# 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. Technical Notes
3. Case Study: /r/uwaterloo
4. General Observations
5. Summary

# 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

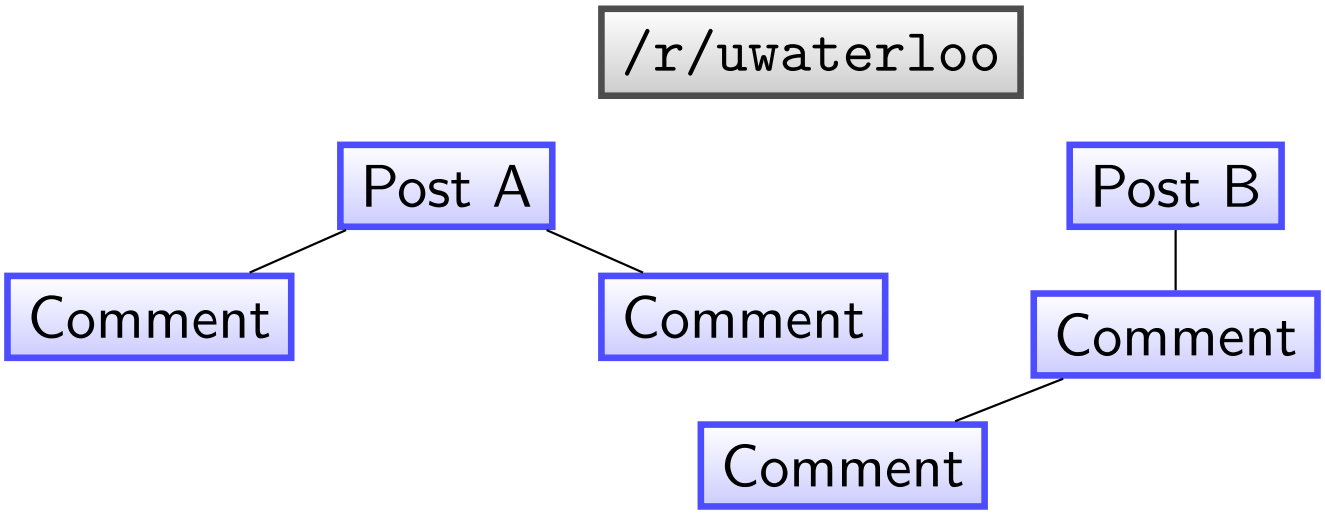
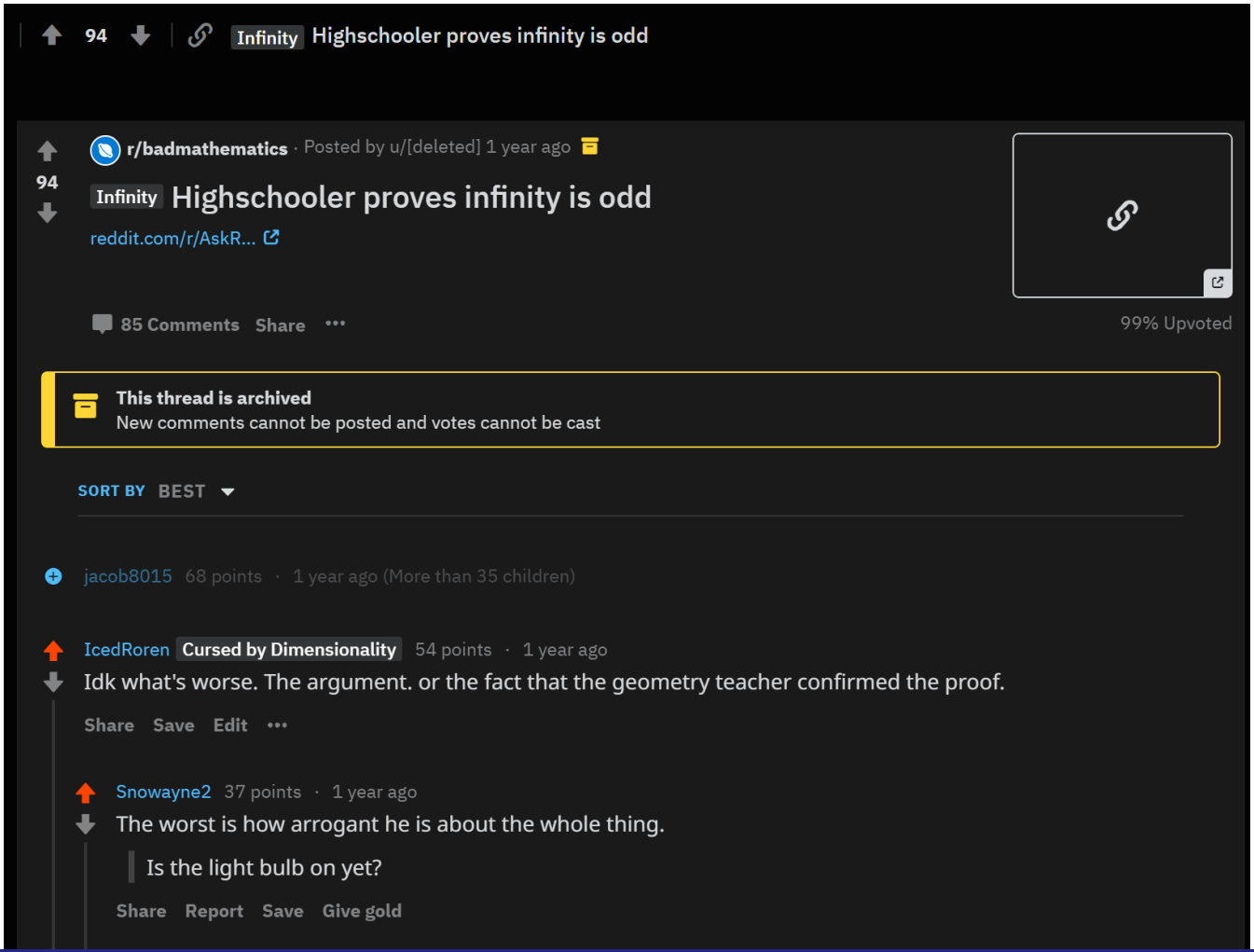



