# 目 录

# 摘 要

近年来比特币发展迅速，是当前使用最为广泛的虚拟货币，吸引了越来越多人关注，人们可以用比特币交易，兑换现实中的货币，购买实物，甚至还有比特币结算的在线商城。考虑到比特币的潜在不安全因素，中国在 2017 年关闭了所有的比特币矿场，并禁止在中国进行比特币交易，但是世界范围内，比特币仍具有较高的关注度。本文通过对比特币的所有交易数据进行分析，进一步发现了比特币匿名性、异常交易等安全隐患。具体这里分析了比特币交易随时间变化的交易特点，讨论了比特币兑换美元的价格与交易数量的关系，对比特币交易进行了定量分析，得到比特币交易输出金额的分布，以及每个地址的交易次数分布；对产生相邻区块的时间间隔进行了分析。此外，本文还对大型交易的流向进行了追踪和分析，绘制了这些交易的流向图。

**关键词：交易次数、比特币兑换美元价格、 定量分析、大型交易、流程图**

# ABSTRACT

In recent years, Bitcoin has developed rapidly and is currently the most widely used virtual currency, attracting more and more people's attention. People can use bitcoin to trade, exchange real money, purchase physical goods, and even have Bitcoin's settlement online mall. . Considering the potential insecurity of Bitcoin, China closed all Bitcoin mines in 2017 and banned Bitcoin transactions in China. However, Bitcoin still has a high degree of concern worldwide. This article analyzes all transaction data of Bitcoin and further discovers security risks such as Bitcoin anonymity and abnormal transactions. Specifically, the characteristics of Bitcoin transactions over time are analyzed, and the relationship between the price of Bitcoin exchanged against the US dollar and the number of transactions is discussed. The quantitative analysis of Bitcoin transactions is performed to obtain the distribution of bitcoin transaction output amounts and each address. The distribution of the number of transactions; the time interval for generating adjacent blocks was analyzed. In addition, this paper also traces and analyzes the flow of large-scale transactions and plots the flow charts of these transactions.

**Key Words: Num of transaction, bitcoin’s exchange rate to USD, quantitative analysis, large-scale transaction, flow chart**

# 正 文

## 1、前言

### 1.1 背景介绍

自从 2008 年中本聪提出比特币，2009 年比特币产生以来，就引发了大众和媒体对其的广泛关注。由于比特币交易不需要真实社会身份，只需要一个公钥生成的比特币地址，所以比特币有匿名的特点。通过对比特币交易数据的分析，可以通过计算每个地址的交易次数来体现每个地址的活跃程度；比特币交易金额的分布；通过对一些大额交易进行分析，可以找到一些比特币的去向和发现一些规律，大部分比特币没有进入流通，而是存储起来了。比特币的匿名性特点也常被用于违法犯罪活动，尤其是很多混币服务和网站的存在[10]，增加了对一些交易进行追踪的困难性。对比特币交易数据的分析的应用，目前有网站[6]做了一些图表，可以很清楚的看到一些变化趋势，像是区块大小、交易数量、交易大小等，对比特币交易的追踪可以找到一些大型交易的去向和来路。[8]区块链的留言功能可以应用于公证防伪、智能合约、银行结算等方面。

未来正式的数字货币发行以后，[7]数字货币交易数据的分析的潜在应用主要是保障新经济运行和金融安全，可以用来选择数字货币的分析指标，与传统货币体系进行关联分析，可以构建数字货币分布云图。

#### 1.1.1 比特币介绍

比特币地址是由一对密钥对里面的公钥通过 hash 函数得到的，密钥对包含一个私钥和一个公钥，私钥通过椭圆曲线乘法得到公钥，公钥再通过哈希函数得到比特币地址，Hash 函数的输入是定长的数字和字母的组合，函数的输入可以是任意长度消息，不同消息的输出一定不一样。比特币系统是一个去中心化的分布式电子货币系统，传统支付系统是依赖于一个中心认证系统，是依靠一个中心机构提供的结算服务来验证并记录和处理传送过来的交易，比特币系统没有中心机构，也没有部分准备金制度来提供比特币，几乎所有的完整节点都有一份公共总账的备份即区块链，每个交易都是需要通过许许多多矿工验证、确认，最终添加到区块中，这笔交易才能成立。整套系统是建立在加密和

数字签名的基础上的，由一个接一个的区块构成了区块链，新生成的区块是链接在历史区块链上的，使用区块链即公开的账簿来防止二次交易的发生。由于比特币系统中没有用户这个概念，只有数字签名的公钥产生的比特币地址来作为接收和发送地址，每个人可以使用公私钥对产生许许多多的地址，所以这个系统具有匿名的特点，Reid and Harrigan[2]关于比特币系统的匿名性研究表明，钱包的开发者是可以通过网络信息例如 TCP/IP 这些信息确定一部分的真实用户的。

### 1.1.2 挖矿的介绍

比特币是电子货币，没有相应的实体，通过挖矿来产生新的比特币，比特币发行的数量和速度是有规律的，数量是大概每 4 年减半，最开始挖出一个区块产生的新的比特币是 50 枚，4 年以后挖出一个区块新产生的比特币数量是 25 枚，即 2012~11 月减半为每个区块奖励 25 个比特币，在 2016~7 月，挖出一个区块的比特币奖励会减半为 12.5 枚。除此之外，还规定了比特币的总量是 2100 万枚，2140 年所有的比特币会被全部挖出来，截止 507999 号区块，产出的比特币总额是 16,850,000 个比特币。

挖矿是通过不断重复的哈希运算找出一个符合相应条件的值，需要很多的计算量，为了鼓励支付大量的计算过程所产生的费用，通过产出新的比特币来作为激励，每产生一个新的区块，就生成新的比特币，所以这个过程被称为挖矿。这种计算的难度是不断调整的，挖出一个区块的时间一般是控制在 10 分钟以内，如果时间多于 10 分钟，就降低挖矿难度，如果时间小于 10 分钟，就提高挖矿的难度。随着挖矿的硬件不断更新，算力以指数级的方式不断提高，相同难度下挖出一个区块所需要的时间越来越短，只能通过提高难度来保证挖出一个矿的时间在 10 分钟左右，总的来说难度是在不断提高的。

## 1.2 主要贡献

翻阅了一些关于比特币数据的分析的文章和论文以后，发现使用的数据基本是文章发表的临近几年的交易数据，有的重点关注了一些网站运营的状况，有的是关注了比特币混币的服务，Ron 与 Shamir[4]对比特币数据的定量分析是 2012 年的，其中的很多数据已经变化了，定量分析最好是现在的，对实时性的要求虽然没有流数据那么高，但是时间跨度也不能太长。最终没有注意到有对目前所有比特币历史数据的定量分析的文章,

没有对交易数据的定量分析，就不能很好的了解目前比特币的发展规模和趋势。

为了更好的认识和了解比特币的历史交易数据和交易状况，掌握一些大型交易的交易流向，本文爬取了所有的比特币历史交易数据--从创世区块 0 到 507999 号区块的所有网页数据，从 2009 年到 2018 年 2 月，时间跨度长，然后对爬取的网页进一步处理以后，从网页中提取出来的比特币数据量大，提取以后的数据有 71.4GB，比同类论文的数据多很多。分析了一些随时间变化的数据，做成了图表，以及着重分析了一些大型交易的流向，还对交易数据进行了一些定量的分析。一些网站已经做好了一些对数据的分析工作，并做成了图表展示出来[6]，这里展示的是大多数网站没有在前端展现的内容。

同类论文主要是针对地址的匿名性特点来进行研究。也有对实体进行分析的论文。但是已经是 2014 年以前的事情了，这次的分析侧重点是定量分析，和追踪一下历史上的一些大型交易，通过追踪这些交易，分析一下交易的特点，并对目前生成实体的一种方法的可行性进行了探讨。

# 2、爬取数据

## 2.1 爬取过程和工具

这里使用的爬虫工具是scrapy，使用的编程语言是Python，工具包是numpy、pandas和scrapy。提取出来的数据的格式是csv文件，对提取的数据进行分析使用的是pandas和大数据处理语言spark。

### 2.1.1 爬取网页

有很多的网站提供比特币交易查询的服务，其中著名的一个就是网站[9],从网站[9]可以爬取所有需要的比特币交易数据网页。爬取过程中使用的语言是 Python，爬取工具包是 scrapy，写好 parse 函数、配置好以后，就可以开始爬取网页。爬取了从创世区块 0 到 507999 号区块的所有网页数据，时间是从 2009 年 1 月到 2018 年 2 月，这里爬取的都是区块链主链的网页，没有爬取支链的网页数据。

### 2.1.2 提取网页数据

对爬取的网页在本地使用 scrapy 处理，也写了一些程序用来提取出了需要的、正确的数据。 由于数据量很多，不能直接保存为一个文件，只好分批处理，每次处理一

部分网页，这样就分成了很多个文件，调整每个文件在 1GB 左右，最后 72 个 CSV 文件总共有 71.4GB。每个文件的开始和最后都是一个完整的区块交易数据。提取过程中只提取了有输入的交易，没有输入的交易没有提取出来。分块的文件方便导入到内存中使用 python 里面的 pandas 包进行相关处理。对于一些无法在内存中完成的操作，只能使用 spark 来操作，为此提取了所有 CSV 文件中相应的列，并进行相应处理合并为一个文件。

### 2.1.3 处理相关数据

处理相关的提取数据使用了 pandas 每次读取一个文件，然后提取和处理需要的信息，最后将处理以后的信息合并起来就可以得到一部分分析结果，其中着重分析了随着时间变化的交易次数和交易的比特币数量特征、比特币兑换美元价格和交易次数的关系。

由于数据量很多，一些操作无法使用 pandas 完成，如果使用的话速度也很慢。对大型交易的分析使用了大数据处理工具 spark 进行了一些操作，得到了所有不重复的地址数量，每个地址的交易次数来反应整体的活跃程度，以及找出了一些大型交易的流向图，最后使用绘图工具画出了交易流向。

## 2.2 比特币的特点

### 2.2.1 交易数据公开

比特币的交易数据是公开的，有很多可以查询比特币历史交易数据的网站，一个页面的区块数据连接着下一个区块所在的页面，交易数据的公开有利于对交易进行验证，能更好的阻止双重交易。一次比特币的交易可以有很多输入和输出，不一样的是挖矿新产生的比特币是没有输入的，只有一个输出地址。输出一般都是多于两个地址的，一个是给接收方，其中一个是找零给自己，一般的输出中会有自己的地址用来接收找零的比特币。输入的比特币大多数是之前某个区块中交易的输出。由于比特币设计的时候是使用公私钥对产生相应的比特币地址，一个用户可以有很多的比特币地址，一定程度上是可以防止社会中的真实身份泄露的。

### 2.2.2 实体的概念

将提取的数据进行进一步的处理以后，就可以得到地址表，地址表是由输入地址、输出地址、输出金额、总金额、时间和区块高度组成的，本文主要分析的是比特币定量

特征和大型交易的特点。

由于只有地址表，没有相应的其他信息，几乎不可能单独通过地址表确定一个用户所拥有的所有比特币地址和交易信息。使用过的比特币地址是指至少在一次交易中作为一个输入，这里定义实体为用户所拥有的一部分使用过的比特币地址的集合，如果一笔交易的输入有很多地址，那么可以将这些比特币地址看做是属于同一个实体的，也就是属于一个用户的，因为要使用一个地址作为输入，需要有相应的私钥才行，而私钥只有用户自己才有的，这个已经在 Reid，Harrigan[2]的文章中论述过了，同样的在 Ron，Shamir[4]的文章中也提到在比特币社区中大部分成员认为多个输入属于同一个用户的概率远大于属于不同用户的概率。实体集合可以看做属于同一个用户的一部分使用过的比特币地址集合，可以看到一般情况下一个用户拥有的使用过的比特币地址是多个实体的集合。一些实体中有相交的地址，就可以合并这几个实体为一个新的实体，因为实体间有公共的交易作为输入，而作为输入需要个人或者机构的授权才可以。有时候用户可能只使用一个地址作为输入，以后不再使用这个地址，导致一个实体表中的地址数量小于这个用户实际使用的地址数量，而且实体数多于实际用户的数量，实体数的总数可以看做用户数量的上限。

