Stack Overflow Analysis

Analysing various graphs,metrics and predicting content quality of questions

Yash Bajaj[1] Vivek Aditya[1]

[1]Department of Computer Science and Engineering
IIT Kharagpur

March 10, 2015

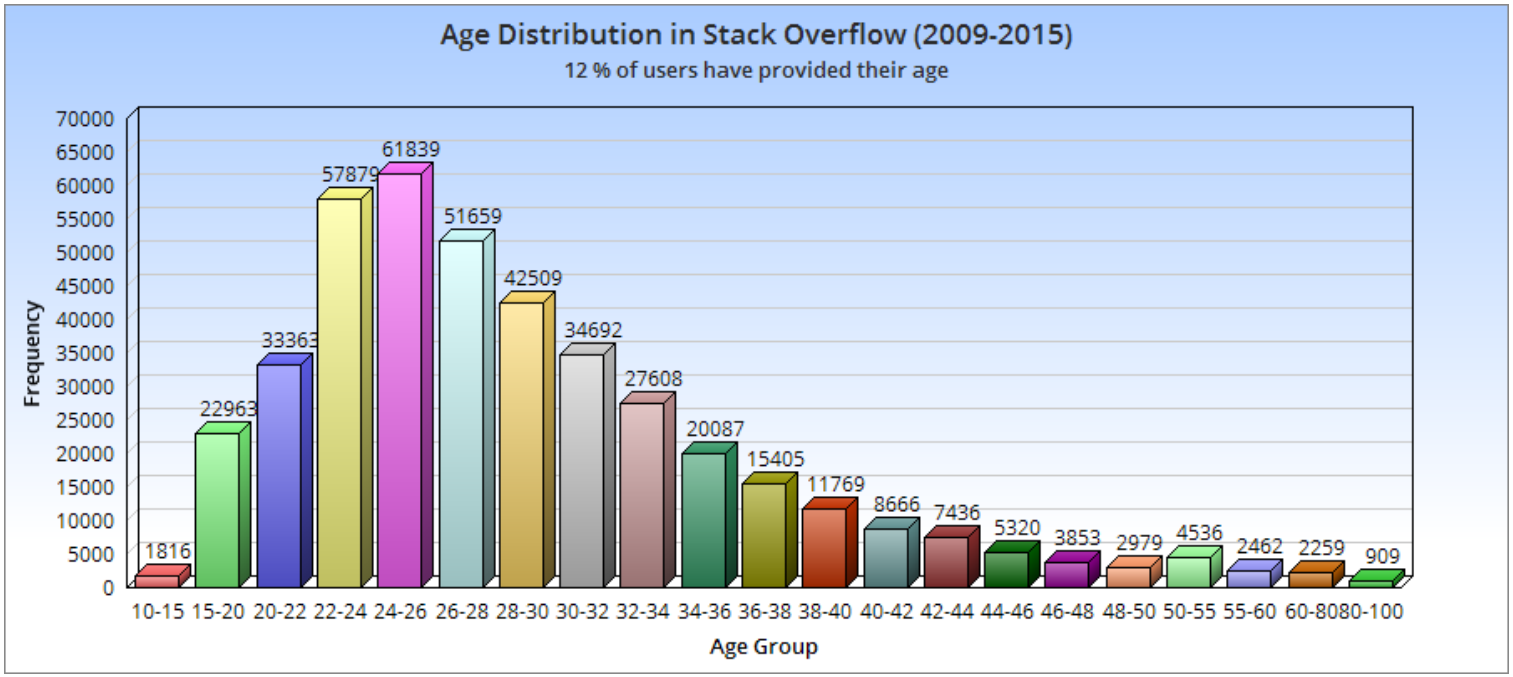
Background

- Stack Overflow - Top QA website
- Took the data from public dump since inception About 32 GB of Users and Posts data.
- Importance of User(Age,Reputation,Location)
- Importance of Topic (Topic Topic Graph - Core Periphery)
- Co-user graph communities from User Topic Graph.

