Recupero di indirizzi bitcoin dal web

Università degli Studi di Perugia
Dipartimento di Matematica e Informatica
Corso di Laurea in Informatica

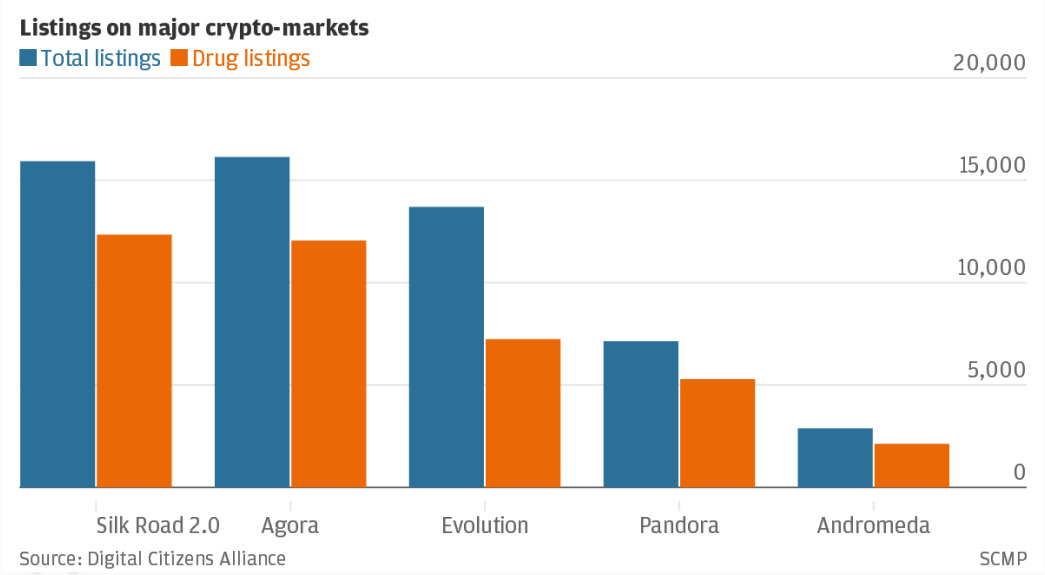
Anno Accademico 2015 - 2016

Laureando
Alessio Santoru

Relatore
Prof. Stefano Bistarelli

Introduzione

I metodi di pagamento anonimi vengono ampiamente utilizzati per **scopi illeciti**: droga, armi o documenti falsi sono solo alcuni dei “prodotti” che è possibile acquistare online con una moneta anonima.



Confronto tra il numero di annunci totali e quelli di droga nei più famosi cripto-mercati

Ransomware

Riscatto da pagare con un metodo di pagamento anonimo: il **bitcoin**

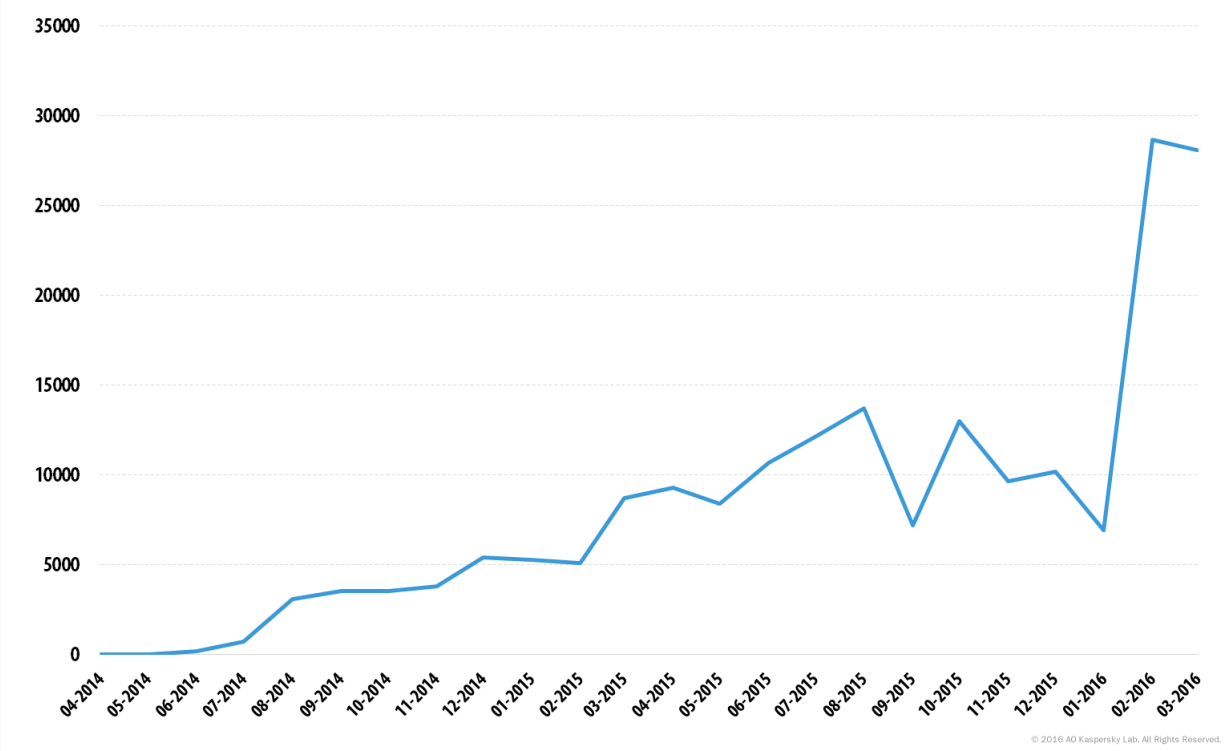


Grafico che mostra il numero di utenti colpiti da un ransomware in un dispositivo mobile (Kaspersky)