一旦一个交易中出现了两个实体的地址，就可以合并这两个实体为一个新的实体，看做是一个用户所拥有的实体，这样用户的实体数减少了，而拥有的地址数量被他人知晓的更多了。作为个人交易者，最好持有几个实体，并且不要使用不同的实体来合并交易，就可以避免被他人找到，增强了匿名性。

在后续的分析结果中，找到了这几年所有的大型交易，通过查看 15 年以后的大型交易，可以看到大额交易的输入几乎都只有一个地址，而且也没有这个地址和其他的地址混合作为输入的交易存在，这样就无法将这个地址和其他地址联系起来，无法加入到实体中。Möser,M. [10]发现，由于混币服务的存在，增大了比特币的匿名性，而且一般的混币服务操作是混合不同人的比特币地址一起作为输入，然后不同的服务商经过不同的操作流程可以得到想要的输出，其中会收取少量的比特币费用，如果将这样的输入作为一个整体，就会导致不同人的实体合并为一个，如果没有额外的信息将其分开，这样的实体就是不准确的。大型交易的输入绝大部分只有一个地址证明了人们为了加强匿名性，尽量的只使用一个输入，避免其他人将自己其他的实体和这个地址联系到一起做的努力。本文中由于数据量过多单机无法处理实体的提取，而且混币服务仍旧存在，即使

进行实体的提取，如果没有额外的信息，也会导致一部分实体的不准确，因为使用了混币服务以后，这些人的实体会被联系到一起。

# 3、定量分析的结果

## 3.1 定量分析总览

从爬取的网页中提取出数据以后，就可以得到很多有用的数据和信息。这里提取出来的交易信息都是有输入的交易，没有交易的输入没有进行提取，因为没有输入的交易一般都是挖矿产生新的比特币，而且获取的网页都是区块链的主链数据，没有支链的数据在内。将所有的地址都提取出来以后，使用 spark 编程进行去重和计数就可以得到所有不重复的比特币地址数量，截止 2018 年 2 月 7 号，总的不重复的比特币地址数量是 367,911,299 个，可以看到目前总的比特币地址数量有 3.6 亿多个。还对比特币交易随着时间的变化进行了分析并绘制了图表，这里随着时间的分析在[6]中已经有很多结果了，本文中主要针对[6]中没有的方面进行了分析，还对每个月一个比特币兑换美元的价格平均值和每个月的交易次数进行了对比分析。还对每个区块产生的时间差进行了计算，可以看到相邻两个区块产生的时间差，以及产生一个区块最短的时间。还对交易输出金额的分布进行了计算。除此之外，还对每个地址的交易次数分布进行了计算，提取了一些交易次数非常多的地址。重点讨论了大型交易的分布和流向，通过分析可以看到一些大型交易是可以联系在一起的。

## 3.2 时间特征

这里以一个月为单位分析比特币随时间的交易特征，由于 2013 年 10 月份以前的比特币数据已经有很多的分析结果了，这里没有重复这些工作，而是分析了 2013 年 11 月以后的数据，有三个分析指标，每个月的交易总次数，简称月总次数，每个月参与交易的比特币数量，简称月比特币数量，每个月一个比特币兑换美元的价格平均值，简称月平均价格。

### 3.2.1 月总次数和月比特币数量

图 1 是 2013-11 月到 2018-01 月期间每个月的比特币交易总量和交易次数的关系图。

每月交易的比特币数量随着每月交易数量变化，可以看到在很多的地方两条曲线的走势基本一致，2013-12 月到 2015-3 月（图 1 蓝色区域）交易次数和交易的比特币数量是同增同减的，14 年 1 月份交易次数增加，交易的比特币数量也增加，13 年 12 月交易数量减少，交易的比特币数量也减少，这里反映了交易的比特币数量受交易次数的影响很大。一个月交易的比特币数量不仅和交易次数有关，还和每笔交易的比特币数量有关，2015-12 月到 2016-6 月、2017-11 月到 2017-12 月（图 2 粉红色区域），两条曲线的走势是相反的，交易次数增加，交易的比特币数量反而比上个月少，交易数量减小，交易的比特币数量反而比上个月多，说明了交易的比特币数量还和每笔交易的比特币数量有关，尽管交易次数较上一月减少，但是其中一些交易的比特币数量远超过前一个月的数量，这个月交易的总的比特币数量就可能比上个月还要多。

观察图 1 里面第一条曲线的变化趋势，尽管有小的波动，但交易次数的规模是逐渐增大的，主要是因为越来越多的人开始关注比特币并进行交易。图 2 展示了 2013-11 月到 2018-1 月这 51 个月期间，每个月交易的比特币数量是 e+07 数量级的有 46 个月，占了 90.196%；e+08 数量级的只有 4 个月 2015-12 到 2016-3 月，占了 7.843%；2016 年 1 月份是 e+06 数量级的。所以绝大部分月份交易的数量都是稳定在 e+07 数量级。可见交易次数逐步上升，但每个月交易的数量是稳定的并没有大幅增加，那就说明每笔交易的比特币数量有限，而每笔挖矿都有新的比特币生产，比特币总量也在增加，说明了长期以来，有很多的比特币是不活跃在市场上的、没有参与交易的，被保存了起来，参与交易的比特币数量有限。

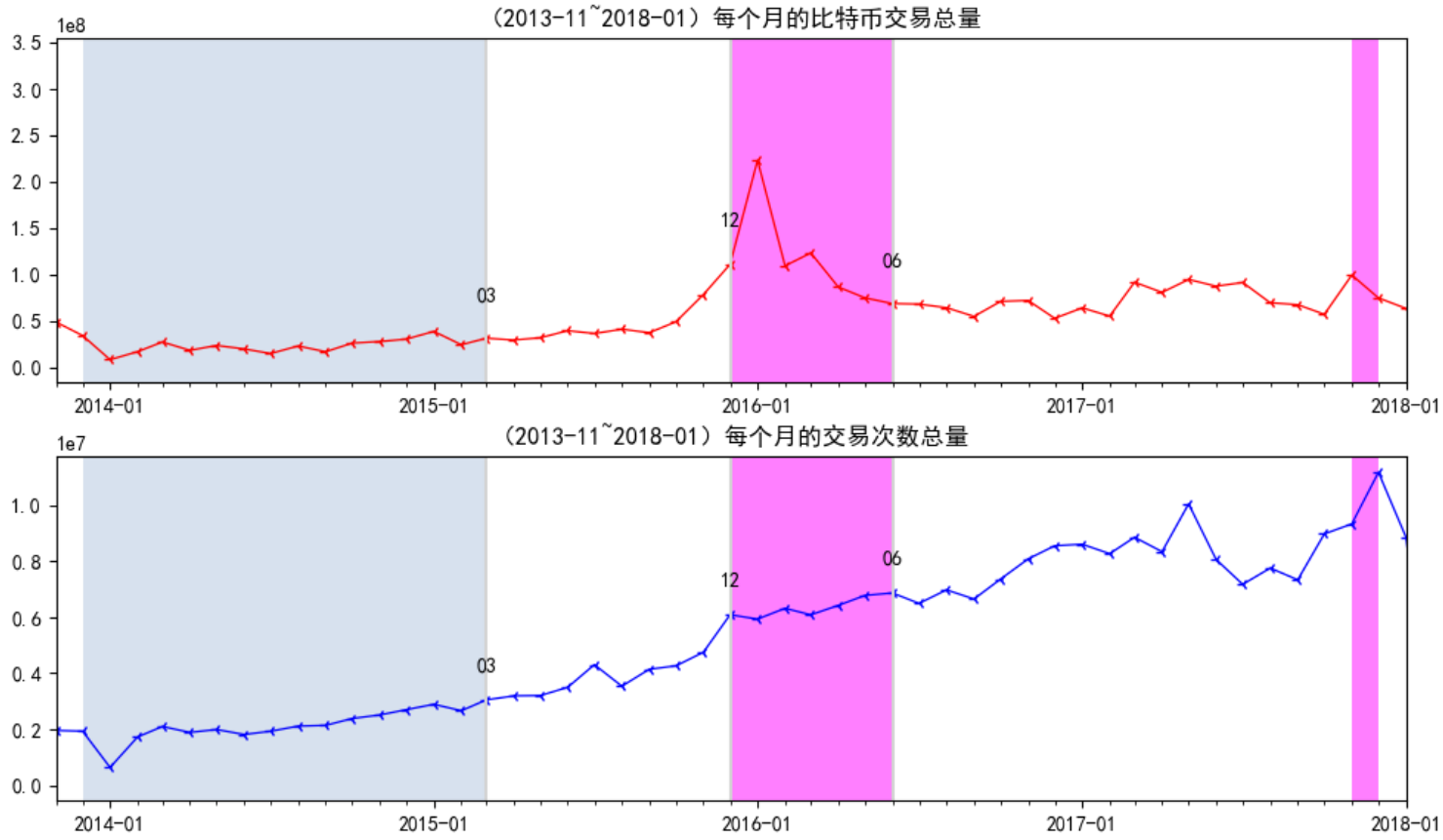


图 1

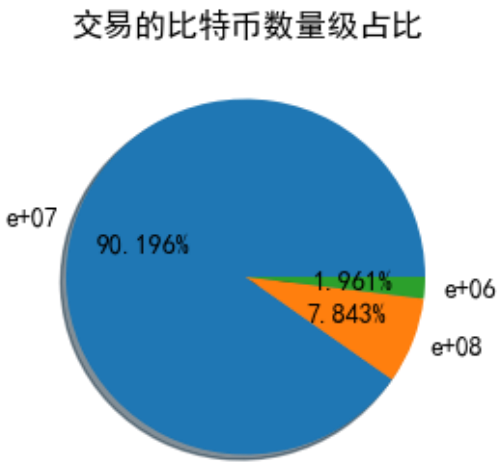


图 2

3.2.2 月平均价格以及月总次数

这里说的价格都是指一个比特币当时兑换美元的价格，在几个特殊的时间阶段，比特币兑换美元的价格急速变化时先增后减，呈现倒 V 形状，交易的比特币数量则先减后增，呈现字母 V 的形状，两者合起来看就是一个菱形。每个月的交易数量受很多因素的影响，其中最重要的因素是比特币的市场价格，由于 2017 年比特币兑换美元的价格大

