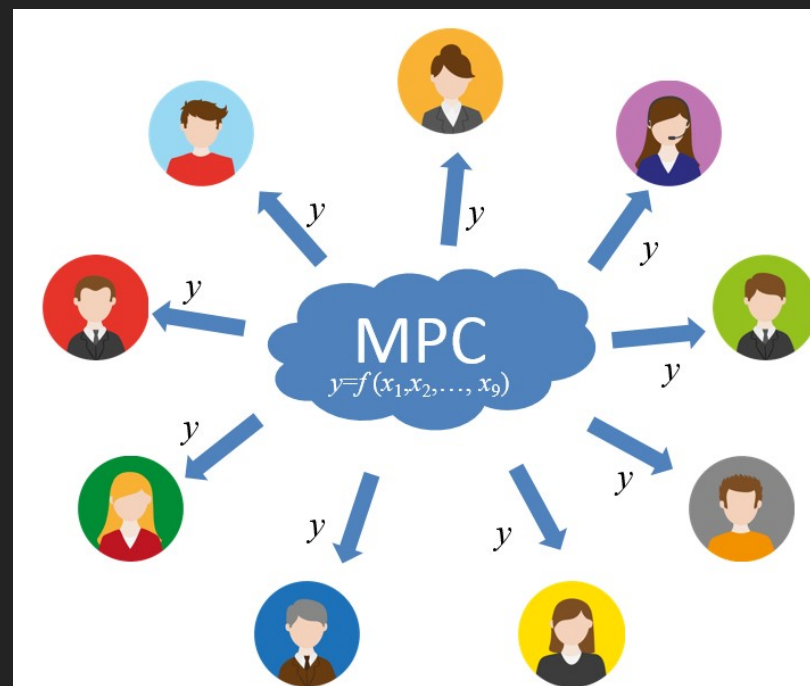
- 盲军棋游戏
 - 双方均不暴露棋面，需要第三个人当裁判
 - 如果没有裁判，这游戏两个人还能玩吗？
- 安全多方计算说：可以

安全多方计算-无限机会

- 两名作者计算彼此文章相似度，但互相不泄露文章内容
- 电商和实体店共同计算最热商品，但互相不泄露数据
- 医生研究疾病概率，但不泄露病人基因隐私
-
- 真正安全的数据交易中心成为可能