Bitcoin: Criptovaluta e Protocollo

Rete bitcoin formata da:

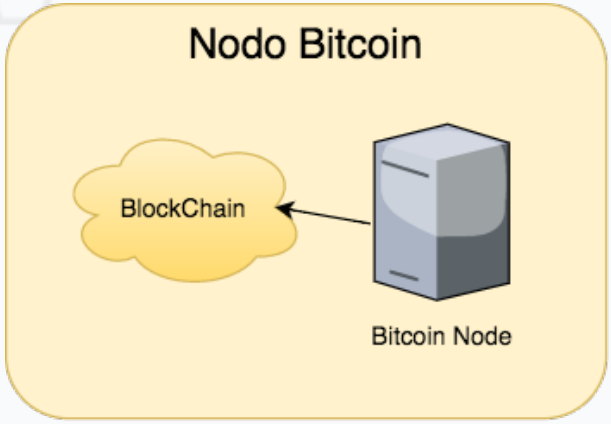
- Blockchain
- Transazioni
- Indirizzi Bitcoin

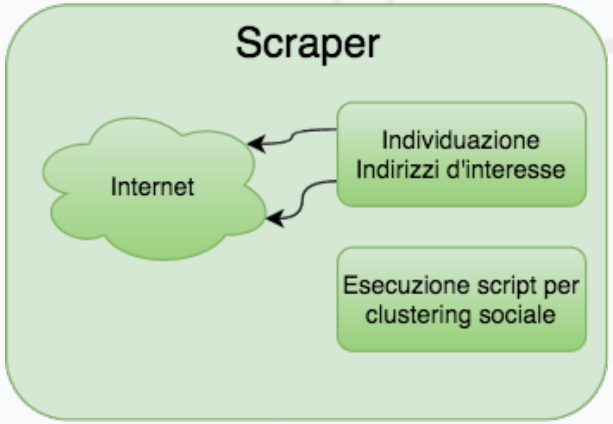
Identificare i pagamenti anonimi

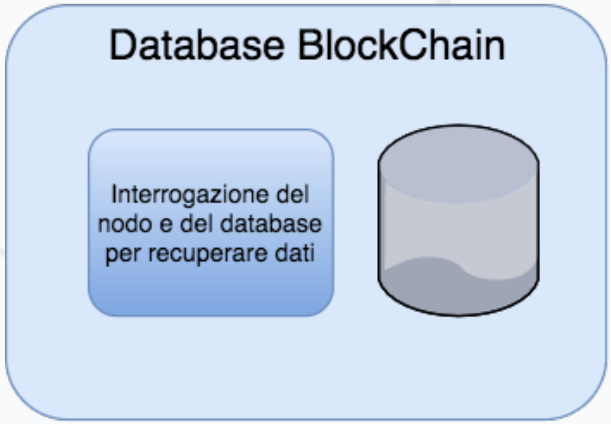
Necessità di:

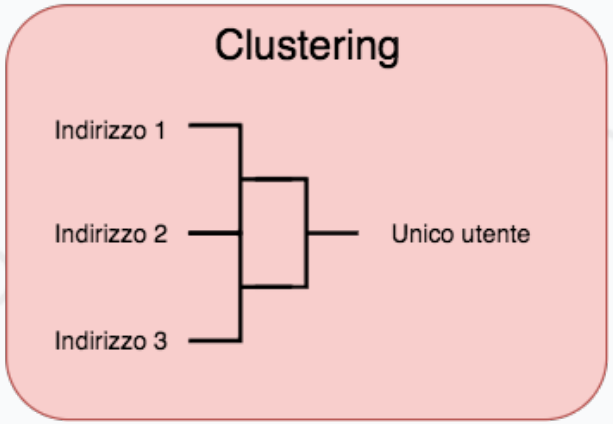
- Recuperare informazioni importanti per l'analisi
- Interrogare le risorse pubbliche disponibili
- Raggruppare sotto un unico utente più informazioni
- Visualizzare le informazioni e i dati ad esse correlate