幅上升，较 16 年上升了一两个数量级，如果将这些数据绘制到一张图上，2017 年之前的价格变化就变成了一条直线，看不到变化了。所以图 3 中只绘制了 2013-11~2016-12 时间段内的每个月一个比特币兑换美元的价格平均值变化、每个月交易的比特币数量，图 3 中的红色区域 2013-12~2014-02、2015-07~2015-10、2016-06~2016-08 这三个时间段，2013 年 12 月、2016 年 6 月价格上浮速度最开始很快呈指数级增长，然后一个月左右以后快速下降，相反的是，每月交易次数是先减少然后再增加，和比特币兑换美元的价格变化趋势正好相反。造成这种现象的原因主要是人的主观因素，人们总是想将自己的比特币卖个好价钱，要想买好价钱就要价格高才行，价格最开始上升很快，绝大部分人认为还没有到卖出的时候，因为比率还可能上升的更高，卖出去一枚比特币的价钱也会更高，就处于观望状态，这样的人很多，就导致了价格快速上升的时候，交易数量反而下降了，人们都在等待一个自己认为合适的时机，就是在价格很高的时候，再卖出比特币，但是到了 2014 年 1 月、2016 年 7 月价格就开始快速下降，人们担心价格继续下降的话，就卖不出好价钱了，这个时候人们就开始大量的交易来卖出比特币，这样就可以看到比率下降，交易数量反而相对上升了很多。价格变化先增后减，呈现倒 V 形状，交易的比特币数量先减后增，呈现字母 V 形状，两者合并起来看就是一个菱形。2015-07~2015-10 也是一个比较有趣的时间段，2015 年 7 月这个时间段价格并没有呈指数级增长，增长相对缓慢，这样最开始交易数量也减少了，但是人们没有等待价格继续上升一个月，一个月左右以后 2015 年 8 月就开始卖出比特币，8 月的交易数量增加了一些，但没有太高，主要是因为价格继续上升没有下降的趋势，2015 年 9 月，这个时候交易数量增长速度较 8 月更慢，应该是人们看到价格还再上升，已经上升 2 个月了，想继续等等再说，更多的人开始观望，9 月到 10 月价格下降了很多，人们开始大量卖出比特币，交易数量明显增加，增加速度也比较快。这幅图表明了交易次数和比特币的市场价格联系紧密。尽管每天比特币兑换美元的价格有波动，但总体来说还是价格急速上升时，处于观望状态的人还是很多的，等到价格开始下降，人们就开始大量地卖出账户中的比特币。

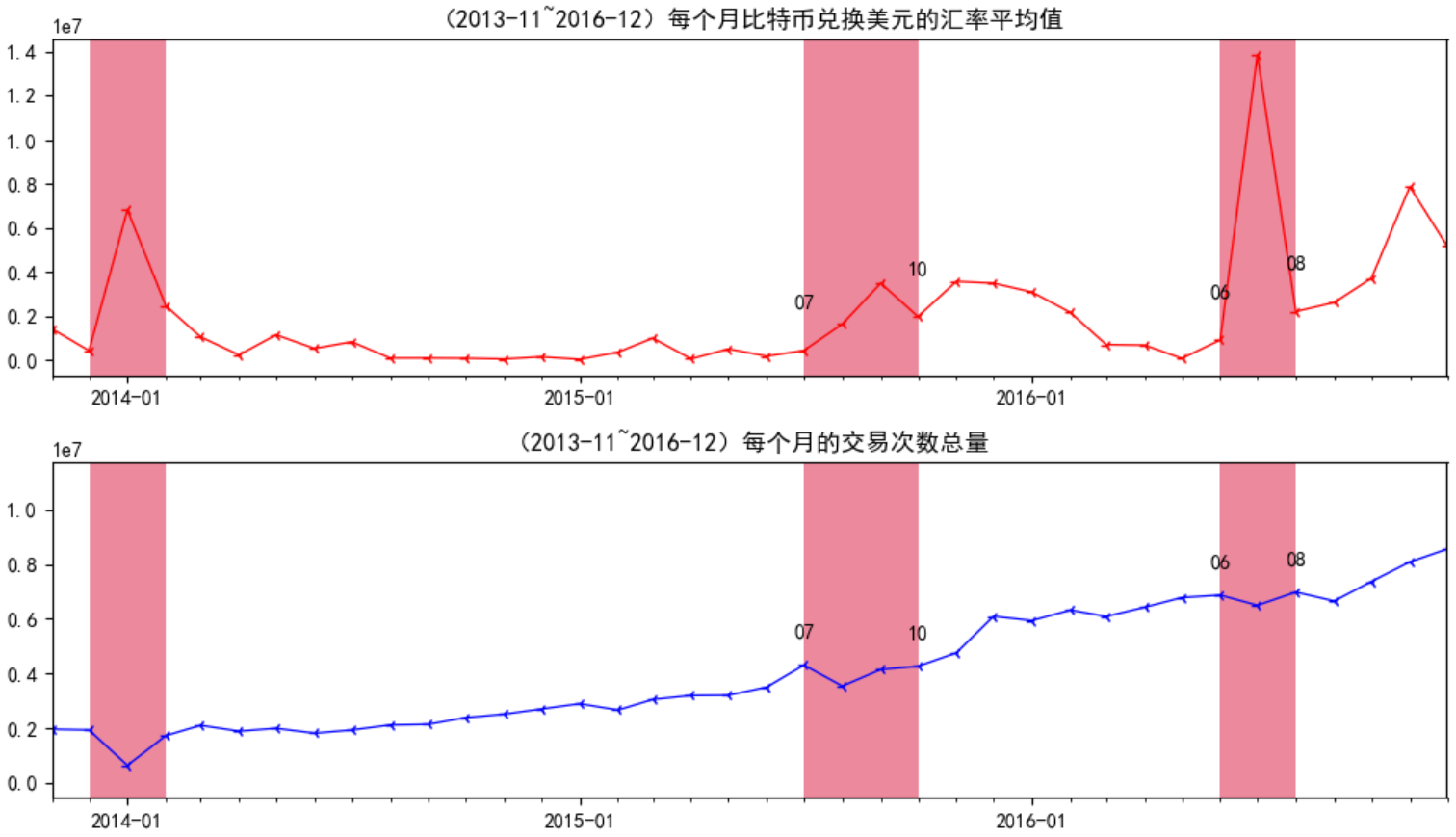


图 3

### 3.3 输出金额分布

对比特币交易规模的分布进行了汇总，这里的地址表是由输入地址、输出地址、输出金额、总金额、时间和区块高度组成的，一个交易也包含了这些内容，比特币允许的最小交易金额是 10-8 比特币，从表 1 可以看到绝大部分的交易都是小额交易，37%的交易总金额小于或等于 0.1 比特币，89%的交易总金额是小于或等于 10 比特币，97%的交易总金额是小于或等于 60 比特币。只有不到 3%的交易总金额大于 60 比特币。只有 524 个交易的总金额超过 6 万比特币，其中只有 259 个交易的总金额超过了 10 万比特币。后面的内容对大型交易做了进一步的分析。

表 2 汇总了每个地址的交易次数分布，一个地址的交易次数反映了这个地址的活跃程度，96%的地址交易次数不高于 10 次，116 个地址的交易次数大于 10 万次。有 6 个地址的交易次数超过了 100 万次。这 6 个地址中有 3 个地址是 SatoshiDICE 网站的，还

有 2 个地址是 LuckyBit 网站的。

**表格 1** 比特币交易规模的分布

| 大于或等于 | 小于 | 地址表中交易的数量 |
|---|---|---|
| 0.000000001 | 0.001 | 7,631,619 |
| 0.001 | 0.01 | 25,127,566 |
| 0.01 | 0.1 | 79,260,390 |
| 0.1 | 1 | 90,946,980 |
| 1 | 10 | 59,026,029 |
| 10 | 60 | 23,803,492 |
| 60 | 100 | 3,453,911 |
| 100 | 600 | 5,429,699 |
| 600 | 6,000 | 834,236 |
| 6,000 | 30,000 | 55,795 |
| 30,000 | 60,000 | 5,785 |
| 60,000 | 100,000 | 265 |
| 100,000 | 1,000,000 | 259 |

**表格 2** 每个地址的交易次数分布

| 大于或等于 | 小于 | 地址数量 |
|---|---|---|
| 1 | 2 | 20,791,848 |
| 2 | 4 | 317,655,378 |
| 4 | 10 | 20,271,422 |
| 10 | 100 | 8,378,605 |
| 100 | 1,000 | 773,135 |
| 1,000 | 5,000 | 37,435 |
| 5,000 | 10,000 | 1,910 |
| 10,000 | 100,000 | 1,451 |
| 100,000 | 500,000 | 101 |
| 500,000 | 1, 000, 000 | 9 |
| 1, 000, 000 | 10, 000, 000 | 6 |

## 4 收到相邻区块时间的间隔

区块链设定的是大约 10 分钟生成一个新的区块，这里对收到相邻区块时间的时间间隔进行了计算，t1-t0=difftime，用收到一个区块时间 t1 减去收到该区块前一个区块的时间 t0，得到时间差 difftime，保留生成时间是 t1 的区块高度 h，对应时间差 difftime，其中 difftime 小于 0 的有 13697 个，占据的百分比是 2.7%。时间间隔小于 0 的情况发生在一个节点接收到一个新的区块，但是在现有的区块链中没有找到它的父区块，这个区块被认为是“孤块”，只能先保存在内存池中，等接收到父区块以后再将其连接到现有区块链上。

图 4 中的横坐标是 t1-t0 中 t1 对应的区块高度 h，纵坐标是时间间隔，单位是 s，横坐标跨度是 125,000，最后一幅图是到 509000 为止，令其中间隔时间大于 10000S 的等于 10000S，时间间隔小于 0 的计数分别是 1747、3865、6628、1457。从图中可以看到 460,000 以后的区块很少有时间间隔小于 0 的情况发生了。对于时间间隔大于 0 小于 30000S，求平均值得到 585S 的时间间隔，总体来说时间是在 10 分钟左右。