Figure: Graph-like structure of the Reddit platform.

# What is Reddit?



# 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,

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

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!*

# About 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`.

# About 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!

# About 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!

Data for: November 2017 — February 2018

# About the Assumptions

In order to do any analysis, we need to make some assumptions.

1. Users are interested (positively or negatively) by content they read.

# About the Assumptions

In order to do any analysis, we need to make some assumptions.

1. Users are interested (positively or negatively) by content they read.
2. Users comment only on content they have read.

# About the Assumptions

In order to do any analysis, we need to make some assumptions.

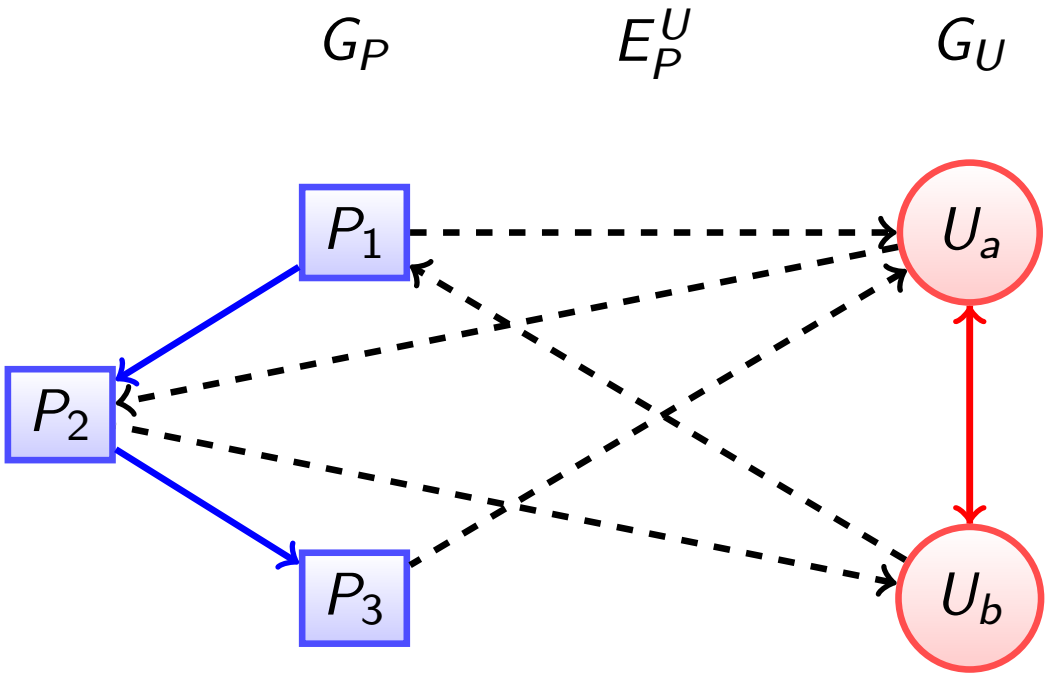
1. Users are interested (positively or negatively) by content they read.
2. Users comment only on content they have read.
3. Deleted content and users are not interesting nor generate interesting content and do not affect other users in a meaningful way.

