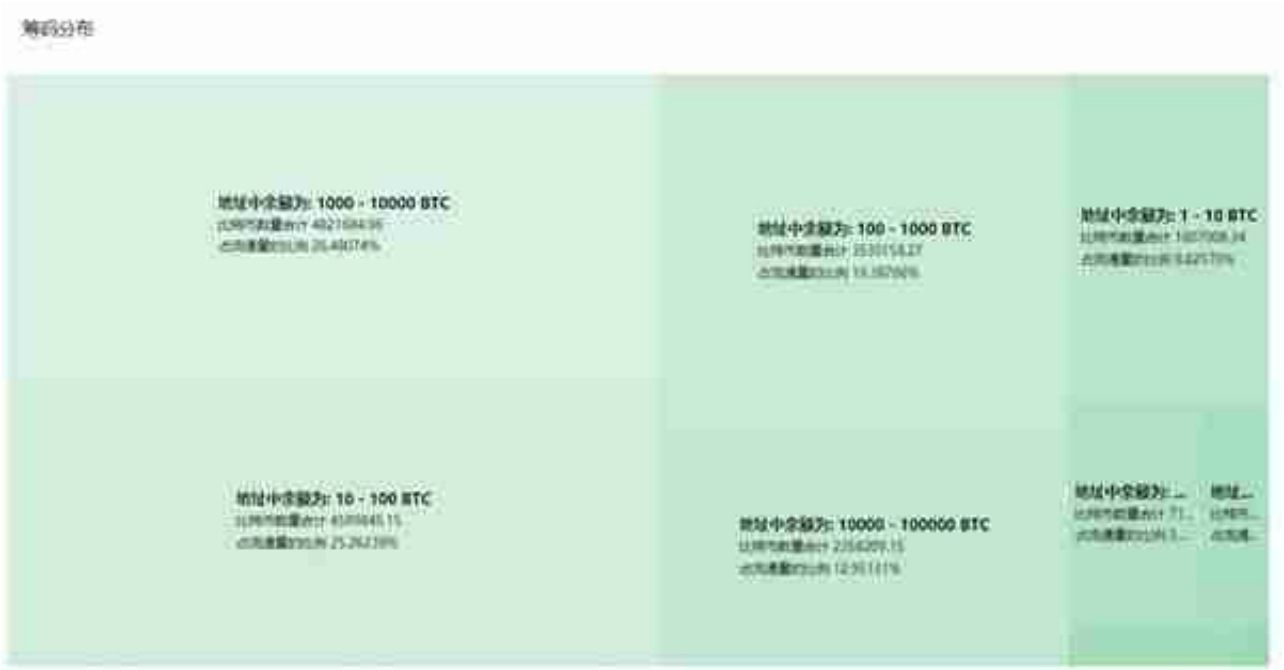
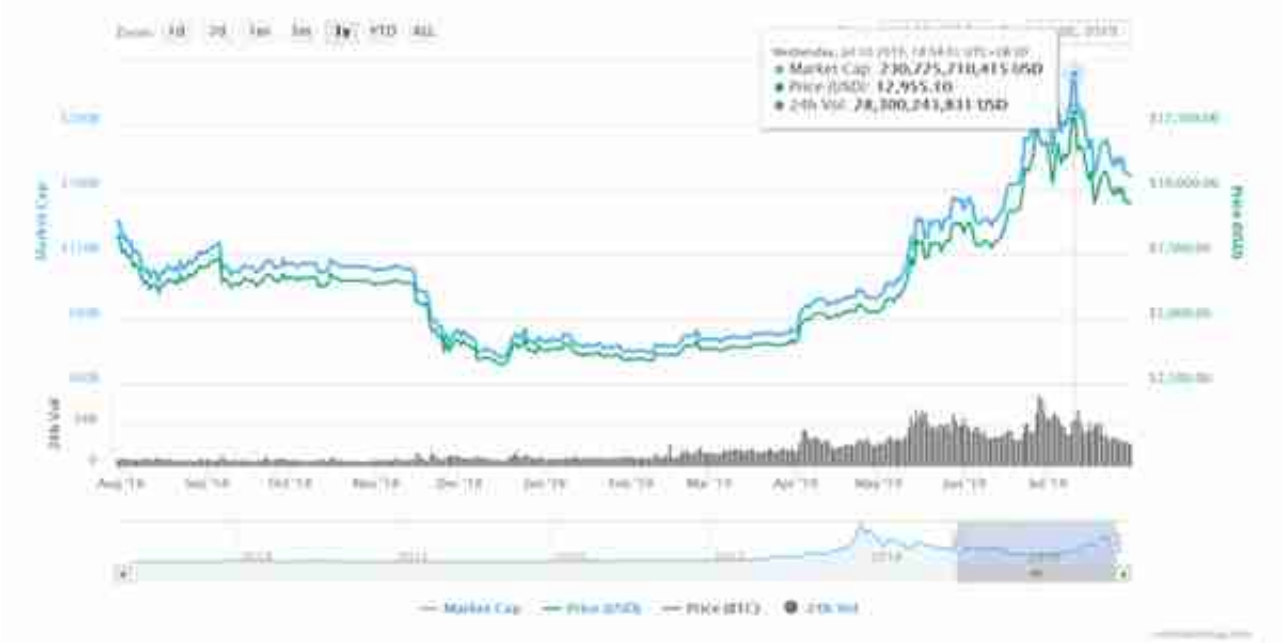
据Coinmarketcap显示，比特币上涨突破一年历史中高位达到12955 USD，目前比特币市值达到1695亿美元 ，在过去的三个月内涨幅达到76.48%。变幻的市场中，加密货币大户“巨鲸”们也在等待着再次吸筹的机会。