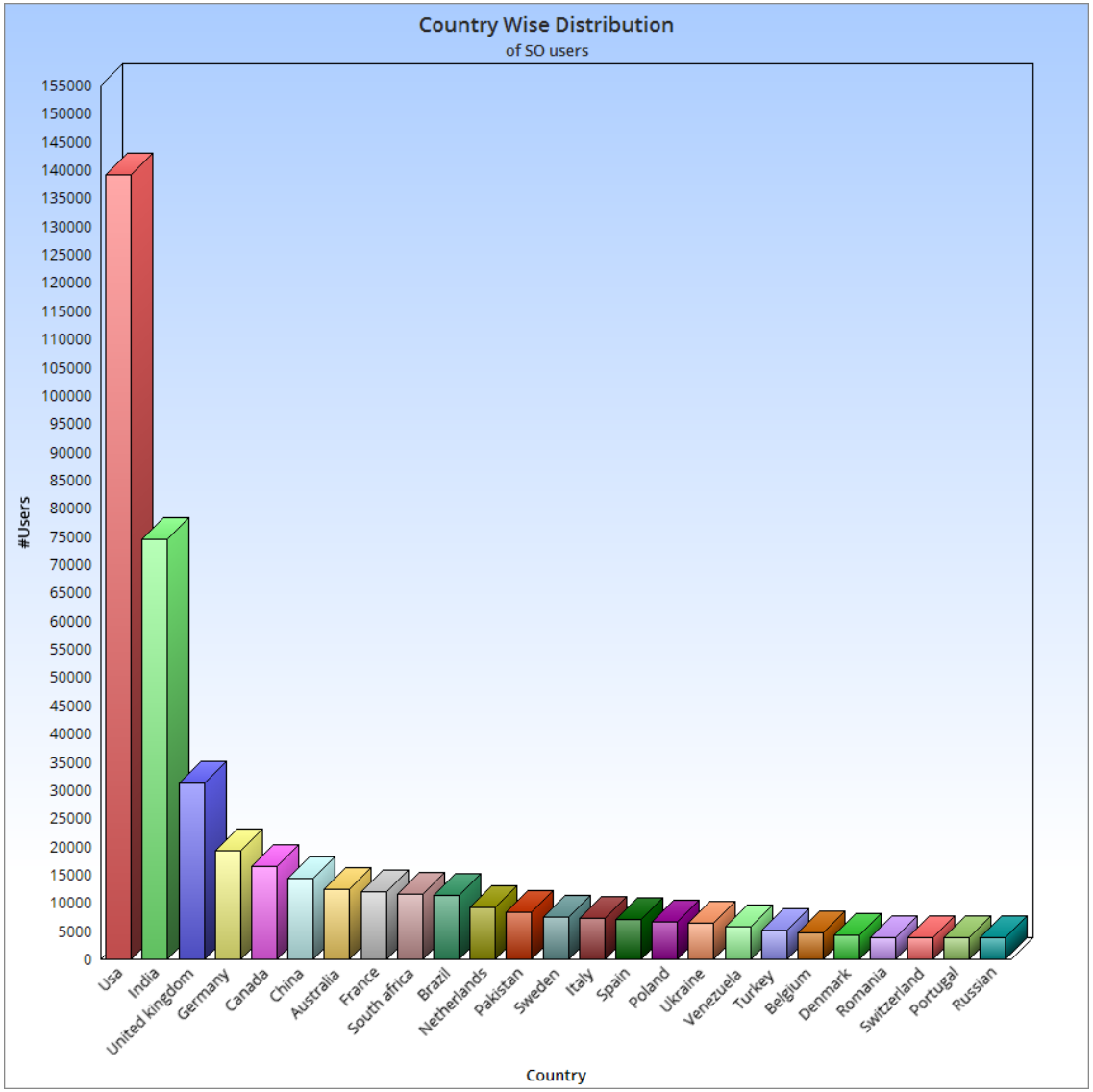
Importance of User

- 34,73,096 users in total
- 4,20,009 have given their age
- 5,60,871 have given their location
- Age Distribution,Location wise users,Location wise reputed and non reputed users.

