Recuperación de Información y Recomendaciones en la Web
2023

Informe Final

Recuperación y procesado de información de una blockchain

Imagen generada con inteligencia artificial generativa (DALL-E 2)

Agustín Amegeiras | Santiago Costa | Santiago Acquarone
Libertad Tansini

Introducción 2
Problema 2
Enfoque de la solución 3
Diseño 3
- **Estructura de un bloque** 3
- **Estructura de una transacción** 4
- **Funciones** 5
  - **get_miners(num_blocks)** 5
  - **get_addrs_with_more_in_trxs(num_blocks, num_addresses) enviaron más** 5
  - **get_addrs_with_more_out_trxs(num_blocks, num_addresses) recibieron más** 5
  - **get_addrs_with_more_btc_recieved(num_blocks, num_addresses)** 6
  - **get_addrs_with_more_btc_sent(num_blocks, num_addresses)** 6

Implementación 6
Evaluación y resultados 7
- **get_miners (obtener mineros)** 7
- **get_addrs_with_more_in_trxs (direcciones que enviaron más transacciones)** 8
- **get_addrs_with_more_out_trxs (direcciones que recibieron más transacciones)** 9
- **get_addrs_with_more_btc_recieved(direcciones que recibieron más bitcoins)** 10

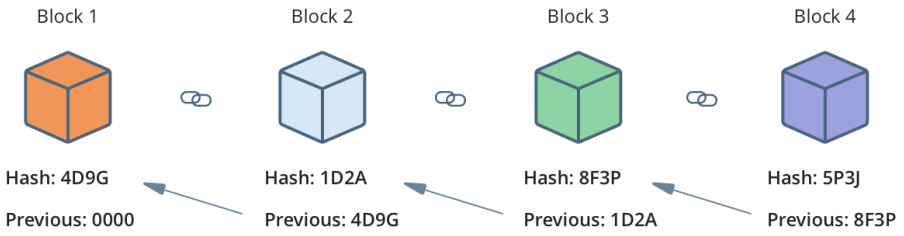
Conclusiones 11
- **¿Cuántos usuarios minaron bloques y cuantos bloques minaron?** 11
- **¿Cuáles fueron las cuentas que más transacciones realizaron?** 12
- **¿Cuáles fueron las cuentas que más cantidad de Bitcoin transfirieron?** 12

Trabajo Futuro 12

Introducción

Bitcoin es una moneda digital descentralizada y un sistema de pago. La tecnología que está detrás de este sistema se conoce como *blockchain*, una estructura de datos que permite almacenar información en bloques a los que se les añade información relativa al bloque anterior. Dentro de cada bloque se almacena la información de un conjunto de transacciones realizadas por usuarios del sistema.

El sistema es una red *peer-to-peer* donde la seguridad está dada por el mecanismo *Proof of Work*. Mediante este, en cada bloque se añade un hash que debe cumplir ciertas particularidades que son computacionalmente difíciles de encontrar. Por ello, varios nodos en distintas partes del mundo compiten por encontrar estos hashes, y el nodo que lo encuentre recibe una recompensa en Bitcoin. Esto es conocido como la minería de Bitcoin.



Problema

El sistema Bitcoin dispone de exuberante información que resulta de interés a ser analizada. Esta se encuentra, a su vez, almacenada en los distintos bloques que componen la estructura interna del sistema. Acceder a esta información no es una tarea trivial y supone el entendimiento del funcionamiento interno de esta red descentralizada así como de su protocolo de seguridad. Recopilar información acerca de este sistema permite el análisis de datos sobre la minería de bloques y transacciones realizadas por los usuarios.
En este sentido, interesa responder a cuestiones como: cuántos usuarios minaron bloques en cierto periodo y cuántos bloques minaron; cuáles fueron las cuentas que más transacciones realizaron y qué suma monetaria transfirieron.
No obstante, la información acerca de estos hitos aludidos es escasa y son pocas las plataformas que disponen de la misma.

En este proyecto, se tiene como eje de investigación la recuperación de información de la blockchain de Bitcoin y procesamiento de la misma, a fin de obtener datos de relevancia que subyacen a la estructura compleja del sistema y sus usuarios. Bajo esta óptica, se anhela arrojar luz sobre la naturaleza de este sistema de pagos, cómo está distribuida la riqueza en el mismo y qué entidades poseen mayor control sobre la red.

Enfoque de la solución

Dada la problemática descrita, se propone a modo de solución la creación de un programa que permita la recuperación de información sobre los bloques del sistema Bitcoin para que luego sean procesados y lograr la obtención de los datos de interés.