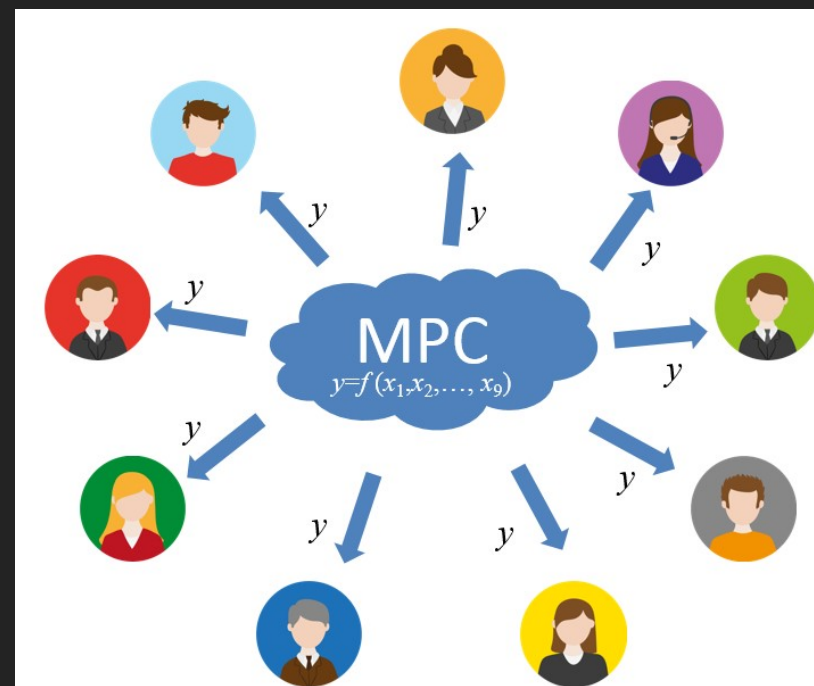
安全多方计算-定义

- 多个参与方各自输入自己的数据，共同计算某个函数 f
- 每个参与方除了 f 之外无法获得其他任何信息



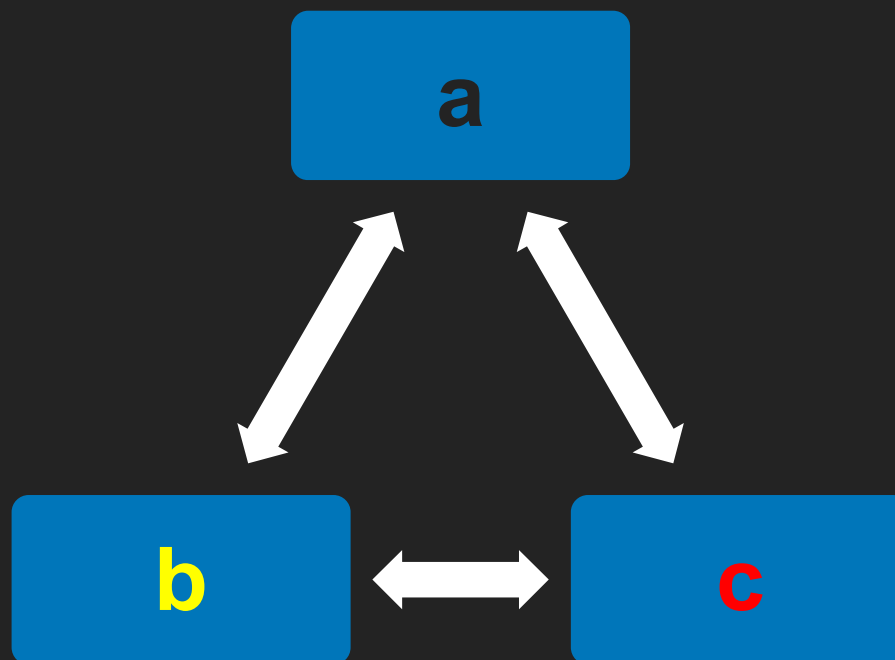
安全多方计算-扩展

- 既可以共享结果，也可以各取所需
- 例：如果结果只想让用户1知道
 - 令 $f = \text{encrypt}(\text{output}, k_1)$ 即可