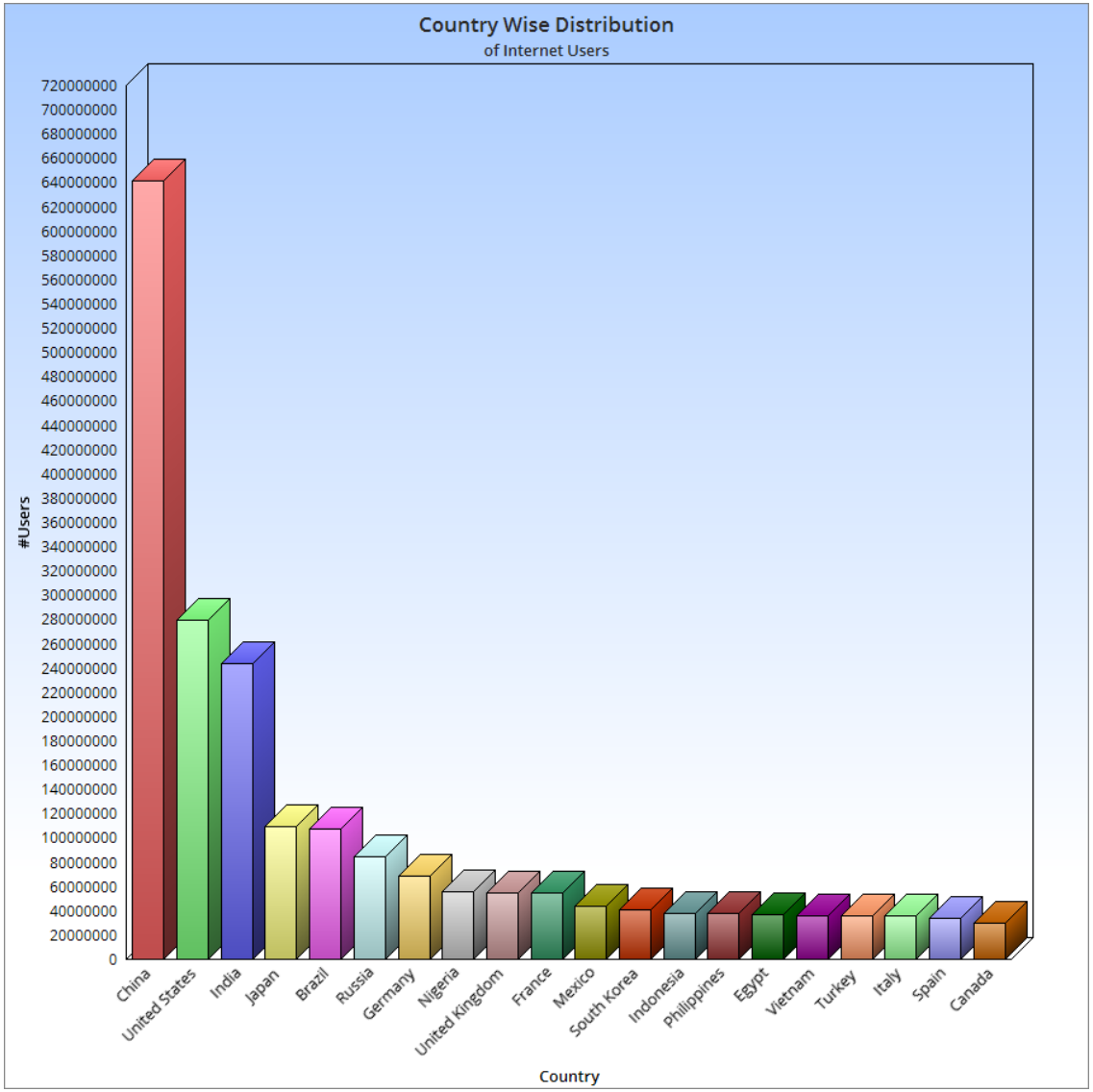
Importance of User - Age



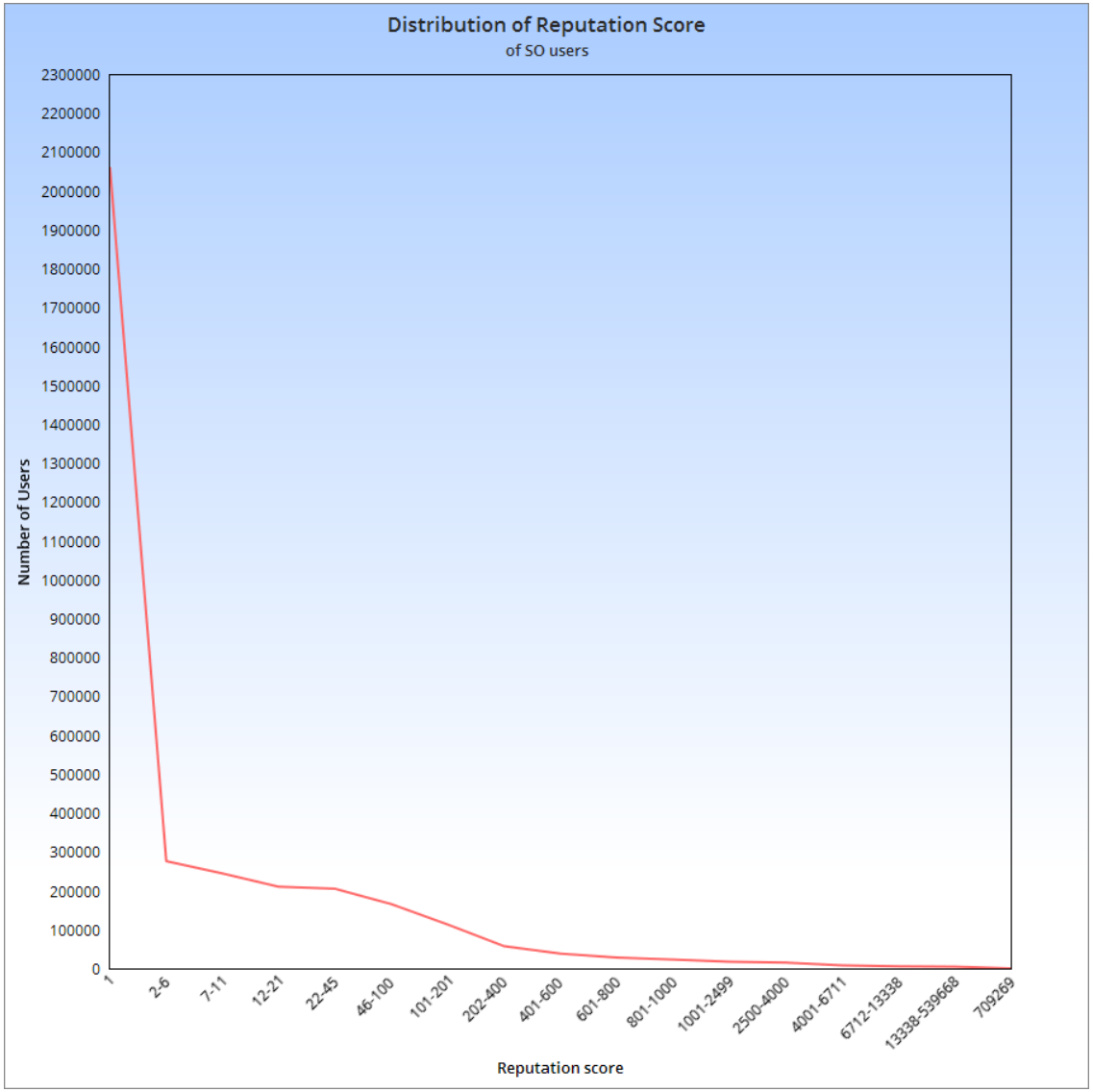
Importance of User- Location