数据截止至2019年7月30日

6月全球研究机构Diar报告显示，自2019年起大户地址累积比特币的数量超过10万枚，数量增长了10%；7月最新比特币富豪榜数据也显示出巨鲸地址的筹码动向。目前，据Coinhills统计，24小时内BTC交易量最大的分别是：BitMEX、bitFlyer、OKEx、COINBIG，而拥有比特币排行除了属于 Binance、Bitstamp、Bitfinex 和 Huobi 的 4 个最大的加密钱包之外，许多比特币地址持有者的身份仍然不为人所知。如何追踪和挖掘这些巨鲸用户？如何及时知道巨鲸用户比特币交易动态？本篇文章将会具体讲述比特币地址挖掘方法以及相关数学原理。



01、背景

比特币是一种广为人知的加密货币，虽然每笔交易都是在链上的，数据都是可查的，但是人们还是不知道地址属于哪个人或者组织。目前，如果对于个人的话，还没有一套行之有效的方法去找出他的地址，但是对于机构来说，地址是可以通过数据挖掘的方法找出来的。

现在有一些网站已经统计了一些公布出来的地址，例如 walletexplorer.com 。这个网站统计了四大类，交易最活跃，持币量最大的网站。它们将比特币地址分成了以下几类：

1. 交易所

2. 矿池

3. 服务机构

4. 赌博网站

但这些机构会经常性的更换地址，如何找到这些地址，或者说挖掘出这些地址，就是本文的主要讨论的问题。

02、技术原理

对于比特币来说，它的地址数据挖掘，主要是依赖于比特币的交易的一些特性。

1.多输入归并

如果在一笔交易中，出现多个输入的地址，那么多个输入的地址，就属于同一个主体。在某个地址的交易中，它出现在了输入一侧，和它一起出现在输入侧的其它地址，可以被认为是属于同一个主体（比如说交易所）。

满足条件：- 输入地址数不为1

这里面的隐含的数学关系，将在后续的文章中详细介绍

例如，下图所示交易中，在输入侧（图中左侧）共有5个地址，通常情况下，可以认为该5个地址属于同一主体。

交易查询链接：https://chain.info/d654064effe87232c30de246eb92732d9313c95e7c08078c7e0551ccb388539d

2.转账与找零

如果一笔交易中，出现了有且只有2个输出地址的时候，并且这两个地址都不是输入地址时,其中一个地址是接收转账，那另一个就是找零地址。那么这个找零地址的主体，应该和输入方是同一个人。

