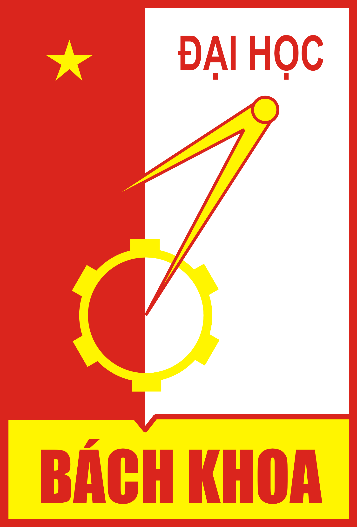
TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI

VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

──────── \* ───────



**BÁO CÁO ĐỒ ÁN TỐT NGHIỆP**

**DATA MINING BLOCKCHAIN**

**Sinh viên thực hiện:**

**Phan Hồng Lĩnh 20142626**

**Giáo viên hướng dẫn: TS. Nguyễn Bình Minh**

*Hà Nội, ngày 29 tháng 08 năm 2018*

**MỤC LỤC**

[I. Những vấn đề đã tìm hiểu được 2](#_Toc523351145)

[1. Tổng quan về bitcoin, blockchain 2](#_Toc523351146)

[2. Phân tích dữ liệu blockchain 2](#_Toc523351147)

[II. Công việc và mục tiêu triển khai 2](#_Toc523351148)

[1. Đề tài 2](#_Toc523351149)

[2. Mục tiêu 2](#_Toc523351150)

[3. Phương pháp triển khai 2](#_Toc523351151)

[III. Danh mục tài liệu tham khảo 3](#_Toc523351152)

# Những vấn đề đã tìm hiểu được

## Tổng quan về bitcoin, blockchain

* Blockchain là gì
* Tổng quan về blockchain, bitcoin
* Cấu trúc của blockchain
* Cách thức blockchain hoạt động
* Cơ chế đồng thuận,..
* …

## Phân tích dữ liệu blockchain

* Dựa trên paper:

**BitIodine: Extracting Intelligence from the Bitcoin Network**

<https://www.ifca.ai/fc14/papers/fc14_submission_11.pdf>

* Từ dữ liệu của blockchain (cụ thể là bitcoin), trích rút thông tin để phân loại người dùng, phá bỏ sự ẩn danh, cảnh báo scam,… thông qua đồ thị giao dịch và đồ thị người dùng trong mạng blockchain.

# Công việc và mục tiêu triển khai

## Đề tài

Khai phá dữ liệu blockchain

## Mục tiêu

Xây dựng hệ thống phân tích dữ liệu blockchain, từ đó phân loại các đối tượng địa chỉ, truy xuất, theo dõi thông tin người dùng thực tế.

## Phương pháp triển khai

* Dựa trên ý tưởng: Blockchain không hoàn toàn ẩn danh:

Mọi giao dịch được công khai. Vì thế, nếu biết địa chỉ của 1 user, ta có thể theo dõi mọi giao dịch của người đó trên mạng blockchain.

* Phân tích dữ liệu dưới dạng đồ thị: Các node là các địa chỉ, cạnh sẽ là các giao dịch.

Thông qua 1 số phương pháp phân tích đồ thị, nhóm các địa chỉ thành từng user.

* Gán nhãn thông qua dữ liệu gán nhãn của 1 số forum về bitcoin như Bitcoin-Talk, Bitcoin-OTC, các scammer, malware đã được xác thực bởi Bitcoin talk trust system.
* Dựa vào 2 heuristic:

+, Một số giao dịch có nhiều input (do trong ví có nhiều địa chỉ, nhưng tổng số dư ko đủ trả cho 1 khoản thanh toán) => Những địa chỉ này đều thuộc cùng 1 users

+, Bitcoin protocol buộc mỗi giao dịch đầu ra sẽ được sử dụng như đầu vào của 1 giao dịch mới. => Những giao dịch chưa tiêu (Unspend output) sẽ được sử dụng làm input => Bitcoin sử dụng shadow address gom các số dư còn lại trả về cho người dùng. => Đoán xem 1 trong những địa chỉ đầu ra sẽ thuộc về cùng 1 người dùng.

**Luồng thực thi:**

* Thu thập dữ liệu blockchain (Có thể lấy từ API của trang nào đó cung cấp)
* Lưu dữ liệu vào database (Có thể dùng Mysql)
* Nhóm dữ liệu địa chỉ thành các cluster, visualize đồ thị.
* Gán nhãn thông qua dữ liệu username, label từ các website, forum Bitcoin.
* Chỉ ra các scammer, malware, 1 số nhóm các địa chỉ nôi bật.
* Module xuất ra 1 phần của đồ thị, tìm kiếm, phân tích thủ công 1 địa chỉ, cluster cụ thể nào đó, có những luồng vào ra nào với các người dùng xung quanh.

# Danh mục tài liệu tham khảo

1. *BitIodine: Extracting Intelligence from the Bitcoin Network*

<https://www.ifca.ai/fc14/papers/fc14_submission_11.pdf>