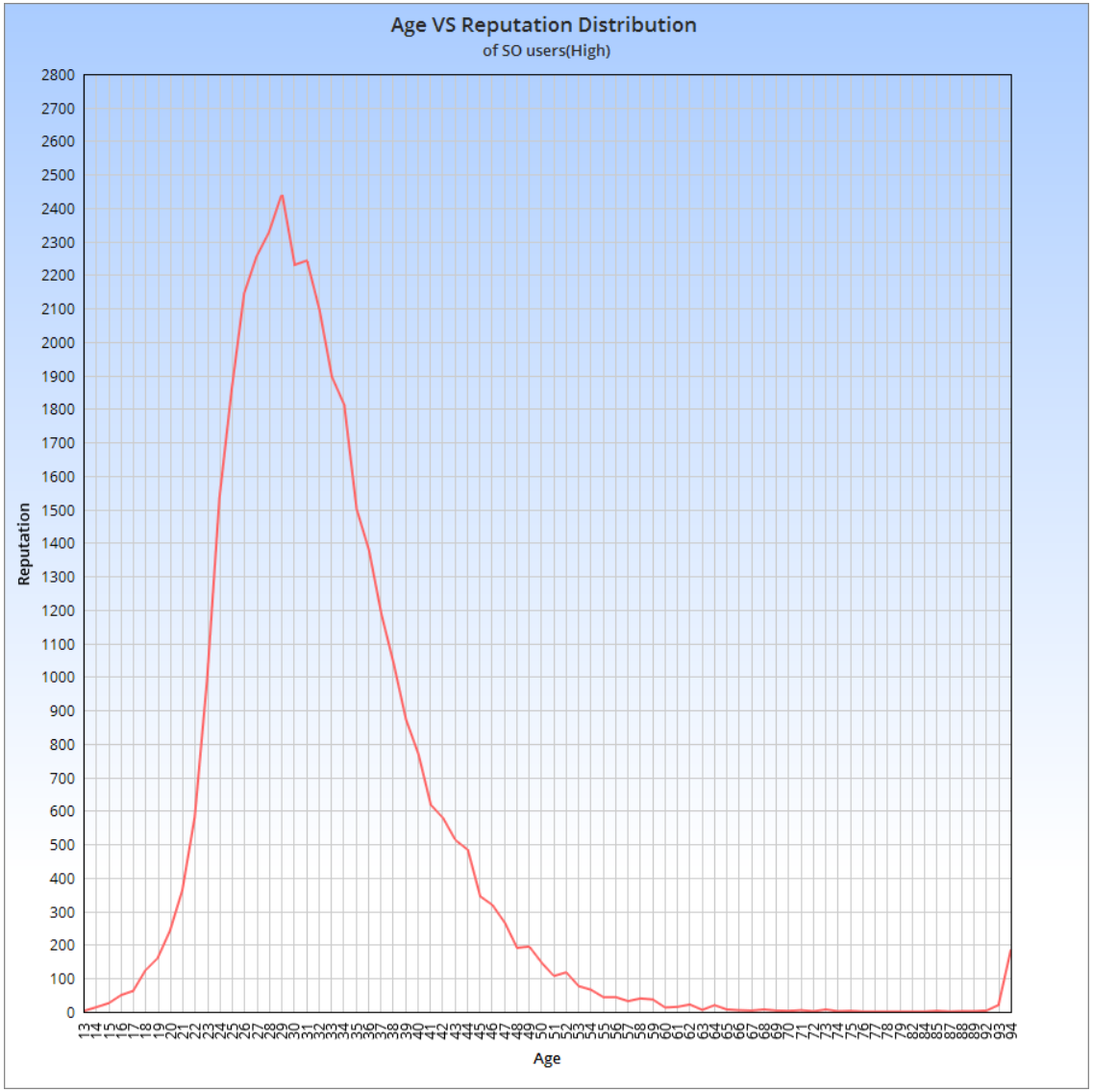
Importance of User - Location



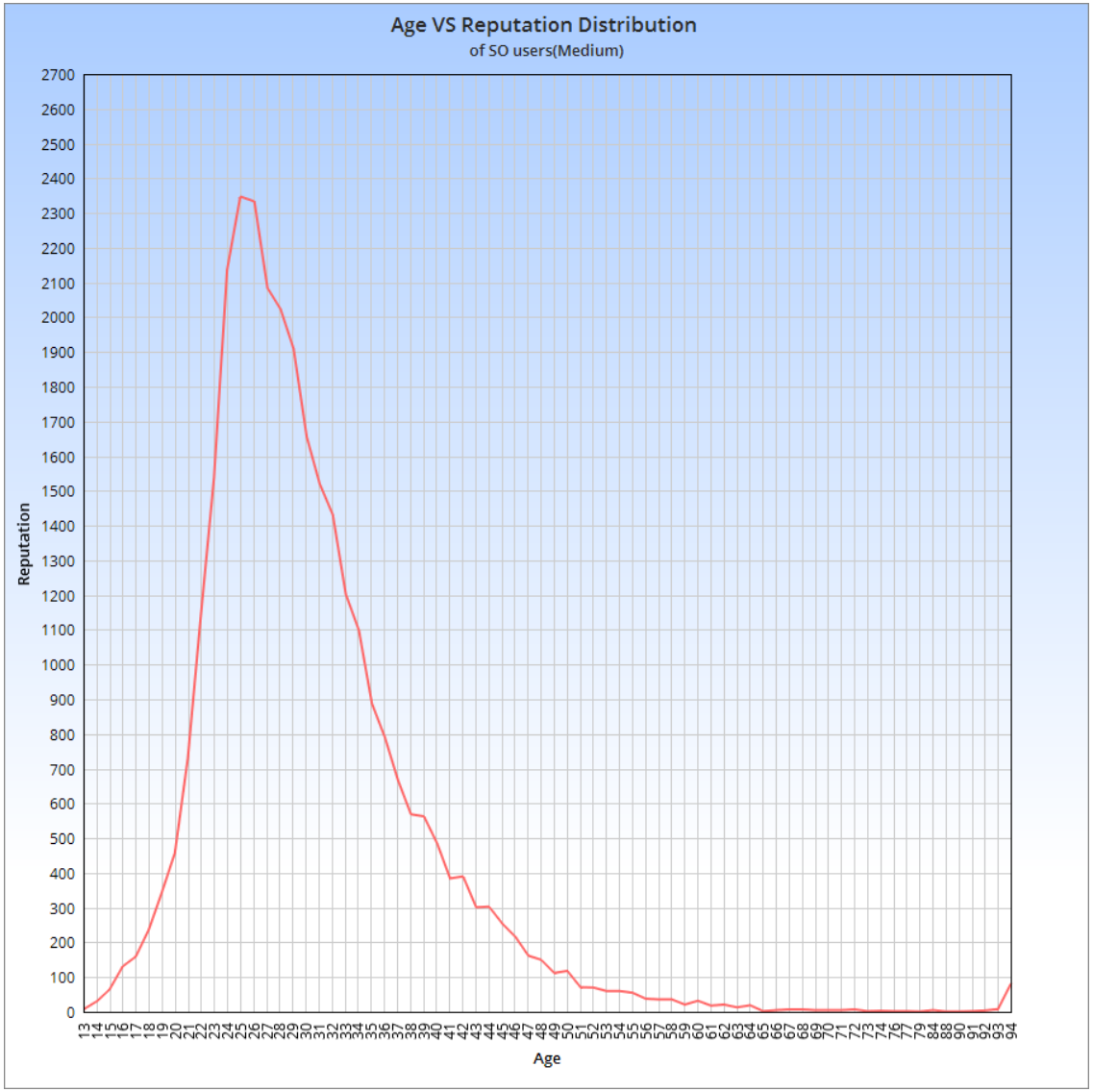
Importance of User - Reputation



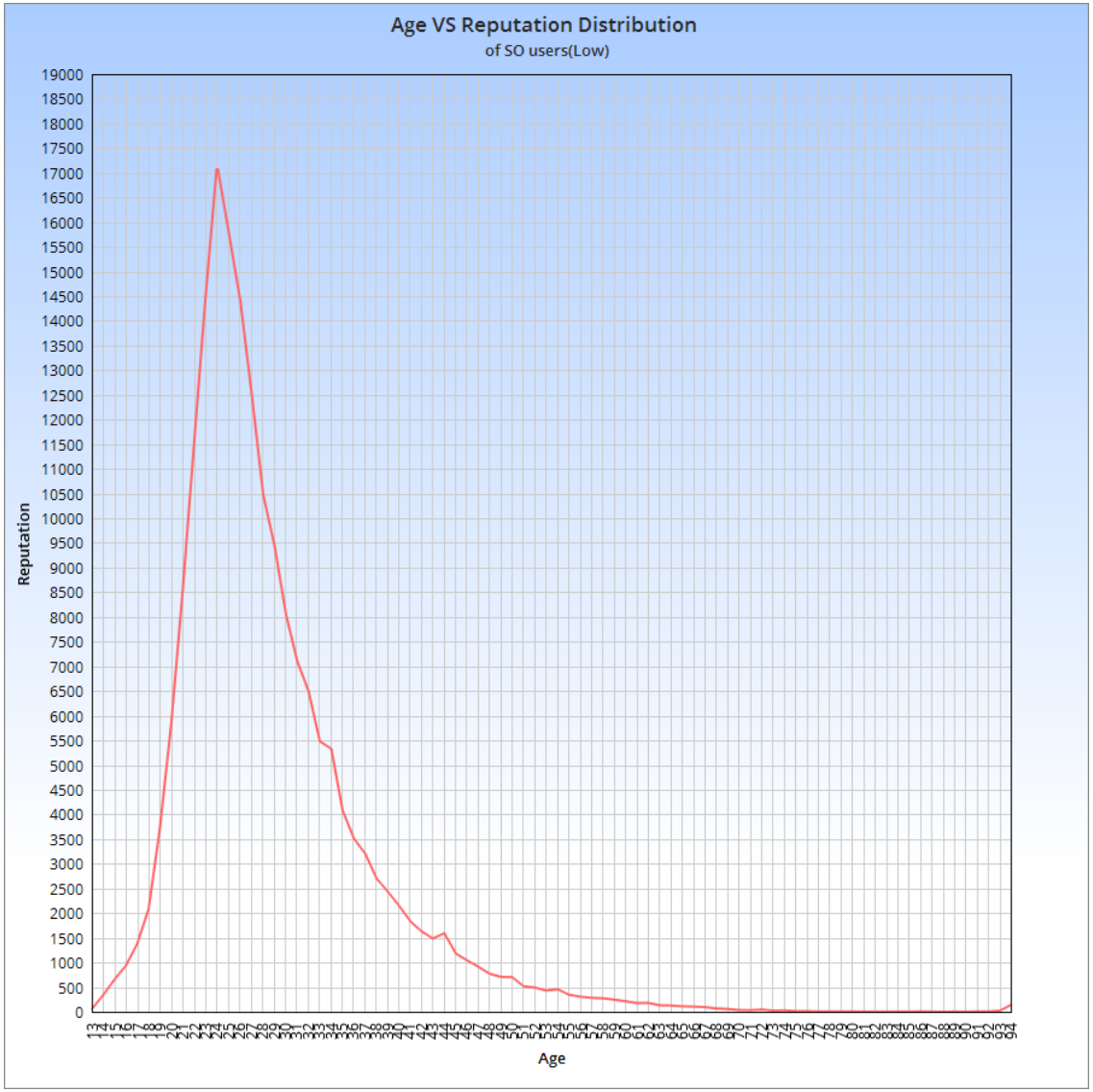