# About the Assumptions

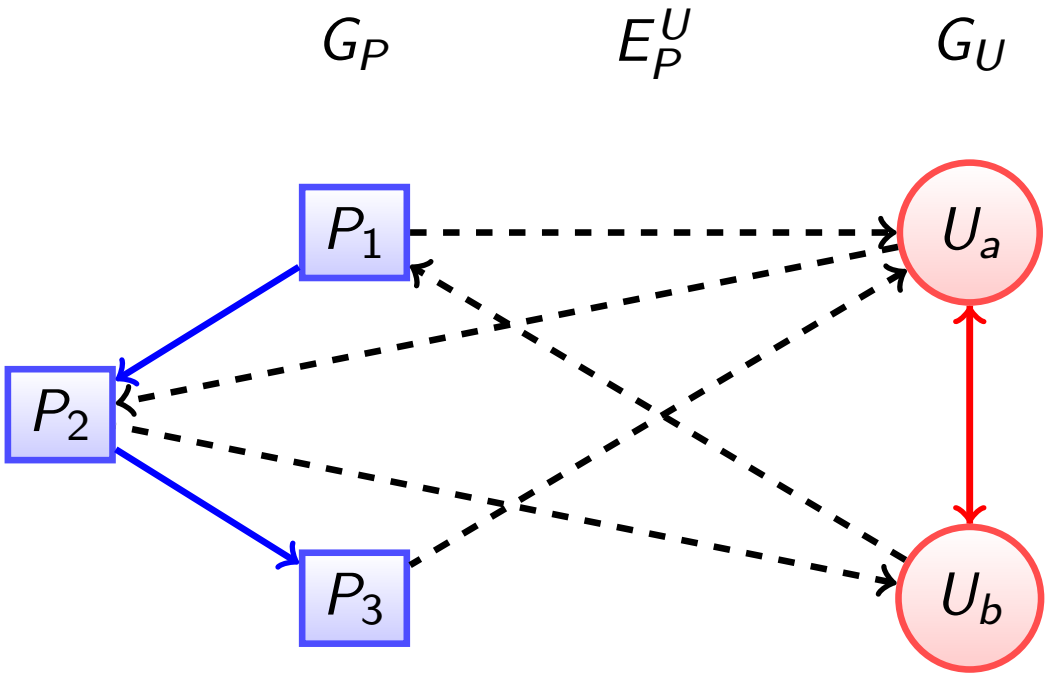
In order to do any analysis, we need to make some assumptions.

1. Users are interested (positively or negatively) by content they read.
2. Users comment only on content they have read.
3. Deleted content and users are not interesting nor generate interesting content and do not affect other users in a meaningful way.
4. Active engagement of users (through comments or submissions) is a good measure of interest.

# Structuring the Graph



# Structuring the Graph




Note that there are atleast 3 interesting graphs here. Where are they?

Applied centrality analysis to reveal the following top nodes,

| /r/uwaterloo | /r/science | /r/CanadianInvestor |
|---|---|---|
| supersonic63 | t3_7e1jo1 | johnnychi |
| kw2002 | AutoModerator | EyesOnTsx |
| **t3_7f2c1d** | mvea | jnf_goonie |
| **t3_7w0dgv** | t3_7vxito | langlois44 |
| **t3_7i045q** | t3_7s6a9z | t3_7ii8af |
| microwavemasterrace | t3_7my58a | t3_7fapx7 |
| uwsmile | t3_7okc7u | t3_7pawu6 |
| **t3_7ld2jb** | t3_7hs6ro | helkish |
| mywaterlooaccount | t3_7sv8vb | Ginhisf |
| honhonhonFRFR | t3_7qs5xz | MarketStorm |