Cada bloque que compone la estructura del sistema comprende la información de un conjunto de transacciones que fueron efectuadas en cierto momento. Los sistemas basados en la estructura de *blockchain* poseen la particularidad de que siempre se agregan (o minan) nuevos bloques pero nunca se eliminan bloques anteriores. En el sistema Bitcoin, se mina aproximadamente un bloque cada diez minutos (esto es 144 bloques en un día), con lo cual la información en este sistema está en constante actualización. Dada la magnitud de información disponible, para el propósito de este trabajo se optó por establecer un lapso de tiempo de 24 horas sobre el cual extraer la información de los bloques.

Diseño

Para obtener la información de la blockchain se utiliza la API pública ofrecida por blockchain.com[1]. Esta API disponibiliza un endpoint[2] que permite obtener toda la información de un bloque en forma de un objeto JSON a partir de su altura en la cadena. De esta forma, sabiendo la altura del último bloque de la red se puede recuperar los *n* bloques anteriores para su posterior procesamiento.

A continuación se describe la información más importante contenida en los bloques de Bitcoin, haciendo énfasis en los campos relevantes para este trabajo.

Estructura de un bloque

Los bloques tienen los siguientes campos:

- **hash**: Es un identificador del bloque. Es generado a través de una función de hash que toma los datos del bloque y calcula este identificador.
- **ver**: Es la versión del protocolo de Bitcoin que se utilizó para crear el bloque.
- **prev_block**: Es el hash del bloque anterior en la cadena de bloques. Esto es lo que enlaza los bloques entre sí en una cadena.
- **mrkl_root**: Es una representación de todas las transacciones en el bloque.
- **time**: Es el tiempo en el que se creó el bloque.
- **bits**: Es un valor que representa la dificultad de la prueba de trabajo que se debió realizar para crear el bloque.
- **nonce**: Es un número que se utiliza en la prueba de trabajo.
- **n_tx**: Es el número de transacciones en el bloque.
- **size**: Es el tamaño del bloque en bytes.
- **height**: Es el número de bloques en la cadena antes de este bloque.

[1] https://www.blockchain.com/explorer/api/blockchain_api
[2] https://blockchain.info/rawblock/$block_height

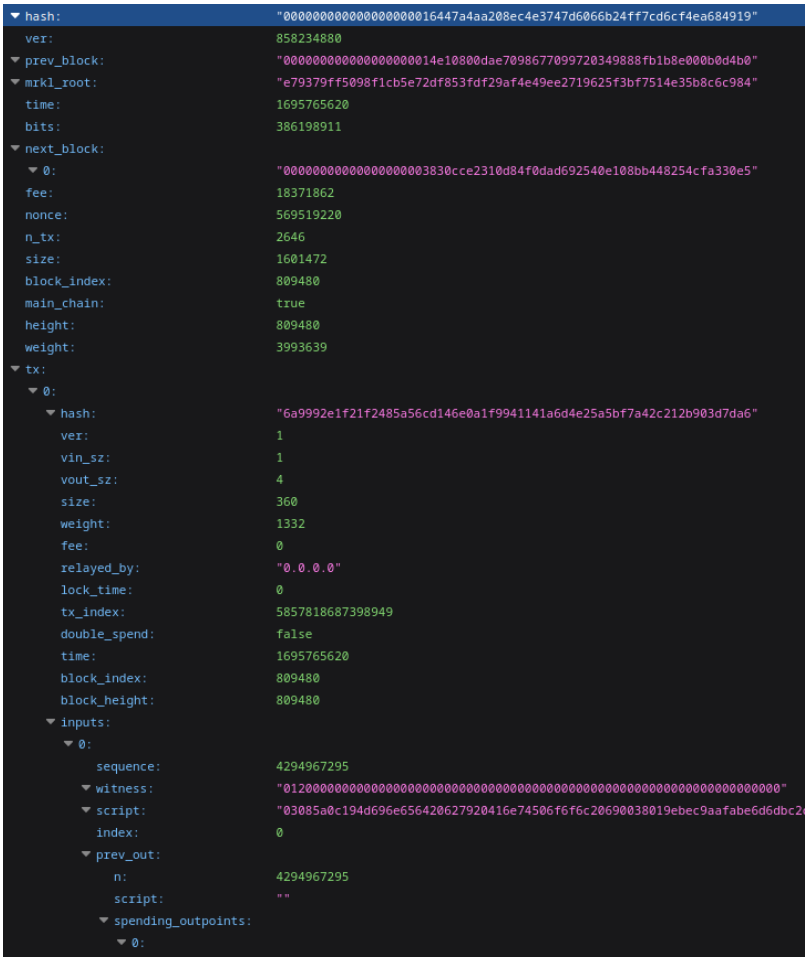
- **tx**: Es una lista de todas las transacciones en el bloque.