这个推理的逻辑其实是，比特币的找零机制。在默认的情况下，找零会出现在一个新的地址中。

满足条件：

1. 输出地址数为2

2. 输入地址数不为2

3. 输入地址和输出地址不能相同

4. 其中一个输出地址的btc数，必须是拥有4位以上小数的值

5. 另一个输出地址，不能在以往的（多输入或者转账与找零地址中）地址的集合中

例如，下图所示交易中，在输出侧（图中右侧）有且只有有2个地址，且输入侧有85个地址。上一个例子中我们已经知道了那输入侧的85个地址属于同一个主体，那么通过这个规则，输出侧中拥有4位小数的地址，和那85个地址属于同一主体。

交易查询链接：https://chain.info/1e4968cac36d91c4a4294810e9d384e4b52bb73695dc23feb9459c5d89ab6e9c

3.数学原理

参考文献[1]中提出了一个概率假设，来代表不同数据源的概率模型。考虑不同类型的模型（我们将其视为独立的，以使其在计算上可解决）：

- 如果某个$t \in T_H$的所有地址$Addr_H(t)$确实属于同一主体为真的概率为p。
- 如果两个地址$\{a_i, a_j\} \in L$属于同一个主体为真的概率为q。

假设概率$P(A, T_H, L \mid p, q)$是聚类A的函数，交易T_H和L:

$$P(A,T_H,L|p,q)=\prod_{t\in T_H} p^{\prod(Addr_H(t)\subset Cl(A))}\times(1-p)^{\prod(Addr_H(t)\not\subset Cl(A))}\times\prod_{\{a,\,a'\}\in L} q^{\prod(\{a,a'\}\subset Cl(A))}\times(1-q)^{\prod(\{a,a'\}\not\subset Cl(A))}$$

其中S为比特币地址的集合。$S \subseteq Cl(A)$表示存在这样一个聚合A_j，使得$S \subseteq A_j$。

下一步，对表达式两边进行对数运算，进而得到

$$\ln P(A,T_H,L|p,q)=\sum_{t\in T_H}\prod(Addr_H(t)\subset Cl(A))\ln(1-p)+\sum_{t\in T_H}\prod(Addr_H(t)\not\subset Cl(A))\ln(p)+\sum_{\{a,\,a'\}\in L}\prod(\{a,a'\}\subset Cl(A))\ln(1-q)+\sum_{\{a,\,a'\}\in L}\prod(\{a,a'\}\not\subset Cl(A))\ln(q)$$

所提出的模型并不是为了捕捉现实世界的概率结构，而是为了系统地研究不同信息来源之间的置信度权衡。对数似然的最大化是离散的优化问题，实际上是NP问题，建议使用贪婪算法来解决它。使用一种启发式方法，对比特币网络中的所有交易进行追溯，并测试上述模型。在每个步骤，我们基于对数似然函数的值来决定是否加入交易t_j的地址$Addr_H(t_j)$的集群。

$$\Delta t_j\Big(\sum_{\{a,\,a'\}\in L}\Big)=\sum_{\{a,\,a'\}\in \hat{A}_j}\prod(\{a,a'\}\subset A_i)-\sum_{i=1}^{m_j}\sum_{\{a,\,a'\}\in \hat{A}_j}\prod(\{a,a'\}\subset A_i)=\Delta\hat{A}_j-\sum_{i=1}^{m_j}\Delta A_i$$

将Δt带入$\ln P(A, T_H, L \mid p, q)$，得到：

$$\Delta P(t_j,A,L|p,q)=ln\left(\frac{p}{1-p}\right)+\left(\Delta\hat{A}_j-\sum_{i=1}^{m_j}\Delta A_i\right)ln\left(\frac{q}{1-q}\right))$$

如果$\Delta P(t_j, A, L \mid p, q)$为正，那么合并对应于$Addr_H(t_j)$的所有集群。

03、意义

Bitcoin地址挖掘有以下几方面的作用：

1.统计各个交易所的资产数，可以更好了解交易所的持币量，和bitcoin的流通量。

2.预测市场变化。一般市场出现变化的时候，对于交易所来说，总会有大额的资金流动。通过监测各个交易所的大额流入流出，可以更好预测市场变化。

3.对于个人用户来说，可以了解机构的资产状况，便于用户作出正确的投资决策。