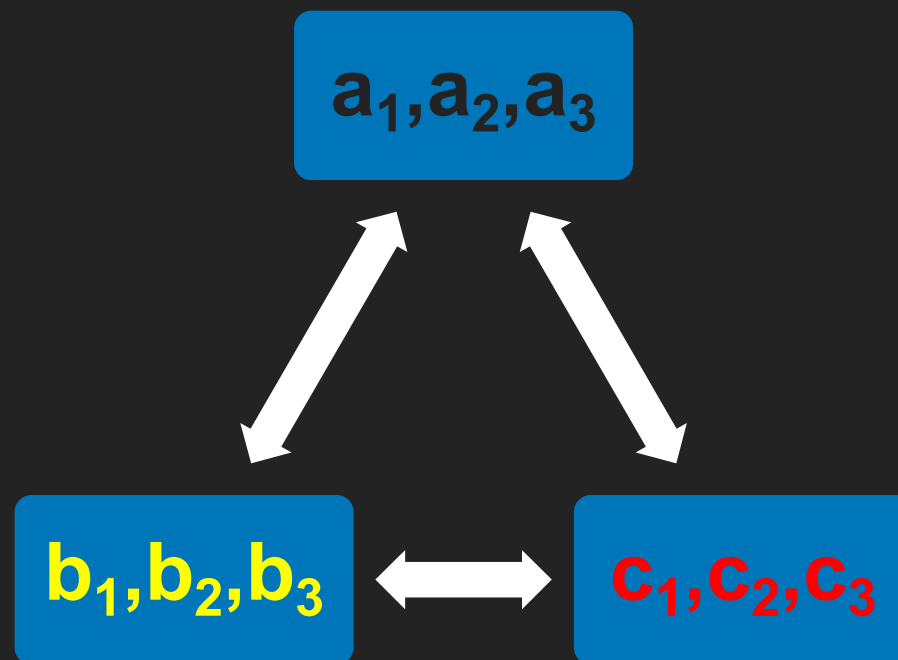
安全多方计算-方法1: Secret Sharing

- 例：三个用户a,b,c
想计算平均工资



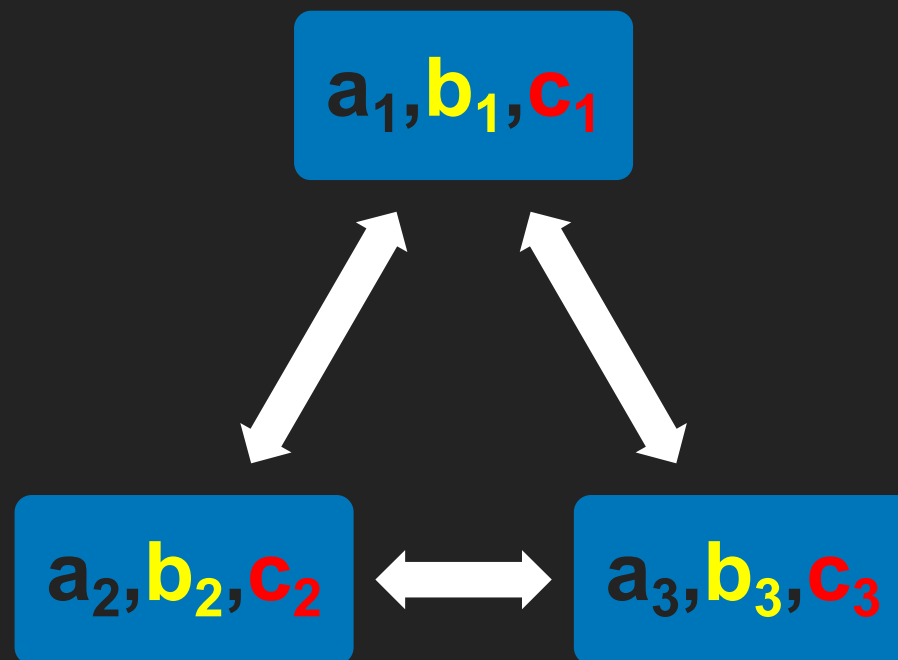
安全多方计算-方法1: Secret Sharing

- 例：三个用户a,b,c
想计算平均工资
- Step 1: Split



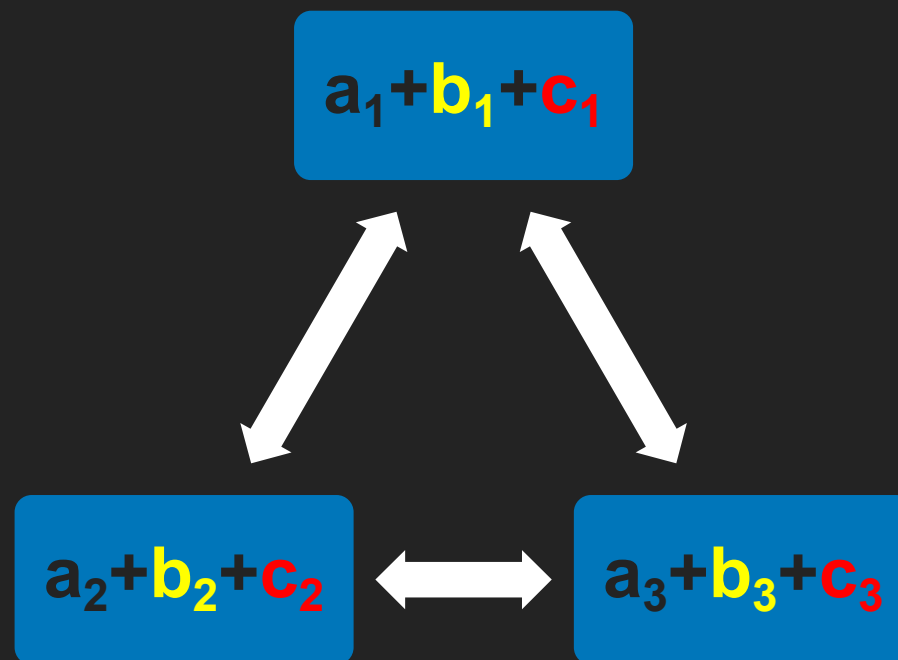
安全多方计算-方法1: Secret Sharing

- 例：三个用户a,b,c
想计算平均工资
- Step 1: Split
- Step 2: Share



安全多方计算-方法1: Secret Sharing

- 例：三个用户a,b,c
想计算平均工资
- Step 1: Split
- Step 2: Share
- Step 3: Compute



安全多方计算-方法1: Secret Sharing

- 加法可以直接本地计算
- 乘法需要额外的通信代价
 - 总的通信次数与计算函数中的乘法深度成正比, 不适合网络延迟较高的场景
- 部分简单场景应用比较容易
- State of art : SPDZ系列

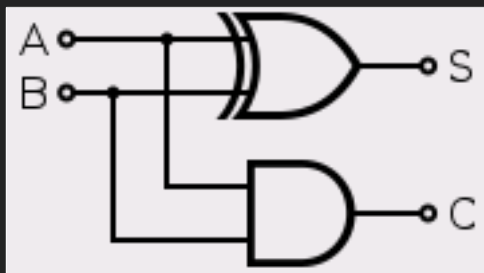
I. Damgard, V. Pastro, N. P. Smart, and S. Zakarias.
Multiparty computation from somewhat
homomorphic encryption.
