区块链数据分析

1. 区块链基础

1.1 区块链架构

描述为分布式数据存储、点对点传输、共识机制、加密算法等计算机技术的新型应用模式。

层次划分：数据层、网络层、共识层、激励层（存在币才需要）、合约层、应用层。

本文从数据分析角度，将区块链描述为下图所示的**三横一纵**的结构。

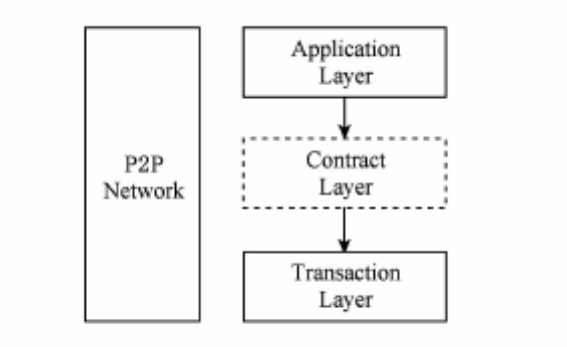

Fig. 1 The blockchain framework

图 1 区块链架构

三横：交易层（区块链1.0，比特币）、智能合约层（智能合约本质上是一段计算机程序，区块链加智能合约被称为区块链2.0）、应用层

一纵：代表区块链的运行环境是分布式的。节点的概念：在区块链分布式网络环境中，存在着大量的节点，这些节点扮演不同的角色，比特币系统中，节点有钱包、挖矿（争夺记账权）、完整区块数据存储、路由。通常一个节点因功能的不同实现不同的角色。同时实现4种角色的节点称为全节点，是整个网络中重要的支持和维护节点。文章中的节点默认为全节点。

1.2 区块链关键技术

区块链本质上是一个去中心化的账本数据库，相比传统数据库，其核心特征是**不可篡改**（可信）。在分布式环境中实现不可篡改的账本，关键问题是数据如何组织以确保不可篡改以及如何在分布式环境中对账本状态达成共识。我们将解决这两个问题的技术概括为**数据结构**和**共识机制**。

数据结构决定了区块链中用户和交易的组织形式。区块链系统中，通常采用Merkle树组织帐本中所有的账户或发生的交易。Merkle树（Hash树），上面所有的值都是Hash值。

比特币系统中，当交易发生，节点根据接收到交易的先后顺序或者手续费高低等条件将交易排在仪器，通过Hash运算得到每个交易的Hash值。这个Hash值就是Merkle树的叶子节点值。然后将这些Hash值两两拼接在一起再次Hash运算得到一个新的Hash值，自底向上不断拼接hash运算就可以得到merkle树的树根节点值（merkle root），这个根节点值代表了一段时间内被打包的所有交易的摘要信息。如果交易发生任何篡改，这个树根节点就会发生变化。

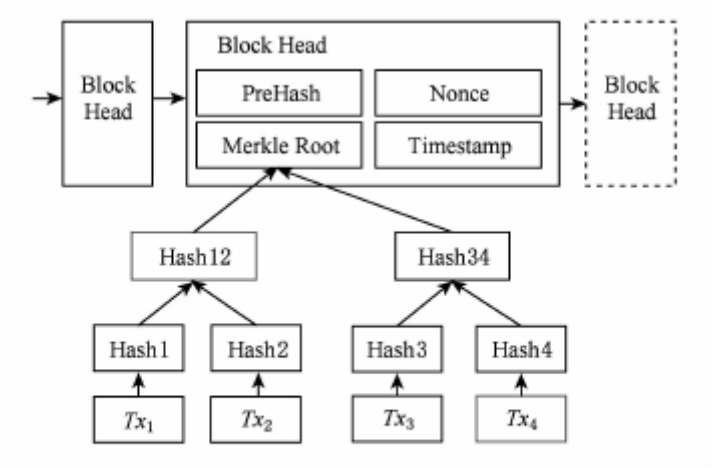

Fig. 2 Bitcoin blockchain data structure

图 2 比特币区块链结构

以太坊等区块链中，引入了账户的概念，为了方便实现智能合约。做法是将账户组织到一个merkle树上，账户代替了交易取作hash运算。当账户状态发生变化时，对应的hash值发生变化，最终导致merkle树根节点值发生变化。

