**Phát hiện đăng nhập bất thường**

1. **Phát biểu bài toán**

Cho tập dữ liệu đăng nhập, xác định các sự kiện đăng nhập bất thường trong tập dữ liệu đầu vào.

* 1. **Mô tả dữ liệu**

Dữ liệu đầu vào gồm 1 tập các sự kiện đăng nhập, một sự kiện đăng nhập

đặc trưng bởi các thuộc tính sau:

- date: ngày đăng nhập

- account id: id của người đăng nhập

- src mask: dải ip chứa địa chỉ ip thực hiện việc đăng nhập

- target mask: dải ip chứa địa chỉ ip được đăng nhập

- src ip: địa chỉ ip thực hiện việc đăng nhập

- target ip: địa chỉ ip được đăng nhập

- src ip: địa chỉ ip thực hiện đăng nhập

- hour of day: thời điểm (giờ) đăng nhập trong ngày {0, 1,…,24}

- day of week: ngày đăng nhập trong tuần {Monday, Tuesday,…, Sunday}

- logon type: kiểu đăng nhập {2, 3, 4, 5, 7, 8, 9, 10, 11}

* 1. **Định nghĩa sự kiện đăng nhập bất thường**

**Định nghĩa 1:** Xét sự kiện đăng nhập là tập các kiện đăng nhập có cùng account id với sự kiện .

**Định nghĩa 2:** Xét sự kiện đăng nhập và thuộc tính là giá trị của thuộc tính trong sự kiện

**Định nghĩa 3:** Xét thuộc tính và thuộc tính v, là tập các sự kiện đăng nhập thỏa mãn:

**Định nghĩa 4:** Xét không gian , thuộc tính , là tập các sự kiện thỏa mãn:

**Định nghĩa 5:** Giả sử là tập các cụm thỏa mãn , gọi là 2 tham số đặc trưng cho ranh giới để xác định cụm bất thường nếu thỏa mãn 1 trong các điều kiện sau:

Trong đó, . Tập các cụm nhỏ được xác định bởi , còn lại là các cụm lớn.

Một sự kiện đăng nhập được coi là bất thường nếu thuộc một trong các trường hợp sau:

TH1:

* Coi một user được đặc trưng bởi cặp (account \_id, src\_ip), khi kẻ tấn công (E) chiếm được tài khoản của user, E đăng nhập vào target\_ip mà user không hoặc rất ít khi đăng nhập (số lần đăng nhập vào target\_ip là ít hơn so với các target\_ip khác) được coi là bất thường

TH2:

* Coi một user được đặc trưng bởi cặp (account \_id, target\_ip), khi kẻ tấn công - E chiếm được tài khoản của user, E vẫn đăng nhập vào target\_ip mà user thường xuyên đăng nhập tuy nhiên việc đăng nhập được thực hiện từ src\_ip mà user không hoặc rất ít khi sử dụng để đăng nhập (số lần đăng nhập từ src\_ip là ít hơn so với các src\_ip khác) được coi là bất thường

TH3:

* Coi một user được đặc trưng bởi bộ 3 (account \_id,src\_ip, target\_ip), khi kẻ tấn công - E chiếm được tài khoản của user, E vẫn đăng nhập vào target\_ip từ src\_ip mà user thường xuyên thực hiện tuy nhiên E đăng nhập bằng logon\_type mà user không hoặc rất ít khi sử dụng để đăng nhập (số lần đăng nhập bằng logon\_type là ít hơn so với các logon\_type khác) được coi là bất thường

1. **Phương pháp giải quyết bài toán**
   1. **Phân cụm**

* Chia thời gian đăng nhập thành 4 khung giờ:

+ 0: tương ứng với 8h – 12h

+ 1: tương ứng với 13h – 18h

+ 2: tương ứng với 19h – 22h

+ 3: tương ứng với các thời gian còn lại

* Với mỗi bộ 3 (account\_id, src, des), xác định vector đặc trưng cho hành vi của user(account\_id, src) và user(account\_id, des)

|  |
| --- |
|  |
|  |
| **…** |
|  |
|  |
| **…** |
|  |
|  |
| **…** |
|  |

**+ :** số lần đăng nhập tương ứng với bộ 3 (account\_id, src, des)**+ ,…,:** số lần đăng nhập vào Monday, Tuesday,….,Sunday

**+ ,…,:** số lần đăng nhập vào khung giờ 0,1,2,3

**+ ,…,:** số lần đăng nhập bằng logon\_type 2, 3, 4, 5, 7, 8, 9, 10, 11

* Sau khi xây dựng xong tập vector tương ứng với (account\_id, src, des), phân cụm sử dụng thuật toán Kmean trên tập vector
* Gọi là k cụm sau khi áp dụng Kmean**,** áp dụng **Định nghĩa 5** để xác định các cụm nhỏ (SC) và các cụm lớn (LS)
* Với 1 sự kiện đăng nhập , xác định vector tương ứng với bộ 3

(p.account\_id, p.src, p.des)

+ gọi là điểm đánh giá độ bất thường của

+ gọi là khoảng cách trung bình của cụm