CRYPTO 2012.

M. Keller, E. Orsini, and P. Scholl.
MASCOT: Faster Malicious Arithmetic Secure
Computation with Oblivious Transfer.
CCS 2016.

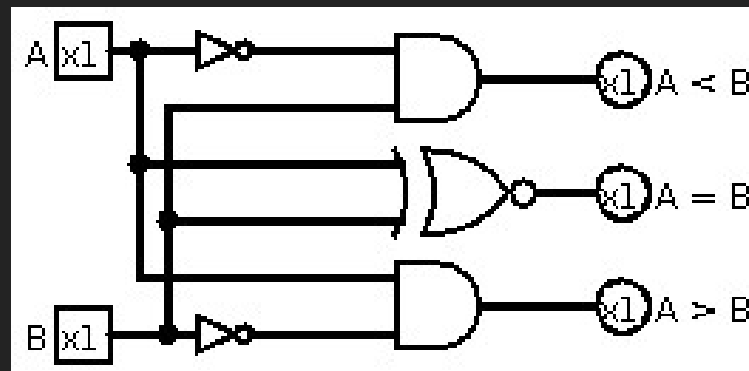
M. Keller, V. Pastro, and D. Rotaru.
Overdrive: Making SPDZ great again.
EUROCRYPT 2018

安全多方计算-方法2: Garbled Circuit

- 任何算法都可以转换为仅用门电路表示



例1: 1 bit 加法



例2: 1 bit 比较

安全多方计算-方法2: Garbled Circuit

- 每个门电路都有一个真值表

输入1 输入2	0	1
0	0	0
1	0	1

例：AND门真值表

安全多方计算-方法2: Garbled Circuit

- 每个门电路都有一个真值表
 - Step 1: 对0/1值进行随机置换

输入1 输入2	a	b
c	e	e
d	e	f

例：AND门真值表

安全多方计算-方法2: Garbled Circuit

- 每个门电路都有一个真值表
 - Step 1: 对0/1值进行随机置换
 - Step 2: 用输入加密输出

输入2 \ 输入1	a	b
	c	d
c	$Enc_{a,c}(e)$	$Enc_{b,c}(e)$
d	$Enc_{a,d}(e)$	$Enc_{b,d}(f)$