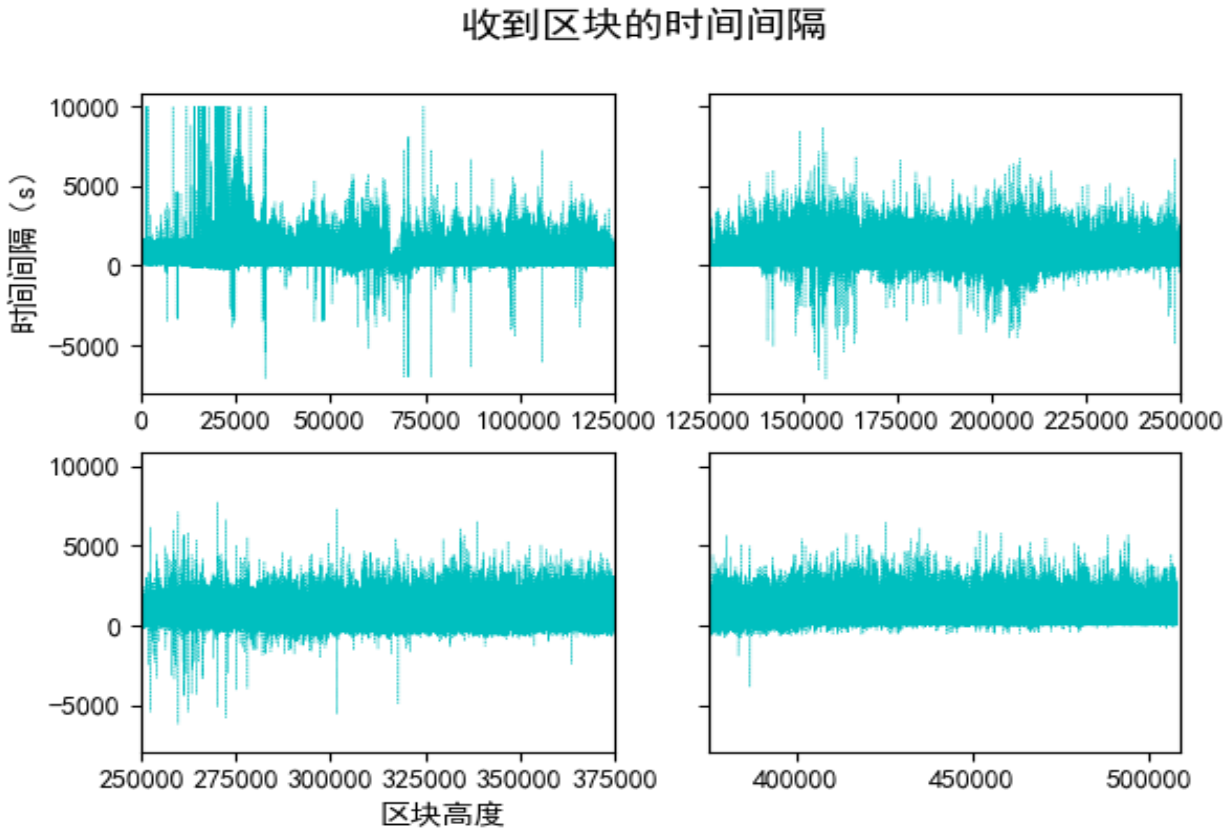


图 4

# 5 大型交易的特点

## 5.1 大于 10 万的交易

由于 2013 年之前的大型交易已经有很多分析结果了，这里分析的大于 10 万的交易是 2013 年 11 月之后的，共有 43 笔交易。最近的交易是 2015 年 11 月的，2015 年 11 月以后就没有大于 10 万的交易了。2014 年有 31 笔这样的交易，2015 年有 8 笔这样的交易。定义有联系的交易：如果一个交易的输出是另一个交易的输入，就称这两个交易是有联系的。使用这种方式来划分这些大型交易，可以将这些交易分为 6 个部分，每个部分的交易都是互相联系的。其中 1 个部分只有 1 个交易，这个地址是丝绸之路网站的地址，已经被查封了。其中交易数不小于 3 个的部分的交易图在里面，绘制的交易流向图为图 5-图 7。

## 5.2 大于 6 万小于 10 万的交易

2017 年有很多笔大于 6 万的交易，而且比较集中联系紧密，这里只分析了 2017 年 3 月以后的大于 6 万的交易。看交易是否有联系划分这些交易，可以将这些交易划分为 5 个部分，这 5 个部分的交易流向图是图 8-图 12。

## 5.3 大型交易的特点

从图中可以看到，这些交易基本只有一个输入，避免这个地址和其他实体联系在一起，通过分析大型交易的流向图，可以发现这些交易具有这些特点：基本流向是线型，尽管有分叉，但是不多；2017 年大于 6 万的交易链条一般都很长；尽管交易的输入比特币很多，但是实际给出去的比特币不多，很多还是存在原先的地址中，或者转移到其他地址中去了。

**大型交易每个部分的流程图：**

- 圆圈的大小代表地址的度中心性；
- 有向边分别指向交易的输入和输出；
- 圆中3位数字和字母的组合是比特币地址的前3位；
- 填充圆圈的颜色深浅代表该地址在这几个交易中的最高交易金额；
- 圆圈的颜色越深代表最高交易金额越多；
- 边的粗细代表输出金额的大小，这里只选取了输出金额大于一定阈值的边和顶点，小于阈值的边和顶点没有画在图中。

图5到图7是交易金额大于10万，交易有联系的部分，这些部分的交易流程图：

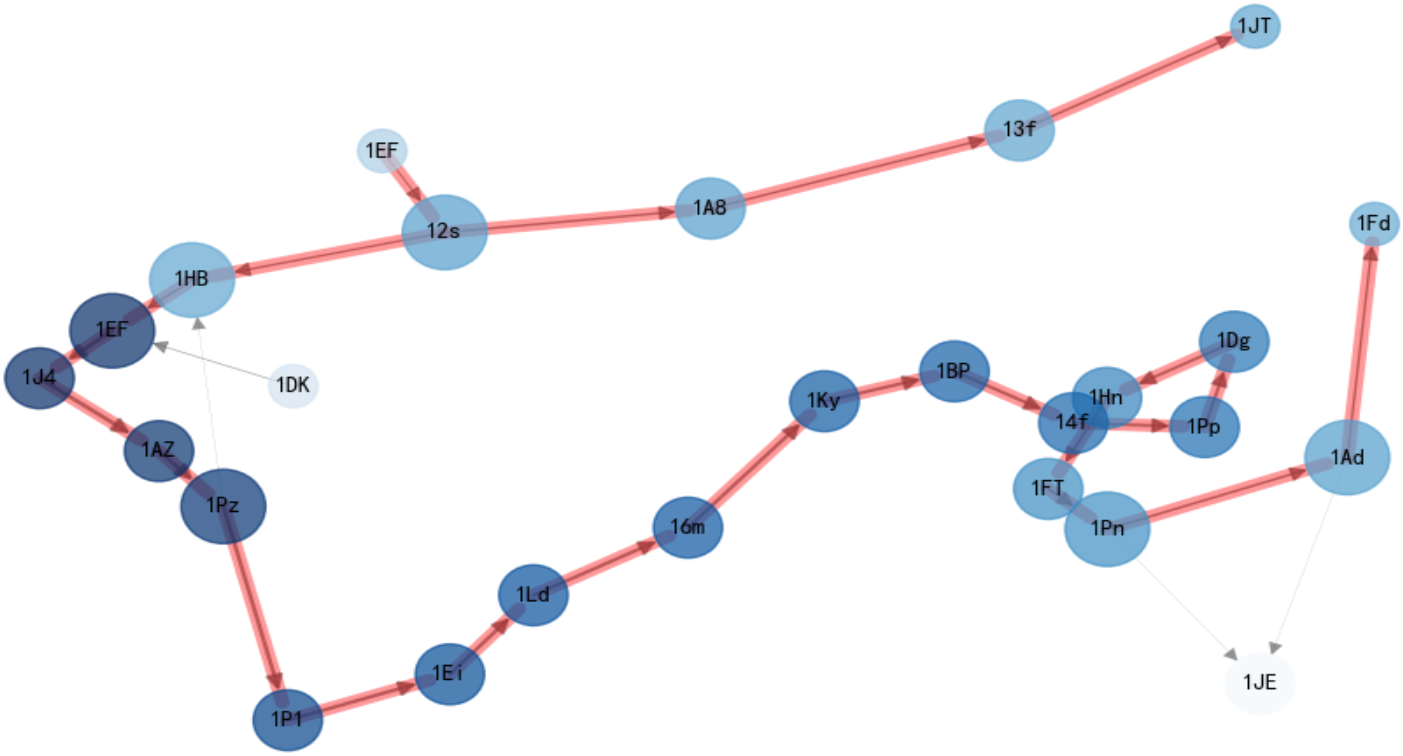


**图 5**

从图 5 中可以看到是这个部分是从地址 1EF 开始的，基本呈线性，1EF 在这个图中出现了两次，但是这 2 个不是同一个地址，只是前 3 位相同。然后分叉了，一条经过 1AB、13f 到 1JT 截止，另一条 1EF 又出现了一次，然后还有其他的地址加入如 1DK，最终到地址 1Fd 为止，最终还有一些比特币流向了地址 1JE。

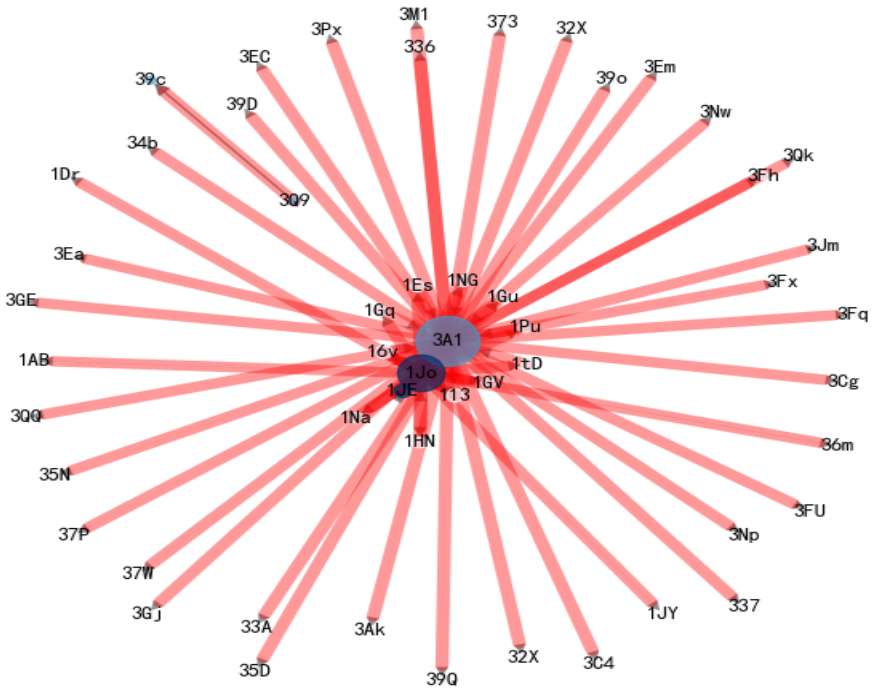


图 6

图6中以地址3A1、1Jo为中心，呈辐射状向四周散发，附近有一些地址，更远的地方还有一些地址，这说明主要是3A1、1Jo做输入，其他的地址接收比特币，但是四周的圆圈颜色很浅，说明这些地址收到的比特币数量不多。