Las transacciones a su vez tienen su propia estructura. Debido a que la información relevante para responder las preguntas planteadas se encuentra en las transacciones, a continuación se describen los campos más importantes de las transacciones.

Estructura de una transacción

- **hash**: Es un identificador de la transacción. Es generado a través de una función de hash que toma los datos de la transacción y calcula este identificador.
- **size**: Es el tamaño de la transacción en bytes.
- **vin**: Es una lista de las entradas de la transacción. Cada entrada hace referencia a una salida no gastada de una transacción anterior.
- **vout**: Es una lista de las salidas de la transacción, cada salida especifica la cantidad de Bitcoin y la dirección del destinatario.

Las entradas de las listas de *vin* y *vout* cuentan con la entrada **addr** la cual tiene las direcciones de las diferentes entidades que participan en el minado, envío y recibo de Bitcoin. Este campo es el más importante para poder hacer un seguimiento de las entidades que participan en la blockchain de Bitcoin y conseguir un panorama general sobre quienes utilizan la red.



Información del bloque 809480

Conociendo la estructura de los bloques y las transacciones de la red, se puede proceder a diseñar funciones que permitan recuperar la información relevante para responder a las preguntas planteadas.

Funciones

Estas funciones diseñadas actúan como un filtro que devuelve la información relacionada a una consulta. Además, por cada función que obtiene la información relevante se creó una función que permite su visualización en un gráfico circular.

Las funciones creadas fueron las siguientes:

get_miners(num_blocks)

Esta función devuelve las direcciones que minaron bloques en los últimos *num_blocks* bloques.
El protocolo de Bitcoin especifica que el minero que mina un bloque puede enviarse una cierta cantidad de dinero como recompensa, y esta recompensa siempre se encuentra en la primera transacción (entrada cero de la lista *tx*) de la lista de transacciones. Aprovechando el patrón que dictamina el protocolo, se implementa una función que para cada uno de los bloques se fija en la primera transacción cuál es la dirección que recibe la recompensa.
De esta forma se genera una lista con las direcciones de todos los mineros que minaron un bloque. Además, estas direcciones se guardan en una estructura *Counter* donde se cuenta cuántas veces se repite una dirección, para saber cuantos bloques minó cada dirección.

get_addrs_with_more_in_trxs(num_blocks, num_addresses) enviaron más

Esta función devuelve las *num_addresses* direcciones que realizaron más transacciones enviando Bitcoin en los últimos *num_blocks* bloques. Para ello se itera sobre la lista de transacciones de cada uno de los últimos bloques, y se agregan las direcciones de la lista *vin* de cada transacción a una estructura *Counter*. Finalmente se devuelven las *num_addresses* más usadas.

get_addrs_with_more_out_trxs(num_blocks, num_addresses) recibieron más

De forma análoga a la función anterior, esta devuelve las direcciones que aparecieron en más transacciones pero recibiendo. De igual forma, se itera sobre la lista de transacciones de cada uno de los bloques, y se agregan las direcciones de la lista *vout* de cada transacción a una estructura *Counter*. Finalmente se devuelven las *num_addresses* que más se repiten.

get_addrs_with_more_btc_recieved(num_blocks, num_addresses)

Esta función es similar a la anterior, pero en lugar de contar las direcciones que aparecen en más transacciones recibiendo, guarda cuánto Bitcoin reciben, y devuelve a las *num_addresses* direcciones que más Bitcoin recibieron en los *num_blocks*.
Para lograr esto se itera sobre la lista de transacciones de cada uno de los bloques, y se agrega a una estructura *Counter* la cantidad de Bitcoin que reciben en cada transacción. Para ello se mira dentro de la lista *vout* el valor *value* asociado a la dirección, que representa la cantidad de Bitcoin enviada.

get_addrs_with_more_btc_sent(num_blocks, num_addresses)

Se considera pertinente mencionar que en principio se intentó implementar esta función para obtener las direcciones que enviaron más cantidad de Bitcoin en los bloques recuperados.

Se encontró que la implementación de una función de este tipo resulta más compleja que las otras funciones implementadas debido al uso de alias por parte de los usuarios del sistema. Un mismo usuario puede tener varias direcciones de Bitcoin asociadas, y al enviar Bitcoin es común que los usuarios referencien a varios alias anteriores.
Se plantea implementar un mecanismo de detección de alias para poder resolver esta función como trabajo futuro.