# Of Interest to /r/uwaterloo

The more generic submissions...

| Link | Description | Score | Comment # |
|---|---|---|---|
| f2c1d | Admissions Megathread | 311 | 12k |
| 7i045q | Mr. Goose for good exams... | 142 | 389 |
| 7ld2jb | Mr. Goose for good grades... | 196 | 371 |

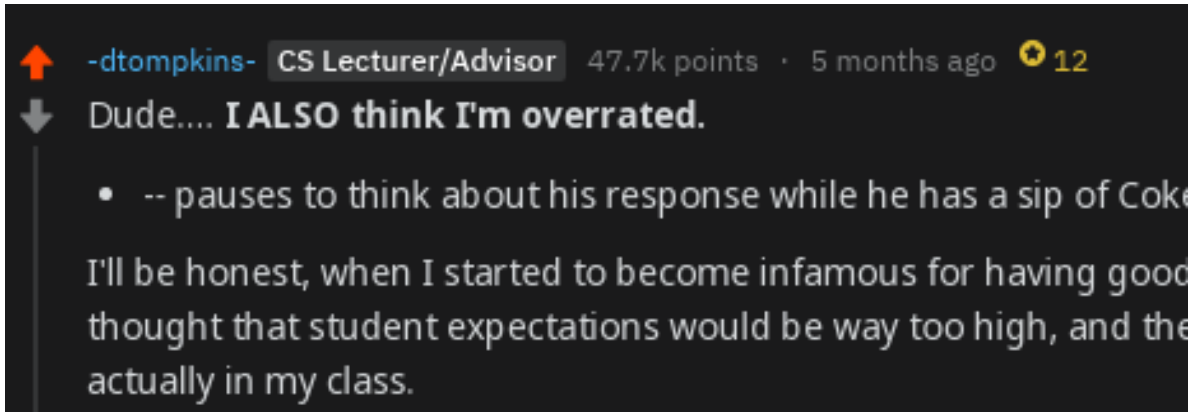
A mass thread on WaterlooWorks also shows up in the top 20.

# Of Interest to /r/uwaterloo

t3_7w0dgv: A much more interesting post...

# Of Interest to /r/uwaterloo

t3_7w0dgv: A much more interesting post...