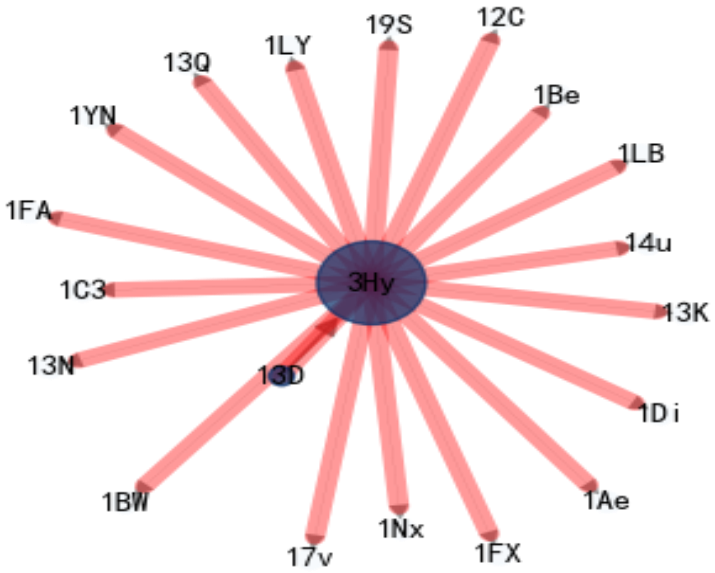


图 7

图7呈辐射状，以地址3Hy为中心，从地址13D开始的，很多比特币是13D给3Hy的，因为四周的地址圆圈基本没有颜色，所以地址3Hy只给了很小一部分给其他地址。

图8到图12是交易金额大于6万小于10万，交易有联系的部分，这些部分的交易流程图：

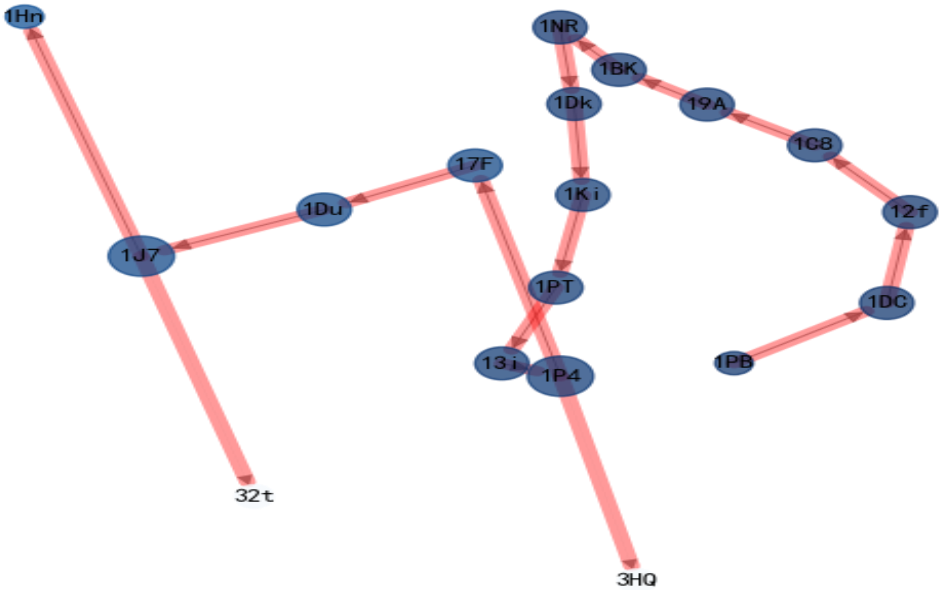


图 8

图8呈线型，从地址1PB开始，然后一直到1PT、13i和1P4，地址1P4分叉了，一些比特币给了地址3HQ，另外一些流向了17F，到地址1J7又分叉了。

图9呈线型从地址18f开始，然后中间没有分叉，最终到地址194截止，圆圈的颜色不断变浅，代表着交易的比特币越来越少。

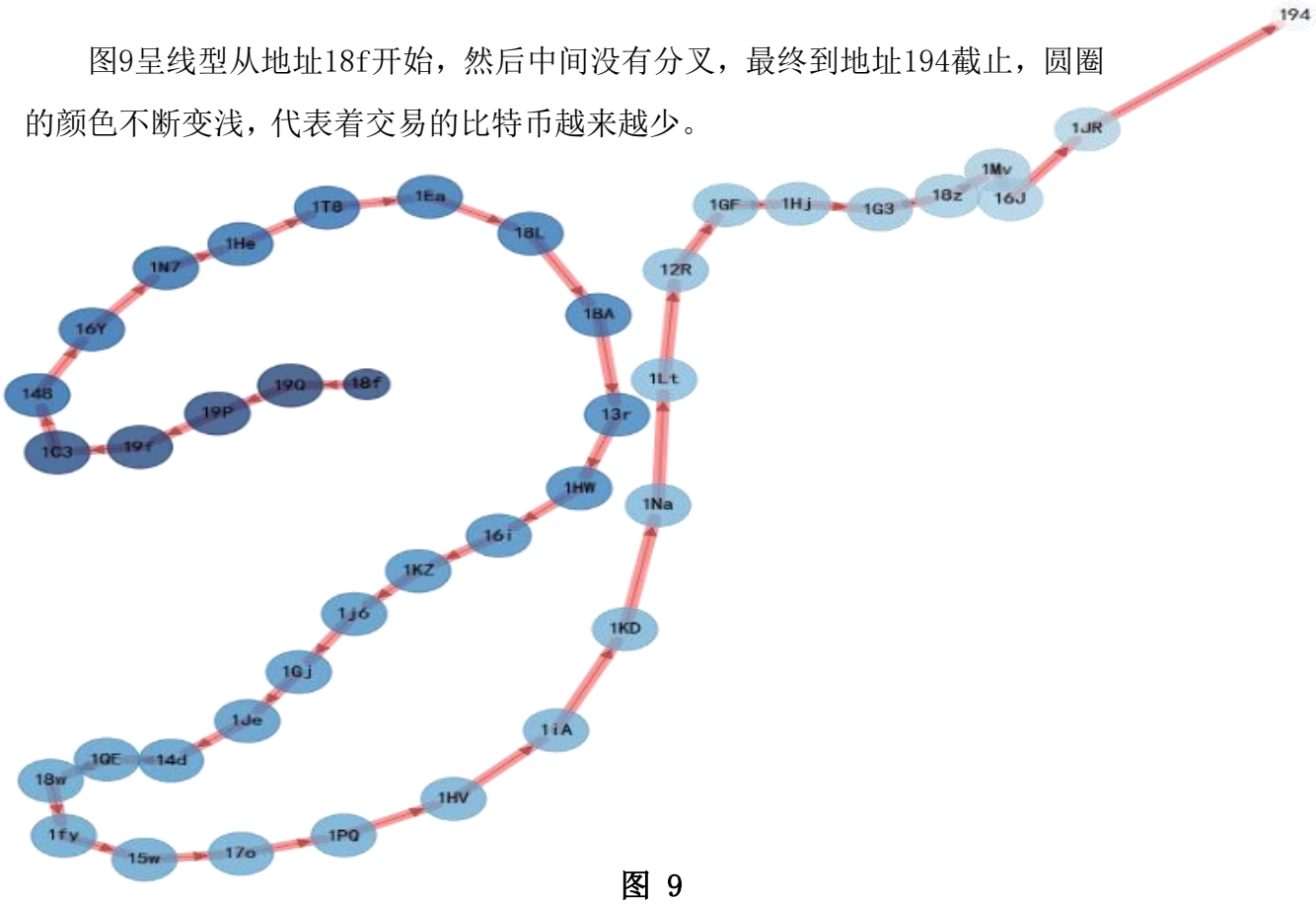


图 9

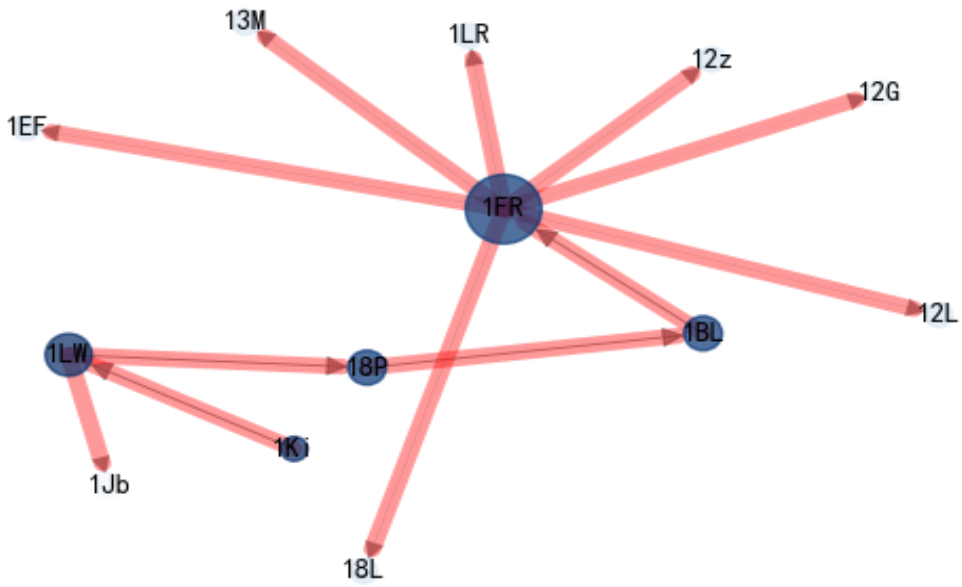


图 10

图10中从地址1Ki开始，然后到地址1FR，地址1FR作为输入，交易出去了很多比特币，圆圈1Ki的颜色比圆圈1LW深是因为还分了一些比特币给地址1Jb

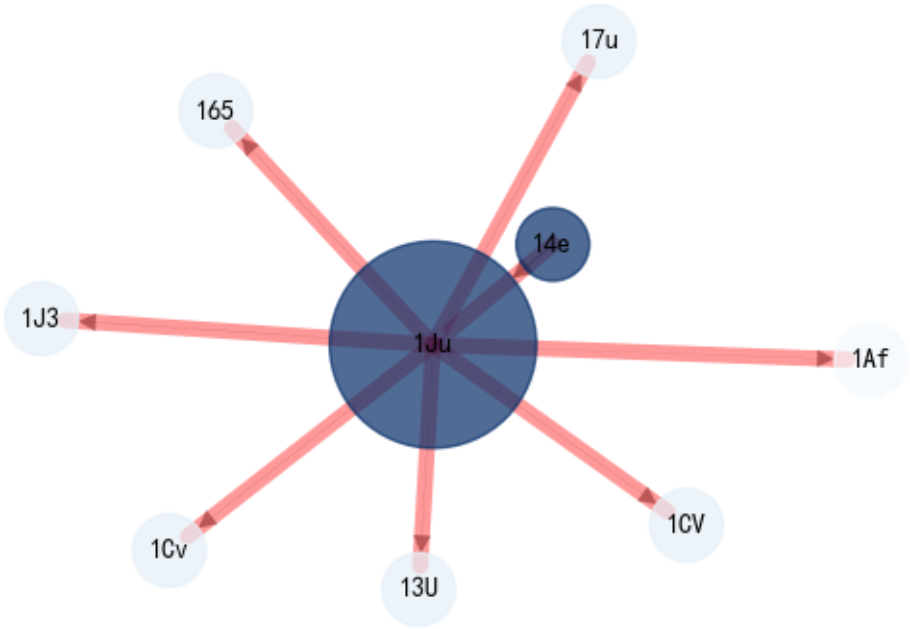


图 11

图11从地址14e作为输入，地址1Ju作为输出开始，然后地址1Ju作为输入，分出去了一些比特币，圆圈的颜色一样深，代表地址14e将所有的比特币都给了地址1Ju。

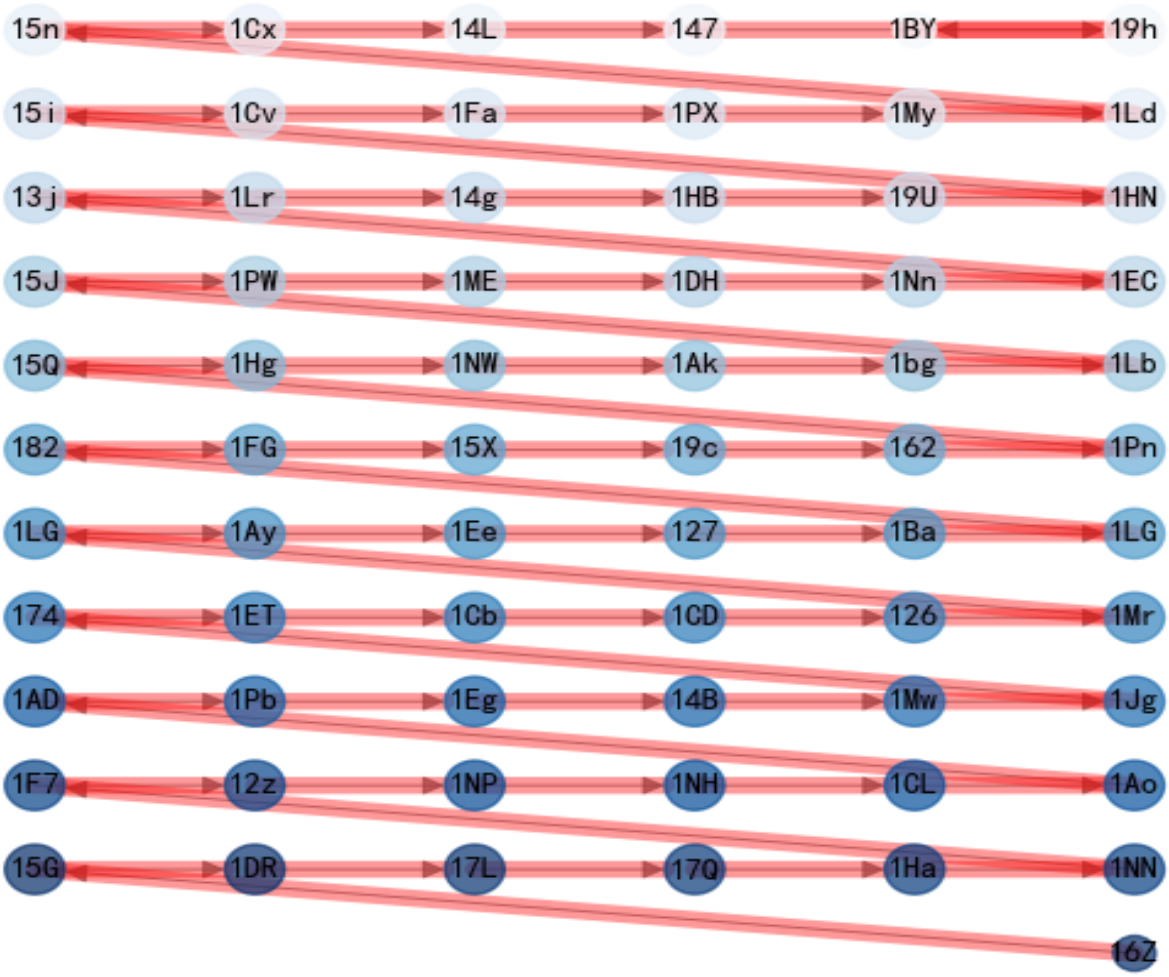


图 12

图12呈明显的线型，从地址16Z开始，然后一直不断的交易，更换地址，最终到地址19h为止，圆圈的大小基本一致，说明这些地址的度中心性基本一致，然后圆圈的颜色是不断变浅的，表明交易的比特币数量是不断减小的，但是减小的幅度很低。

# 6 结论

比特币受到越来越多的关注，通过分析比特币的交易数据和兑换美元的价格数据，并进行一定的对比，可以更好的了解比特币的整体情况。

分析得到了比特币的随着时间的交易特点，月交易比特币数量和月交易次数是紧密联系的，月交易比特币数量不仅和交易次数有关，还有每笔交易的比特币数量有关。月交易次数和月平均比特币兑换美元的价格关系很大，如果比特币兑换美元的价格快速上升，这段时间内交易的数量反而会下降，等价格开始下降的时候，就是人们大量卖出比特币的信号。还分析了收到相邻区块的时间间隔，其中有一部分是小于 0 的，总体来说生成一个区块的时间是控制在了 10 分钟左右。通过分析大型交易，可以看到很多交易都只有一个输入，这是为了避免其他人将这个大额交易地址和自己其他的地址实体联系到一起，增强匿名性。分析这些大型交易的流向，可以看到很多交易都是线型和放射状、长链的，分叉很少，而且拥有大额比特币的地址交易出去的比特币很少，很多比特币还留存在原地址中，或者转移到其他地址中去了。

# 谢 辞

时间过去的很快转眼间论文已经完成了，非常感谢我的指导老师魏普文。在论文的写作过程中，魏老师在确定论文选题、搜寻资料、撰写、修改一直到最后定稿，都给予了很多的帮助。

## 参考文献

[1] Antonopoulos,A.M.: Mastering Bitcoin[M],薄荷凉幼，陈萌琦，陈姝吉等译，第1章-第10章,pp.10-169.

[2] Reid,F.,Harrigan,M.: An Analysis of Anonymity in the Bitcoin System,arXiv:1107.4524v2 [physics.soc-ph] 7 May 2012.

[3] Nakamoto, S.: Bitcoin: A Peer-to-Peer Electronic Cash System, 2008.

[4] Ron,D.,Shamir,A.: Quantitative Analysis of the Full Bitcoin Transaction Graph; IACR Cryptology ePrint Archive,2012, p.584.

[5]Koshy,P., Koshy,D., AND McDaniel,P.: An analysis of anonymity in bitcoin using p2p network traffic. In Financial Cryptography and Data Security. 2014.

[6] 比特大陆：https://btc.com/stats

[7] 姚前,李连三, 大数据分析在数字货币中的应用[J]. 中国金融, 2016年, 17期, 37-38

[8]巴比特：http://www.8btc.com/

[9] blockchain: https://blockchain.info

[10] Möser,M.: Anonymity of Bitcoin Transactions: An Analysis of Mixing Services. In Proceedings of Münster Bitcoin Conference, 2013.

[11]冯登国，裴定一.密码学导引.北京：科学出版社.1999.4

[12](加)Alfred J.Menezes等著.应用密码学手册.胡磊，王鹏等译.北京：电子工业出版社.2005.6

[13]Ober,M.,Stefan,Katzenbeisser,S.,Hamacher,K.: Structure and Anonymity of the Bitcoin Transaction Graph; Future Internet 5,no.2(2013)

[14] scrapy: https://docs.scrapy.org/en/latest/

[15]bitcoincharts :http://api.bitcoincharts.com/v1/csv/

[16] networkx: https://networkx.github.io/documentation/stable/

# 附 录

## 原文

**Abstract.** The Bitcoin scheme is a rare example of a large scale global payment system in which all the transactions are publicly accessible (but in an anonymous way). We downloaded the full history of this scheme, and analyzed many statistical properties of its associated transaction graph. In this paper we answer for the first time a variety of interesting questions about the typical behavior of users, how they acquire and how they spend their bitcoins, the balance of bitcoins they keep in their accounts, and how they move bitcoins between their various accounts in order to better protect their privacy. In addition, we isolated all the large transactions in the system, and discovered that almost all of them are closely related to a single large transaction that took place in November 2010, even though the associated users apparently tried to hide this fact with many strange looking long chains and fork-merge structures in the transaction graph.

### 1 Introduction

Bitcoins are digital coins which are not issued by any government, bank, or organization, and rely on cryptographic protocols and a distributed network of users to mint, store, and transfer. The scheme was first suggested in 2008 by Satoshi Nakamoto [1], and became fully operational in January 2009. It had attracted a large number of users and a lot of media attention [2] [3] [4], but so far it was difficult to get precise answers to simple questions such as: How many different users are there in the system? How many bitcoins are typically kept in each account, and how does this balance vary over time? Are most bitcoins kept by a few large users? Do they keep their bitcoins in “saving accounts” or do they spend them immediately? How many users had large balances at some point in time? What is the size distribution of bitcoin transactions, and how

many of them are micropayments?

In this paper we answer all these (and many additional) questions. We use the fact that all the transactions ever carried out in the Bitcoin system are available on the internet (in an anonymous way). On May 13th 2012 we downloaded the full public record of this system in one of its two major forms1, which consisted of about 180,000 HTML files. After parsing and processing these files, we built a graph of all the Bitcoin addresses and transactions up to that date. We then used the methodology described in the next section in order to try to identify which addresses are likely to belong to the same entity, and used this information to contract the transaction graph by merging such addresses, in order to get a more accurate picture of the full financial activity of each user. We then analyzed many statistical properties of both the original and the contracted transaction