

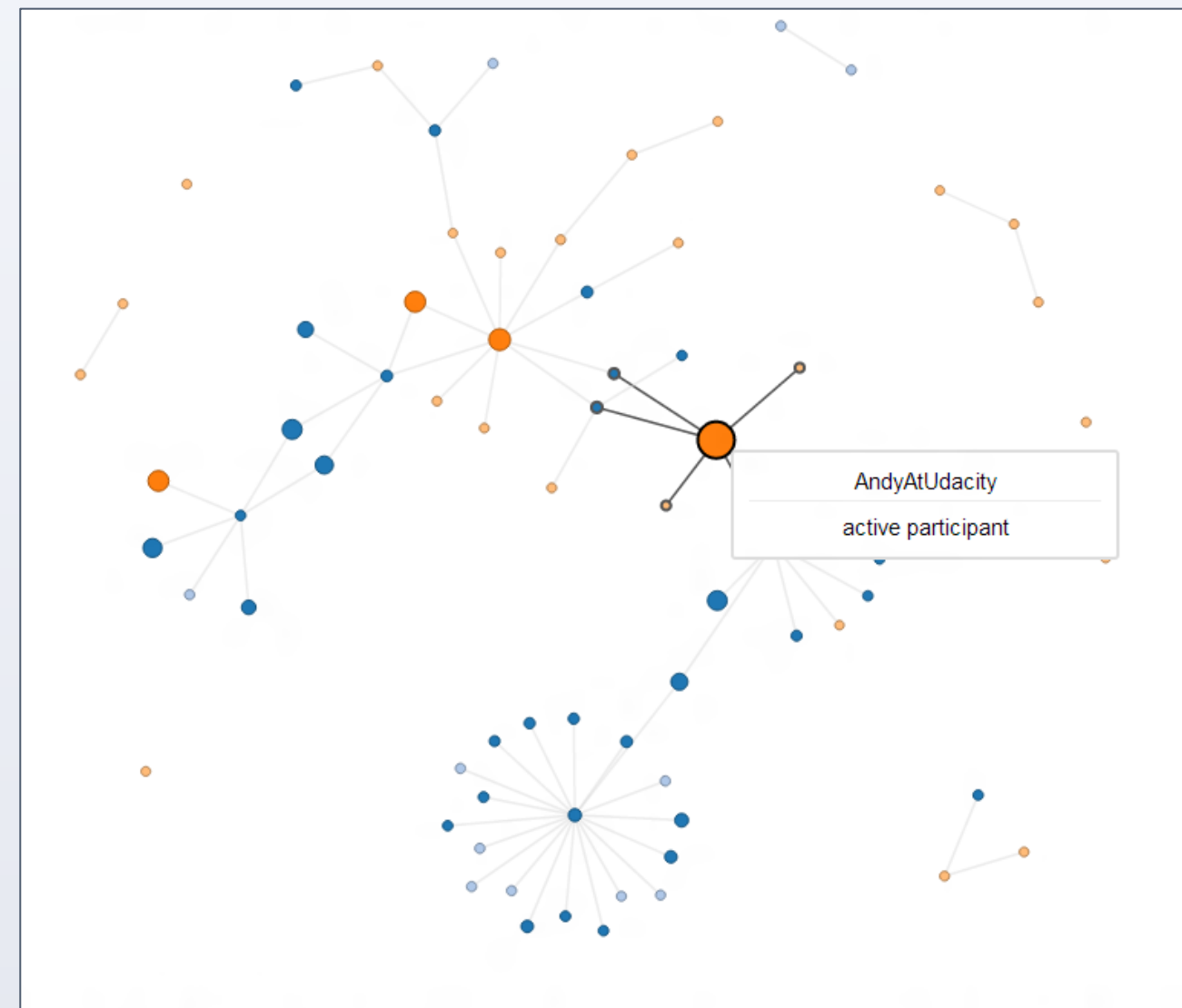
Text Analytics in Udacity Forums

Sai Pinapati

sai.pinapati@ufl.edu



INTRODUCTION



LOGISTIC REGRESSION

MATH MODELING - I

Define a hypothesis $h(x) = \text{sgn}(w^T x)$. We need to fit the weights to the training data. **online** updates to i^{th} user-model parameters. Each response and question pair is used to update the model.