Importance of User - Age V/S Reputation



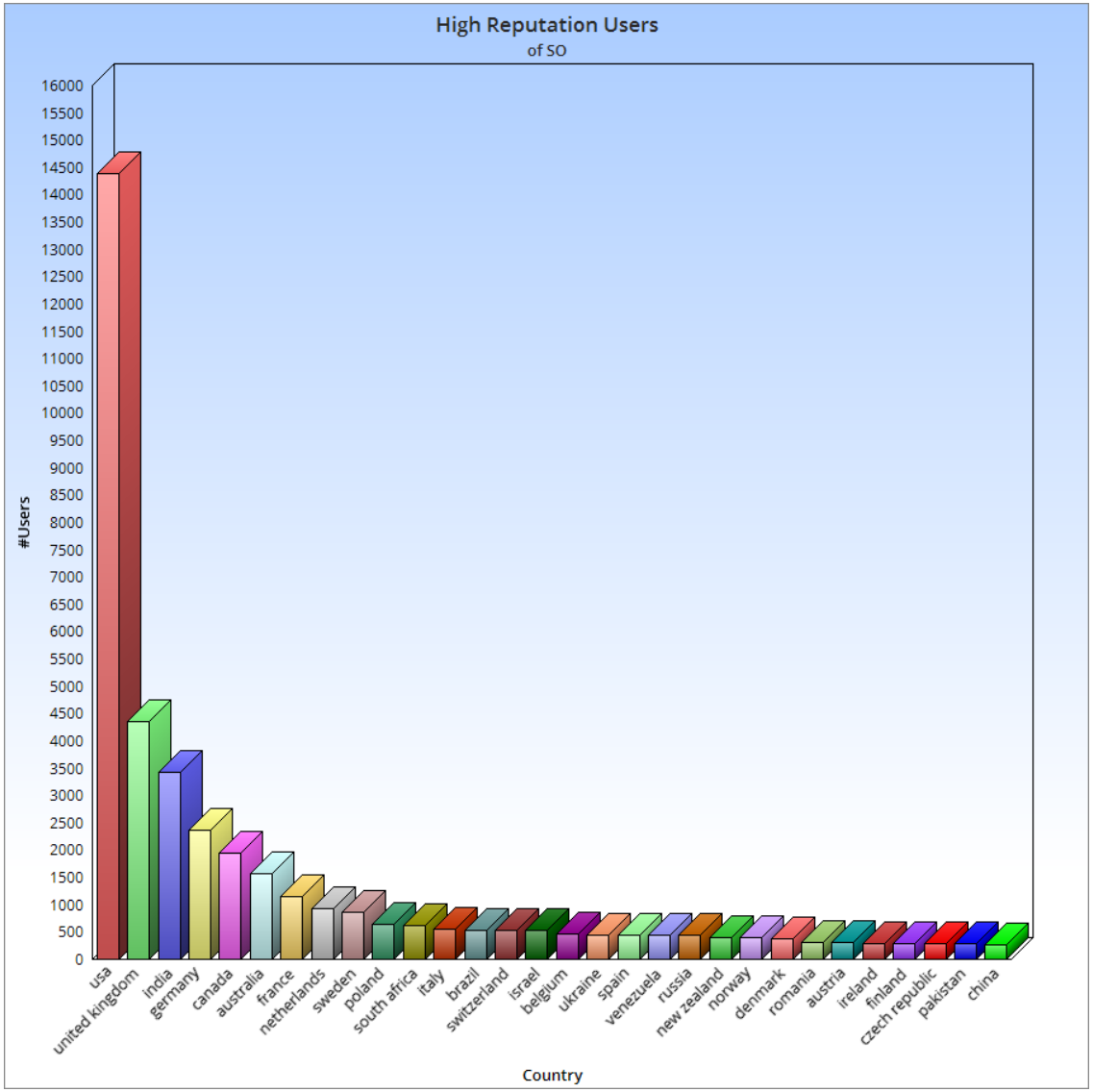
Importance of User - Age V/S Reputation



