Présentation des données

Le boson de Higgs est une particule élémentaire dont l’existence est postulée en 1964. De nombreuses recherches pour prouver son existence ont été conduites pendant près de 50. En 2012, deux expériences conjointes menées par ATLAS et CMS ont réussi à prouver l’existence du boson de Higss. L’annonce de cette découverte a été faite via les media sociaux notamment, elle a généré des discussions, des post, des mentions. Toutes ces informations, disponibles publilquement, ont été rassemblées sur une période de 6 jours : du 1er juillet 2012 au 7 juillet 2012 (l’annonce ayant été faite le 4 juillet 2012).

Les données ont été concaténées sous forme de quatre bases de données :

1. re-tweeting (retweet network)
2. replying (reply network) to existing tweets
3. mentioning (mention network) other users
4. friends/followers social relationships among user involved in the above activities
5. information about activity on Twitter during the discovery of Higgs boson

Il s’agit de graphes orientés.

Problématique :

Dans quelle mesure l’analyse et la visualisation de graphes peuvent permettre :

* d’identifier les leaders d’opinion dans le domaine de la physique quantique (boson de Higgs)
* de comprendre la dynamique de diffusion de l’information sur Twitter

On notera que les données ont commencées à être enregistrées avant l’annonce officielle de la découverte du boson de Higgs. On pourra étudier la dynamique de la diffusion de la rumeur puis de la diffusion de l’information.

Présentation des graphes :

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Socia Network | Retweet Network | Reply Network | Mention Network |
| Nodes | 456626 | 256491 | 38918 | 116408 |
| Edges | 14855842 | 328132 | 32523 | 150818 |
| Nodes in largest WCC | 456290 (0.999) | 223833 (0.873) | 12839 (0.330) | 91606 (0.787) |
| Edges in largest WCC | 14855466 (1.000) | 308596 (0.940) | 14944 (0.459) | 132068 (0.876) |
| Nodes in largest SCC | 360210 (0.789) | 984 (0.004) | 322 (0.008) | 1801 (0.015) |
| Edges in largest SCC | 14102605 (0.949) | 3850 (0.012) | 708 (0.022) | 7069 (0.047) |
| Average clustering coefficient | 0.1887 | 0.0156 | 0.0058 | 0.0825 |
| Number of triangles | 83023401 | 21172 | 244 | 23068 |
| Fraction of closed triangles | 0.002901 | 0.0001085 | 0.0001561 | 0.0002417 |
| Diameter (longest shortest path) | 9 | 19 | 29 | 18 |
| 90-percentile effective diameter | 3.7 | 6.8 | 10 | 6.5 |

Approche :

Le graphe que nous allons construire représente un réseau social dont les nœuds sont les individus et arcs orientés les interactions (retweet, reply ou mention) qui lient ces individus.

Il faut définir sur quels critères on va identifier les leader d’opinion. On se base sur les critères de Alloing et Haikel-Elsabeh. Selon eux, un leader d’opinion se définit par ses caractéristiques :

* structurelle : volume des contacts et centralité
* énonciative : volume et rythme de production (nous ne disposons pas de ces données)
* informationnelle : volume de reprise d'information (nombre de retweet)

Les leaders opinion - au sens d'Alloing, sont repérables par leurs caractéristiques structurelles (), caractéristiques énonciatives), caractéristiques informationnelles (

Le degré d’un sommet : nombre d’arcs entrants ou sortants de ce sommet.

Centralité d’un sommet : position relative d’un sommet par rapport aux autres sommets

Construction d’un graph property

Les sommets sont caractérisés par :

Les arcs sont caractérisés par :

Pour idnetifier les leaders d’opinion, on crée un indice de popularité composé. Sur la bae des critères utiilisés pour le score de Klout. Le score Klout est élaboré autour de trois mesures spécifiques :

* la « portée réelle » : nombre d’abonnés (liens sociaux entrants)
* la valeur d' « amplification » : nombre d’interactions provoquées (retweet, mention, réponse entrants)
* l' « impact sur le réseau » : ratio entre abonnés et abonnements (liens sociaux entrants vs liens sociaux sortants). Ce critère reflète l'influence de l'audience interagissant avec le contenu posté par l'utilisateur évalué.

L’indicateur que l’on construit est la concaténation des indices de Page Rank issu de chacun des graphe.

//1.2 - Leaders d'opinion - on construit un graphe qui ura des sommets : (Id, followers) et des arcs définis par (source, destination,interactions). "interactions" correspond au nombre de rewteet + nb de réponses + nombre de mentions. Le graphe est orienté.

//(Mon ordi n'a pas assez de mémoire pour traiter séparément les différents types d'arcs. Je les regroupe donc sous un même libellé "interactions"

un sommet est défini par son Id (VertexId), et ses caractéristiques : nombre de liens sociaux -degré-, nombre de followers-degré entrant-, nombre de fois où il est retweeté, nombre de fois où il a retweeté,nombre de fois où il a répondu, nombre de fois où il a été mentionné

// Popularité d'un individu - on ajoute le critère du PageRank

//PageRank algorithm - individus les plus populaires sur base du nombre de relation sociales, du nb de followers, du nombre de mentions et de réponses

// on calcule le Page Rank pour chacun des graphe d'interaction puis on additionne les différents indicateurs

remarque : l'individus 1503 n'a que des liens sortants. Il se peut qu'il ne s'agisse pas d'une personne physique mais d'une "personne morale" : compte twitter relié aux publications d'un journal par exemple."

Répartition des degrés

Partie 1. Etude des graphes – leader d’opinion

Partie 2. Etude de la dynamique de graphe

4 juillet

Nombre de tweet par heure en fonction des jours

Partie 3. Visualisation