graphs (most of our statistical results were very similar for the two graphs, within a factor of 2). The most interesting and informative distributions we found are described in a series of tables. In addition, we isolated all the large ($\geq$ 50; 000 bitcoins) transactions which were ever recorded in the system, and analyzed how these amounts were accumulated and then spent. We discovered that almost all these large transactions were the descendants of a single large transaction involving 90,000 bitcoins which took place on November 8th 2010, and that the subgraph of these transactions contains many strange looking chains and forkmerge structures, in which a large balance is either transferred within a few hours through hundreds of temporary intermediate accounts, or split into many small amounts which are sent to different accounts only in order to be recombined shortly afterwards into essentially the same amount in a new account.

There was one previous reported attempt [5] to download and analyze the full Bitcoin history, which also used the same methodology to try to contract all the addresses which are believed to belong to the same user. They created the graph of transactions on July 12th 2011, which was before the scheme really caught

on. Thus, the total number of bitcoins participating in all the transactions in our graph is about three times larger than in their graph. In addition, we expect the transactions in our more mature graph to better represent typical use of the system, whereas their graph represents primarily the experiments run by early adopters. However, the biggest difference between our papers is that they were primarily interested in privacy issues, whereas we are primarily interested in the statistical properties of the Bitcoin transaction graph. Another analysis of the Bitcoin transaction graph was presented at the Chaos Computer Club Conference in Germany in December 2011 [6]. Again, they were primarily interested in how to defeat the anonymity of the network, but also included some interesting comments about the economic principles behind the scheme, the effect of lost coins on its operation, weaknesses in its protocols, and the general topological properties of this transaction graph.

The paper is organized as follows. In Section 2 we describe the Bitcoin scheme. In Section 3 we summarize the main statistical distributions we extracted from the downloaded transactions, which describe many interesting and even surprising properties of the scheme. Finally, in Section 4 we present the graph of the largest transactions and analyze its strange structure.

## 2 The Bitcoin Scheme

Bitcoin is a decentralized electronic cash system using peer-to-peer networking to enable payments between parties without relying on mutual trust. It was first described in a paper by Satoshi Nakamoto (widely presumed to be a pseudonym) in 2008. Payments are made in bitcoins (BTC's), which are digital coins issued and transferred by the Bitcoin network. The data of all these transactions, after being validated with a proof-of-work system, is collected into what is called the block chain.

Participants begin using bitcoin by first acquiring a program called a Bitcoin wallet and one or more Bitcoin addresses. Bitcoin addresses are used

for receiving bitcoins, in the same way that e-mail addresses are used for receiving e-mails. Even though Bitcoin is considered to be an experimental payment system, it is already deployed on a large scale (in the sense that the current value of all the coins issued so far exceeds 100,000,000 USD) and attracts a lot of media attention. Its proponents claim that it is the first truly global currency which does not discriminate its users based on citizenship or location, it is always running with no holidays, it is easy to secure with very low usage fees, it has no chargebacks, etc. On the other hand, its detractors claim that it is widely misused to buy illegal items and to launder large sums of money, and that it is too easy to steal bitcoins from wallets via cyber attacks.Unlike fiat currency, which has been declared to be legal tender by a government despite the fact that it has no intrinsic value and is not backed by reserves, the Bitcoin scheme has no centralized issuing authority. The network is programmed to increase the money supply in a slowly increasing geometric series until the total number of bitcoins reaches an upper limit of about 21 million BTC's. Bitcoins are awarded to Bitcoin "miners" for solving increasingly difficult proof-of-work problems which confirm transactions and prevent doublespending. The network currently requires over one million times more work for confirming a block and receiving an award (currently 50 BTC's) than when the first blocks were confirmed.

The exchange rate of bitcoins has fluctuated widely over the years, from merely $0:01 to over $30 per BTC. Today (October 2012) it is worth a little over $12 per BTC. The entire activity in the Bitcoin network is publicly available through the internet in two major forms, and the one we decided to download appears as a block chain, starting at block 0 [7] (created back on the 3rd of January 2009). Each block reports on as little as a single transaction to as much as over a thousand transactions, and provides hyperlinks to other blocks and to other activities of each address.