Implementación

Se decidió desarrollar la solución utilizando el lenguaje Python. por la flexibilidad que este permite para prototipar rápidamente, además de la experiencia previa del equipo trabajando con el lenguaje y sus librerías.

Las librerías usadas fueron las siguientes:

- Requests[3]: se utiliza para comunicarse con la API usando el protocolo HTTP.
- Matplotlib[4]: se utiliza para realizar gráficos que permiten visualizar la información obtenida.

Además se utilizó la estructura *Counter* de la librería estándar *collections*, la cual es una estructura de datos performante que facilita el procesamiento de la información.

Se buscó hacer la implementación lo más parametrizable posible para facilitar el análisis en distintos rangos de tiempo. Por ello a todas las funciones se les puede pasar la cantidad de bloques que se desea analizar, y la cantidad de direcciones que se desea incluir en los resultados.

[3] https://github.com/psf/requests
[4] https://github.com/matplotlib/matplotlib

Al experimentar con la API se descubrió que cuando se realizaban muchas solicitudes de forma repetida, el servidor dejaba de responder por unos minutos. Para evitar esto, se implementó una función auxiliar *get_block* la cual se utiliza siempre que se desea recuperar un bloque de la blockchain. Esta función siempre guarda los bloques recuperados a través de la API y evita que se realicen dos o más solicitudes por un mismo bloque. De esta forma se logró evitar el problema para cantidades pequeñas de bloques (144).

Si se deseara trabajar con cantidades mayores de bloques se podría agregar lógica a la función *get_block* para realizar las solicitudes de forma más espaciada en el tiempo y así evitar entrar en la lista negra del servicio.

La solución implementada se encuentra en GitHub[5].

Evaluación y resultados

Los resultados obtenidos se corresponden a los generados por cada una de las funciones definidas previamente.

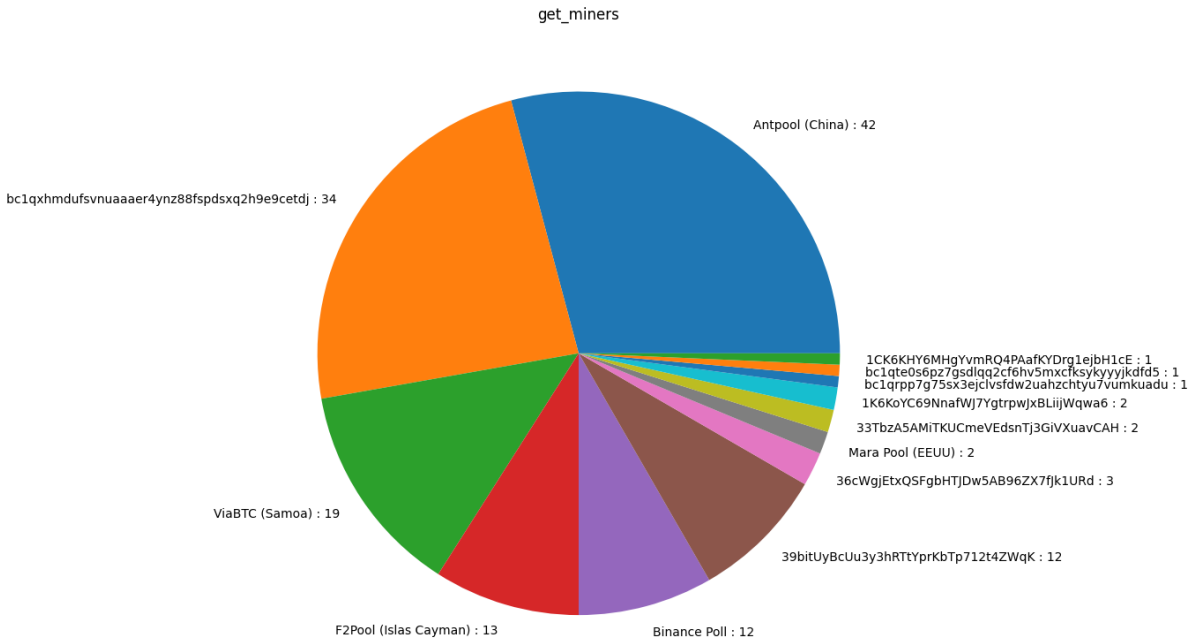
get_miners (obtener mineros)