Un framework modulare

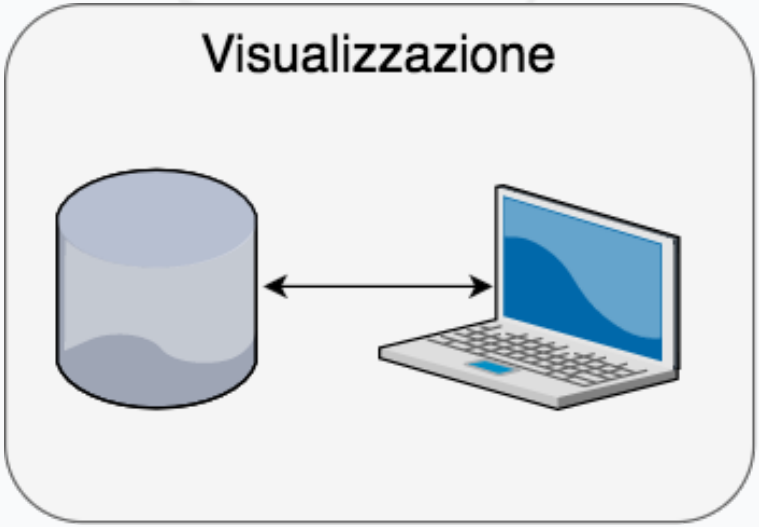


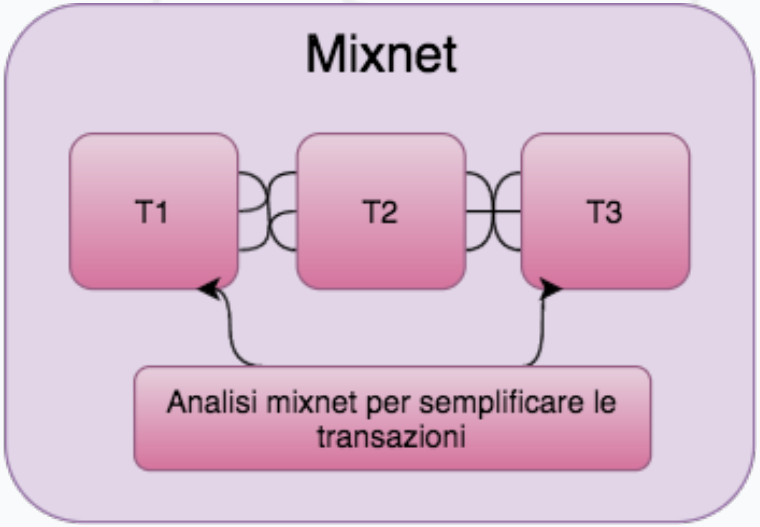




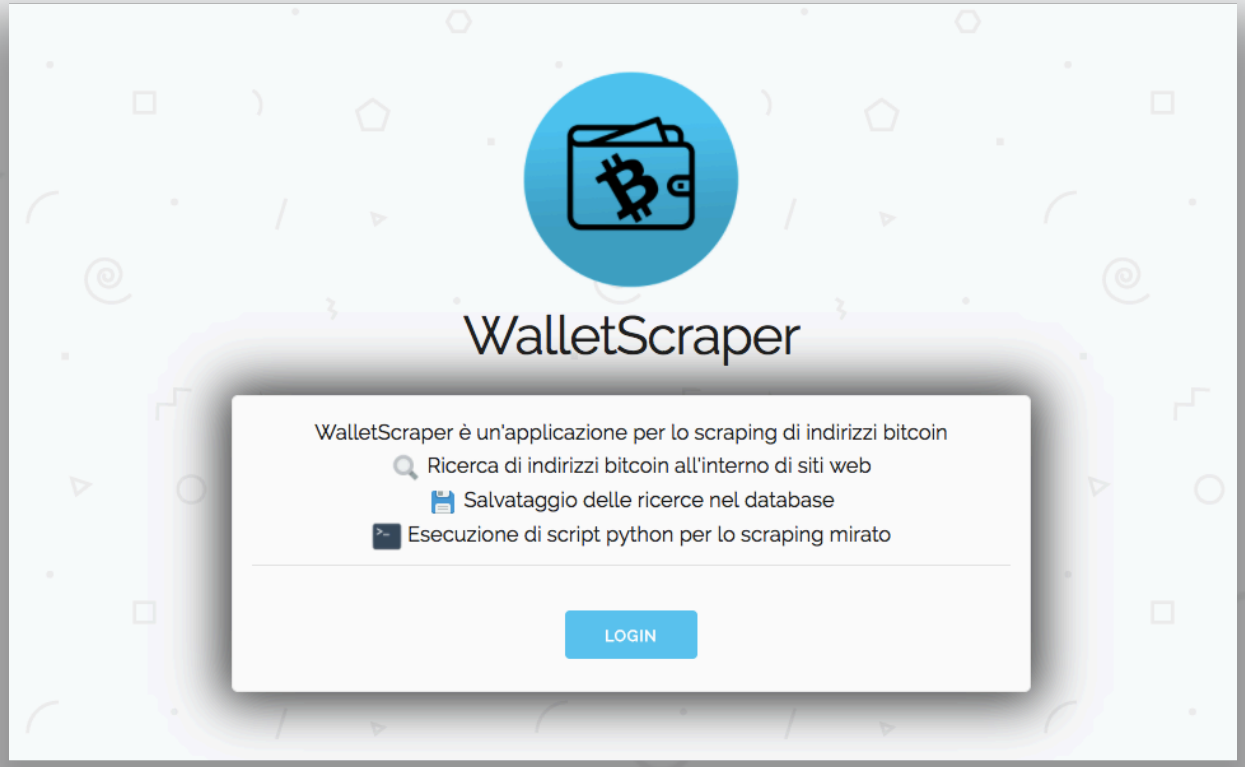


Un framework modulare





Un framework modulare: Scraper



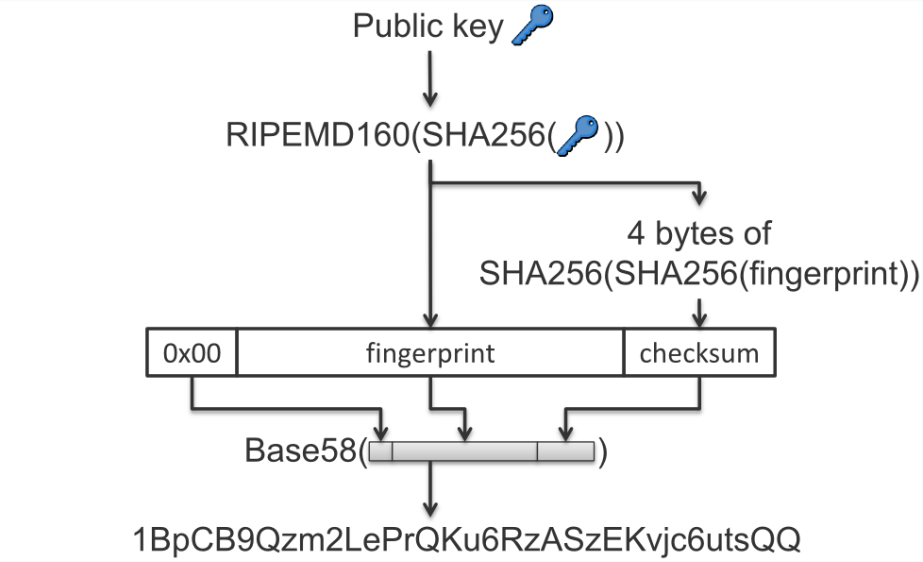
Indirizzi bitcoin

- Stringhe lunghe da 26 a 35 caratteri
- Primo carattere: 1 o 3
- Non contengono caratteri ambigui (O 0 I l) nella codifica utilizzata (base58)

Esempio:

`1BvBMSEYstWetqTFn5Au4m4GFg7xJaNVN2`

Creazione di un indirizzo bitcoin



Coppia di chiavi pubblica/privata ECDSA (Elliptic Curve Digital Signature Algorithm)

