

Interesting Content and User Discovery without Upvotes

An application of centrality measures to Reddit

Rollen S. D'Souza
rollen.dsouza@uwaterloo.ca

University of Waterloo

2018

Outline

- 1 Introduction
 - About Reddit
 - Mathematical Preliminaries

- 2 Analysis
 - Technical Details

What is Reddit?

Purposes

- Social news aggregation

What is Reddit?

Purposes

- Social news aggregation
- Discussion platform

What is Reddit?

Purposes

- Social news aggregation
- Discussion platform
- Community for help or relaxation

What is Reddit?

Basic Structure

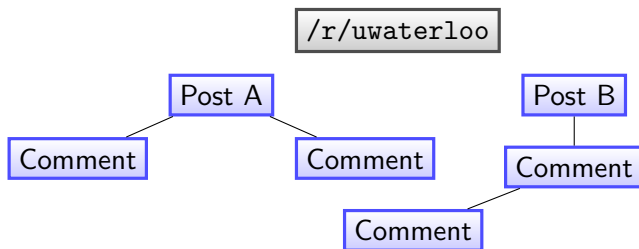


Figure: Graph-like structure of the Reddit platform.

What is Reddit?

The screenshot shows a Reddit post interface. At the top, there are up/down arrows, a score of 94, a link icon, and a tag 'Infinity'. The post title is 'Highschooler proves infinity is odd'. Below the title, it says 'r/badmathematics · Posted by u/[deleted] 1 year ago'. There is a link to 'reddit.com/r/AskR...' and a box with a broken image icon. Below the post, it says '85 Comments', 'Share', and '99% Upvoted'. A yellow banner states 'This thread is archived' and 'New comments cannot be posted and votes cannot be cast'. The sorting is set to 'BEST'. The first comment is by 'jacob8015' with 68 points, 1 year ago, and more than 35 children. The second comment is by 'IcedRoren' with 54 points, 1 year ago, titled 'Cursed by Dimensionality', and the text 'Idk what's worse. The argument, or the fact that the geometry teacher confirmed the proof.' Below this is a comment by 'Snowayne2' with 37 points, 1 year ago, titled 'The worst is how arrogant he is about the whole thing.', and the text 'Is the light bulb on yet?'. At the bottom right, there are icons for menu, play, list, search, and refresh.

↑ 94 ↓ | 🔗 **Infinity** Highschooler proves infinity is odd

↑ **r/badmathematics** · Posted by u/[deleted] 1 year ago 🗨️

94 **Infinity** Highschooler proves infinity is odd

↓ reddit.com/r/AskR... 🔗

💬 85 Comments Share ... 99% Upvoted

This thread is archived
New comments cannot be posted and votes cannot be cast

SORT BY BEST ▾

👤 jacob8015 68 points · 1 year ago (More than 35 children)

👤 IcedRoren **Cursed by Dimensionality** 54 points · 1 year ago

👇 Idk what's worse. The argument, or the fact that the geometry teacher confirmed the proof.

Share Save Edit ...

👤 Snowayne2 37 points · 1 year ago

👇 The worst is how arrogant he is about the whole thing.

Is the light bulb on yet?

Share Report Save Give gold

Mathematical Preliminaries

What I assume:

- Centrality Measures
- Properties of Laplacians

The former is how I sort through content and users.

Katz Centrality Measure

Definition

The *Katz Centrality Measure* for a node i in graph G with adjacency matrix A is defined as,

$$c_i = \sum_{k=1}^{\infty} \sum_{j=1}^n \alpha^k (A_{j,i})^k,$$

where $0 < \alpha \leq \rho(A)$.

Katz Centrality Measure

Why this measure?

- Has larger reach in the graph (looks at all paths to a node.)
- Easy to compute.
- Loose requirements on A .
- Intuitive meaning.

On Laplacians

Laplacians can be used for clustering. Recall,

$$L = D_{out} - A.$$

Use eigenvalues near 0 for clustering (eigenvectors associated with nearly disconnected components).

On Laplacians

Laplacians can be used for clustering. Recall,

$$L = D_{out} - A.$$

Use eigenvalues near 0 for clustering (eigenvectors associated with nearly disconnected components).

So...finding eigenvalues is easy right?

On Laplacians

Different Formulation

Problem: The graph can have as many as 10^4 nodes!

On Laplacians

Different Formulation

Problem: The graph can have as many as 10^4 nodes!

Definition

The *normalized Laplacian* of a graph is defined as,

$$L_{rw} = I_{n \times n} - D_{out}^{-1}A$$

On Laplacians

Different Formulation

Problem: The graph can have as many as 10^4 nodes!

Definition

The *normalized Laplacian* of a graph is defined as,

$$L_{rw} = I_{n \times n} - D_{out}^{-1}A$$

Need to ensure the graph has no sinks!

Technical Details

The Code

Code was implemented using Python 3 and MATLAB. Code is available on my Github. Other than that,

- redis for storing data in-memory.
- numpy, scipy for computation outside of MATLAB.

Technical Details

The Data

Sourced thanks to Jason Baumgartner who uploaded data dumps of the JSON objects gathered from the Reddit API. Structures contain,

- Submissions with author (id and flair), tags, score, text or link,
- Comments with author (id and flair), text
- and much much more!