# MOOC Data Week 1 Kicking Tires

AI Planning Forums / General

MOOC Data Week 1 Kicking Tires 1

Overarching Questions: 2

Basic Stats: 2

Network Analysis Techniques on forums 2

Content Analysis 3

IP Data 5

Content Based Analysis 5

At this stage the techniques are quite ‘two dimensional’, with an aim to start thinking about the questions we may ask of the data and building reusable scripts/techniques/code examples. This document has split the techniques around three questions:

1. SNA/Graph based techniques

Coming up with a process to create network data that follows the workflow of Query->Adjacency Matrix->Graph. Where can we find ties in the data set?

1. Location based

Based upon IP data. Used my own dataset.

1. Content Based

Can we analyse the text of posts to make observations about the type of post and actions of the poster

At this stage only AI planning data is used.

## Overarching Questions:

Q) Difference between forum\_posts and forum\_comments?

From Figure 1.1. Why so many node on their own? Comments seem like they are a different type of reply to a post.

Q) Do the Anon\_user\_id and forum\_user\_ids match up?

A) Hash table natch.

## Basic Stats:

Unique users: 29894 from Edinburgh estimation, 26775 in users users table?

26,775

Unique users in General database:

Unique users in forums database: 918 unique users in AI planning forum. This includes subscribers and ratings(?) 671 who posted?

Not sure if this query is dodgy:

select count(distinct forum\_user\_id) from (

select forum\_user\_id from forum\_comments

UNION

select forum\_user\_id from forum\_posts

UNION

select forum\_user\_id from forum\_threads

) as users;

Unique threads: 436 (of which deleted?)

Unique posts 1846 (of which deleted?)

Unique comments: 697 (of which deleted?)

## Network Analysis Techniques on forums

fig 1.1: Forum\_posts tables: All Forum users (nodes) conversing through comments on a post (edges) including deleted.

SQL: select post\_id, `forum\_user\_id` from `forum\_comments` ;

Adjacency matrix generated through R iGraph package.

## Content Analysis

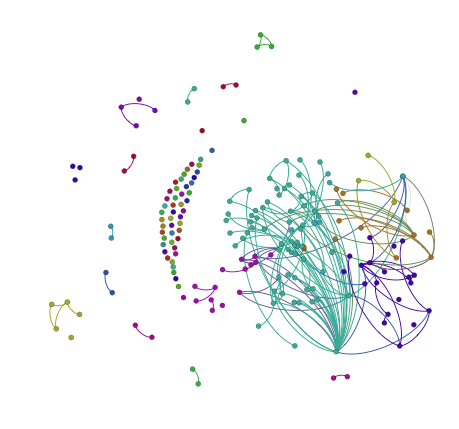


fig 1.1: All Forum users (nodes) conversing through comments on a post (edges) including deleted from forum\_comments table

Fig 1.1 Remarks

Fig 1.1 Questions:

Who are the people who aren’t linked? Has nobody answered their post? How does this effect their experience?

SQL: select post\_id, count(\*) as cnt, `comment\_text` from `forum\_comments` group by `post\_id` order by cnt ASC;

Seem to suggest that single posts are replies. Why don’t we have the full thread? Perhaps I misunderstand posts vs comments vs threads

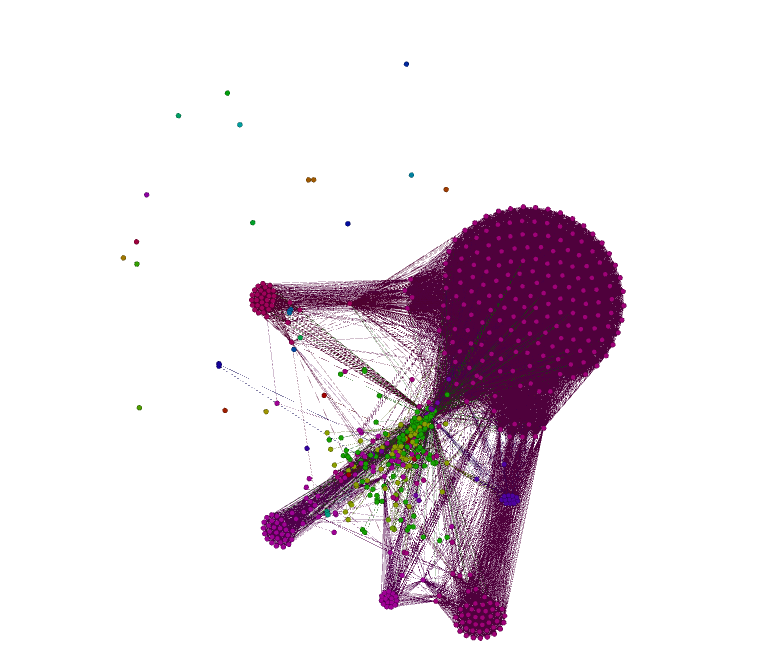
Who are the people who have only talked to one or two people?

How does their interaction with the

Fig 1.2:

SQL: select thread\_id, `forum\_user\_id` from `forum\_posts` ;

Adjacency matrix generated through R iGraph package.



Comments:

Tutors and Pupils as same nodes.. (check access groups)

Fig 1.2:

General Table.

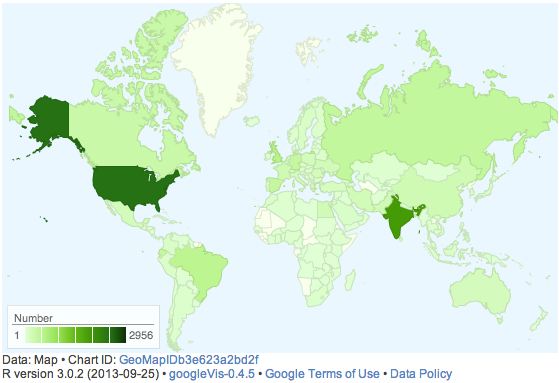
## IP Data

Comments:

I used my own IP lookup before realising A-M has already got a lookup table. Fortunatly we seem to have used the same data source but this needs double checking.

Only \*\*\* have data for last used IP address, does this mean they signed up, but never signed in?

Interactive File location: Aiplan\_Forum/GeoMap.html

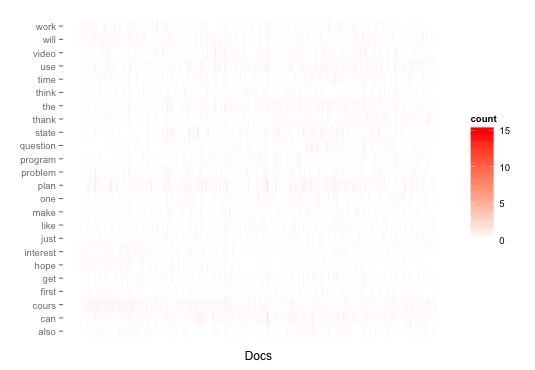


Comment

## Content Based Analysis

File location: Aiplan\_Forum/heat\_of\_freq\_terms.png

Script location: R\_workspace/heatmap.R



Topics Modelling

Comments: Too many documents doesn’t lend itself to heatmaps, perhaps comparable heatmaps of students X and Y.