例: AND门真值表

安全多方计算-方法2: Garbled Circuit

- 每个门电路都有一个真值表
 - Step 1: 对0/1值进行随机置换
 - Step 2: 用输入加密输出

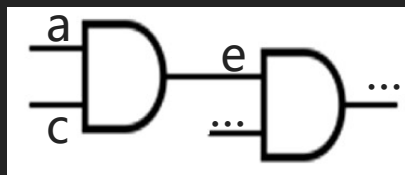
输入2 \ 输入1	a	b
	c	d
	$Enc_{a,c}(e)$	$Enc_{b,c}(e)$
	$Enc_{a,d}(e)$	$Enc_{b,d}(f)$

数据使用者输入a和c可以解密出e
但是a,c,e究竟代表0还是1是保密的

安全多方计算-方法2: Garbled Circuit

- 每个门电路都有一个真值表

- Step 1: 对0/1值进行随机置换
- Step 2: 用输入加密输出
- Step 3: 每个门的输出又作为下一个门的输入



输入1 \ 输入2	a	b
c	$Enc_{a,c}(e)$	$Enc_{b,c}(e)$
d	$Enc_{a,d}(e)$	$Enc_{b,d}(f)$

数据使用者输入 a 和 c 可以解密出 e
但是 a, c, e 究竟代表 0 还是 1 是保密的

安全多方计算-双人盲军棋解决方案

- A,B双方设计棋子比较函数 $F(a, b) \rightarrow \{\text{胜/负/平}\}$
- A将 F 转换为门电路形式
- 第一层门电路：双方共同输入加密的棋子值 a 和 b
- B从第一层开始依次运行整个电路：对每个门都使用输入解密得到输出，此输出又作为下一个门的输入
- 最后一层门电路输出即是战斗结果：胜/负/平
- **Nothing Else is revealed !**

输入协商细节此处略去

安全多方计算-方法2: Garbled Circuit

- 开山鼻祖是姚期智院士，30年来不断的改进
- 最新成果的计算代价比明文计算代价高~2个数量级，已经可以实用

Yao	Point & Permute	Row reduction	Free XOR	Fixed key AES-NI	Half gate
1986	1990	1999	2008	2013	2015

安全多方计算-方法2: Garbled Circuit

- 通信次数较少
 - 适合网络延迟较高的场景
- 通用性强, 但学习成本较高
- State of art :
Authenticated Garbling

X. Wang, S. Ranellucci, J. Katz.
Authenticated Garbling and Efficient
Maliciously Secure Two-Party Computation.
CCS 2017

J. Katz, S. Ranellucci, M. Rosulek, X. Wang.
Optimizing Authenticated Garbling for
Faster Secure Two-Party Computation.
CRYPTO 2018

- 数据可用不可见-技术背景
 - 安全多方计算(Secure Multiparty Computation, MPC)
- MPC在阿里大数据中的应用
 - 私有交集 (Private set intersection, **PSI**)
 - 机器学习 (Machine learning)

PSI 需求背景

天猫

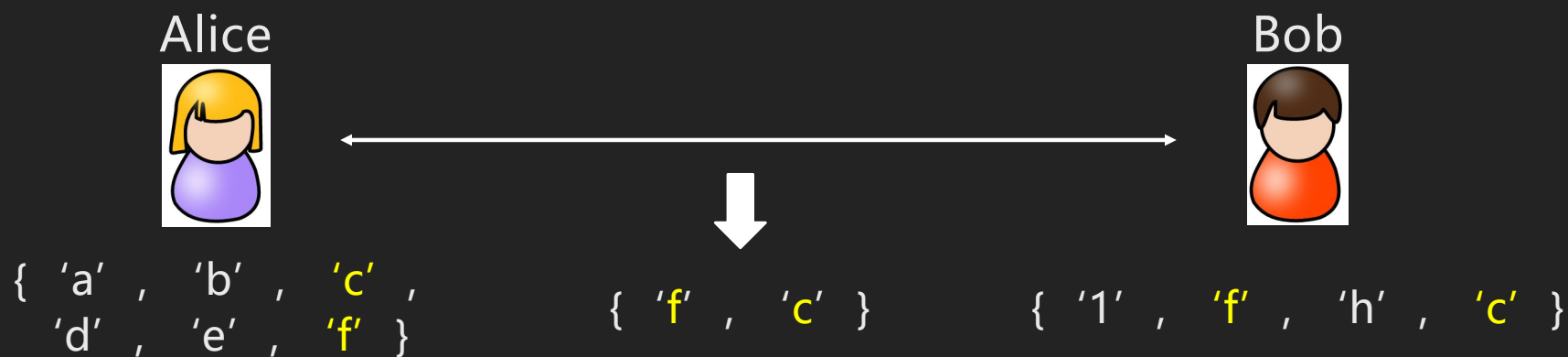
用户	购买记录
peter3@gmail.com	***
alice@163.com	***
bob@hotmail.com	***
kate@yahoo.com	***
...	***
oscar@sina.com	***