En el contexto del funcionamiento del sistema Bitcoin, existen usuarios que transaccionan sumas de dinero, haciendo uso de la condición natural de este sistema, y existen otros que se abocan a la tarea de la minería de bloques (también llamados mineros) recibiendo una recompensa monetaria por la misma. Cada usuario posee un identificador o dirección única asociada.
Adicionalmente, existen organizaciones que se dedican exclusivamente a la minería de bloques. Estas están compuestas por un grupo de mineros, eventualmente de distintas locaciones geográficas, que actúan como una única entidad y por ende responden a una misma dirección. A estas organizaciones también se las conoce como *pool de mineros*. El fundamento de estas agrupaciones es aumentar la probabilidad de minar un bloque y distribuir la recompensa económica entre sus mineros. Cabe destacar que la minería de bloques es una tarea que requiere de un alto poder de cómputo y por tanto la probabilidad de que un usuario aislado logre minar un solo bloque, es muy baja. En este sentido, los bloques son minados en mayor proporción por organizaciones del estilo mencionadas. Algunas de ellas poseen un nombre propio mientras que otras permanecen en el anonimato.

El gráfico de abajo expone las organizaciones que más bloques minaron en 24 horas. Es fácil notar que la minería de bloques está principalmente distribuida entre pocas *pools*. Observando la gráfica, la organización China, *Antpool*, mina casi un tercio (42 bloques) de un total de 144 minados en un día. En segundo lugar, le sigue una entidad anónima con 34 bloques minados en un día seguido por otras cuatro que minan entre 19 y 12 bloques.

[5] https://gist.github.com/santiacq/0d70deb0eaf09eea7b2e9ca7b423cf7c

La diferencia entre los dos primeros con respecto a los siguientes cuatro es llamativa e induce a cuestionar la descentralización del sistema, pues en vista de los resultados obtenidos podría significar una cierta concentración acerca del control de la minería de bloques y por ende el control sobre el sistema.



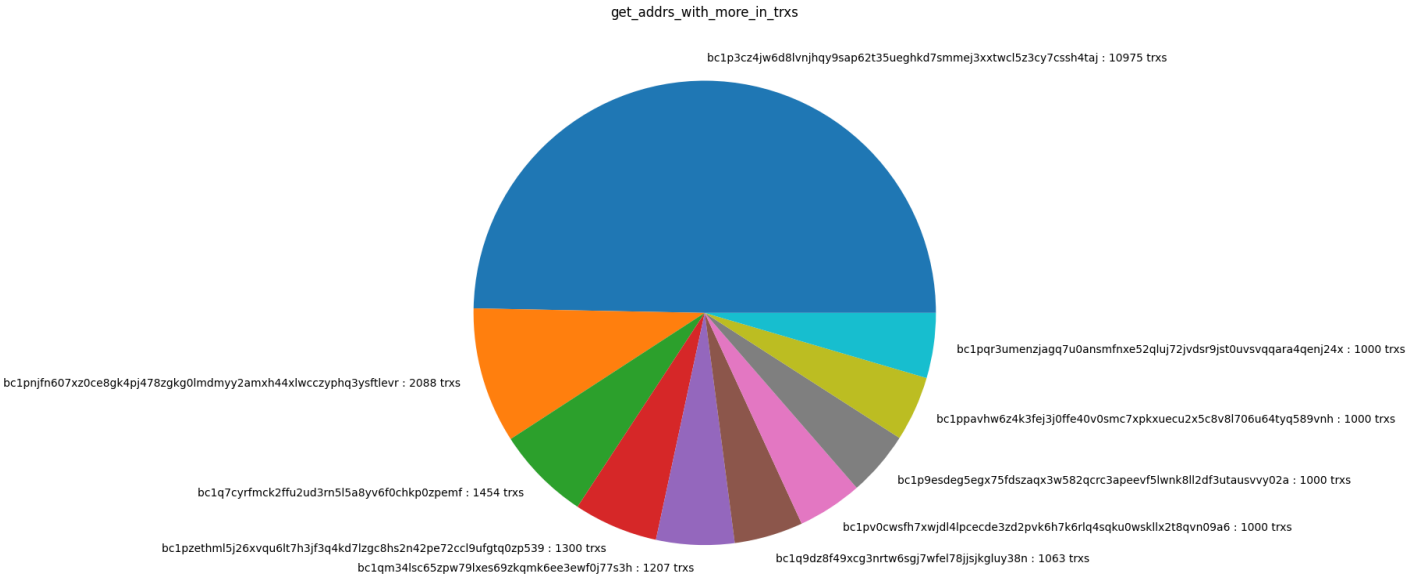
Esto es de suma relevancia dado que dilucida acerca del dominio que podrían disponer ciertas entidades sobre el mecanismo de funcionamiento de este sistema. No obstante, ninguna entidad controla más del 50% de la minería de bloques lo cual es crucial para un correcto funcionamiento del sistema. Existe un posible ataque denominado “ataque del 51%” donde una entidad posee dominio sobre más de la mitad de la minería de bloques, hecho que atenta contra la seguridad el sistema dado que dicha entidad al tener bajo disposición tal poder de cómputo, eventualmente, podría minar una mayor cantidad de bloques sobreponiéndose a las pruebas de trabajo (*proofs of work*) realizadas por otras entidades, pudiendo desencadenar en un acto ilícito donde la entidad maliciosa se hace de dinero no correspondido.