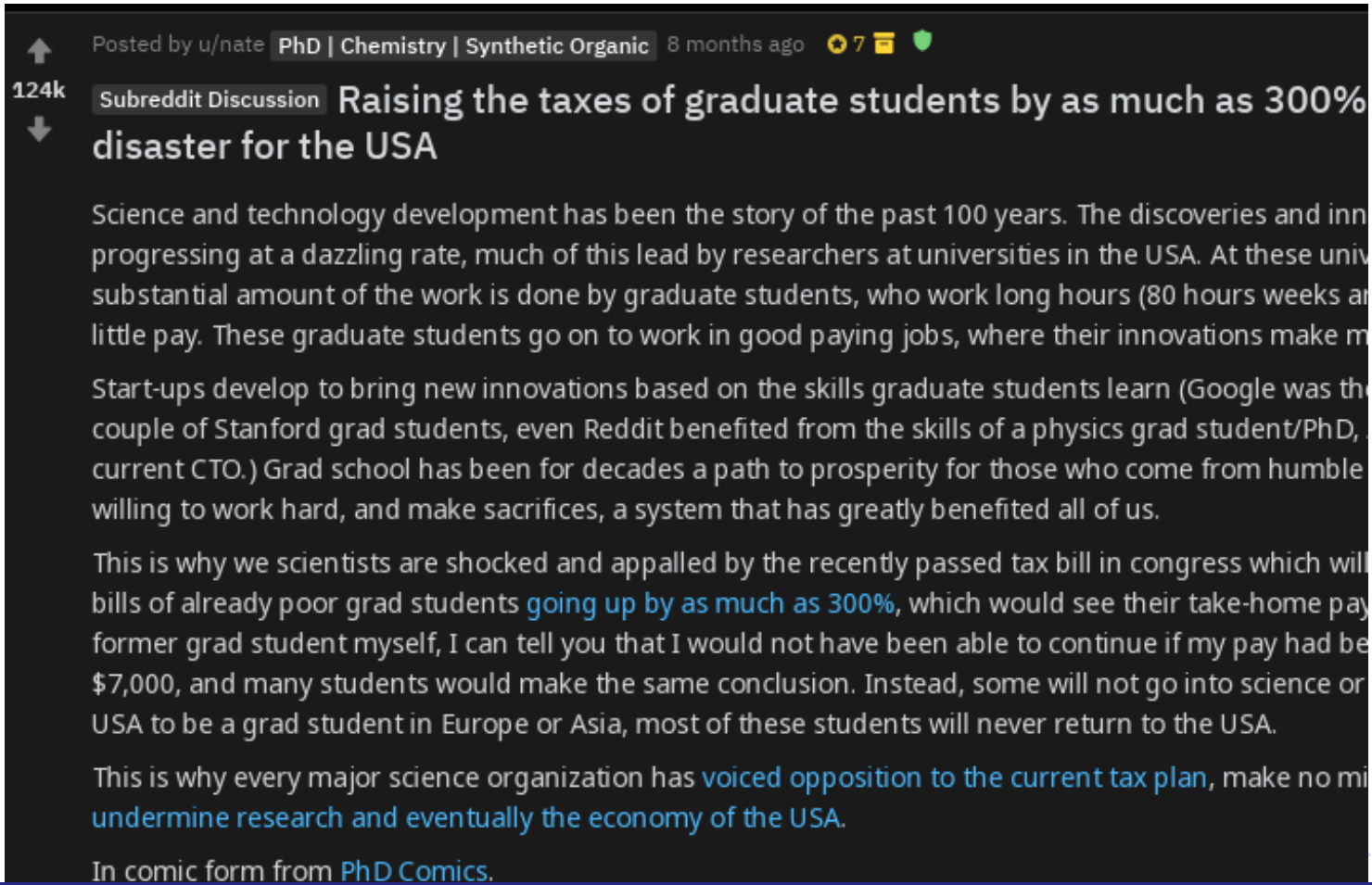
# General Observations

## On /r/science

| /r/uwaterloo | /r/science | /r/CanadianInvestor |
|---|---|---|
| supersonic63 | **t3_7e1jo1** | johnnychi |
| kw2002 | AutoModerator | EyesOnTsx |
| t3_7f2c1d | mvea | jnf_goonie |
| t3_7w0dgv | t3_7vxito | langlois44 |
| t3_7i045q | t3_7s6a9z | t3_7ii8af |
| microwavemasterrace | t3_7my58a | t3_7fapx7 |
| uwsmile | t3_7okc7u | t3_7pawu6 |
| t3_7ld2jb | t3_7hs6ro | helkish |
| mywaterlooaccount | t3_7sv8vb | Ginhisf |
| honhonhonFRFR | t3_7qs5xz | MarketStorm |

# General Observations

## On /r/science



Posted by u/nate PhD | Chemistry | Synthetic Organic 8 months ago 7

124k

Subreddit Discussion **Raising the taxes of graduate students by as much as 300% disaster for the USA**

Science and technology development has been the story of the past 100 years. The discoveries and inn progressing at a dazzling rate, much of this lead by researchers at universities in the USA. At these univ substantial amount of the work is done by graduate students, who work long hours (80 hours weeks ar little pay. These graduate students go on to work in good paying jobs, where their innovations make m

Start-ups develop to bring new innovations based on the skills graduate students learn (Google was th couple of Stanford grad students, even Reddit benefited from the skills of a physics grad student/PhD, current CTO.) Grad school has been for decades a path to prosperity for those who come from humble willing to work hard, and make sacrifices, a system that has greatly benefited all of us.