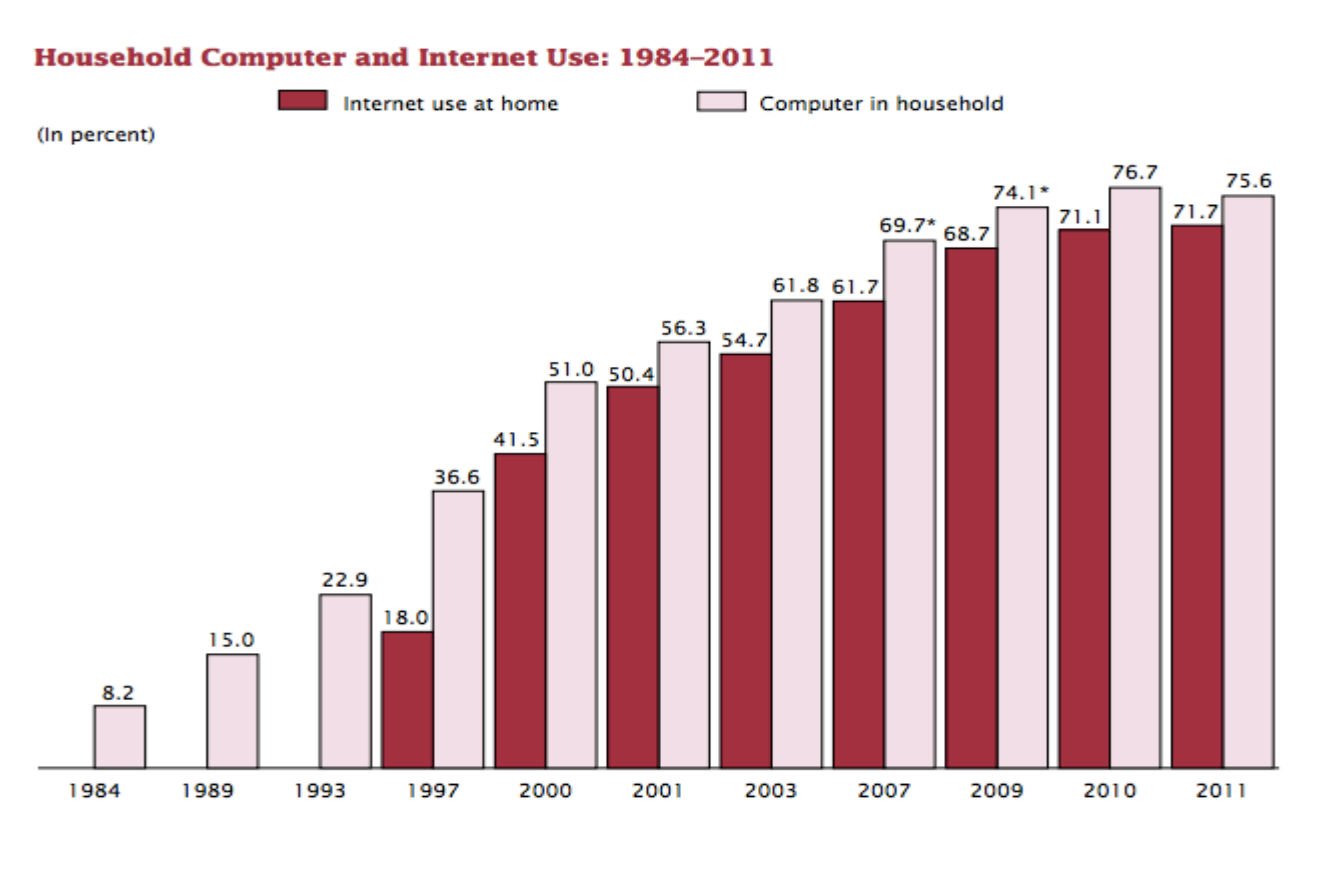
Chiave pubblica

- Compressa: primo byte 0x02 o 0x03
- Estesa: primo byte 0x04

Indirizzo bitcoin (**fingerprint**) ottenuto da

- algoritmo di hashing **ripmed160** eseguito sull'
- algoritmo di hashing **sha256** eseguito sulla chiave pubblica.

Recupero di informazioni dal web



Presenza di computer e connessioni ad internet nelle famiglie americane nel corso degli ultimi 30 anni

Recupero di informazioni dal web

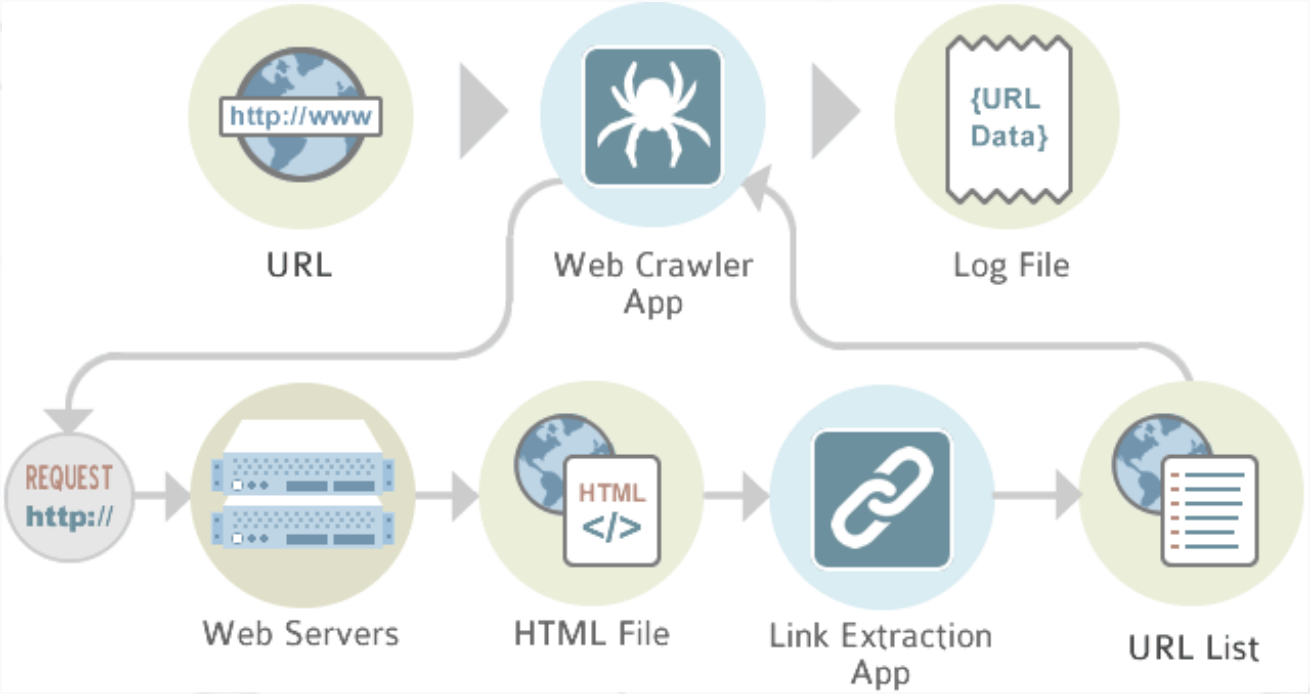
Utilizzi:

- Indicizzazione da parte dei motori di ricerca
- Calcolo statistico
- Marketing

Tecniche più utilizzate:

- Web crawling
- Web scraping

Web crawling

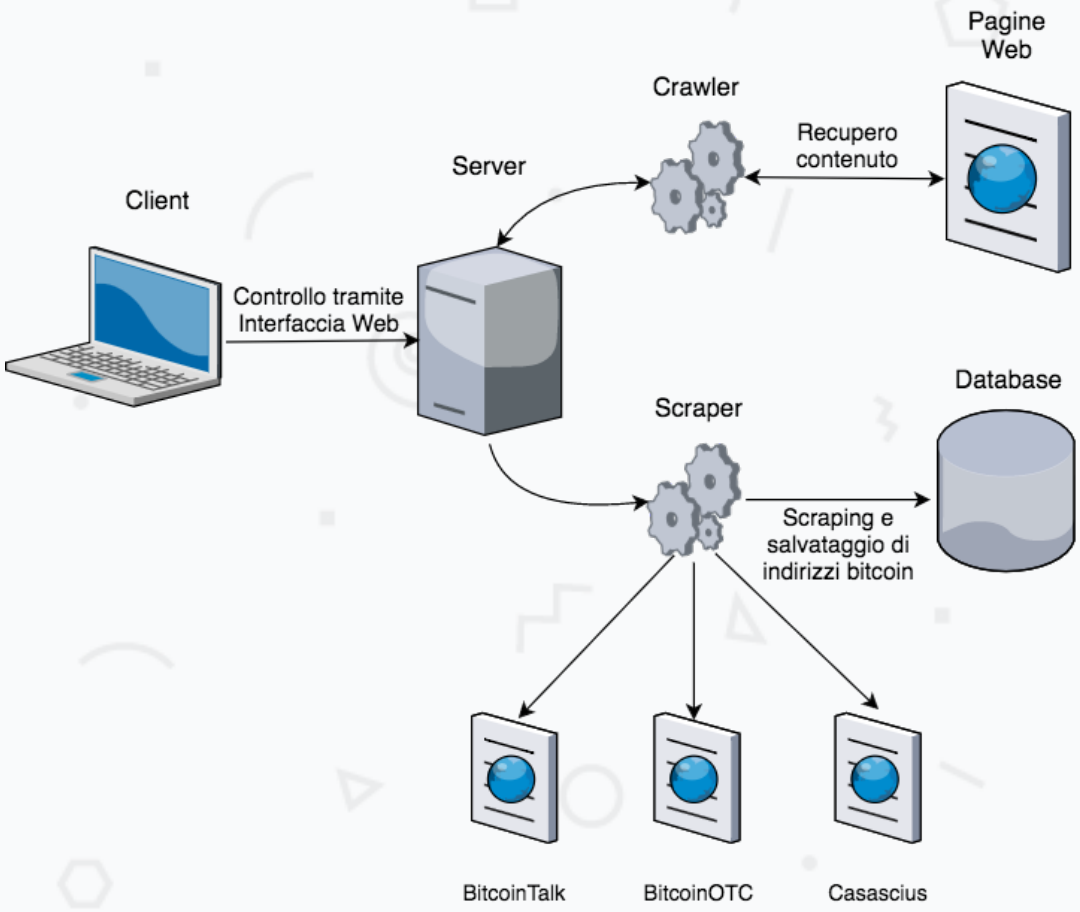


Il web crawling è il processo di analisi di siti web utilizzato per indicizzarne tutti i contenuti. Un crawler, anche conosciuto come spider, è un software specializzato per prelevare tutto il contenuto di una pagina web e seguirne i vari link per analizzare siti web collegati o pagine secondarie.

Web scraping

Il web scraping, chiamato anche in diversi modi tra cui *web data extraction*, è una tecnica, di solito automatizzata, che consiste nel prelevare singoli dati da un insieme di pagine web, per collezionarli all'interno di database o file per un'analisi futura.

Realizzazione di uno scraper



Uno scraper “dinamico”

Implementazione: Dinamica per permettere il funzionamento su ogni sito web.

Come?

Soluzione: basare la ricerca sulla struttura del dato, piuttosto che sulla struttura della pagina.

Ricerca degli indirizzi bitcoin

Il codice sorgente della pagina web viene considerato come una stringa.

Due passaggi fondamentali per la ricerca:

- Utilizzo di un’espressione regolare per ottenere dal codice sorgente i possibili indirizzi bitcoin.
- Utilizzo di un algoritmo di validazione per verificare l’esatta presenza di un indirizzo bitcoin tra quelli possibili.

Espressione regolare

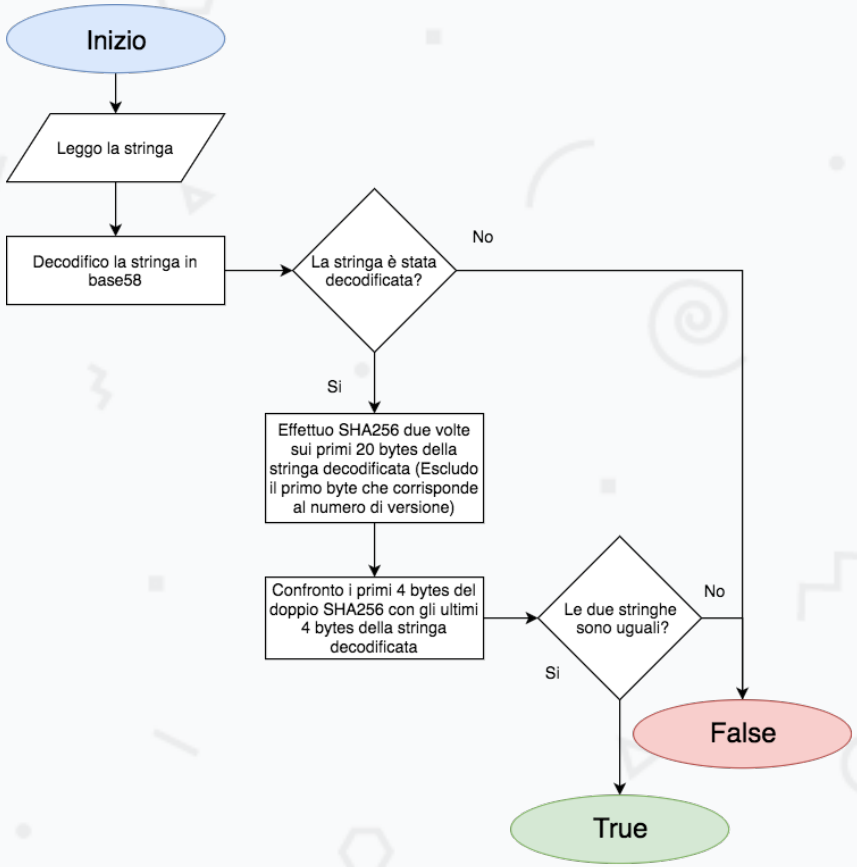
L’espressione regolare

```
$regex = ’/[13][a-km-zA-HJ-NP-Z1-9]{25,34}/’;
```