Lazada

apple@yahoo.com	***
oscar@sina.com	***
alex@yahoo.com	***
bob@hotmail.com	***
...	***
tyler@163.com	***

- 不同部门（公司）数据规范和安全等级不同，直接共享容易引发数据泄露
- PSI 可用来发现共同客户，为后续数据挖掘做准备

Private Set Intersection



任一方看不到对方交集以外其他的元素

- Alice 不知道 Bob 拥有 $\{ 'l', 'h' \}$
- Bob 不知道 Alice 拥有 $\{ 'a', 'b', 'd', 'e' \}$

单纯Hash的解决方案

天猫

用户	购买记录
peter3@gmail.com	***
alice@163.com	***
bob@hotmail.com	***
kate@yahoo.com	***
...	***
oscar@sina.com	***

Lazada

apple@yahoo.com	***
oscar@sina.com	***
alex@yahoo.com	***
bob@hotmail.com	***
...	***
tyler@163.com	***

单纯Hash的解决方案

天猫

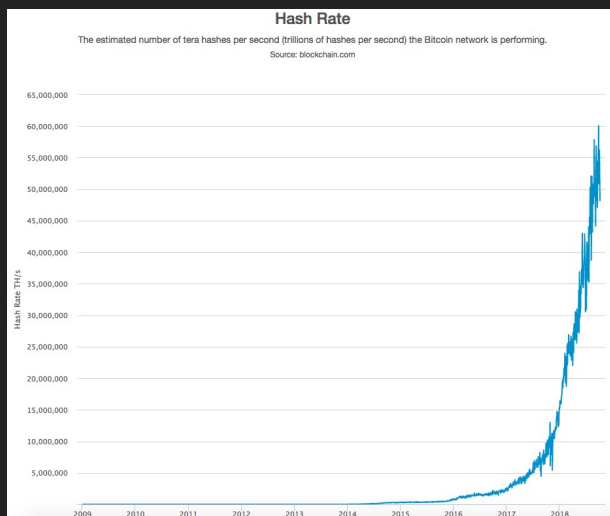
用户	购买记录
Hash_x1	***
Hash_x2	***
Hash_x3	***
Hash_x4	***
...	***
Hash_xN	***

Lazada

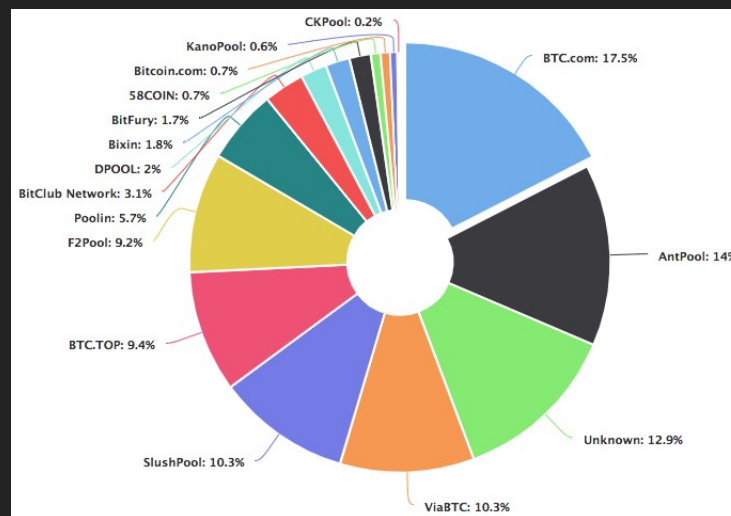
Hash_y1	***
Hash_y2	***
Hash_y3	***
Hash_y4	***
...	***
Hash_yM	***