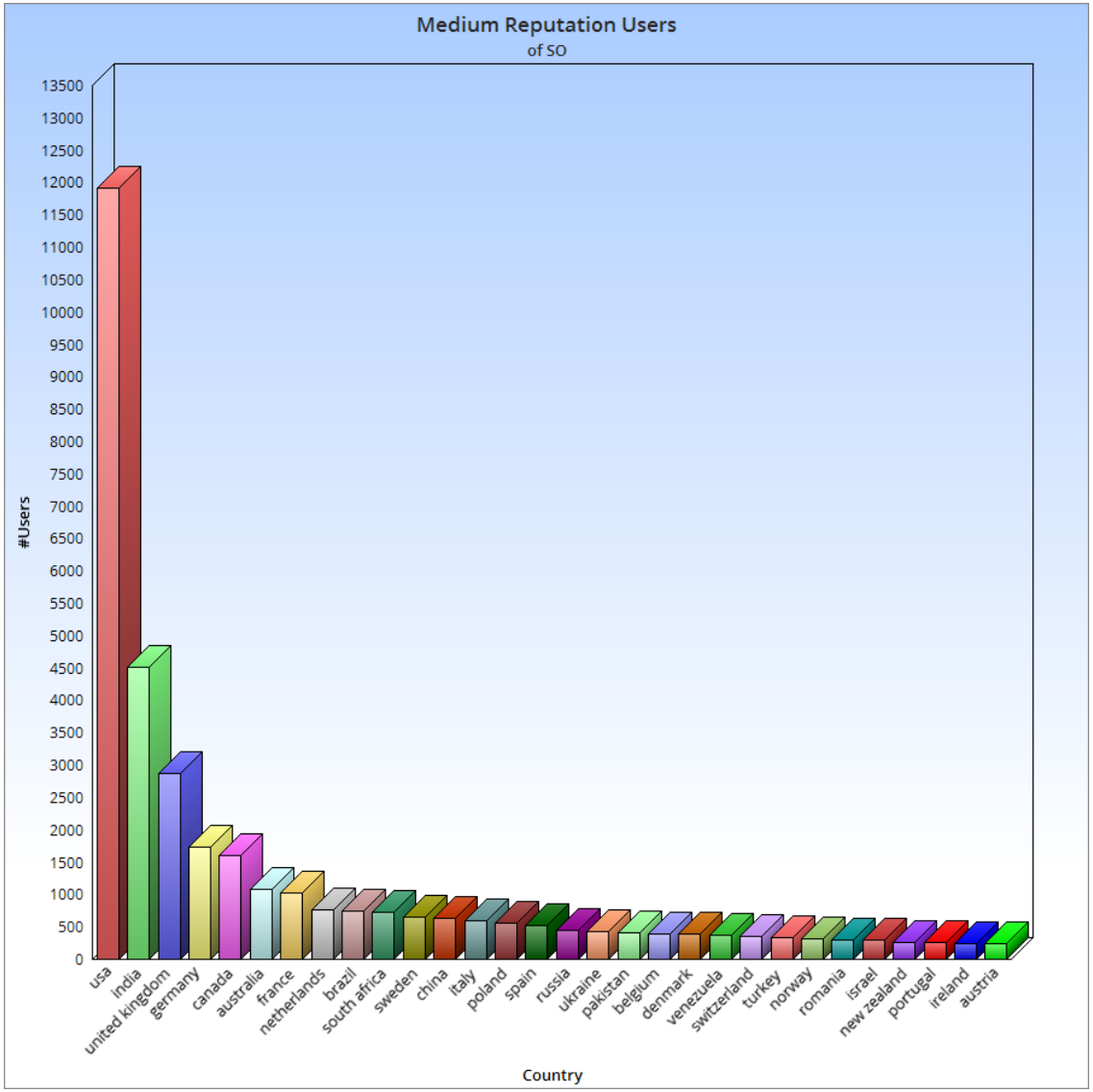
Importance of User - Age V/S Reputation



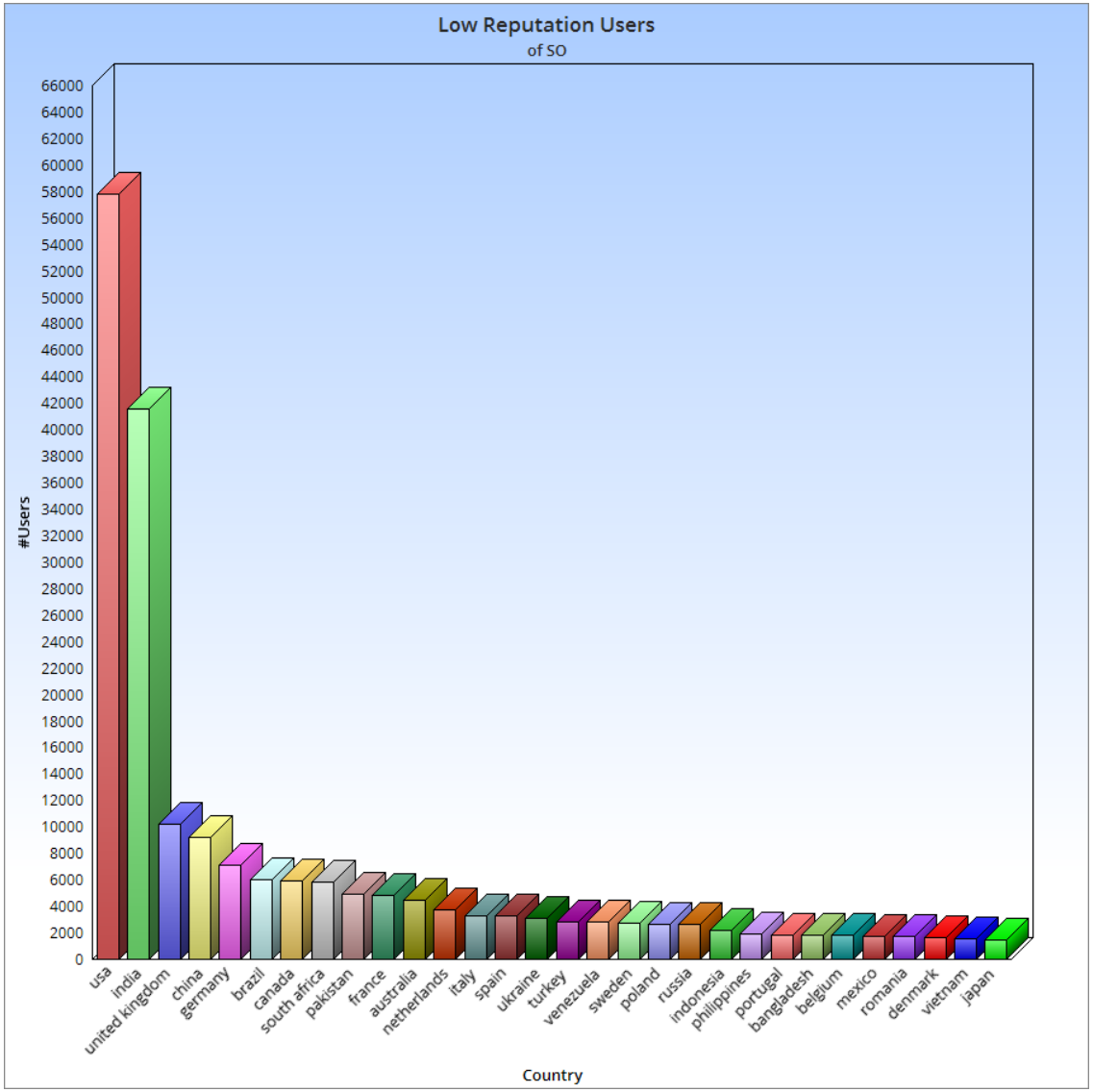
Importance of User - Location V/S Reputation