Identifica le stringhe che:

- Iniziano con 1 o con 3 ([13])
- Sono formate solo da caratteri ammessi ([a-km-zA-HJ-NP-Z1-9])
- I caratteri ammessi sono presenti dalle 25 alle 34 volte.

Algoritmo di validazione



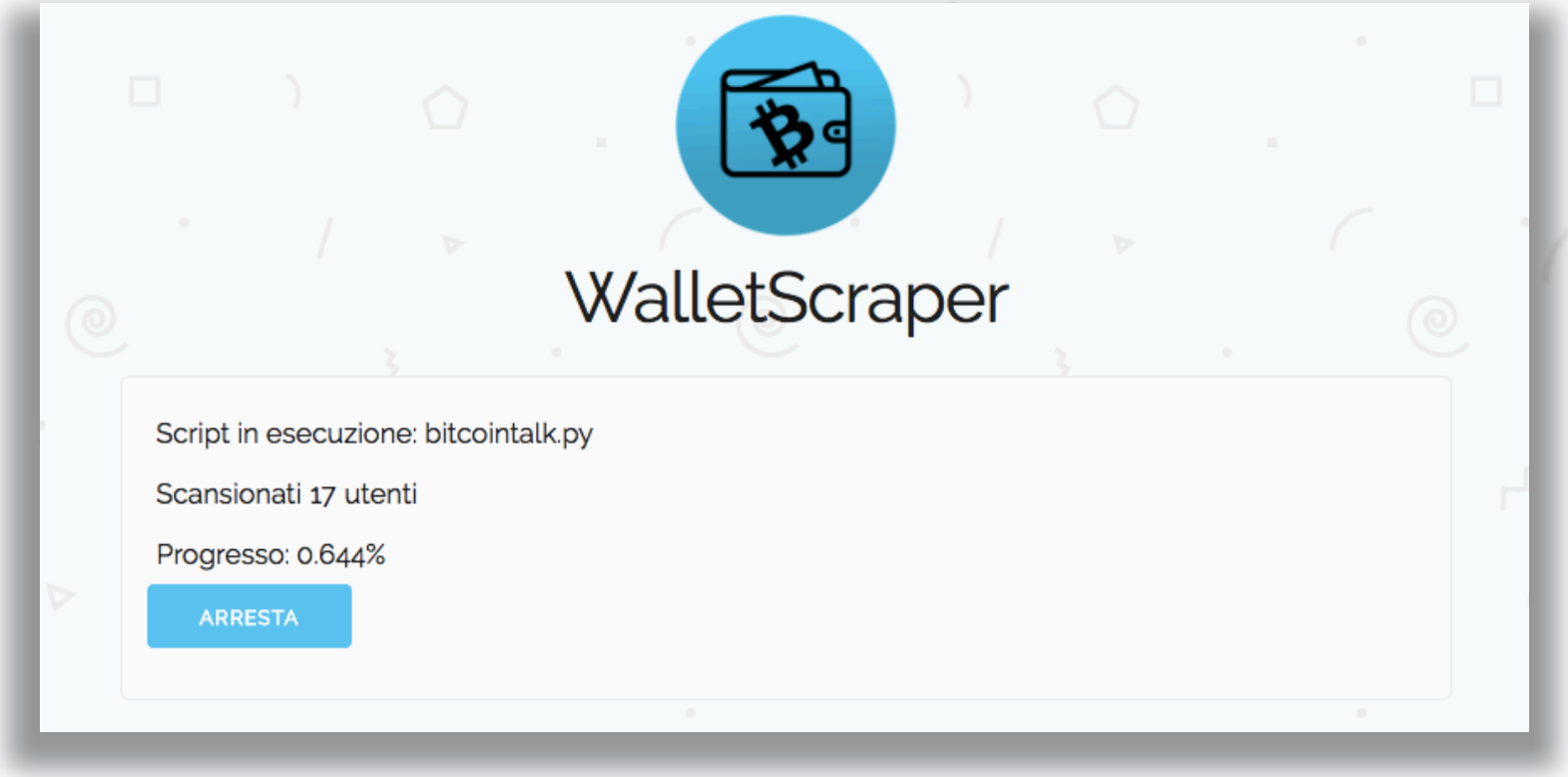
Il processo di validazione svolge un'operazione inversa rispetto alla creazione di un indirizzo bitcoin.
Dato un indirizzo bitcoin.
L'indirizzo viene decodificato da base58.
La stringa risultate è formata da:

- 1 byte che identifica il numero di versione
- 20 bytes che identificano la chiave pubblica
- 4 bytes per il codice di controllo

Siccome il codice di controllo è generato prendendo i primi 4 bytes del doppio algoritmo di hashing sha256 effettuato sui 20 bytes della chiave pubblica, l'algoritmo provvede a controllare che il codice di controllo da lui generato sia lo stesso di quello presente nell'indirizzo.