get_addrs_with_more_in_trxs (direcciones que enviaron más transacciones)

Los resultados obtenidos para esta consulta reflejan los usuarios del sistema que efectuaron mayor cantidad de transacciones hacia a otros. Esta información manifiesta qué usuarios realizan un mayor uso de este sistema como medio de pago.

El gráfico circular de abajo expone las diez entidades que enviaron mayor cantidad de transacciones. Observando los resultados obtenidos se aprecia la existencia de una entidad (en azul) que realiza casi la misma cantidad de transacciones (10975) que las nueve entidades

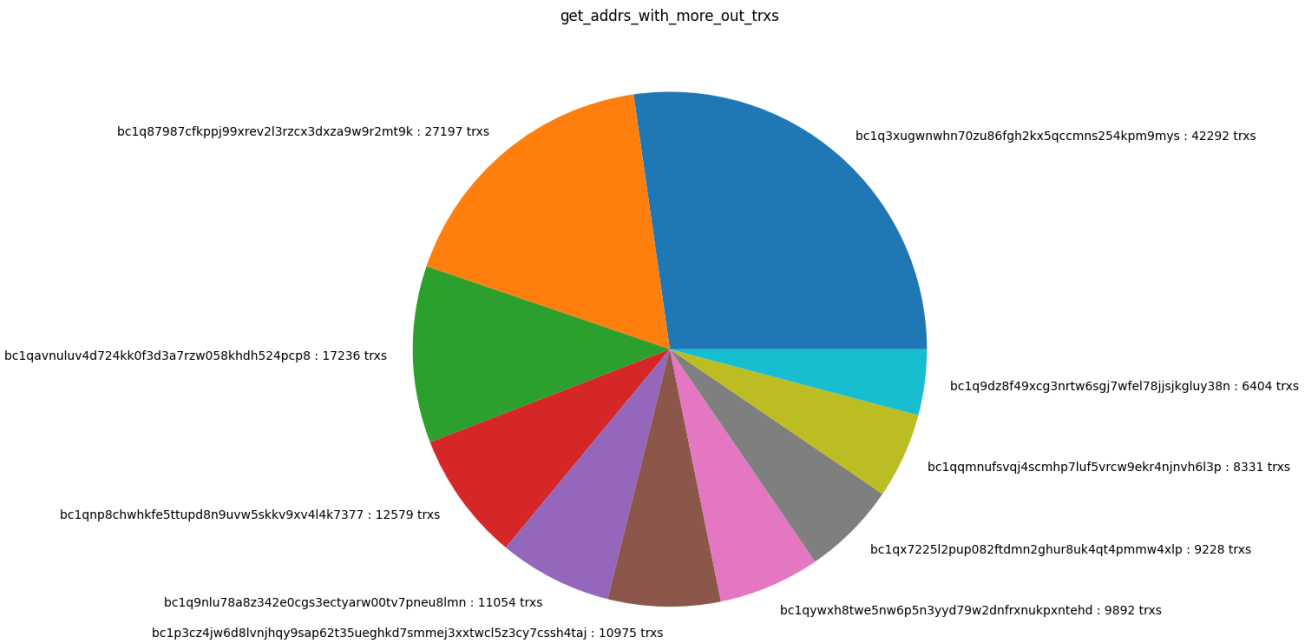
restantes juntas. Esto sugiere que dicha entidad se corresponde probablemente con una organización financiera donde distintos usuarios individuales realizan transacciones por medio de la misma.



get_addrs_with_more_out_trxs (direcciones que recibieron más transacciones)

A diferencia del caso anterior, aquí interesan los usuarios que recibieron mayor cantidad de transacciones. Observando el gráfico de abajo, se nota una distribución más equitativa con respecto al anterior. No obstante, existe cierto predominio por parte de la entidad en azul la cual recibió un total de 42292 transacciones en un día.