Near real-time ranking of responses by the **probability** that the user will upvote the response. Inferring the how helpful a response is to a user without explicit user labeling.

FEATURES

- Social* - measures interaction between users
- Content* - attempt to identify terms that are highly correlated with the user acting (or not) on the response.
- Thread* - user's interaction with the thread.
- Label* - user applied topic-filters.
- Spam* - filter-out irrelevant data-points.
- tags*: 'off-topic', 'not constructive'.

MODELING TOPIC FLOWS

consider topics as a special kind of information that can "spread" through in the process of discussions. The evolution of topic discussions as Random Walks of topics on graphs where users are nodes.

The stationary-state probability for each node of the random walk represents how likely a topics flowing in the network will arrive at a certain user, or the importance and willingness of a user in participation to the discussion of a certain topic.

TOPIC FLOW - EQUATIONS

R_{ij}^d - frequency of user u_i replied by u_j in a thread d .

$$w_{ij} = \sum_{d \in D} R_{ij}^d$$

$$\mathbf{p}_{(t+1)} = \beta \mathbf{S}^T \mathbf{p}_{(t)} + (1 - \beta) \mathbf{q}$$

S - transition probability matrix
P - participation rank of user

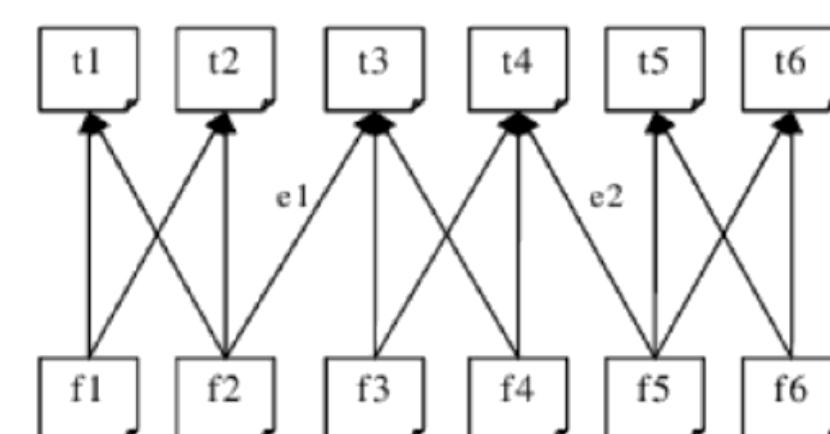
classification: lurkers, drop-ins, passive participants and active participants.

When is a student falling behind or not getting response in the forum?

Finding similar students.