Esecuzione dello scraper

Esecuzione di script python



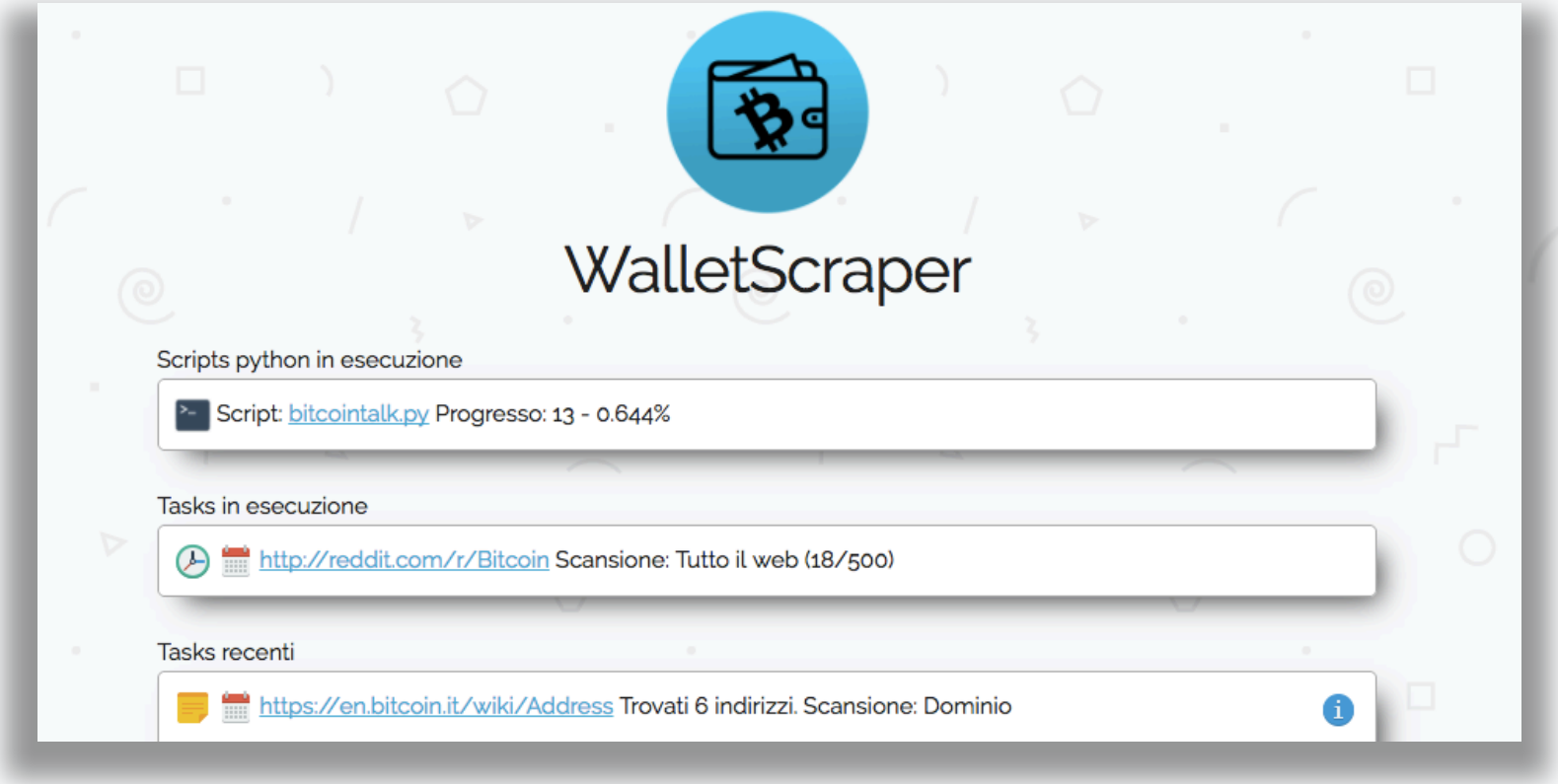
Scraping mirato su siti web “interessanti”

- BitcoinTalk
- Bitcoin-OTC
- Casascius

Controllo degli script python

- Gli script vengono lanciati in maniera asincrona
- La loro esecuzione viene controllata tramite dei contatori inseriti nel database.
- È possibile recuperare l'esecuzione di uno script lanciato in background.
- Gli script possono salvare il loro stato nel database, così in caso di arresto anomalo possono riprendere la loro esecuzione senza dover ripartire dall'inizio.

Controllo dei task



Task recenti: lista di task terminati recentemente con un riassunto sul risultato.
Task e script python in esecuzione: mostrati insieme al loro progresso, aggiornato in tempo reale.

Possibili miglioramenti futuri

- Migliorare l’esecuzione di task particolarmente pesanti per non incorrere in blocchi imposti dal PHP (max_execution_time, memory_limit).
- Migliorare la visualizzazione dei dati inseriti nel database, così come la ricerca dei dati stessi.

Recupero di indirizzi bitcoin dal web

Laureando

Alessio Santoru

Relatore

Prof. Stefano Bistarelli