Esta función junto a la anterior expresan el uso que efectúan ciertas entidades del sistema. Particularmente, son pocas las entidades que concentran grandes volúmenes de transacciones que se cree que corresponden a organizaciones financieras, por ejemplo, una bastante popular es Binance, la cual ofrece como servicio a sus usuario el cambios de criptomonedas.



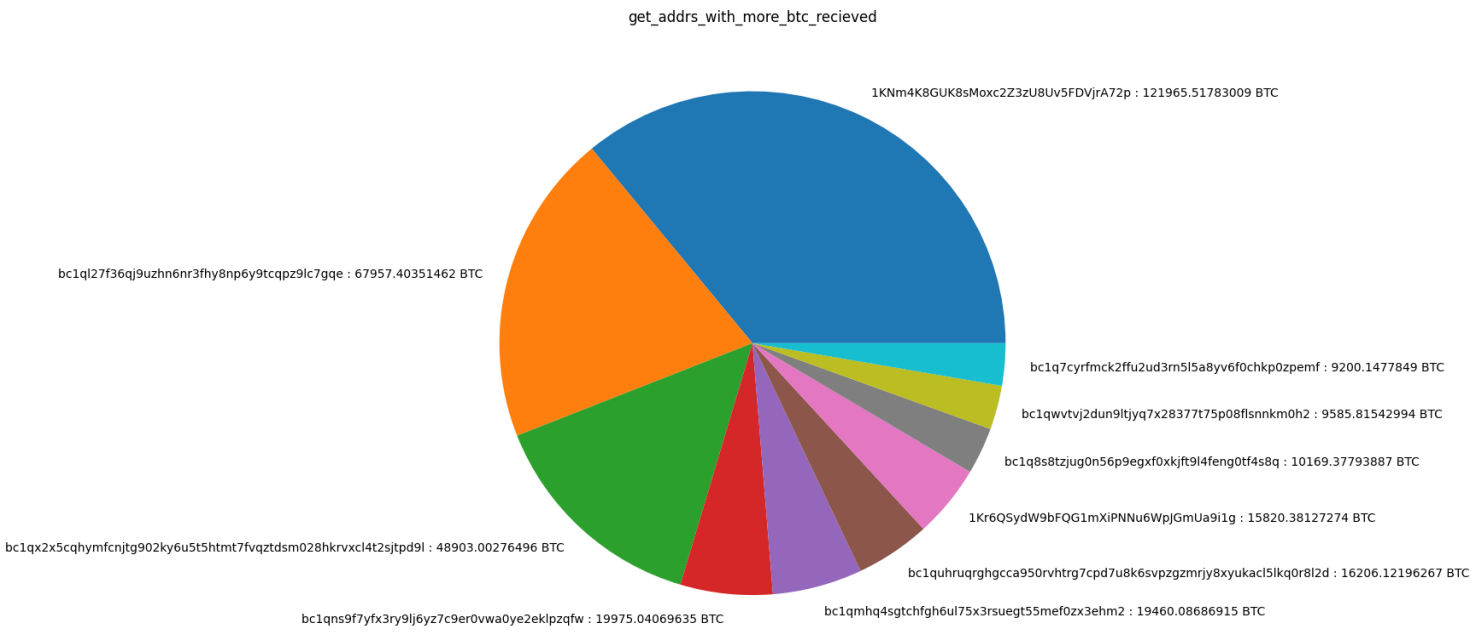
get_addrs_with_more_btc_recieved(direcciones que recibieron más bitcoins)

Resulta de interés visualizar las sumas monetarias que transaccionan los usuarios en la red. En este sentido, la presente función brinda información sobre este aspecto la cual ayuda a comprender un poco más qué tan distribuida se encuentra la riqueza en este sistema.

En el gráfico de abajo se dispone de las diez entidades que mayor cantidad monetarias recibieron. Nuevamente, existen entidades con mayor predominancia que otras (marcadas en azul, anaranjado y verde) que concentran la mayor cantidad de dinero.
La organización en azul recibió un total de 121965 BTC (bitcoins), la cual, tomando el cambio actual, corresponde a 44 mil millones de USD, una cifra exorbitante que dilucida acerca del dinero que se maneja en este sistema. A modo de referencia y comparación, el PBI de Uruguay en 2021 fue de 59 mil millones de USD, teniendo en cuenta que el dinero que recibe la entidad azul corresponde a una única organización y solo en 24hs.

Este resultado parece mostrar que la riqueza en este sistema no está equitativamente distribuída sino que solamente unas pocas entidades concentran la mayor parte de la misma. Cabe destacar que no se tratan de usuarios individuales sino que constituyen organizaciones financieras que como se aludió previamente, ofrecen servicios a otros usuarios para la

transacción de la moneda bitcoin, hecho que explica en gran medida la sumas de dinero que se manejan.