树根节点值可以看成是对账本当前状态形成的一个“快照”，是区块链系统防篡改的第1步。

采用merkle树组织交易的另一个重要作用是简化支付验证（Simplified Payment Verification）。

区块链数据防篡改的第2步也是最重要的一步，是链式结构、共识机制。

因为系统存在网络时延，不同节点接收交易的顺序可能不同，由于交易支付的手续费差异，不同的节点打包交易的顺序可能不同，所以不同节点计算的系统当前状态的快照可能都不一样，那么如何在分布式环境下维护一个完整和唯一的账本，形成一个唯一的快照？普遍的做法是：根据某种条件，在分布式网络中挑选一个具有“优先记账权”的节点，然后网络中所有节点都和他保持同步。这种选择具有优先记账权的机制称为共识机制。

比特币系统中，共识的过程可以划分为三个阶段。

2. 区块链数据类型

2.1 交易数据

区块链系统中，为了匿名性，通常用地址代表账户，地址是由字母、数字组成的字符串。一个地址可能锁定了一定数据量的币或者拥有自己的存储空间和代码。不同的区块链系统可能采用不同地账本组织方式。

在比特币系统中，没有通常意义上下的账户概念，作为一种强化匿名的手段，一个用户可以拥有任意多个地址，一个交易也可能涉及多个地址。

一个地址上未被花费的交易输出数额称为一个未花费交易（Unspent Transaction output，UTXO）。通常一个地址上所有的UTXO都对应着此前某些交易的输出额，所以比特币地址上存储的UTXO都可以按照产生该UTXO的交易回溯，一直挖矿得到。挖矿所得是经过全网验证和接受的，这样，比特币系统实现了币的防伪造。

比特币系统的数据组织方式保留了交易过程中的所有细节信息，通过这些信息可以方便验证一些交易中币的来源，但是这种机制导致在比特币系统中比如验证账户余额这种简单操作变得低效。这种低效，使得基于比特币系统实现智能合约变得复杂。所以以太坊等实现智能合约的平台引入了账户的概念，账户在形式上仍是地址，但是账户引入了存储空间，用来记录账户余额、交易次数、代码等（以太坊等账户可以看成银行卡账户的一个类似物）。

以太坊系统中，账户分为两类：普通账户（和银行账户类似，用来记录用户参与交易的账户月、交易次数等信息）、智能合约账户（记录着合约的字节码等信息？）

以太坊的交易形式：交易仅发生在普通账户之间时，简单。当涉及到智能合约用户的时候，比较复杂（触发交易或内部交易，为什么会触发其他交易？）。

交易包含的部分：输入地址、输入额、输出地址、输出额、交易发生的时间戳。

2.2 合约数据

智能合约本质上是一段根据预先指定的条件被触发执行的代码。目前的区块链技术中，以太坊是最流行的智能合约平台。

智能合约涉及2类数据：实现合约相关的代码数据、合约在运行过程中被触发的交易数据。

代码数据由两种存在形式：源代码（高级语言编写，可以通过阅读了解智能合约功能的文本数据）和字节码（只对虚拟机有意义的数字串）。在以太坊中部署一个智能合约只需要提供相应的字节码即可，提供源代码可以方便使用者验证智能合约的内容。

智能合约只有相当少的一部分是可以查看源代码的，使得基于代码数据存在大量研究问题。由于代码数据中，大量存在的是字节码数据，除了通过字符相似度角度来挖掘合约间可能的关系外，能用的方法很少。一个通常的做法是将字节码通过工具反编译未虚拟机的操作码。