单纯Hash足够安全吗？

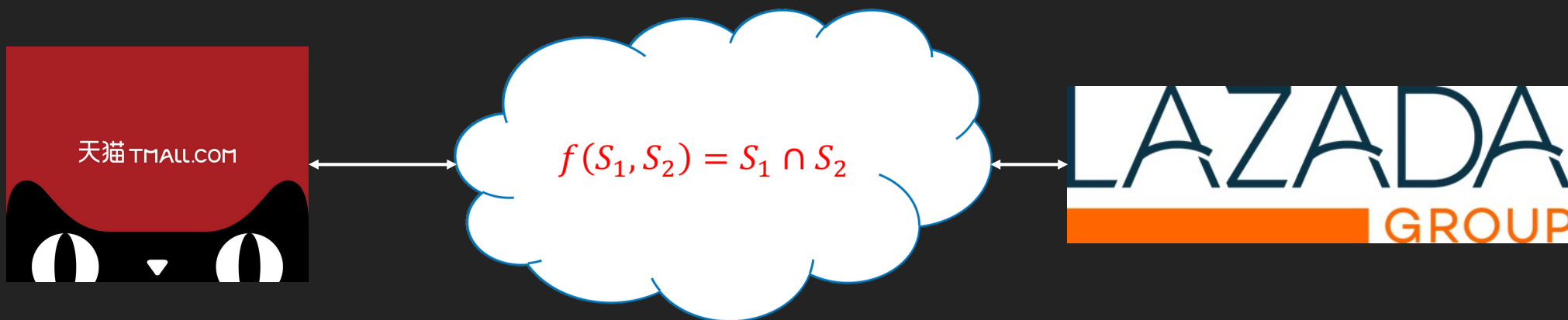
- 例子：比特币网络不断增长的hash算力
当前总算力 $\approx 2^{45}$ hash/second, 最大矿池占17.5% $\approx 2^{42}$ hash/second
- 假设用户用6位小写字母做邮箱名, 可能性： $26^6 \approx 2^{28}$
- 原文空间的熵 (entropy) 远远比不上hash算力



图片来源: <https://www.blockchain.com>



PSI-方案1: Garbled Circuit



- 一个通用的解决方案
- 可以优化的地方：降低计算复杂度，通信开销... [PSZ16]

[PSZ16] B. Pinkas, T. Schneider, M. Zohner. "Scalable Private Set Intersection Based on OT Extension" , 2016

PSI-方案2: ECDH (Elliptic Curve Diffie-Hellman)