Many users adopt the Bitcoin payment system for political and philosophical

reasons. Each user can have an unbounded number of addresses (which are characterized by their public/private key pairs) owned by him. A transaction in bitcoins is a generalization of a regular bank transaction in the sense that it allows multiple sending addresses and multiple receiving addresses in the same transaction. It specifies how many bitcoins were taken from each sending address and how many bitcoins were credited to each receiving address, without the details of who gave how much to whom. An address may receive bitcoins which are either newly minted or have a specific sending address. Another important difference between Bitcoin transactions and regular bank transactions is the notion of change, which is related to the fact that bitcoins are kept in (possibly fractional sized) chunks which have to be transferred in an all or nothing way. For example, a user can have three chunks of 10 bitcoins each. A transaction can spend 12:5 bitcoins by transferring the first full chunk plus 2:5 bitcoins from the second chunk, and then the 7:5 bitcoin change should be sent to a new address owned by the same user with new public and private keys. The user then has the option of either transferring the third chunk to the new address, or leaving it in the old address. In fact, it is considered good practice for a user to generate a new address, i.e., public-private key-pair, for every transaction even if this is not necessary. To better protect their identity, users are advised to take the following steps: they do not have to reveal any identifying information in connection with their addresses; they can repeatedly send varying fractions of their BTC's to themselves using multiple (newly generated) addresses; and/or they can use a trusted third-party in the form of a shared e-wallet to mix their transactions with those of other owners.

[译]These operational and privacy policies of the Bitcoin scheme make it desirable for us to try to contract the transaction graph in order to get a more informative picture of the total assets and financial activities of users which are associated with many addresses, and to try to distinguish between "internal" and "external" transfers of bitcoins in it. Performing this contraction in

a completely accurate way seems to be extremely difficult, but we can use the available data in order to try to find a good first approximation. Since many transactions have multiple sending addresses, we can make the reasonable assumption that all these addresses have the same owner. We then compute the transitive closure of this property over all the transactions. For example, if there is one transaction in which 1 and 2 are used as sending addresses, and another transaction in which 2 and 3 are used as sending addresses, we conclude that all three addresses are jointly owned. This can lead to two types of errors: We can underestimate the common ownership of some addresses because there was no evidence for it in the available data, and we can overestimate it if several users decided to pool their activities and to send a single transaction to which each one of them contributes some of the sending addresses. Discussions with several members of the Bitcoin community lead us to believe that at the moment there are likely to be very few overestimation errors of this type, but quite a few underestimation errors. For example, when we tried to use all the available transactions to merge the addresses of a particular large user, we were told that we managed to identify with our methodology only about one quarter of his real addresses. Note that the link ability of the addresses does not imply that the identity of the user becomes known. However, if we have any external information about the real ownership of any one of the merged addresses, we can get a fuller picture of the Bitcoin activity of that particular individual or organization. For example, since WikiLeaks publicly advertised one of its addresses when it asked for donations, we can estimate with our methodology that WikiLeaks owns at least 83 addresses, that it was involved in at least 1088 transactions, and that it had an accumulated income in all these addresses of 2605.25 BTC's.

We acquired the complete state of the Bitcoin transaction system on May 13*th* 2012, which contained all the transactions carried out in the system since its inception on January 3*rd* 2009 until that date. This required downloading 180,001

separate but linked HTML files, starting from block number 180,000 [8] and following the links backwards to the zeroth block initiating the system in January 2009. Each file was parsed in order to extract all the multisender/multireceiver transactions in it, and then the collection of transactions was encoded as a standard database on our local machine. We then ran a variant of a Union-Find graph algorithm [9] in order to find sets of addresses which are expected to belong to the same user. We merged all the nodes and combined all the transactions which can be associated with him (without eliminating the internal transfers, which become self loops in the new graph). We call the original transaction graph the *address graph*, and the contracted transaction graph the *entity graph* (we avoid using the word "owner" with its complex legal connotations since we do not really know who owns each address, and instead use the neutral word "entity" as our best approximation to the common owner of multiple addresses). All the statistics described in Section 3 are derived from both the address graph and the entity graph, as indicated in the tables. In most (but not all) cases, we expect the statistics to change monotonically as we move from the address graph to the entity graph and then to the (unknown) owner graph, since each entity is typically the union of several addresses which we managed to merge,and each real owner is typically the union of several entities that we failed to merge. For example, since the average balance of an address is 2.4 BTC's and the average balance of an entity is 3.7, we can argue that the average balance of an owner is likely to be larger than 3.7 BTC's. This monotonicity can thus be used to provide plausible upper or lower bounds for the statistical properties of the real ownership graph, even though we do not know it. [译]

## 3 Statistics Calculated Over the Bitcoin Transaction Graph

At the time we downloaded the graph there were 3,730,218 different public keys, each associated with a different address: 3,120,948 of them were involved

as senders in at least one transaction, while the additional 609,270 appear in the network only as receivers of BTC's. By running the Union-Find algorithm, we were able to associate the 3,120,948 addresses with 1,851,544 different entities. Since the other 609,270 addresses were never used as senders, they could not be merged with any other addresses by the Union-Find algorithm, and thus they all remained as entities with a single address. By adding these singletons, we get a total of 2,460,814 entities, which implies that each one of them has on average about 1.5 addresses. However, there is a huge variance in this statistics, and in fact one entity is associated with 156,722 different addresses. By analyzing some of these addresses and following their transactions, it is easy to determine that this entity is Mt.Gox, which is the most popular Bitcoin Exchange site (responsible for almost 90% of all the exchange operations in the network). The full distribution of the number of addresses per entity is given in Table 1.

In our reduced entity graph, each m-to-n transaction has a single sender (since the m sending addresses necessarily belong to the same entity) and at most n receivers. It can thus be decomposed into at most n different transactions from the single entity associated with the m senders to the entities associated with the n receivers. In case some of the receiving addresses are identified as belonging to the same entity, their amounts are accumulated to create a single common transaction, and if some of the receivers are identified with the single sender, we create a single self loop with the combined amounts. The resulting entity graph has 7,134,836 single sender and single receiver transactions, out of which 814,044 (about 11%) involve Deepbit (the largest Bitcoin mining pool),and 477,526 (about 7%) involve Mt.Gox. About 10% of the transactions are self loops. The entity graph is not connected as it is composed of 133,742 different connected components, many of size one. For instance, there are as many as 43,710 components (about 33%) consisting of a single address which are used only for accepting (one or several batches of) freshly minted bitcoins,

and which have never participated in any incoming or outgoing transactions. Note that the address graph has a larger number of 13,734,847 transactions of lower values, since a single transaction with 2 sending addresses and 3 receiving addresses is represented in the address graph as 6 single-sender and single-receiver transactions.

There are many types of statistics and graphs about the Bitcoin network which can be readily downloaded from the internet [10] [11]. However, these types of statistics tend to describe some global property of the network over time, such as the number of daily transactions, their total volume, the number of bitcoins minted so far, and the exchange rate between bitcoins and US dollars. We can go much further than that, since the entire transaction graph can be used to determine the financial history of each entity including all of its sending/receiving activities along with the daily balance of bitcoins in its various addresses and how they vary over time. Having this entity graph at hand enables us to study various statistical properties of the network, which are not easy to determine by following a small number of online links in the Blockexplorer representation of the Bitcoin network. In the rest of this section, we describe some of our findings so far.

Here is our first surprising discovery, which is related to the question of whether most bitcoins are stored or spent. The total number of BTC's in the system is linear in the number of blocks. Each block is associated with the generation of 50 new BTC's and thus there are 9,000,050 BTC's in our address graph (generated from the 180,001 blocks between block number zero and block number 180,000). If we sum up the amounts accumulated at the 609,270 addresses which only receive and never send any BTC's, we see that they contain 7,019,100 BTC's, which are almost 78% of all existing BTC's. Due to the way bitcoins can be repeatedly moved to fresh addresses, some of which can be very recent, we can not claim that all these bitcoins are out of circulation. However, 76.5% of these 78% (i.e., 59.7% of all the coins in the system) are "old

coins”,defined as bitcoins received at some address more than three months before the cut off date (May 13th 2012), which were not followed by any outgoing transactions from that address after they were received. One can also argue that very old dormant bitcoins were simply abandoned or lost by users who experimented with the system in its early days, when it was very difficult to buy anything or to exchange bitcoins into dollars. To be even more cautious with our estimation of dormant bitcoins, we decided to ignore all the transactions which took place prior to July 18th 2010, when Mt.Gox started its exchange and price quoting services. The sum of the balances of all the addresses which have not been active since that date is 1,657,480 bitcoins. Clearly, by considering all these bitcoins as “lost” rather than “hoarded” we are underestimating the number of bitcoins which are kept dormant in “saving accounts”. By ignoring these very old bitcoins and repeating the same calculation, we found that 73% of all the remaining BTC’s were accumulated at addresses which only receive and never send bitcoins, and that 70% of these 73% (i.e., 51%) are dormant bitcoins in the sense that they were received more than three months before our cutoff date but after it became easy to exchange them. If instead of summing the transaction values we sum the final balances of all the addresses that were active after July 18th 2010 but became inactive in the last three months, we get that 55% of all coins in the system are dormant in this sense. This is strong evidence that the majority of bitcoins are not circulating in the system, and since it is based on the address rather than the entity graph, this conclusion is not affected by possible inaccuracies in the way we associate addresses with users. Note that the total number of bitcoins participating in all the transactions since the establishment of the system (except for the actual minting operations) is 423,287,950 BTC’s, and thus each coin which is in circulation had to be moved a large number of times to account for this total flow.

Another interesting finding is that the total number of bitcoins received by most entities and addresses is negligible. In the rest of this section, we

use unparenthesized numbers to indicate values derived from the entity graph, and parenthesized numbers to indicate values derived from the address graph. For example, as can be seen from Table 2, 36% of all entities (and 40% of all addresses) received fewer than one BTC, currently worth about 12 USD, throughout their lifetime, 52% (59%) received fewer than 10 BTC's and 88% (91%) fewer than 100. At the other end of the distribution there are only four entities (and one address) which received over 800,000 BTC's, and 80 entities (129 addresses) which received over 400,000.

Similarly, as can be seen in Table 3 the current (on May 13th 2012) balance of almost 97% (98%) of all entities (addresses) was less than 10 BTC's. This number decreases to 88% (91%) if instead of looking at one specific moment,we look at the maximal balance ever seen throughout an entity's (address'es) lifetime. This statistics is summarized in Table 4. In addition, it can be seen that there are only 78 entities (70 addresses) with current balance larger than 10,000 BTC's. This number grows to 3,812 (3,876) when looking at the maximal balance ever seen.

Another measure that may indicate the level of activity of an entity (address) is the number of transactions it has been involved with. Its distribution is presented in Table 5. It is remarkable that 97% (93%) of all entities (addresses) had fewer than 10 transactions each, while 75 entities (80 addresses) use the network very often and are affiliated with at least 5,000 transactions.

We have also calculated the distribution of the size of the transactions in the two graphs as summarized in Table 6. Again, it is evident that many transactions are very small, and 28% (47%) are smaller than 0.1 BTC each. The Bitcoin scheme actually enables sending micro transactions, which are of the order of 10−8 BTC (this is the smallest fraction into which a BTC can be broken, and is called a satoshi). When we also consider midsize amounts, we see that 73% (84%) of the transactions involve fewer than 10 BTC's. On the other hand, large transactions are rare at Bitcoin: there are only 364 (340) transactions

larger than 50,000 BTC's. We have carefully inspected all these large transactions and describe our findings in the next section.