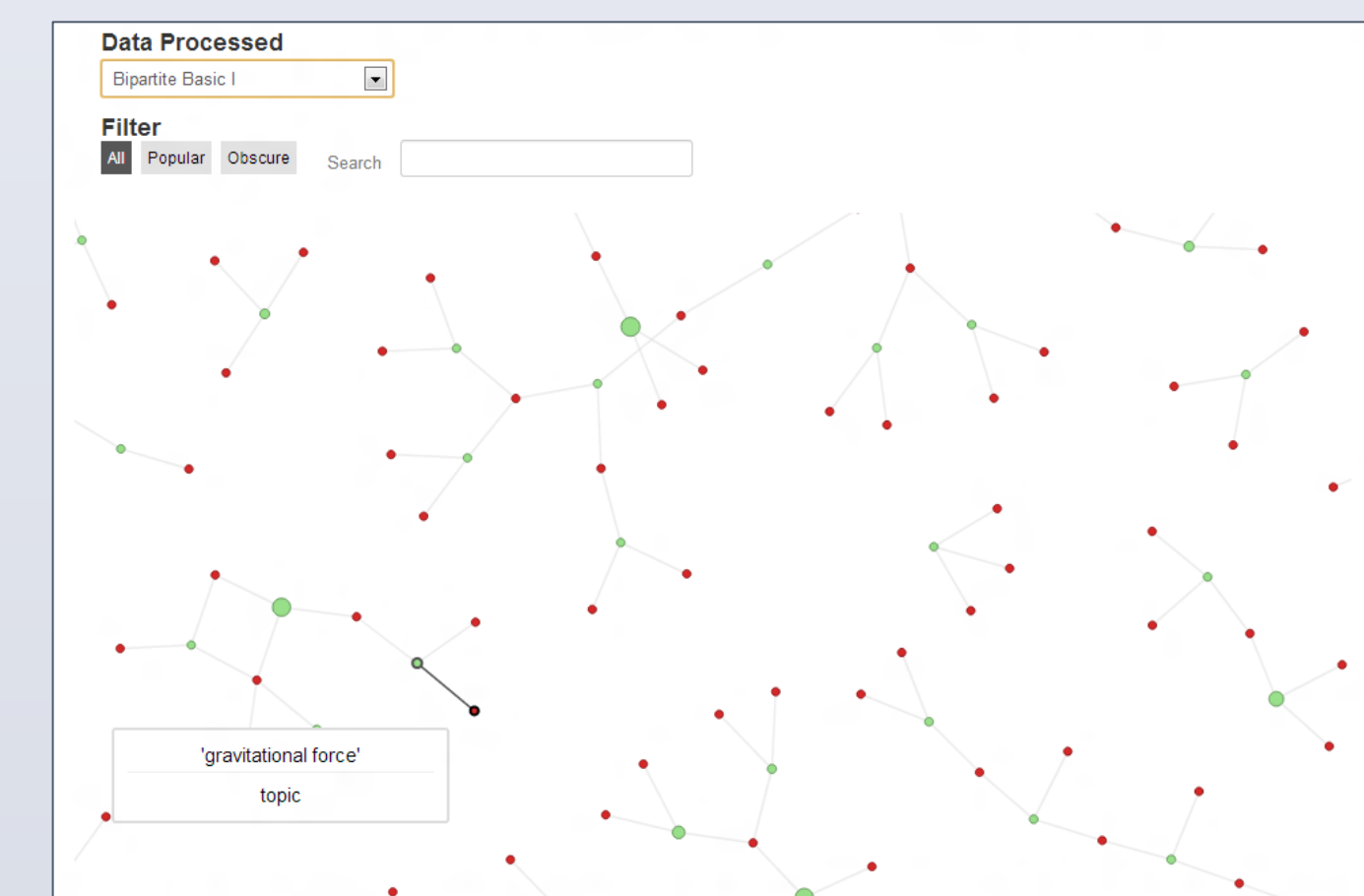