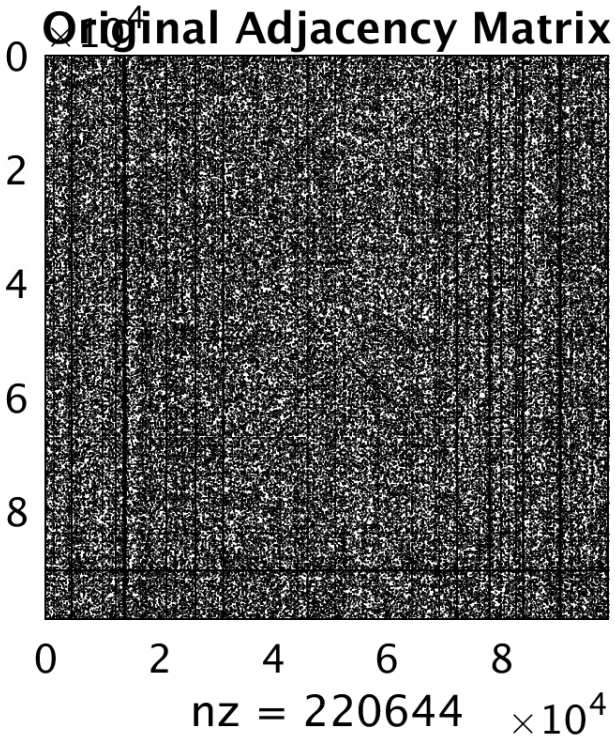
This is why we scientists are shocked and appalled by the recently passed tax bill in congress which will bills of already poor grad students going up by as much as 300%, which would see their take-home pay former grad student myself, I can tell you that I would not have been able to continue if my pay had be $7,000, and many students would make the same conclusion. Instead, some will not go into science or USA to be a grad student in Europe or Asia, most of these students will never return to the USA.

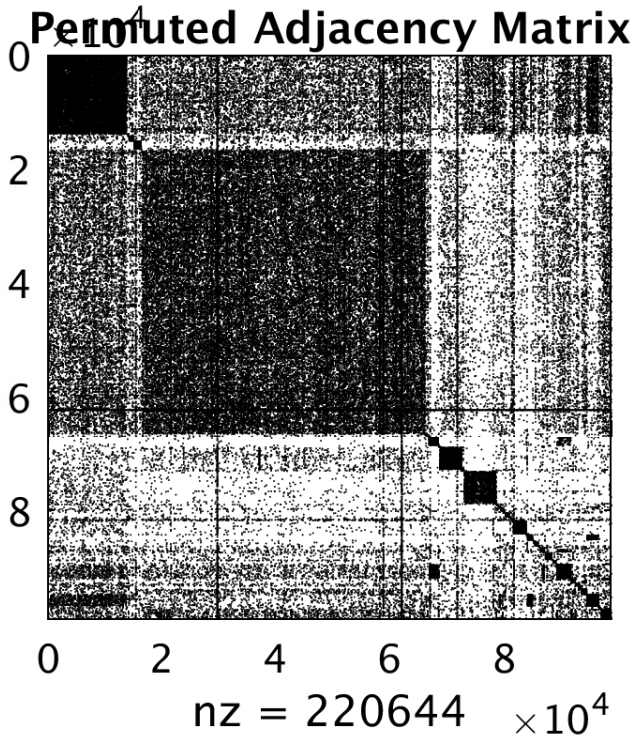
This is why every major science organization has voiced opposition to the current tax plan, make no mi undermine research and eventually the economy of the USA.

In comic form from PhD Comics.