It is interesting to investigate the most active entities in the Bitcoin system, those who have either maximal incoming BTC's or maximal number of transactions. 19 such entities are shown in Table 7 sorted in descending order of the number of accumulated incoming BTC's shown in the third column. The leftmost column associates the entities with letters between A to S out of which three are identified: B is Mt.Gox, G is Instawallet and L is Deepbit. Eight additional entities: F, H, J, M, N, O, P, and Q are pointed out in the graph of the largest transactions (Fig. 1) which is presented in the next section. The second column gives the number of addresses merged into each entity. The fourth column presents the number of transactions the entity is involved with.

Table 7 shows that Mt.Gox has the maximal number of addresses, but not the largest accumulated incoming BTC's nor the largest number of transactions.Entity A in the first row of Table 7 owns the next largest number of addresses,about 50% of those of Mt.Gox's, but received 31% more BTC's than Mt.Gox.Deepbit had sent 70% more transactions than Mt.Gox. It is interesting to realize that the number of addresses of 13 of these entities is a fifth or more ofthe number of transactions they have executed, which may indicate that each address is indeed used for just a few transactions. It is also clear that six out of the 19 entities in the table have each sent fewer than 30 transactions with a total volume of more than 400,000 BTC's. Since these entities were using large transactions, we were able to isolate them and to follow the flow of their transactions, see Section 4 below. On the other hand, entity A had never sent any large transactions and thus it has not been included in our graph of the largest transactions.

## 4 The Graph of the Largest Transactions in Bitcoin

[译]We have identified and analyzed all the largest (≥ 50; 000 BTC's)

transactions in the entity graph, (there were 364 such transactions as described in the last column of Table 6), and followed their flow. We started with the *earliest* such large transaction, the one of 90,000 BTC's made on November 8*th* 2010. By tracing each of the other 363 large transactions in this category, we were able to show that 348 were actual successors of this initial transaction. The resulting directed graph is depicted in Fig. 1. This graph reveals several characteristic behaviors of the flow in the Bitcoin transaction graph: long consecutive chains of transactions, fork-merge patterns that may include self loops, setting aside BTC's and final distribution of large sums via a binary tree-like structure.

**Long Chains.** A common prominent practice of Bitcoin users is to create chains of consecutive transactions. Some of these chains can be explained by the change mechanism in which small payments are accompanied by the creation of a new address, into which the user transfers the difference. Such chains can be found in Fig. 2, Fig. 4, Fig. 5 and Fig. 7, with lengths of 3, 15, 26, 80, 88 and 350 transactions. However, the behavior seen in Fig. 3 deviates significantly from this pattern, since the same amount of 5,000 bitcoins is repeatedly split off the main sum and put into accounts which have no additional transactions associated with them.

**Fork-Merge Patterns and Self Loops.** Another frequent scenario in Bitcoin is transferring a large number of BTC's from one address to another via several intermediate addresses, each receiving part of the entire amount and then sending it, mostly in full, to the same destination whether directly or via other mediators. Examples can be seen in Fig. 6, Fig. 8 and Fig. 9. A harder to follow fork-merge pattern is presented in Fig. 5: An entity is sending 90,000 BTC's to itself three times in self loops. Each time it splits it into different amounts, 76+14, 72+18 and 69+21. It uses the same address for the small amounts and different addresses for the large amounts. Then it exchanges the entire 90,000 BTC's at Mt.Gox. Finally, the 90,000 BTC's are being transferred via a chain of 90 transactions using 90 different addresses (which may or may not belong

to the same owner), where at each one of them 1,000 BTC's are sent back to the first entity, recombined into essentially the very first amount of 90,000 BTC's.[译]

**Keeping Bitcoins in "Saving Accounts".** Another long chain of transactions from the beginning of March 2011 can be seen in Fig. 3. This chain is different from the above ones, since at 28 out of its 30 steps, it puts aside 5,000 BTC's in what seems to be "saving accounts". The accumulated sum of 140,000 BTC's has never been sent since. These bitcoins are an example of our discovery that most of the bitcoins are not circulating in the system.

**Binary Tree-Like Distributions.** Often amounts of BTC's are distributed among many addresses by splitting it into two similar amounts at each step. This results in a binary tree-like structure as depicted in Fig. 10 and in Fig. 4.

## [译]5 Conclusions

The Bitcoin system is the best known and most widely used alternative payment scheme, but so far it was very difficult to get accurate information about how it is used in practice. In this paper we describe a large number of statistical properties of the Bitcoin transaction graph, which contains all the transactions which were carried out by all the users until May 13*th* 2012. We discovered that most of the minted bitcoins remain dormant in addresses which had never participated in any outgoing transactions. We found out that there is a huge number of tiny transactions which move only a small fraction of a single bitcoin, but there are also hundreds of transactions which move more than 50,000 bitcoins. We analyzed all these large transactions by following in detail the way these sums were accumulated and the way they were dispersed, and realized that almost all these large transactions were descendants of a single transaction which was carried out in November 2010. Finally, we noted that the subgraph which contains these large transactions along with their neighborhood has many strange looking structures which could be an attempt to conceal the existence and relationship

between these transactions, but such an attempt can be foiled by following the money trail in a sufficiently persistent way.[译]

**Acknowledgments.** This research was supported by the Citi Foundation. We would like to thank Ronen Basri, Uriel Feige, Michal Irani, Robert Krauthgamer, Boaz Nadler, Moni Naor and David Peleg from the Computer Science and Applied Mathematics Department of the Weizmann Institute of Science for many interesting and informative discussions. We would also like to thank Aharon Friedman for his major help in acquiring and processing the Bitcoin data base. Finally, we would like to thank all the members of the Bitcoin community, and in particular Meni Rosenfeld and Stefan Richter, who sent us excellent comments, criticisms and suggestions. We revised the original version of the paper in order to respond to their input.

**译文**

比特币方案的操作和隐私政策使得我们希望能尽力提取出交易图，以便得到一个包含更多关于多地址用户的资金和比特币活动信息的图，并且从交易图中努力区分内部和外部比特币交易。以完全正确的方式执行提取操作是很困难的，但我们可以使用可用的数据来尝试找到一个好的第一个近似值。因为很多交易有多个输入地址，我们能做出合理的假设：所有这些地址为一个人所有。我们计算这个属性在所有交易中的传递闭包。例如，如果这里有一个交易 1、2 作为输入地址，另一个交易中 2、3 作为输入地址。则我们得出结论：这三个地址都是共同拥有的。这可能导致两种类型的错误：我们可能低估了某些地址的共同所有权，因为在可用数据中没有证据表明我们可以高估它，如果多个用户决定集中其活动并发送单个事务给其中每个人都贡献一些发送地址。与比特币社区的几名成员进行的讨论使我们相信，目前这种类型的高估错误可能非常少，但是存在相当多的低估错误。例如，当我们试图使用所有可用的事务合并特定大用户的地址时，我们被告知我们只能确定他的实际地址的四分之一。请注意，地址的链接能力并不意味着用户的身份变得已知。但是，如果我们有任何关于任何一个合并地址的真实所有权的外部信息，我们可以更全面地了解该特定个人或组织的比特币活动。例如，由于维基解密在要求捐款时公开发布了其中一个地址，我们可以用我们的方法估计，维基解密拥有

至少 83 个地址，至少涉及 1088 笔交易，并且它在所有交易中都有累计收入这些 2605.25 BTC 的地址。

我们获得了比特币交易系统的完整状态在 2012 年 5 月 13 日，其中包含系统自 2009 年 1 月 3 日成立至今的所有交易。这需要下载 180,001 个单独但互相链接的 HTML 文件，从 180000 号区块开始，反向追踪链接到 2009 年的 0 号区块。每个文件都被解析以提取其中的所有多发送者 / 多接收者事务，然后将事务集合编码为标准数据库我们的本地机器。然后，我们运行了 Union-Find 图算法[9]的一个变体，以便找到预期属于同一用户的地址集。我们合并了所有的节点并合并了所有可以与之关联的交易（没有消除内部转移，这在新图中变成了自循环）。我们将原始事务处理图称为地址图，将互相有关系的交易图称为实体图（我们避免使用具有复杂法律内涵的”所有者”一词，因为我们并不真正知道谁拥有每个地址，而是使用中性词”实体”作为我们对多个地址的共同拥有者的最佳近似）。如表中所示，第 3 节中描述的所有统计数据均来自地址图和实体图。在大多数（但不是全部）的情况下，我们预计统计数据在从地址图转移到实体图，然后转到（未知）所有者图时单调变化，因为每个实体通常是我们管理的多个地址的联合合并，每个真正的所有者通常是我们未能合并的几个实体的联合。例如，由于地址的平均余额为 2.4 BTC，实体的平均余额为 3.7，所以我们可以认为所有者的平均余额可能大于 3.7 BTC。因此，这种单调性可以用来为真实所有权图的统计特性提供可信的上限或下限，尽管我们不知道它。

我们已经识别和分析了实体图中所有最大（≥50，000 比特币）的交易（表 6 最后一列中描述了 364 笔交易），并且追踪了他们的流向。我们从最早的这样的大宗交易开始，发生在 2010 年 11 月 8 日交易比特币是 9 万中的一笔交易。通过跟踪其他 363 次这一类别的大型交易，我们能够证明 348 次是该次交易的延续交易。由此产生的有向图如图 1 所示。该图揭示了比特币交易图中流量的几个特征行为：长期连续的交易链，可能包括自行循环的分叉合并模式，通过二叉树状结构来分散比特币。长链。比特币用户常见的突出做法是创建连续交易链。其中一些链条可以通过改变机制来解释，在这种改变机制中，小额付款伴随着新地址的创建，用户将差额转入其中。这些链可以在图 2，图 4，图 5 和图 7 中找到，长度分别为 3,15,26,80,88 和 350 个交易。然而，图 3 所示的行为偏离了这种模式，因为相同数量的 5000 比特币会重复分离出主要数额并投入到没有与其相关的额外交易的账户中。

叉合并模式和自循环。比特币的另一个常见情况是通过几个中间地址将大量 BTC 从一个地址转移到另一个地址，每个地址都接收部分全部金额，然后将其大部分全部发送到相同的目的地，无论是直接还是通过其他中介。例子可以在图 6，图 8 和图 9 中看到。图 5 给出了更难遵循叉合并模式：一个实体向自己发送 90,000 个 BTC 自身三次自循环。每次它分裂成不同的数量，76 + 14,72 + 18 和 69 + 21。它使用相同的地址来处理大量的小数量和不同的地址。然后它交换 Mt.Gox 的全部 9 万比特币。最后，90,000 比特币将通过 90 个交易链使用 90 个不同的地址（可能属于或不属于同一个所有者）进行传输，其中每个地址都发送 1000 比特币到第一个实体，重组为基本第一笔 90,000 比特币。

结论：

比特币系统是最广为人知和最广泛使用的替代支付方案，但到目前为止，很难获得有关在实践中如何使用它的准确信息。在本文中，我们描述了比特币交易图的大量统计特性，该图包含了所有用户在 2012 年 5 月 13 日之前进行的所有交易。我们发现，大多数铸造的比特币在从未使用过的地址中保持休眠状态并未作为输入参与到任何的交易中。我们发现有大量微小交易只能移动一个比特币中的很小一小部分，但也有数百个交易移动了超过 5 万个比特币。我们详细分析了所有这些大额交易的累积方式和分散方式，并认识到几乎所有这些大宗交易都是 2010 年 11 月进行的单笔交易的延续。最后，我们注意到包含这些大宗交易的子图及其附近有许多奇怪的外观结构，可能试图掩盖这些交易之间的存在和关系，但这种尝试可以通过以足够持久的方式追踪金钱痕迹来攻破。