3. 研究现状与进展

将当前区块链数据分析的研究概括为：实体识别、隐私泄露风险分析、网络画像、网络可视化、交易模式识别、市场效应分析、非法行为检测与分析等。

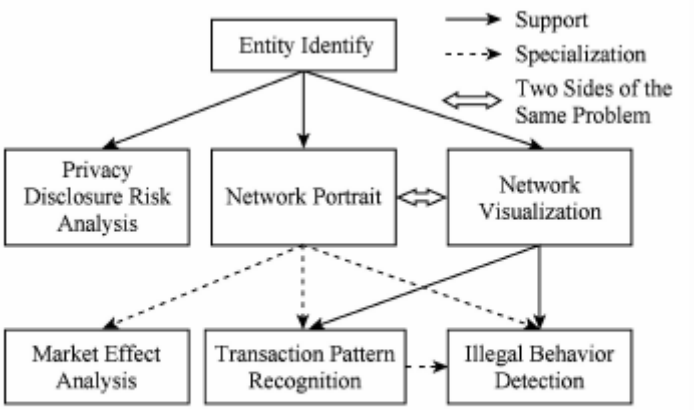

Fig. 5 Research problems and their relationship

图 5 研究问题及其相互关系

3.1 实体识别

用户是匿名的，能否从交易记录中识别出用户，哪些地址是属于同一个用户的。实体可能是用户或者机构。

文献中，通常采用启发式方法识别潜在的实体：共同输入法、找零地址法。

共同输入：在同一个交易中，将输入端的地址识别为属于同一个实体。有一种方法将比特币中的许多问题分析转化为对矩阵的分析，易于理解和实现，缺陷是随着交易的增多，矩阵的维数过大。

找零地址的识别有多个启发式方法，最直接的始把一个交易的2个输出地址中唯一的新地址视为找零地址。

3.2 隐私泄露风险分析

如果我们能够获得某个用户的一个地址信息，那么该用户的其他地址信息以及其在系统中的交易行为和账户余额等隐私信息就可能全部泄露。

隐私泄露风险分析的目标在于回答，我们如何将系统中的实体对应到真实的实体，以及如果我们拥有了用户的一些额外信息，能在多大程度上获悉用户的地址？

3.3 网络画像

面对大量的交易数据，数据中包含多少用户？用户有什么特征？这个巨大的支付网络是否具有一般的复杂网络的特征？比特币作为一种资产，它是如何在用户之间分配的，是否满足一般的经济学规律等。我们将这类研究整个网络的一些特征的研究概括为网络画像。

3.4 网络可视化

3.5 市场效应分析

加密货币价格的极强的波动性，数据分析研究的问题是解释这种极端波动性的背后驱动因素是什么。

目前广泛使用的价格影响因素分为6类：矿工因素、系统因素、用户因素、政策事件因素、网络因素、竞争替代因素。

3.6 交易模式识别

“剥离链”，“伪刷屏交易”

3.7 非法行为检测与分析

区块链技术的洗钱和诈骗的相关研究。

洗钱：使得非法所得无法追踪。“混币”服务是洗钱的一个重要工具。混币服务的想法是多个人同时输入某个交易使得基于共同输入的启发式方法失效。

诈骗：

4. 趋势与挑战

区块链是一项新兴的技术，具有颠覆许多行业的可能性，尽管目前除了加密货币以及智能合约两个典型的应用场景外，尚缺乏有足够影响力的应用，但未来区块链技术将在广阔的领域发挥基础性作用。

4.1 趋势

1. 网络特征与规律。目前利用复杂网络分析交易网络是区块链数据分析中采用较多的方法，但是将交易数据简单建构为一个复杂网络损失了大量的有用信息。所以未来在区块链数据分析的建模选择上需要考虑更多的信息，如交易的方向、数额、交易时间等，即通过交易数据构造有向网络、加权网络、时间网络等来研究各种网络特性。

   区块链数据分析的一个典型问题：网络的生成机制。另外，比特币的发行方式对网络的形成与发展有什么影响，货币在系统中的流动有什么规律，和现实金融系统相比其货币流动规律有什么特别的地方。
2. 区块链监管与价值挖掘。
3. 区块链数据分析+行业需求。

4.2 挑战

1. 基于区块链数据的网络分析和传统的网络分析有着明显的不同。典型的网络研究中，节点和连边的含义是相对明确的，但在区块链数据中则不然。构造网络时，只能选择2类节点：地址、实体。因

为一个用户拥有多个地址，所以很难识别出一个用户拥有的所有地址。所以以地址和实体作为节点，很难反映出真实的用户网络状态。

2. 区块链的去中心化和用户匿名特征让基于区块链数据分析的监管和价值挖掘充满挑战。
3. 如果区块链成为一种“底层设施”，数据实现了全行业流通，数据分析人员将面临全新的挑战：数据的全行业流通必然导致数据意义的多样化，为了通过数据分析得出有价值的结论，需要人员更加透彻的理解数据背后的实际意义。由于数据的可信和规范化，许多传统情况下需要人工完成的工作，可能会被智能合约实现的AI取代。所以未来的数据分析人员，不仅需要深刻的理解全方面的数据意义，同时需要了解智能合约的相关知识。

5. 总结