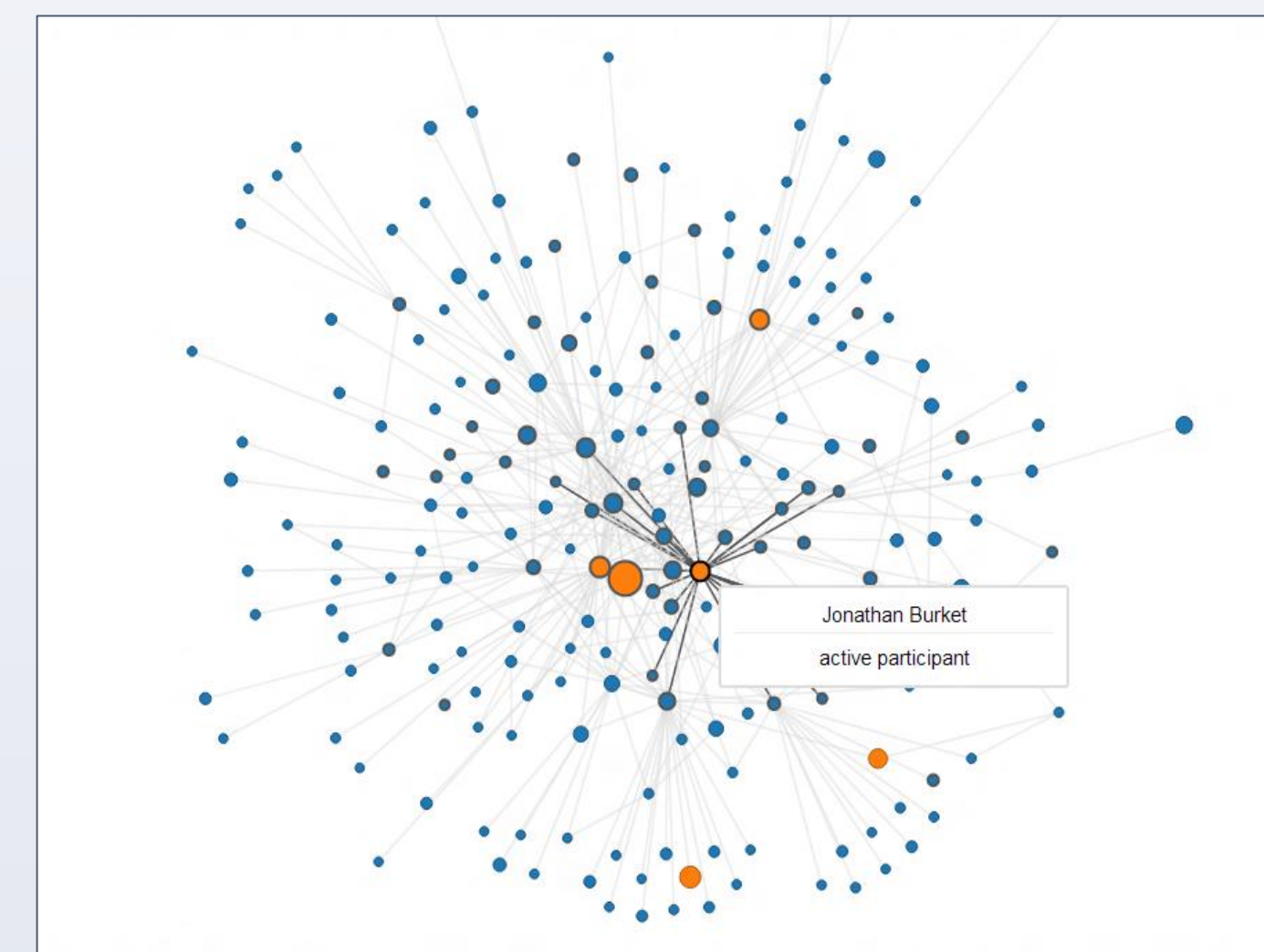
BIPARTITE GRAPH