$$X = \{x_1, x_2, \dots, x_{n_1}\}$$

Power of hash

$$H(x_i)^\alpha$$

$$\{tx_1, tx_2, \dots, tx_{n_1}\}$$

$$Y = \{y_1, y_2, \dots, y_{n_2}\}$$

DH: 从power of hash
推测不出hash和power

Power of Power of hash

$$\left(H(y_j)^\beta\right)^\alpha$$

$$\{a_1, a_2, \dots, a_{n_2}\}$$

$$\{a'_1, a'_2, \dots, a'_{n_2}\}$$

Power of hash

$$H(y_j)^\beta$$

Power of hash

$$\left(\left(H(y_j)^\beta\right)^\alpha\right)^{1/\beta} = H(y_j)^\alpha$$

Intersection: $y_j \mid ty_j \in \{tx_1, tx_2, \dots, tx_{n_1}\}$

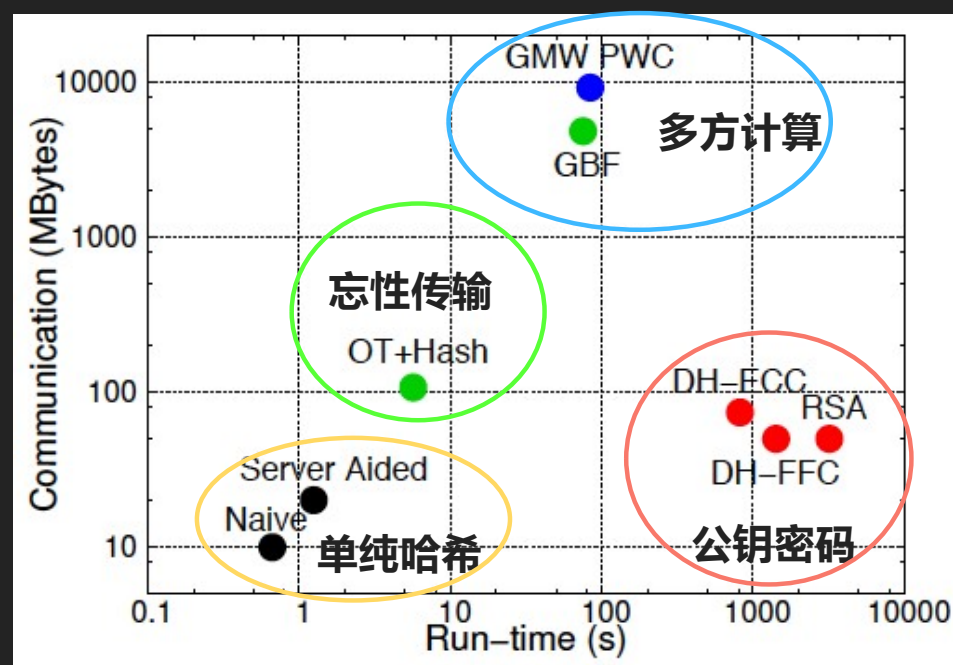
PSI 总结

● 多种不同的方案

- 单纯哈希 (原文空间熵不足时容易被破解)
- 基于多方计算 (e.g., garbled circuit)
- 基于公钥密码 (e.g., ECDH)
- 基于不经意传输 (Oblivious transfer)
- . . .

● 阿里应用场景

- 跨部门数据匹配
- 跨公司数据匹配



图片来源: [PSZ16]

- 数据可用不可见-技术背景
 - 安全多方计算(Secure Multiparty Computation, MPC)
- MPC在阿里大数据中的应用
 - 私有交集 (Private set intersection, PSI)
 - 机器学习 (**Machine learning**)

双方机器学习-需求背景



- 根据用户在天猫的消费记录，为用户在Lazada推荐类似产品
- 结合两边消费记录，更好为用户消费行为建模

双方机器学习-线性回归 [MZ17]

天猫

x : 某用户在天猫购买某些商品的数量
 y : 上述用户在Lazada购买某个商品的数量

Lazada

$x: (3, 5, \dots, 6)$

$y: 3$

$x_A: (200, 38, \dots, 4) \bmod 2^8$

y_A

$y_A: 43 \bmod 2^8$

$x_B: (59, 223, \dots, 2) \bmod 2^8$

x_B

$y_B: 216 \bmod 2^8$

x_A, y_A

模型: $Y = Xw$

训练: $w := w - \alpha X^T(Xw - Y)$

都是加法和乘法

多方计算技术: Secret Sharing

x_B, y_B

w_A

w_B

$$w = w_A + w_B \bmod 2^8$$

双方机器学习 总结

- 多种不同的方案

- 基于多方计算 (e.g., secret sharing, garbled circuit)
- 基于同态加密
- 基于安全硬件
- . . .

- 基于多方计算的试验模型

- 线性回归
- 逻辑回归
- 神经网络

安全多方计算-总结

- 基于密码学技术，将数据使用权和数据所有权分离
- 阿里巴巴安全部、蚂蚁创新技术实验室均已研发相应产品
 - 内部 - 跨部门数据共享
 - 外部 - 合作伙伴数据交流

安全多方计算-总结

- 前沿技术，目前刚开始从理论走入现实
 - 欧美相关创业公司大量涌现
 - 国内尚处于起步阶段
- 需要更多的生态实践来促进创新，推动技术进步



不止于计算 更在乎安全