Conclusiones

Para concluir, se intenta responder a las preguntas planteadas inicialmente y se analiza los resultados para ver qué dicen sobre la naturaleza de este sistema de pagos, la distribución de la riqueza en el mismo y qué entidades poseen mayor control sobre la red.

¿Cuántos usuarios minaron bloques y cuantos bloques minaron?

En las 24 horas analizadas se crearon 144 bloques, los cuales fueron analizados por la solución desarrollada. Se encontró que hubo 13 direcciones distintas que minaron bloques, y que de ellas, más del 50% de los bloques fueron minados por dos direcciones.
Como se comentó en la sección de resultados, esto último llama a cuestionar la descentralización de la red, ya que hay dos entidades que juntas concentran más del 50% del poder de minería, y que si lograran coordinar un ataque podrían comprometer la seguridad del sistema.
Sin embargo, esto **no necesariamente** significa que el sistema no sea descentralizado. Las entidades que concentran cantidades tan grandes de poder de minería son las agrupaciones de mineros conocidas como *pools*. Lo que sugiere el resultado obtenido es que hay pools que concentran mucho poder de cómputo, por lo tanto para dar un veredicto sobre la descentralización del sistema sería necesario analizar puntualmente que tan descentralizadas están estas entidades, y cual es su capacidad de coordinar un ataque.

¿Cuáles fueron las cuentas que más transacciones realizaron?

Se observó que hay algunas entidades que realizan un enorme volumen de transacciones. Con una apareciendo en casi la mitad de las transacciones como emisor. Y otras ocupando un rol muy significativo también. Esto se vio tanto en emisores como receptores de las transacciones.
La interpretación que se le da a este resultado es que hay algunas entidades importantes con las que muchos usuarios interactúan. Estas podrían ser instituciones que prestan servicios financieros como cambios, que son utilizados por millones de personas.

Es importante aclarar que este resultado no está relacionado con la descentralización del sistema, ya que eso pasa por la minería y no por las transacciones. Lo que sí se puede intuir de este resultado es que hay algunas instituciones que son actores muy importantes en el ecosistema de Bitcoin, y estas seguramente tengan un rol central en el desarrollo y la evolución del sistema debido a su influencia sobre sus usuarios. Estas instituciones podrían influir en decisiones de una naturaleza más política como lo es la evolución del protocolo.

¿Cuáles fueron las cuentas que más cantidad de Bitcoin transfirieron?

Se observó que una sola dirección recibió 121.965,51 BTC en los 144 bloques analizados. Tomando el cambio actual 1 BTC = 37.803,50 USD equivale a unos 44 mil millones de dólares, esto es aproximadamente el 75% del PBI anual de Uruguay. Esta única dirección recibió más del 25% de los fondos transferidos en las 24 horas analizadas, además entre las tres direcciones que más recibieron se observa que reciben más del 50% de los fondos recibidos en ese periodo de tiempo.
Se concluye que hay algunas entidades dentro del sistema que manejan cantidades enormes de capital dentro de este sistema, comparables incluso con el PIB de un país como Uruguay.

Si bien este resultado sugiere que la riqueza puede estar centralizada en este sistema, para hacer conclusiones definitivas sería necesario ejecutar el programa realizado para todos los bloques de la historia de Bitcoin (un poco más de 800.000) y así poder hacer un conteo de las cuentas más ricas de todo el sistema. Se plantea como trabajo futuro la modificación de la solución para obtener este resultado.

Trabajo Futuro

Como trabajo futuro se plantean dos adiciones a la solución que ayudarían a poder obtener resultados de mayor calidad para tener más información para el análisis de la naturaleza del sistema.
Por un lado sería de interés implementar un mecanismo para la detección de alias en el sistema, lo que permite detectar cuando un usuario o entidad utiliza diferentes direcciones en

el sistema, y así poder ver quienes son las direcciones o entidades que envían mayor capital en la red.
Por otro lado sería pertinente considerar la implementación de una función que realice un recorrido exhaustivo de la cadena de bloques, calculando la riqueza asociada a cada dirección. Este enfoque permitiría obtener una comprensión detallada de la distribución de la riqueza en la red, identificando a los poseedores más prominentes de activos.

No obstante, es crucial destacar que esta propuesta conlleva importantes desafíos técnicos y de infraestructura. El procesamiento de numerosos bloques y la necesidad de almacenar saldos asociados a cada dirección implican una carga considerable en términos de recursos computacionales.