Mapping between topics and students.

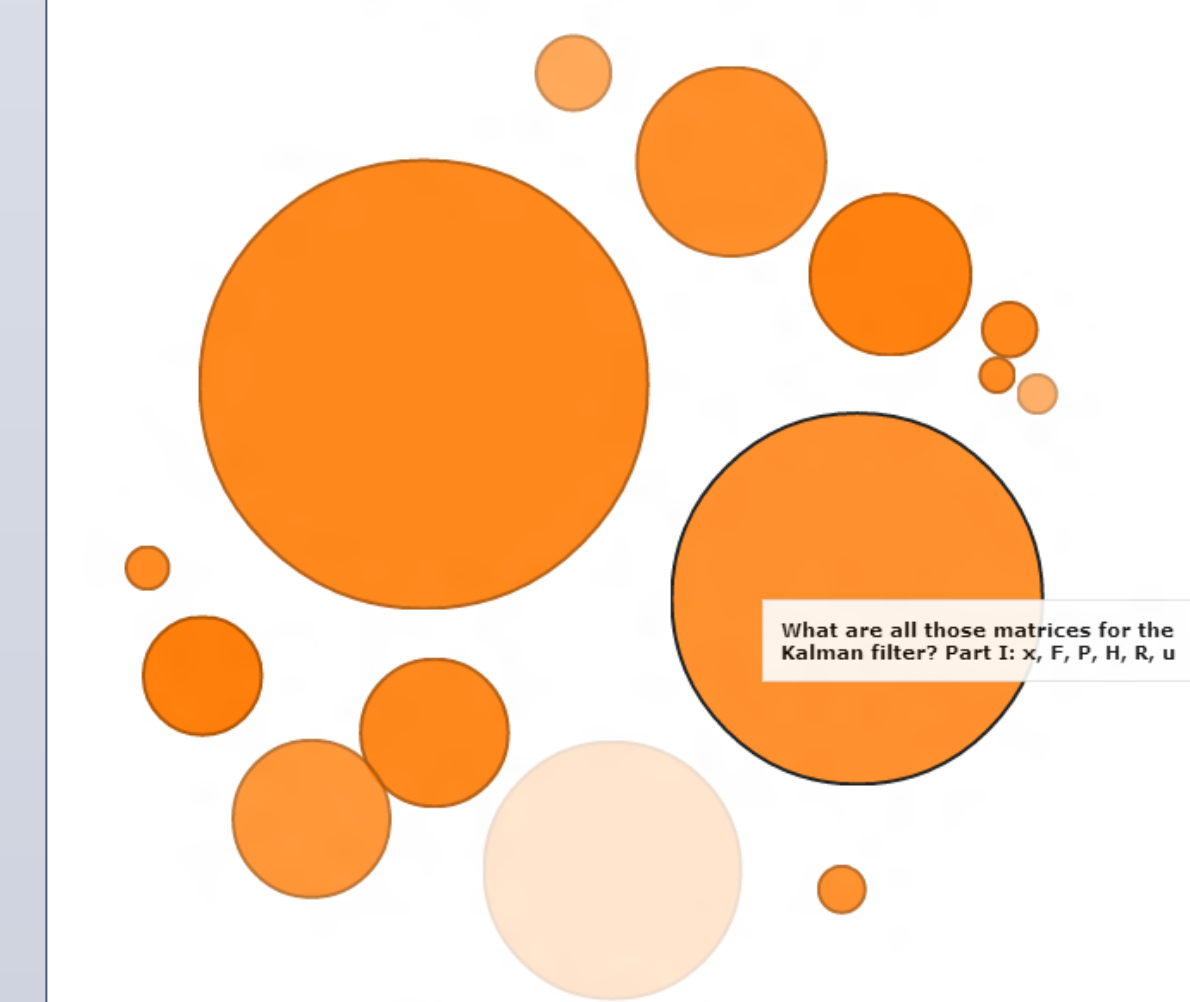


VISUALS

Distillation of complex data into easily digestible visuals.



Surfacing Interesting Discussions



INFRASTRUCTURE

Hadoop Jobtracker

Cluster Summary (Heap Size is 145.88 MB/3.56 GB)

Running Map Tasks	Running Reduce Tasks	Total Submissions	Nodes	Occupied Map Slots	Occupied Reduce Slots
1	73	173	15	1	73

Map % Complete	Map Total	Maps Completed	Reduce % Complete	Reduce Total	Reduces Completed	Job Scheduling Information
100.00%	7476	7476	55.58%	499	276	NA
100.00%	45	45	98.00%	3	0	NA
0.00%	1	0	0.00%	1	0	NA

APACHE PIG

event format in csv files after crawling:

```
user1, user2, thread_id, votes, views, timestamp
actions = LOAD '/crawled_data' using PigStorage(',')
-- grouping actions
grp = GROUP actions BY user1;
-- aggregating views
actions_by_user = FOREACH grp GENERATE group
as user_id, COUNT(actions) as score;
```

Data Pipeline

I wrote Apache Pig scripts to process crawled-data to find popular discussion threads.

Developed a decay-based algorithm to rank them.

The results were written into a Redis store so that they can be fetched during the page-load.