Importance of User - Location V/S Reputation



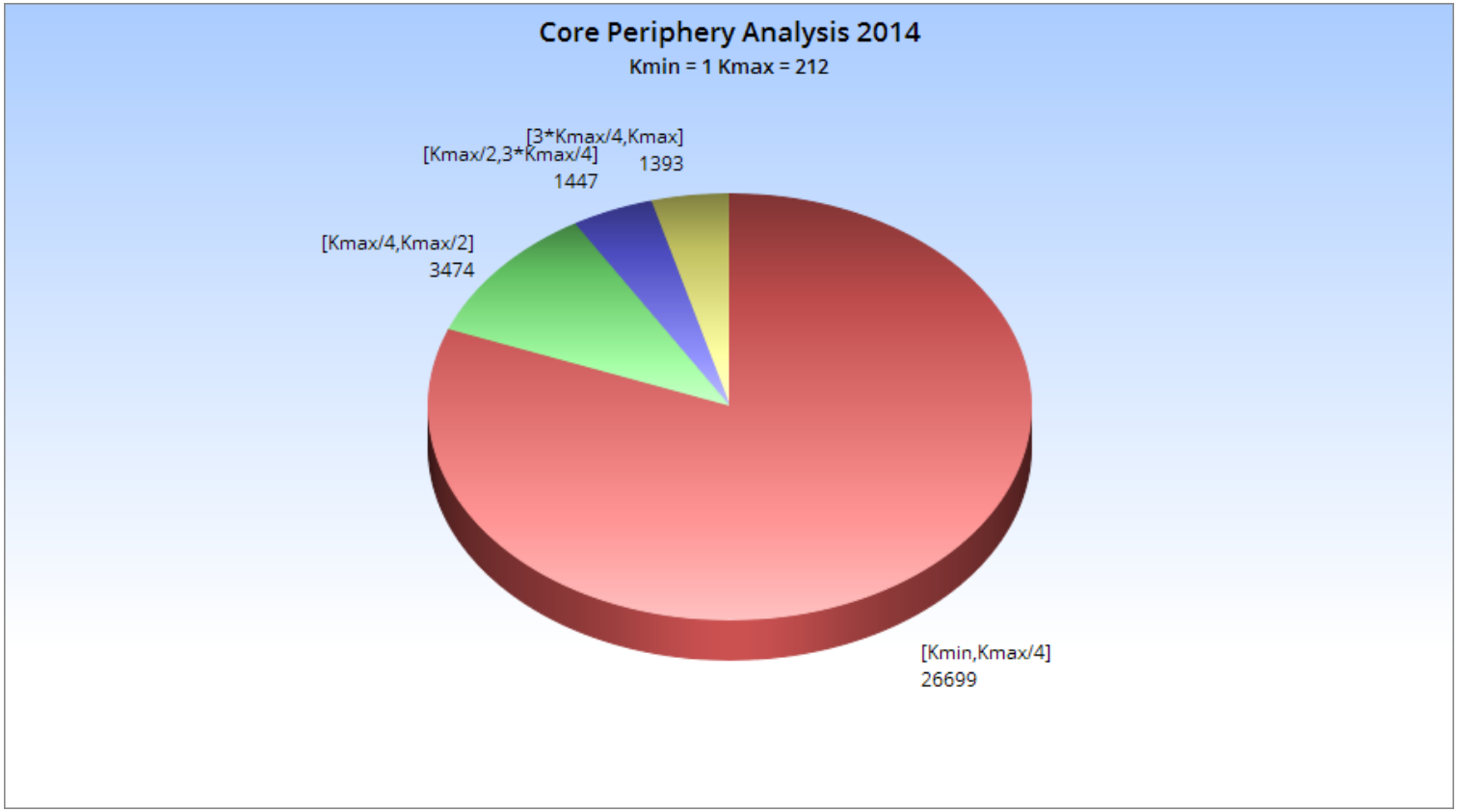
Importance of User - Location V/S Reputation