# General Observations

## On /r/science



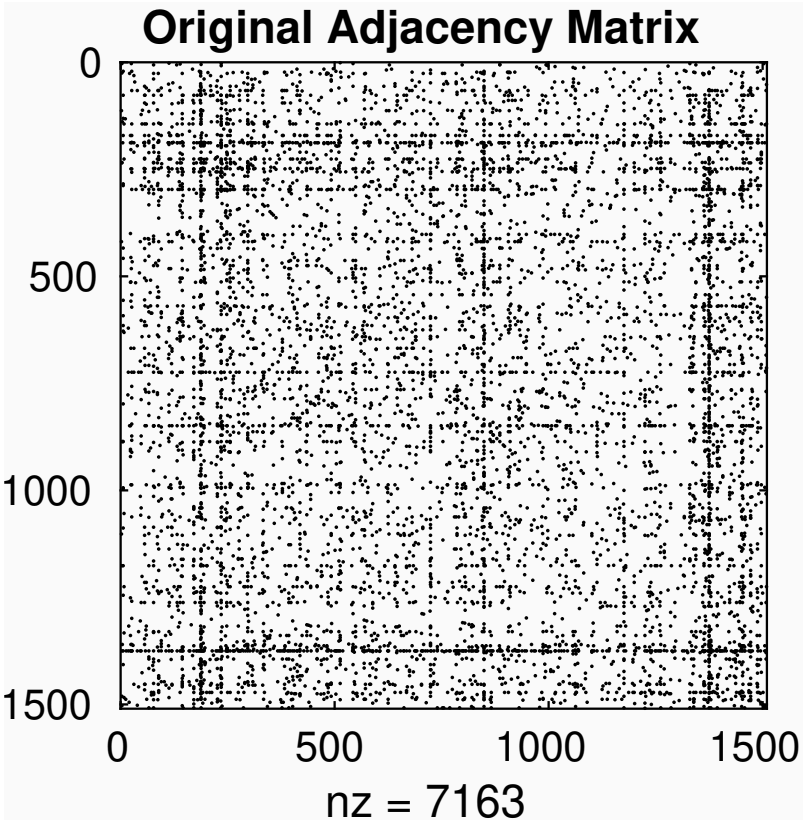
# General Observations

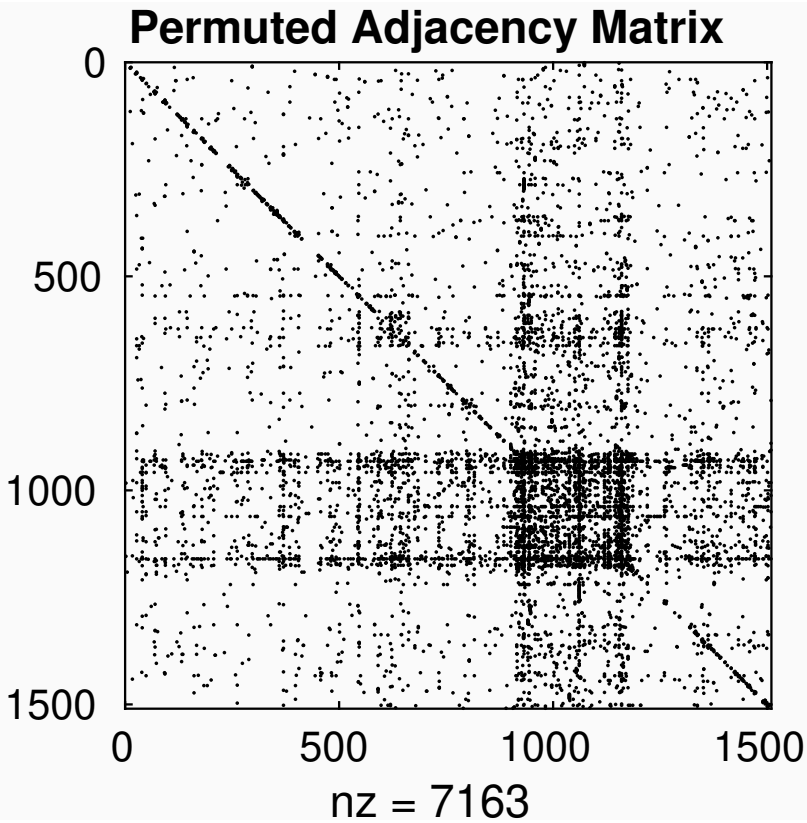
## On /r/CanadianInvestor

| /r/uwaterloo | /r/science | /r/CanadianInvestor |
|---|---|---|
| supersonic63 | t3_7e1jo1 | johnnychi |
| kw2002 | AutoModerator | EyesOnTsx |
| t3_7f2c1d | mvea | jnf_goonie |
| t3_7w0dgv | t3_7vxito | langlois44 |
| t3_7i045q | t3_7s6a9z | t3_7ii8af |
| microwavemasterrace | t3_7my58a | t3_7fapx7 |
| uwsmile | t3_7okc7u | t3_7pawu6 |
| t3_7ld2jb | t3_7hs6ro | helkish |
| mywaterlooaccount | t3_7sv8vb | Ginhisf |
| honhonhonFRFR | t3_7qs5xz | MarketStorm |

# General Observations

## On /r/CanadianInvestor



# Summary

## What did we learn?

- Infinity is odd.

# Summary

## What did we learn?

- Infinity is odd.
- Users totally read what they react to.

# Summary

## What did we learn?

- Infinity is odd.
- Users totally read what they react to.
- Dave Tompkins is awesome.

# Summary

## What did we learn?

- Infinity is odd.
- Users totally read what they react to.
- Dave Tompkins is awesome.
- Obligatory reference to Trump.

# Summary

## What did we learn?

- Infinity is odd.
- Users totally read what they react to.
- Dave Tompkins is awesome.
- Obligatory reference to Trump.
- Importance of investing expertise.