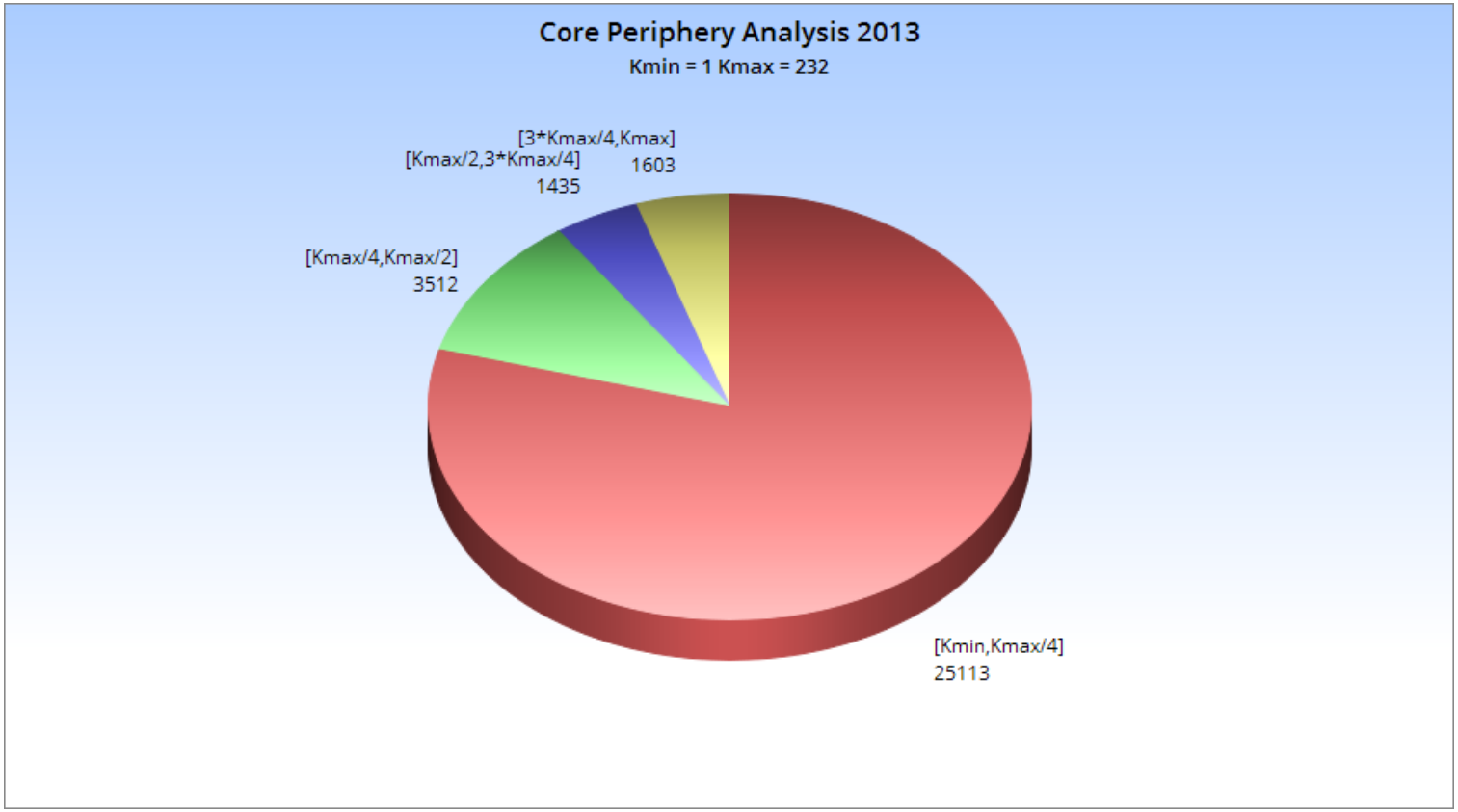
Importance of Topic

- Parsed all questions.
- If 2 tags appear in a question- add an edge
- Created topic-topic graph from 2008-2014 for each year
- Implemented Core Periphery analysis for all years.
- Divided into four equal slots from Kmin to Kmax.
- Example : android from Periphery to Core. erp - ambiguous. gps - steady.

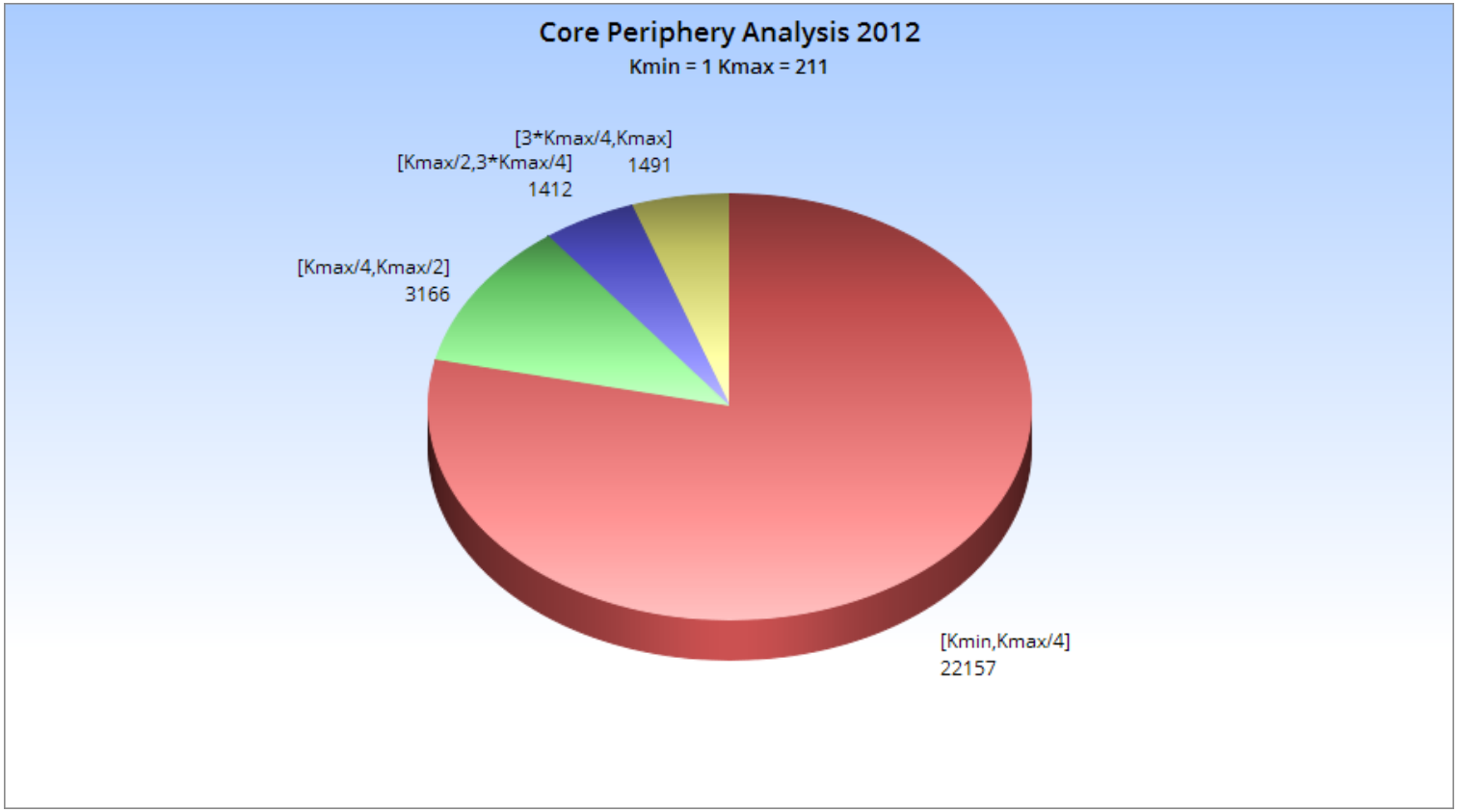
Importance of Topic- CP2014



Importance of Topic - CP2013



Importance of Topic - CP2012



User-Topic Graph

- Parsed all questions. Extracted userid and tags.
- Made a distribution of any random user asking x number of questions regarding any tag.
- Found out average to be 3.02.Thresholded an edge between user and topic for more than equals 3 vertices
- Did Bipartite projection and found out the co user graph
- Ran community detection (Louvain)

Co-User graph parameters

- Year - 2008 . Vertices - 2950/13247 . Edges - 633006 . Modularity - 0.390981 . Community Structure (10,25,2950)
- Year - 2009 . Vertices - 15754/58679 . Edges - 17343030 . Modularity - 0.418547 . Community Structure (9,21,15754)

Future Work

- Big data sampling for plotting communities.
- Making a webapp for trending topics over years
- Designing metric for implementing follower-followee network in Stack Overflow
- Analysing content quality of answers of Super users by R metric.

References I

Wisdom in the Social Crowd : an Analysis of Quora
Gang Wang , Konark Gill , Manish Mohanlal , Haitao Zheng , Ben Y Zhao
2013

The evolution of interdisciplinarity in physics research
Raj Kumar Pan ,Sitabhra Sinha ,Kimmo Kaski ,Jari Sarama
2012

Quantifying social group evolution
Gergely Palla ,Albert-Lszl Barabsi ,Tams Vicsek
2013